The University of Sheffield
Department of Chemical and Biological Engineering

The University Of Sheffield.

Improving proteomic methods and investigating H_2 production in *Synechocystis sp.* PCC6803

Andrew R. Landels

Thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

November 2016

Improving proteomic methods and investigating H_2 production in *Synechocystis sp.* PCC6803

Andrew Landels

28/11/2016

Contents

Acknowledgements

'Try not to worry, no one will read it anyway...'

By opening this book, according to academic conventional opinion, you have just made your way into the null set of people. Like you, this thesis probably shouldn't exist – I generally lack the sticking power to complete anything this involved, and have a bad habit of getting bored or easily distracted. That being said, it does exist; and whilst I am largely responsible for that fact: (*this book is most likely built out of some demonic combination of my now fleeting youth, sanity and mental acumen – which is a shame because it probably isn't even that good*) there are also a huge host of others without whom it would simply not have existed.

First and foremost though, I'm dedicating this book to Emma and Lee – you guys fed me, clothed me, and housed me; it's 4 years on and I'm still alive. Without you I never would have made it. To my parents (and extended parents), I love you. Thank you for giving me exactly the right set of features and continued support to get to this point. I know it hasn't always been easy, especially given the number of spare children in our ever-growing family, but when times got dark I've always been spurred on by the knowledge that you're always there for me – reinforced by the universal constant that is 'socks at Christmas'. To my siblings: Gerard, Alasdair, Leon and Maddie – I was the first one of us to write a book! Maybe now that I'm finally out of school, I can visit home once in a while – thanks to all of you for holding down the fort in my extended absence and reminding me that there's a place in the world full of people who are just as strange as me.

To my wonderful array of supervisors: true, you did all leave, but the impact each of you made will stay with me forever. Joss, a Bayesian at heart, you taught me right from the start to question my priors – exemplified by my surprise when you turned out *not* to be a woman! The long hours you invested in teaching me to be vaguely competent with a computer will stay with me forever, as will the memories of 'decent' burgers and leffe when we needed a *'manger a trois'*. Paul, only briefly a supervisor, but an excellent mentor and solid friend – the ability to bind the practical world with scientific knowledge that you ingrained in me will likely be the key to my future success. Phil, possibly the wisest and most prescient scientist I know – *'I'm not saying I told you so, but...'* – I asked for a technician job, you offered me a PhD. Thank you for having complete faith in my abilities, even when I didn't. Hopefully whatever potential you saw in me at that early stage will come to fruition over the coming years!

A shout out to silent majority, through a combination of brute force, gentle persuasion, scathing wit and casual hedonism; you are the heroes of this story. You all know who you are, we've laughed and cried, fought and taught, sung and danced, and the morning

after wondered how the hell we got there. Each and every one of you made the last four years of my life not just bearable, but special, and for that I am eternally grateful. To the wonderful folk at BSAC36 (and one particular diving goddess), thank you for opening up an entirely new world to me, giving me a precious escape from the stress-inducing clutches of academia. Special thanks to Caroline – my mentor in the world of mass spec. I hope that in my career as researcher I can demonstrate a fraction of the knowledge, wisdom, patience and kindness that you've shown me. To my fellow PhD-ers, thanks for always being there with a cup of tea, a glass of wine, a pint on a Friday, cake on a Wednesday, a party for every occasion, an irrational fear, an award-winning jumper, a night of karaoke, or a wedding; to raise my spirits when times got tough.

In the end it all comes down to money. From a financial point of view, I would personally be a lot poorer if it wasn't for the generous funding provided by The University of Sheffield. My materials and jet-setting lifestyle was kindly funded by the CyanoFactory consortium – EU FP7 grant No 308518. Being involved with that project gave me access to some of the finest minds in the EU; along with more than a few new friends to commiserate Brexit with. Finally, thanks to Mike and Team V for taking me in at the end of it all and hauling my withered husk over the finish line – it might not always have seemed it these past few months, but I really am grateful.

For many of you reading this, this is where the story ends. For those who plan to continue, I hope you have as much fun reading it as I had writing it! I think I'll close with a solid piece of advice to the prospective PhD student, that I received during my undergraduate course and seem to have ignored ever since:

'There are an infinite number of possible mistakes you can make, stop trying to learn through the process of elimination and just do it the way I told you to!'

Abstract

The annual EU consumption of energy is approaching 3 Terawatt.hr^{-1}, but the majority of this is powered by fossil fuel. Burning of fossil fuels has produced a global catastrophe, climate change, and carbon-free replacement technologies are urgently required to prevent this from becoming worse in the coming years. The CyanoFactory consortium worked to optimise the organism *Synechocystis sp.* PCC6803 (herein *Synechocystis*) to produce industrially relevant levels of bio-hydrogen as one such potential solution. This thesis discusses aspects of this ambitious project, focusing on understanding and optimising the internal protein network of the organism to engineer a functional and efficient system.

Synechocystis is a model cyanobacteria – and so has a significant body of research associated with it compared with other cyanobacteria, but is nowhere near as well studied as the other major model organisms such as *E. coli* or *S. cerevisiae* – particularly in protein-level studies, although this is changing with time. Whole-proteome studies are highly advanced in medical applications, however bioengineering using proteomics still lags behind studies which directly measure individual proteins, metabolic outputs, or nucleic acid studies. A number of proposals emerge from the literature as the most effective way to move forward, part of which is filling the gaps in the literature for *Synechocystis* and production strains in general. The major improvement missing from this field is the broad-spectrum inclusion of broadly applicable bioengineering techniques, such as synthetic biology, being integrated with whole-proteome studies, rather than just focusing on individual pathways. This gap is likely to be filled in the near future, with the recent improvements to proteomic technologies and the increasing popularity of the methodology – which has seen a sharp increase since the start of 2015.

The current gap between the medical studies and production strains provides an opportunity to test a variety of different approaches, that look more at general whole-cell level responses rather than targeted observations. These gaps in knowledge are assessed herein, and new methods for analysing *Synechocystis* specifically are proposed. These proposals cover both alterations to the practical protocol, including physically lysing cells based on meta-analysis of the literature with experimental verification, more accurate methods of determining protein levels – which are generally complicated by coloured compounds found in cyanobacteria; and computational protocols for improving the quantity, quality and relevance of the data obtained, including better observation of low-abundance proteins in a complex background, assessment and recommendations for expanding the number of different samples that can be measured simultaneously, and simpler tools for identifying broad-sweeping changes, where metabolic-network derived investigations are unsuitable.

Isobaric tags are popular methods for analysing the relative quantity of proteins observed in a cell-wide sample, however there are different technologies for this method. The two most popular tag-based quantification technologies – iTRAQ and TMT – are directly compared, to determine which method is more suitable for analyses in *Synechocystis*. The study was focused on *Synechocystis*, however the observations are also more generally applicable to other investigations. To perform this study, a modelled assessment of the 'proteomic background' of *Synechocystis* was carried out, providing an impression of the internal proteome distribution – a valuable set of information for carrying out more accurate engineering of the internal mechanisms with technologies, such as Synthetic Biology. The study found that whilst TMT tags generally produced more quantifications, the iTRAQ tags were more accurate over a greater range – however to take advantage of this would require a larger number of repeated injections of the iTRAQ samples, producing a relatively inflated cost for better quality data.

Combining these tools, a direct assessment was carried out of the systemic changes that occur in *Synechocystis* under hydrogen-producing conditions, along with an assessment of a media proposed for optimised H_2 production. This experiment first carried out with the methods used more widely at the start of this analysis, and the second was conducted afterwards, utilising many of the methodological improvements proposed in this thesis. Ultimately, an increase in data quantity and quality was observed. As hydrogen production is a response to a change of conditions, the pathway-level assessment of the proteome changes show a concordant switch between 2 very clear states under the experimental conditions used. This suggests that finding a way to produce hydrogen directly – under normal growth conditions in light – will be extremely challenging as it fundamentally competes with the growth and function of the organism; however an integrated approach, merging the production of high-value side products during the day, coupled with hydrogen production at night for generating power to run the bioreactor system, has a much greater chance of success. A decision on which products should be targeted to make the system economically viable will dictate further analysis of the data.

The major conclusions of this work show that the suggested improvements are beneficial to proteomic studies in *Synechocystis*, producing an improvement to quantity, quality and accessibility of proteomic data. These observations have been applied to hydrogen production systems, demonstrating that whilst bio-hydrogen is unlikely to be the white knight that will save the world from climate change, it can be integrated into large-scale production systems to improve energy efficiency – where the energy saved can reduce costs and power-inputs required from carbon-based fuels. The methods suggested here, whilst ultimately adding little to the assessment of H_2 production, have huge potential when integrated into future project focused on the production of more economically viable complex organic molecules or fine chemicals.

Chapter 1

Introduction

1.1 Energy, Biofuels and High Value Products

1.1.1 Fossil Fuels

In 2010, European countries consumed over 2.77 Terawatt.hr^{-1} of energy (Energy Information Administration, 2013), with the major provision of this generated from fossil fuel sources. Fossil fuels are convenient for use: they are high-energy density hydrocarbon molecules that can be transported great distances and stored stably for long periods of time. This makes fossil fuels ideal for transportation such as automobiles, where electrically powered solutions are limited by battery technologies or lack of a constantly availabile of a power supply. Furthermore, a stockpile of fossil energy resources can be collected and pooled, enabling the provision of stable outputs of electricity from a powerstation over long periods of time, generally without dramatic short-term fluctuations in either price or supply. This means that the energy output can also be tuned to the operational requirements at the time, and so during winter months when energy demand is generally higher, the output of the power stations can be altered to provide what is needed.

Beyond this convenience of use and flexibility, the technology has been in widespread use for over a century and is implemented in every country across the modern world. As a result, there is huge financial backing of the fossil fuel industry, who have been investing in carbon capture and efficiency-improving technologies; such as clean-burning liquefied natural gas plants (Metz et al., 2005). These are a significant improvement on the classic coal-fired power plants, which produced harmful environmental pollutants such as sulphur dioxide derived acid rain, and are now largely decommissioned - although these coal-fired power stations are still notably used widely in developing countries such as China and India, which together make up over 75% of the currently planned coal-fired power stations (Yang and Cui, 2012). Ultimately, the strongest support for fossil fuels is that there is a great deal of infrastructure already in place, along with widespread public acceptance of pre-existing technology powered by coal, oil and gas.

1.1.2 Limitations of Fossil Fuels

Despite these key advantages, there are several world-altering drawbacks to the widespread use of fossil fuels. Supplies are limited because creation of these fuels takes hundreds of thousands of years and extraction techniques are becoming more expensive as easily obtained oil is consumed. We are now beyond the era referred to as 'peak oil', where the amount of energy required to draw the oil from the ground was significantly lower than the energy that could be obtained by burning it as fuel (Bardi, 2009). The ratio was initially 1:100 input to output, but recent estimates place it at a range closer

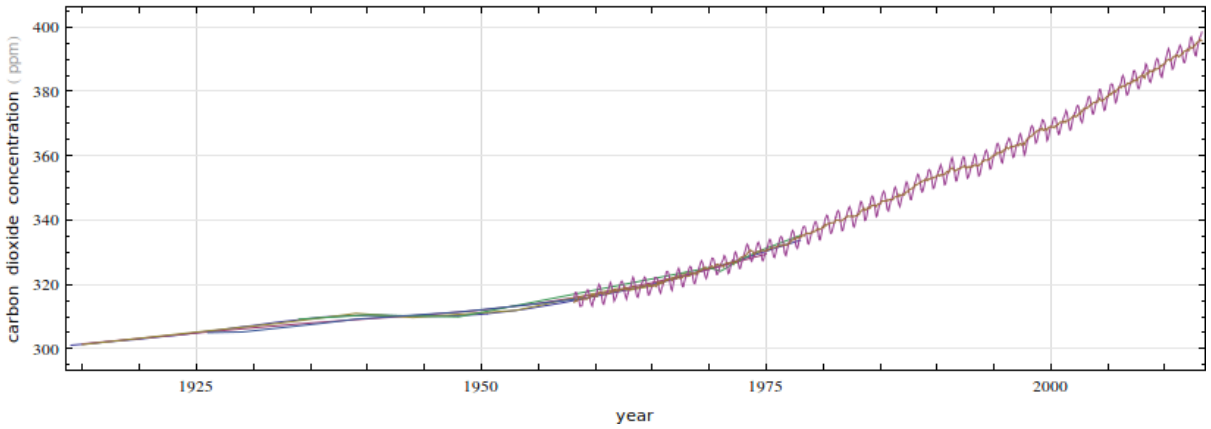


Figure 1.1: A graph charting global CO_2 concentrations in parts per million over the last century. (Image created in Wolfram Mathematica.)

to 1:8. Both of these factors drive the price of fossil fuels up, which – given the heavy requirement for it in modern society – has potentially catastrophic economic and social consequences.

Burning fossil fuels releases a number of pollutants into the environment that have been long-dormant. Numerous world-wide investigations have been conducted, and a compiled report published by the Intergovernmental Panel on Climate Change (IPCC) has shown that over the past 60 years, human activity is responsible for most of the systematic warming effect. The report goes on to say: 'Emmissions of CO_2 due to fossil fuel burning are virtually certain to be the dominant influence on the trends in atmospheric CO_2 concentration during the 21st century' (Metz et al., 2005). CO_2 is a known greenhouse gas that has been strongly linked to changes in global climates and levels of CO_2 have been rising at an accellerating rate in recent decades - as seen in figure 1.1 (p. 15). Europe alone produces over 4 billion metric tonnes of CO_2 each year (Energy Information Administration, 2013).

Furthermore, in 2014 the EU imported 1.4×10^6 ktoe (kilotonne oil equivalent) of energy, whilst producing 7.7×10^5 ktoe and exporting 5.3×10^5 ktoe. This resulted in a gross consumption of 1.6×10^6 ktoe, with more than 80% of that energy being supplied from outside the EU (Eurostat, 2016). This influx of energy puts the EU energy supply at risk of disruption due to external political instability, or world-wide events – and the risk is compounded when the climate effects produced by the burning of fossil fuels trigger that instability. Two key recent examples of major climate-change driven disruptions, which have impacted the global energy price and supply, are hurricane Katrina and the drought that precipitated the Syrian war (Janković and Schultz, 2016; Fröhlich, 2016). The first led to the closure of a major oil refinery and resulted in a large spike in oil prices, whilst the second resulted in a rapid decline in oil prices: triggered by the decreased stability in

the middle east resulting from the war led to in a large influx of oil into the market from Saudi Arabia, to undermine oil sales funding factions with opposing interests. Both of these events had world-wide consequences that could have been mitigated if not for the dependence on fossil fuels.

It is worth noting at this point that large-scale electricity generation methods, such as hydroelectric and nuclear, are ready-to-go alternatives for producing large-scale electrical power; however without radical increases in battery power they are insufficient to continue to power the world the same way as fossil fuels can. There is a great deal of research currently being carried out in the field of energy storage, which has been greatly boosted by the rise of cellular powered mobile telephones and personal devices. So whilst energy storage is an unsolved problem, and batteries and charge-station infrastructure are improving at a rapid rate, they are unable to offer an out-of-the-box replacement to fossil fuels at this time – although they may in the future.

In terms of issues with these forms of power generation, whilst hydroelectric can be tuned to match output requirements, and energy can be stored in the form of water held against gravity; construction of new hydroelectric plants can be highly disruptive to the environment, as it requires flooding of areas and re-engineering of the landscape. The major issues in the case of nuclear power is that it is generally distrusted by the public and seen as a dangerous form of energy. As such, it must undergo a major re-brand if it is to become more prominent – particularly in the west. There is also an issue of long-lived, very harmful waste; although new generation reactors, such as propagating wave reactors, can overcome these limitations.

1.1.3 Renewable Energy and Biofuel

There has been large-scale investment in renewable energy that has been rising year on year over the last decade; the majority of investment in 2015 was in solar panels (80% of investment) and wind turbines (10% of investment). In 2015, renewable energy (excluding large hydro) made up 16.2% of the established power capacity - up from 15.8% in 2014, but only contributing to 10.3% of the global electricity supply – the difference between these figures resulting from the part-time energy generation ability of solar and wind power (McCrone et al., 2016). Whilst these technologies are vital for electricity generation and provide part of a solution to weaning the world off of fossil fuels, they do not address the issues like the need for automotive fuel or long term energy storage. This is where biofuels have potential to contribute to the solution.

Biofuel is any fuel generated from biological processes, although it usually refers to liquid fuels that are terrestrially generated when referred to in public reports. These are usually

Figure 1.2: A probable cause for the increasing CO_2 concentrations in parts per million over the last century. (Image by Z. Weinersmith - SMBC (Weinersmith, 2016))

classified as first and second generation biofuels, where first generation uses food crops as a feedstock to make fuel, whilst second generation uses non edible, woody or grass-like plants. These biofuels can either be utilised directly, in cases such as oil seed rape, or else processed by microbial action to produce organic molecules such as bio-ethanol (Ducat et al., 2011). Direct photosynthetic creation of biofuel using photosynthetic microbes is referred to as 3rd generation biofuel. This can produce oils, alkanes or even hydrogen gas. Like fossil fuels, biofuel has the advantage of being an energy-dense carrier molecule and is considered to be a viable replacement for fossil fuels (Savage et al., 2008). All fossil fuels that are used today were originally biologically generated, and so it follows that biofuels would be a suitable substitute.

Biologically, processes can broadly be divided into two categories, heterotrophic growth – or those deriving energy from a chemical supply, and photo-autotrophic growth – or those deriving energy from a solar supply. When considering a solution to the current energy crisis, it is important not to generate a solution which creates foreseeable problems in the near future. The majority of biofuel being produced currently is derived from first generation biofuels, with the bulk of remaining production coming from second generation biofuels (Doshi et al., 2016). The three most prominent first generation biofuels are maize (USA), sugarcane (Brazil), and oilseed rape (Europe)(Hill et al., 2006; Escobar et al., 2009; Ajanovic, 2011; Gasparatos et al., 2013). These biofuels are supported financially by public bodies to make them competitive with fossil fuels, however this has artificially raised the prices as a food crop, and has also contributed to a broader food price increases of between 5 - 15% over the last decade by raising the prices of agricultural feed (Rosegrant, 2008; Mitchell, 2008; Drabik, 2011). The most prominent example from these is maize, where around 40% of the recent increases in price are as a direct result of fluctuations in the biofuels market (Fischer et al., 2009). As both first and second generation biofuels utilise terrestrial space, they both have issues of competing with agricultural land needed to produce food. Where crop-bearing land is not used, conversion of land into cause CO_2 release through the clearing of land; and these biofuels take a long time to produce a return on investment – either financially when discounting the government subsidies, or even from CO_2 reduction given the dependence on fossil fuels for harvesting or requirements for blending, although this 'green paradox' can be mitigated to a certain extent by simultaneously subsidising bio-ethanol but imposing counter-balancing taxes on fossil fuel use (Galinato and Yoder, 2010). It has been predicted that reaching a carbon neutral state with land-clearing can take around 100 years, with a maximal prediction of 700 years when considering natural carbon sinks like peat rainforest land (Doshi et al., 2016). Finally, it is not just land use but also water use that generates issues for crop-based biofuel production. Agricultural processes and energy production currently utilise around 90% of global fresh water. In addition, treating and recycling

that water uses about 3% of the annual electricity generated.

1.1.4 Third generation biofuels and high value bio-products

Third generation biofuels, where energy is generated directly from photosynthetic inputs, have a more promising outlook. It has been estimated that 20 - 30% of global bioproduction takes place in the sea by photosynthetic algae, microalgae and cyanobacteria (Hall and Rao, 1999); and it is possible to grow certain these photosynthetic organisms in environments that would otherwise be unsuitable for crop production, such as locations with extreme soil pH or high salinity (Ducat et al., 2011). Potable water resource utilisation is less of an issue as well; as water salinity is a local rather than global phenomenon, and so microalgae tend to be resistant to ranges of salinity, from fresh water, to brackish water, to full marine and even hyper-saline environments such as the dead sea for *Dunaliella* (Evans and Kates, 1984; Kirst, 1990). Some photosynthetic algae are also particularly fast-growing, high-producing strains, such as *Chlorella* can reach productivities of up to 30 $g/m^2/day$ with a starch content of 37% (Hirano et al., 1997).

Certain species of photosynthetic microalgae, such as *Spirulina*, are also considered 'superfoods' and as a result, have the potential to add to the global food supply, rather than detracting from it (Ciferri, 1983). In addition to this, 3rd generation biofuels have the potential to produce a wide variety of high value products (HVPs) along side the production of oils for fuel, such as carotenoids (Wijffels and Barbosa, 2010) or astaxanthin from *Haematococcus* (Lorenz and Cysewski, 2000) – which retail from $10 kg and $100 kg respectively (price estimates taken from alibaba.com on 09/2016). These HVPs can provide a grown-in financial support system, which is usually provided by government subsidies for previous generations of biofuels (Heidorn et al., 2011; Stephens et al., 2010).

Whilst the prospects for third generation biofuels appear to be quite promising (Duffy et al., 2009), there are also a large number of limiting factors that need to be considered. Firstly, cyanobacteria and algae represent a huge variety of diverse species, and whilst there are a variety of very advantageous features, these are spread across 3 phyla – cyanobacteria, green algae and red algae. As a result, many of the projections about their potential productivity have been greatly exaggerated, as to date no single organism has ever existed containing all the key features often described when considering the future of the field; Klein-Marcuschamer quite humorously states:

"*lignocellulosic ethanol could probably be a widespread product today if one could design and build a process that used a microbe with the growth rate of Escherichia coli, the ethanol tolerance of Saccharomyces cerevisiae, the thermophilic nature of Sulfolobus acidocaldarius, and the cellulolytic activity of Trichoderma reesei.*"

(Klein-Marcuschamer et al., 2013)

Whilst photosynthetic growth has been suggested as a good method of carbon capture and a promising biofuel production facility for the future (Chisti, 2010; Chisti and Yan, 2011), it is important to be realistic about the capabilities. There is an issue with providing the required amount of CO_2 and other nutrients where algal growth parameters are also optimal (Pate et al., 2011). It has been estimated that to produce 100 t algal biomass requires approximately 183 t CO_2, at a concentration approximately 100 fold higher than that found in ambient air (Chisti, 2007). From a practical point of view, this would require a dedicated industrial source within an area that is conducive to large scale algal growth. It has been estimated 35 mt of algal oil is the feasible limit at which freely available CO_2 resources in suitable locations are able to supply in the US (Pate et al., 2011), utilising around 10% of the CO_2 generated from electrical production (Klein-Marcuschamer et al., 2013).

The technology is still in its infancy, and whilst there are a number of pilot-scale studies detailing extensive research and promise (Wijffels and Barbosa, 2010), there are relatively few functioning large-scale algal biotechnology companies worldwide, and fewer still pursuing biofuels (Schmidt, 2012). The largest risk associated with generating photosynthetic output comes from the initial capital expenditure (capex) of the project vs the rate of return through operation expenditure (opex) (Bosma et al., 2014). The plant set ups are therefore an important consideration when looking at scale-up. There are two major designs for large scale algal production, open topped 'raceway ponds' – also known as Oswald ponds after their creator (Oswald, 1988) and enclosed photobioreactor sysytems (PBRs). PBRs have both higher capex and opex costs than raceway systems, however, PBRs also have higher productivity (1% vs 3% in vertical column designs) and the protection against contamination that is observed in raceway systems (Bosma et al., 2014; Klein-Marcuschamer et al., 2013). In addition, a direct comparison between the two methods has shown that whilst there is higher initial expense for the set up and operation of a PBR, the biomass is axenic (only the intended culture is present) and can grow up to 30 fold higher density, which reduces costs for downstream processing steps, such as removing the algae from the growth media (Chisti, 2007). More recent developments in this field have found that separation of biomass from the liquid culture is less of an issue (Chisti, 2013), at least when considering biomass harvesting for crude production. Ultimately, when considering a production platform it should be matched accurately to the production system required. Raceway systems are by far the more prevalent system for biomass production, however a single, highly profitable system of tubular bioreactors operates in Israel producing high value (>\$100,000/t) *H. pluvialis* biomass (Klein-Marcuschamer et al., 2013), demonstrating that when the value of production is high enough it can offset the capex and opex costs – although this is a very

specific case.

It is important at this point to also consider the socio-economic effects of wide-spread biofuel production. The locations where the environmental conditions are optimal for algal biofuel production: good solar light availability, high ambient temperatures, nearby water resources and a built up industry for CO_2 influx, are also frequently in the developing world and could be subjected to political 'land-grab' events (Comelli, 2012). In addition, some of the products that are planned for algal development are currently produced in the developing world, such as palmitate (palm oil), and so could cause economic disruption to communities that rely on the production of such resources. A similar argument was made when the company Amyris planned to perform microbial production of the anti-malarial artemisinin, which is extracted from the plant *Artemisia annua* (Thompson, 2012). Amyris initially claimed that their efforts were to reduce the costs of this important drug, however later this was revisted to increasing stability of the supply. As of 2011, the synthetic variant of the drug made up approximately 1/3 of annual global production, and Amyris re-branded and became publicly listed as a biofuels company (Thompson, 2012).

1.1.5 Hydrogen and biohydrogen

Many of the issues mentioned above relate directly to algal oil production, however there are other, less prevalent forms of biofuel that have promise. Biohydrogen is a clean-burning waste product, that can either be produced by a single organism directly or through fermentative processing of biomass (Vignais and Billoud, 2007; Saifuddin and Priatharsini, 2016). H_2 is light, burns in air to produce H_2O, has the highest energy density of all known chemical molecules by mass, and can be used in fuel cells to directly produce high-efficiency electricity.

Around 80% H_2 generated is produced chemically, through steam reforming of natural gas (Friedrich et al., 2011); although this requires a fossil fuel input and produces CO_2 as a side product. Potentially carbon neutral chemical processes, such as electrolysis, are well established but generally require high temperatures and pressures to operate efficiently. As a result, these processes are typically coupled to large-scale centralised energy production facilities, such as nuclear power stations, where such conditions are readily available. Other methods, such as direct photovoltaic separation are also possible, however the costs of such processes are currently an order of magnitude above a financially viable level (Tributsch, 2008). Through biological action, the initial activation energy required to conduct this process is greatly reduced, as it is performed at biologically relevant temperatures, and so biological processes have the potential to produce H_2 in an efficient manner. Hydrogen infrastructure is currently limited, however, it is important

to note that despite this, H_2 driven technologies are developing. The first H_2-powered vehicles already in existence, a device that runs on low temperature hydrogen fuel cells (Eberle and von Helmolt, 2016) and is powered by compressed H_2 gas (von Helmolt and Eberle, 2014). When comparing the cost of energy storage for a hydrogen car vs an electric car, the H_2-powered car costs around 1/10th of a similar Li-ion battery system (US$3000 vs US$30,000 - 50,000) (Eberle et al., 2012); making the hydrogen car a cheaper option.

Fermentation of biomass to produce H_2, typically with bacterial organisms such as *Clostridium*, has a wide variety of applications and is far more prevalent in the literature; however it faces many of the same issues as algal oil production in terms of requirements for biomass. Biohydrogen production within a single organism, such as through direct photolysis of water, atmospheric nitrogen fixation and light/dark fermentative production within a photosynthetic system (Karthic and Joseph, 2012; Tamagnini et al., 2002), are a collection of much more promising options for avoiding some of these issues. The product can be collected without the need for harvesting the biomass, where the hydrogen is produced through either photolysis of water or through fermentative processes, which can potentially reduce operational costs significantly for the requirements of nitrate, phosphate, and carbon; as the biomass doesn't need to be harvested completely. As the technology is still largely undeveloped, there are numerous technological impediments to its utilisation: the costs involved are very high, the yields are very low, the infrastructure for supporting hydrogen as a fuel sources doesn't exist in a particularly large scale at the moment. Whilst pilot scale plants operating at between $0.035m^3 - 100m^3$ (Ren et al., 2006; Morra et al., 2014; Zhi et al., 2010; Vatsala et al., 2008) have demonstrated fermentative production of H_2 (Ren et al., 2006), there are no large-scale tests of a single-organism biohydrogen production system currently in operation.

There are issues with H_2 storage. Due to its small size, H_2 has a particularly prevalent issue with permeation – where a gas or liquid can pass directly through a solid. This is not an issue when using industrial level aluminium coated storage tanks (Korinko et al., 2001), but when considering low cost polymer-based tanks as found in many motor vehicles, the permeation can be as high as 12 orders of magnitude higher than in aluminium tanks (Stodilka et al., 2000). This poses efficiency and safety issues, as there can be explosion risks when considering such tanks being placed within an enclosed space, such as when a car is stored in a garage (Saffers et al., 2011). While the need for this advanced tank does add to the price of a hydrogen powered vehicle, as stated above the cost is still far lower than an equivalent rechargeable battery storage device.

The need for specific storage tanks adds to the capex of any biohydrogen production plant planning on storing hydrogen. In addition, as hydrogen is a gas, storing it will require a compression step which reduces efficiency that is not present with liquid fuels.

As an example, a $50m^3$ pilot-scale hydrogen production plant in China, attached to a citric acid production plant, showed that whilst chemical waste could be converted into hydrogen and stored. As an efficiency comparison, the loss associated with hydrogen storage was compared to directly converting the output into electricity and charging a vanadium battery, which showed that the stored H_2 had both a lower efficiency and higher associated costs (Zhi et al., 2010) – although the energy outputs between the systems were small. The efficiency in this case was affected by several factors, the output gas from fermentative processing also contained CO_2, which made up around 60% of the gas. To purify the H_2 to >99.99% purity, it was carbon-scrubbed through an alkaline absorption tower, which reduced the energy efficiency. Compression was also needed for storage, and so the efficiency was reduced by a further 30%. In terms of final storage systems, the H_2 produced by the process was stored in 11 660 L storage tanks. These additional systems brought the total cost to around 60% more than the battery system (Zhi et al., 2010).

When considering biohydrogen by direct photolysis or light-fermentation, the issue of CO_2 removal from the outputs are removed as fermentative processes are not active within the cell. Unfortunately, there is the much more significant problem of oxygen inactivation of the process. O_2 is already a limiting factor for microbial photosynthetic production – when the O_2 levels get too high they inhibit photosynthesis and cause oxidative damage to the cells; which impacts on PBR design, necessitating gas exchangers and limiting PBR pipe lengths to a maximum of around 80 m (Chisti, 2013; Chisti and Yan, 2011; Wijffels and Barbosa, 2010; Bosma et al., 2014). Hydrogenases are generally highly O_2 sensitive (discussed in more detail later), and so when producing H_2, either the O_2 levels need to be kept at a minimum within the system (a significant challenge during photosynthesis), or the cell must have a method of excluding O_2 around the hydrogenase machinery. This is not an insurmountable task, although will require technological development. There are some O_2-tolerant hydrogenases, which still function in the presence of O_2 although operate at a lower energy density (Lenz et al., 2010). Beyond this, intracellular oxygen consuming devices have been created for photosynthetic organisms (Tamagnini et al., 2002, 2007), and hydrogenases that hijack the photosynthetic process to directly channel electrons into hydrogen production, preventing O_2 production are also under development (Friedrich et al., 2011).

1.1.6 CyanoFactory

As shown, there are many issues with the production of third generation biofuels, however a large number of these are technical. For example, a reduction in the cost of PBRs and improvements in operating efficiency will significantly reduce the related costs whilst maintaining the benefits. In addition, it is possible to make an operation feasible by

increasing the value of the outputs of production, such as with the utilisation of HVP side products as mentioned above.

Ultimately, the creation of a cost-effective photosynthetic microbial production system spans a number of disciplines, ranging from genetic and molecular, to analytical modelling, to large scale process design and scale-up. As a project of this nature is very interdisciplinary, it requires a few principles to ensure cohesive flow: firstly a suitable organism or chassis should be selected and the entire process should engineered specifically around the features of that organism. This avoids problems of incompatible modules being developed by different individuals. Secondly, it is essential that the different modules of work feed back into each other through regular communication. This enables a broad point of view to be taken over the whole project at each step, to ensure that drift doesn't occur during progression of the project. Finally, when a terminal product is the desired output, a clear plan for final integration of the modules should be presented at the outset and re-evaluated for feasibility at regular intervals. This ensures that the project has a global final aim, which enables benchmarking of progress and allows contextual demonstration of issues that arise during the project.

This thesis was written in concordance with such a project – CyanoFactory. CyanoFactory (CyanoFactory, 2012) was a European consortium of 10 groups in 7 countries working towards optimising the model cyanobacteria *Synechocystis* sp. PCC 6803 (herein referred to as *Synechocystis*) for the photosynthetic production of H_2 as a biofuel. The consortia was funded over a three year period by the 7th Framework Program as a future emerging technology (FET) project in energy development under grant agreement number 308518.

CyanoFactory was the second stage of a longer-standing project on hydrogen production in *Synechocystis*, where it was built upon the FP6 EU pathfinder project Biomodular H2. Biomodular H2 was a pioneering study into genetic modification of the cyanobacteria *Synechocystis* and development of Synthetic Biology tools – it was one of 18 such projects funded by the EU, yet received 8.3% (1/12) of the total funding awarded to Synthetic Biology projects under the 6th Framework program NEST (Pei et al., 2011). During this project, the Ni−Fe hydrogenase was investigated in detail, and the functional subunits of the hydrogenase-active 'hox' cluster were identified. This was verified through the production of a hydrogenase-inactivated mutant, which was found to be sensitive to reductive stress environments such as those found during anaerobic fermentative production. This mutant was one of several resources that were created, with others including biological models to provide and an understanding of the background genetic interactions, an oxygen consuming device construct, and the groudwork for developing a number of molecular tools (Heidorn et al., 2011; Pinto et al., 2009; Ow et al., 2011; Huang et al., 2010; Camsund et al., 2011; Noirel et al., 2008).

Where Biomodular H2 looked mainly at the molecular aspects, CyanoFactory aimed to both further develop a system based on those tools, and then pull them together for the first time to identify a possible complete system that could be taken forward for further development. Cyanofactory therefore approached the problem of biofuel production on many different levels simultaneously, as shown in figure 1.3 (p. 26):

- Four groups focused on working on modifying the organism, where investigations into improving the H_2 output were carried out.
- A Synthetic Biology toolbox was produced, to enable much simpler and standardized genetic modification of the organism, using synthetic biology techniques (Huang and Lindblad, 2013; Lindblad et al., 2012).
- *Synechocystis* was stress-tested for salinity survival and ab investigation was carried out into how the organism was affected by the typical extremes of temperatures experienced within a PBR system during annual operation - with temperatures ranging from 4 - 40 °C.
- The genome of the organism was assessed for genetically neutral sites (Pinto et al., 2015) is also being optimised for genetic transformation and production of H_2.
- A light harvesting mutant referred to as 'olive', with better permittivity of light through the culture, was analysed to determine how it could improve the efficiency of the system (Kwon et al., 2013).
- There was also work performed on a biological safety mechanism, as a fail-safe in the event that the organism escapes into the environment; which was coupled with an investigation on the growth of the lab strain in local waters.
- Two industrial partners designed and produced flat-panel PBR systems. This PBR design has been shown to be the most efficient PBR designs for maximising cell dispersal and light harvesting; whilst minimising extreme temperature and light conditions and reducing costs for harvesting (Bosma et al., 2014).
- Two different flat-panel PBR systems were designed, each at different scales. The first was a lab-scale 5 L system, whilst the second was 2000 L outdoor reactor. The outdoor system was tested and investigations into predation management systems and daily operational effects were investigated (Touloupakis et al., 2016).

Information collected within the consortium was fed back to the other members through a central electronic storage facility known as the 'Data Warehouse'. Information in the Data Warehouse was presented as a central interactive biological model, which was annotated and maintained by a consortium partner (Wünschiers, 2016).

Throughout this project, the Sheffield partners were responsible for assessing how modi-

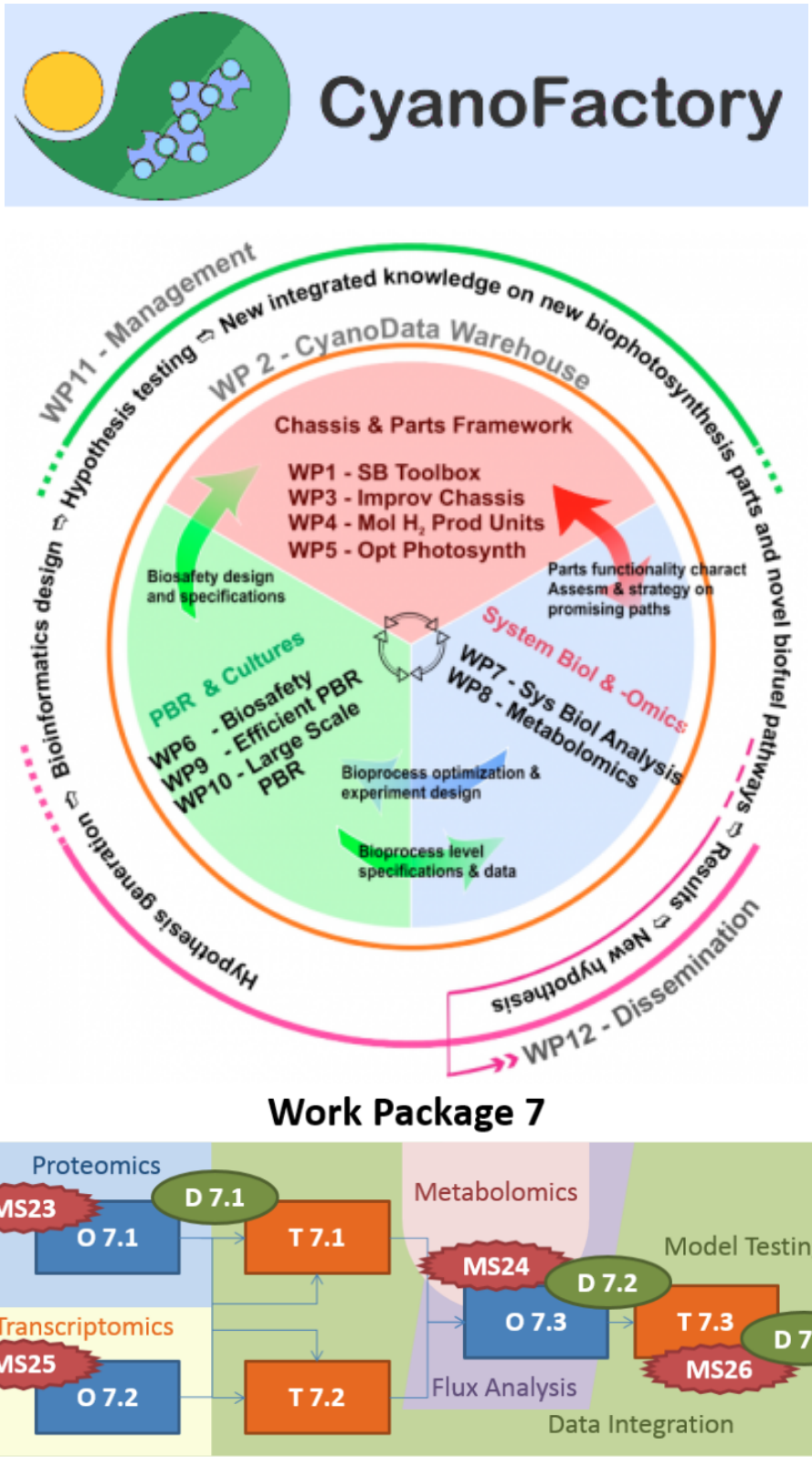


Figure 1.3: Within CyanoFactory, work conducted at Sheffield made up work package 7. This work package integrated a variety of different 'omics approaches for understanding the systems-level changes occurring within the organism. The three deliverable reports – highlighted as green ovals labelled D7.1, D7.2 and D7.3 – made up the core returns throughout the project and are available as an appendix to this thesis.

fications to the genetics and reactor environment affect the molecular state of the cells. This was done through a combination of different 'omics techniques, initially proteomics, followed by transcriptomics and metabolomics. Data generated in this manner was passed on to other partners as direct feedback, but also included in biological models for understanding the features of the system. In addition to this, the Sheffield partners worked on a number of developments to improve the quality of the data being fed back to the group. As a final action, a general summary of all compiled data was prepared to highlight issues arising from the metabolic level and to understand the remaining targets for converting the complete body of work into a series of practical goals. The experimental work in the thesis is driven by CyanoFactory investigations, with the final chapter containing the summary of key recommendations for the progression of this technology.

All the different 'omics approaches described here were conducted at Sheffield during CyanoFactory, combining transcriptomics and transient C^{13} labelled metabolic flux into the investigations. As these investigations were spread across a number of different applications – investigating the effects of salinity on the organism, and how the internal metabolic fluxes were affected in a mutant with altered light-harvesting capabilities – combining all these together made the thesis too broad-reaching. The proteomic improvements were chosen as the main focus, however the other investigations are alluded to within the thesis. If the reader is interested, the project deliverable reports are included as an appendix to this report, giving a summary of the overall assessments that were conducted throughout the CyanoFactory project.

1.2 Scale-up and large scale processing

In the previous section, there was a large amount of discussion regarding the features of photosynthetic microbes and how they were suitable for biofuel production, but there were very few examples of taking that photosynthetic output and scaling it to a larger size (Bosma et al., 2014). Taking any process and making it large scale is non-trivial, however it is important to explain some key terms here that are referred to later; and also to explicitly state some of the major pitfalls associated with the scaling up of biological processes. The following sections use ethanol as an example, since it provides a clearer example for describing a biological process being described in a chemical engineering manner. The pitfalls for scaling bio-hydrogen production have been discussed previously in section 1.1.5.

1.2.1 Biotechnology markets and scale

Biology has played a major role in human industry for centuries. The majority of the industrialised biological processes are fermentative, where sugar is converted to acids, alcohols and other organic components in the absence of air. These fermentative processes were the earliest examples of biotechnology, and were mainly used in the production of foods and beverages, like cheese, yoghurt and wine. The same processes are still used today, albeit on a much larger scale (Liese et al., 2006). Fermentation is still the most widely used biotechnological production method for anything that needs to be produced in a substantial quantity, ranging from antibiotics (Strohl, 1997), to vitamins (Leeper, 2000), to speciality chemicals that are difficult to synthesise chemically, such as isoprenoids (Schreiber, 2000). Isoprenoids are a family of high-value molecules synthesised from isoprene, and are the basis for a wide variety of aromatic and flavour enhancing molecules, as well as a number of human drugs, such as the anti-cancer agent lupeol.

To give an idea of the scale of production: in 2005, fermentative processes produced 26,000,000 tons of ethanol, 1,000,000 tons of both MSG and citric acid, 80,000 tons of vitamin C and 35,000 tons of antibiotics (Gavrilescu and Chisti, 2005). The antibiotics market alone exceeded US$30 billion, with the total pharmaceutical market being well in excess of US$400 billion, and all of these processes rely on micro-organism production (Gavrilescu and Chisti, 2005). Over the past 40 years the biotech market has grown dramatically, with concerns about limitations on products previously derived from fossil fuels pushing biological replacements back into the production chain. This is especially true for speciality and life science chemicals, where harmful chemical processes can be made cheaper and more efficient by introducing a biological agent into the process (Gavrilescu and Chisti, 2005).

1.2.2 Basic principles of chemical engineering from a biological perspective

In many of the cases listed above, it is convenient to think of a biological organism as a type of catalyst – a substance added to a reaction to reduce the energy input needed to carry out a chemical reaction, with a specific outcome. This broad-stroke overview of a biological organism is a useful analogy when considering a process at scale, despite the actual organism being far more complicated in reality than providing a simple catalytic function. This enables the underlying chemical process taking place to be evaluated simply, for example, *Saccharomyces cerevisiae*, or brewers yeast is well known for the ability to convert sugar into ethanol and carbon dioxide, resulting in a widely appreciated carbonated beverage. This process can be summarised as:

$$\text{glucose} \xrightarrow{Saccharomyces} \text{ethanol} + \text{carbon dioxide}$$

or

$$(1)\ \mathrm{C_6H_{12}O_6} \xrightarrow{Saccharomyces} 2\,\mathrm{CH_3CH_2OH} + 2\,\mathrm{CO_2}$$

This balanced chemical observation shown here is referred to as stoichiometry, where all the inputs or reactants are balanced on a molar level to all the outputs or products. This technique is useful for calculating the concentration of the final products, so in the above example each mole of glucose would produce 2 moles of ethanol and 2 moles of carbon dioxide. This gives an overview of a system, without delving into the specifics.

Whilst fermentative processes, like equation (1) are widely used in industrial production, aerobic or respirative (in the presence of oxygen) processes have application for cases where biomass accumulation is the major aim. One such case is in sewage treatment, where the principle aim is removal of nitrate, phosphate and organic molecules from a wastewater supply, to protect downstream waterways from eutrophication. Eutrophication is a process that occurs when high levels of nutrients, such as nitrate or phosphate, are deposited into a waterway, resulting an a 'bloom' of algae. When the nutrient spike is exhausted, the algae usually die back, providing conditions for rapid growth of bacteria. This rapid growth removes the dissolved oxygen, which results in death of other aquatic life in the waterway (Schindler, 2006). Within wastewater treatment, the sewage is aerated to encourage bacterial growth in a controlled environment; this depletes the nutrients and provides biomass, which can then be harvested and converted into useful products or energy for the system. The amount of oxygen consumed during this process is referred to as biological oxygen demand (BOD), which is a common feature used to assess sewage. When describing this system, the mass of the organism, or the 'biomass', can be considered as a product rather than a catalyst

$$\text{wastewater} + \mathrm{O_2} \xrightarrow{bacteria} \text{biomass} + \text{treated wastewater}$$

As no matter has been created or destroyed during either the fermentative or water treatment process (ie. neither involved the use of a nuclear reaction), all the mass that is put into the process must either be retained within the system or released from the system – in accordance with the first law of thermodynamics (conservation of energy). This can be described by the simple equation:

$$\text{reactants - products} = \text{mass change in the system}$$

This forms the basis of a chemical engineering principle referred to as (unsteady state) mass balance, or material balance (Himmelblau and Riggs, 2012). This can be either positive, where the system is gaining mass; negative, where the system is losing mass; or 0, where the mass of the system doesn't change. From a bioreactor point of view, a positive value indicates growth of the biomass, whilst a negative value indicates 'wash out'

and 0 indicates a state of continuous culture. The wastewater example listed above could be re-written from an operational view-point, where the reaction it monitors is a large holding tank for growing biomass with an aeration nozzle, an in-flow for waste water and an out-flow for treated water. In this case, the inputs are air (or oxygen) and wastewater, and the outputs are treated/depleted wastewater. Assuming the biomass remains within the holding tank, the mass of the reactants will be more than the products, and so the system gains mass during operation. As the system has a finite volume, this process is unsustainable unless a biomass 'harvesting' step is added into the reaction.

1.3 Key biological principles

1.3.1 A simple model of a cell

Previously, we considered the cell to be both a catalyst for producing an output, and an output itself in the form of biomass. To resolve this dichotomy, we need a new model for understanding a biological cell. Instead of considering it as a catalyst for a single reaction, it is more accurate to consider a cell as a partially-permeable living 'bag' containing a large number of individual catalysts and intermediates. Selected reactants can freely pass into the bag, and selected products can pass out, although this is controlled to prevent the bag from either over-filling and bursting, or draining completely and losing all contents. One final additional point is needed for the bag model, not only does it have a wide array of catalysts and intermediates contained within it, but the number and properties of these catalysts changes dynamically in response to the environment as well. This generally gives a hint of the complexity involved in predictably engineering a biological organism.

At this point, before becoming overwhelmed by the complexity of the model, it is worth tearing the bag open and re-examining our previous fermentation reaction in the light of this discovery. In reaction (1), there are actually a number of different steps taking place. Expansion of this process is important for understanding the concepts of 'intermediates' and 'side reactions'. The actual process goes through 3 individual catalytic 'stages': glycolysis, which converts glucose into pyruvate; pyruvate decarboxylase, which converts pyruvate into acetaldehyde and carbon dioxide; and finally alcohol dehydrogenase, which converts acetaldehyde into ethanol (Berg et al., 2006). Glycoloysis is one of the most ancient and highly conserved pathways observed in nature (Romano and Conway, 1996), and is thought to have occurred originally in oceans in archean times (Keller et al., 2014).

$$\text{glucose} \xrightarrow{glycolysis} \text{pyruvate}$$

can be considered as two separate reactions:

$$\text{(2) } C_6H_{12}O_6 + 2\,NAD^+ \xrightarrow{glycolysis} 2\,CH_3COCOO^- + 4\,H^+ + 2\,NADH$$

and

$$\text{(3) } 2\,ADP + 2\,P_i \xrightarrow{glycolysis} 2\,ATP + 2\,H_2O$$

This reaction is slightly more complicated when considered chemically. The reason for this is that whilst glycolysis has been summarised into a single step here, it is actually made up of 10 separate catalysed reactions, each with intermediates. It is important to note here that this reaction contains a 'side reaction', denoted as (3). The energy made available from the exothermic (energy liberating) reaction of splitting the glucose molecule in reaction (2) is used to power the synthesis of adenosine triphosphate (ATP) from adenosine diphosphate (ADP) and inorganic phosphate (P_i) – this reaction is referred to as a condensation reaction, as it results in the production of a water molecule; such reactions are usually stable and do not usually spontaneously reverse. ATP is used in a wide range of biological reactions and will be discussed in more detail later.

In addition, the conversion of glucose to pyruvate is an oxidation reaction – it liberates electrons. For this reaction to move forward, another agent needs to be present to drive this – in this case the electrophile nicotinamide adenine dinucleotide (NAD^+) acts as an oxidising agent. During this process, NAD^+ consumes 2 electrons ($2\,e^-$) and 1 proton (H^+) from the reaction and is converted to its reduced form (NADH). This reaction is reversible, and so when NADH is converted to NAD^+ the electrons are liberated again. NAD^+, and its phosphorylated cousin NADP, play an important role as electron carriers for oxidation and reduction reactions within the cell, and are extremely important in hydrogen synthesis. In this reaction there are also 4 protons (H^+) produced which were not present in reaction (1). These are intermediates, products that are created in one stage of a process but then consumed in another before the entire reaction is complete. Intermediates are ubiquitous in biology, and while they can often be bundled into a process and ignored to keep the general system simple, they frequently contribute in additional side reactions of their own and should be considered when the system performs in an unexpected manner.

$$\text{pyruvate} \xrightarrow{pyruvate-decarboxylase} \text{acetaldehyde} + \text{carbon dioxide}$$

$$\text{(4) } 2\,CH_3COCOO^- + 2\,H^+ \xrightarrow{pyruvate-decarboxylase} 2\,CH_3CHO + CO_2$$

For clarity, the reactants and products from (3) have been excluded, as have any products from (2) that do not participate at this stage. Two of the protons (H^+) generated as a product in (2) have been used as a reactant in (4), and so have been stoichiometrically 'neutralised'. This balance between reactants and products results in no net change in the global reaction, and is the reason they were not visible in (1).

$$\text{acetaldehyde} \xrightarrow{alcohol-dehydrogenase} \text{ethanol}$$

$$(5)\ 2\,CH_3CHO + 2\,NADH + 2\,H^+ \xrightarrow{alcohol-dehydrogenase} 2\,CH_3CH_2OH + 2\,NAD^+$$

Finally, in (5) the NADH and NAD^+ occur on the opposide sides of the reaction compared with (2). As with the protons in the previous stage, there is no net global change in the reaction and so they are also intermediates, which can be excluded in (1).

So to summarise, the intermediates from the catalytic bag cell model are the outputs of individual reactions, driven by the different catalysts present within the bag. Due to the close proximity of all the bag contents, it is a good general rule of thumb to assume that any intermediate produced within the bag will be immediately exposed to all other catalysts present, which can result in side reactions. A side reaction is displayed in reaction (3), in this case the production of ATP from ADP and P_i. In this case, energy is taken from the reaction and stored by the cell, but material products from a reaction can become participants in side reactions as well. Side reactions can cause branching-chain effects, and as a result, can make producing a desired product, or understanding a reaction pathway, challenging within the complexity of a cell background, especially when the catalytic properties can change in response to the environment!

1.3.2 Proteins

In the semi-permeable catalytic bag cell model, the most important aspect is clearly the catalysts, since they drive all the reactions taking place within the bag. Within a real cell system, these catalysts are called proteins. Proteins play a wide role of functions within a cell, ranging from structural roles, to catalytic functions, to storage and controlling the flow of materials into, around and out of the cell.

At its most basic level, a protein is a polymer of individual building blocks called amino acids. Amino acids are a class of chemical molecules, which vary greatly in their individual chemical properties, but are all composed of an amino ($-NH_2$) and carboxylic acid ($-COOH$) functional group – hence the name. The general chemical structure of an amino acid is $H_2N-CHR'-COOH$, where the R' group can be one of 21 commonly occurring side chains. The carbon that the R' group falls on is referred to as the α carbon, and this is also a chiral centre. In living organisms, all amino acids are L-form enantiomers; this is an important feature for the structural properties of proteins. There are other, rarer side chains, however these will not be discussed here. A complete table of the most commonly used amino acids is given in figure 1.4 (p. 33). These amino acids all have a single letter code, as shown in figure 1.4, which is used as short-hand when describing the sequence of amino acids in a protein.

During protein synthesis, the the amino acid units (monomers) are built into a chain

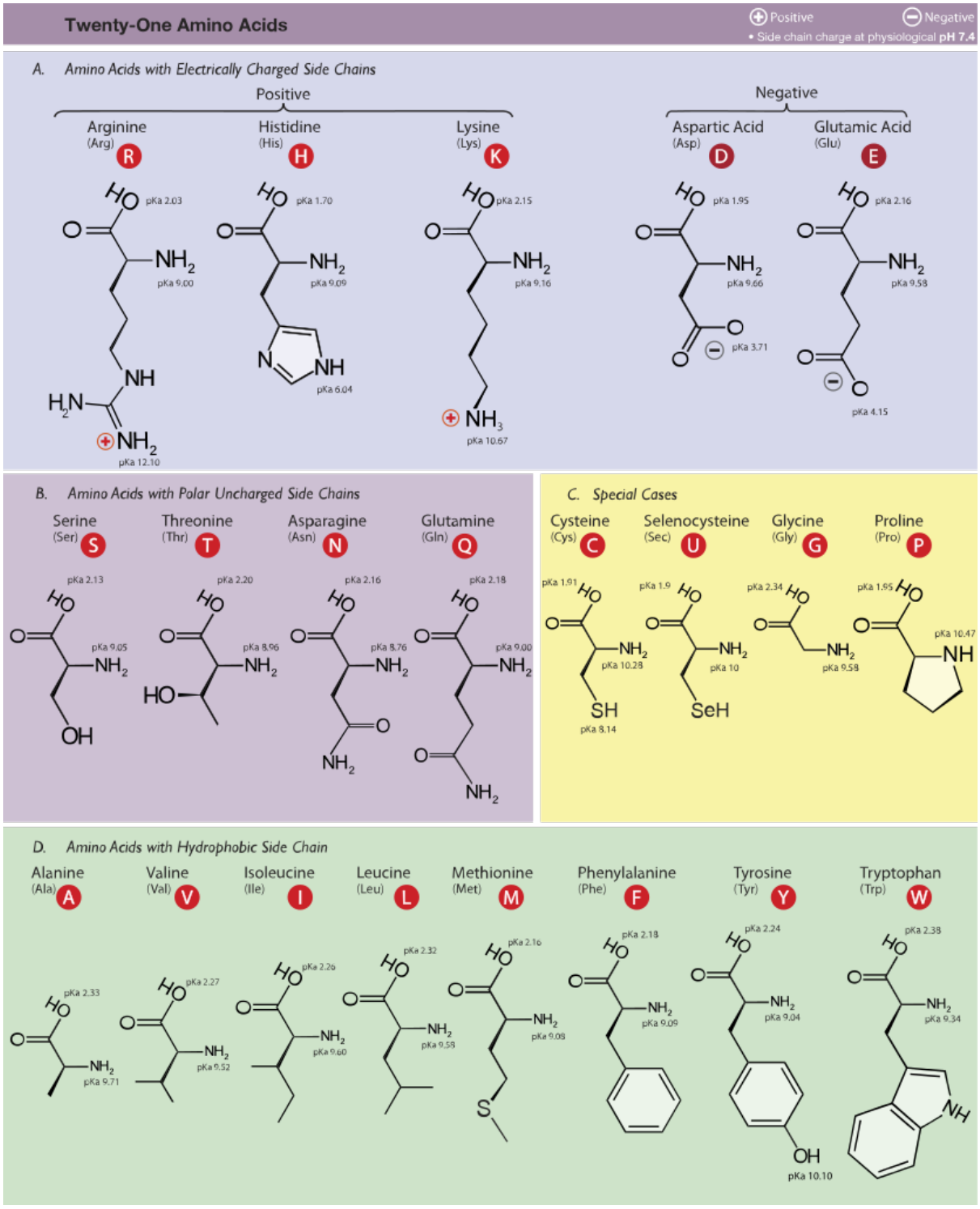


Figure 1.4: A graphical representation of the 21 most frequently occurring amino acids, grouped by their general features. All amino acids presented here are shown with charge states based on pKa values at physiological pH (7.4). This figure was produced by (Cojocari, 2016), and is freely available for reuse under the creative commons licence, via Wikimedia Commons

(polymer) sequentially through a series of amide (or peptide) bonds; this is where the carboxylic acid group reacts with the amino group in a condensation reaction, to form a covalent linkage between the two monomers. As the polymer grows, the protein has a regular repeating chain referred to as the protein 'backbone', with the R groups spanning out from it. Due to physical space requirements (also referred to as steric hindrance), these groups usually arrange themselves on alternating sides of the backbone in the absence of external forces.

Chirality is a description of the ordering of the atoms or groups within the molecule around a central point in 3 dimensions, which is typically a carbon atom as it readily forms 4 bonds with tetrahedral conformation. What this means is that around each of these 'chiral carbon' centres, there are 2 distinct conformations the surrounding atoms can take – and in one arrangement there is no amount of rotation that can be performed to reach the other arrangement. In order for a carbon to be chiral, it must be surrounded by 4 different atoms or groups of atoms. These two forms are called enantiometers, and are referred to as l and d – signifying the left- or right- handedness of the direction rotational of atomic ordering ordering.

All amino acids have uniform l-chirality, with the exception of glycine, which lacks a chiral carbon as the R′ group is H. As a result, the order in which the subunits are joined has an effect on the structural properties of the protein. When joining 2 amino acids, $A + H \neq H + A$; as a result, a protein sequence is always read beginning from the amino side and ending at the acid side by convention. It should be noted that it is possible to form palendromic sequences that ignore this directionality (the peptide 'RACECAR', for example), and whilst these are generally rare, they can cause problems with proteomic data processing in some cases (discussed later). Polymers of amino acid residues are referred to as peptides when they are 'short' and proteins when they are longer, although the cut-off boundary for this is fairly arbitrary. In this thesis, a peptide will refer to either any polypeptide molecule that is less than 4000 da (generally <36 amino acids long on average), or any protein that has had its backbone severed.

Shapes and structures are vitally important in biology, and are the driving force for specificity of reactions and the catalytic activity of the proteins. Protein structure is controlled at 4 levels to produce the final functional structural conformations. These are referred to primary, secondary, tertiary and quaternary structure. The precise ordering and type of amino acid residues that makes up the peptide backbone is referred to as the 'primary structure'. This primary structure is fundamentally important for structural properties of the protein, as it controls the order of the different R′ groups and is generally fixed at the point when the protein is created. A major exception to this is the cysteine. When 2 of these sulphur-containing amino acid residues are spatially close together, a spontaneous oxidation reaction occurs, resulting in a covalent bond forming between the

two. This reaction can either happen between 2 cysteine residues in the same protein, forming a loop in the primary protein structure, or between 2 separate protein molecules, forming a quaternary covalent bridge between different proteins. It is also worth noting that the amino acid proline has a significant effect on the structure of proteins and peptides. Proline is the only one of the 21 common amino acids that contains a secondary amine instead of a primary amine – this occurs due to the R′ group reacting with the amine, and so the regular repeating structure of the peptide backbone is kinked at an unusual angle due to steric hindrance. This kink is also pronounced in short peptides that would normally otherwise be linear.

'Secondary' protein structure consists of local structures called α-helices and β-sheets. These regular repeating structures are stabilised by hydrogen bonding from the peptide bonds in the backbone, and are observed repeatedly across all proteins. Glycine and proline residues are referred to as 'helix breaking', and can disrupt these structures; as proline has a secondary amine it is uncharged and therefore doesn't contribute to the hydrogen bonding motif, whilst in the case of glycine as the R′ group is simply a hydrogen atom, the residue is non-chiral and therefore too unconstrained to contribute to a regular repeating pattern. 'Tertiary' protein structure refers to the global packing of the whole protein. This structure is stabilised by a range of different forces:

- Hydrophobic interactions – a form of packing that takes place where water is present in the environment. This is partially secured by van der waals forces, although is largely an expression of hydrogen bonding effects of water excluding hydrophobic residues. These forces are supported by hydrophobic amino acids such as tryptophan.

- Hydrogen bonding – an induced dipole effect that is responsible for the interactions between water and other charged species. Hydrogen bonds are responsible for stabilising the majority of biological interactions and can form on any amino acid residue that is not hydrophobic.

- Salt bridges – a permanent dipole effect. A salt bridge forms between 2 residues with opposite charge signs, such as lysine and aspartic acid. They are affected by the surrounding pH, as raising or lowering the pH beyond a pKa value can determine if an R′ group is charged or not.

- Disulphide bonds – as described above, a covalent bond that forms spontaneously in oxidising conditions. These bonds can be reversed through exposure to a reducing environment, however unless the cysteine residues responsible are chemically blocked – typically through alkalation – then the bonds will spontaneously reform when exposed to an oxidising environment again.

- Steric hindrance – The forces associated with stopping two atoms occupying the same physical space. Also referred to as Pauli exclusion, this occurs when the negatively charged electron clouds of two atoms attempt to occupy the same physical space.

These features are a direct result of the different amino acid residues that make up the protein and how they pack together. Whilst this process is currently too complex to accurately predict a protein structure from its primary sequence alone, two proteins with the same primary sequence will fold to produce exactly the same terminal structure, and so controlling the primary sequence of the protein gives control over the final 3 dimensional structure. Finally, 'quaternary' structure is stabilised by all the same features as tertiary structure, however it refers to interactions between 2 or more distinct proteins. These proteins can be identical subunits, or two distinct proteins – the important part of this interaction is that two separate primary protein chains are interacting together to produce a functional final product.

Proteins can also be modified after their creation, in a process called post translational modification (PTM). The most common PTM is the addition of a phosphate (P_i, or a mix of $HPO_4^{2-} + H_2PO_4^{1-}$) onto an alcohol containing residue, such as serine, threonine or tyrosine. This addition is performed by proteins known as kinases, and is reversed by proteins known as phosphorylases. This can have a number of effects on a protein, firstly by adding a large charged molecule to the residue it becomes charged, enabling it to participate in hydrogen bonding and salt bridge formation. Adding phosphate can disrupt entire hydrophobic sections of a protein and result in dramatic re-shuffling of the protein structure. As a result, phosphorylation is the most frequently method for feedback and cascade control in living organisms, and so phosphorus levels are typically a limiting factor for biological growth. Beyond the charge aspect, the modification is relatively large, particularly for a residue like serine, and so addition of phosphate can physically change the shape of a protein by generating a steric hindrance effect. Both of these features can occur in differing amounts, but can enabling participation, or exclusion, of the modified protein in quaternary interactions by attaining a different shape. Dozens of other known PTMs exist, include glycosylation, namely the addition of sugars onto proteins; ubiquitination, which can attach smaller proteins to a protein to signal for its destruction; and alkylation, where hydrophobic hydrocarbons are added to charged residues, blocking their activity (such as cysteine forming disulphide bridges) or burying them within the hydrophobic core of the protein.

The same way that the structures of proteins are stabilised by these different forces, the arrangement of charges and flexibility of protein molecules are also what provide their catalytic properties. For example, some proteins have 'pockets' that certain shaped molecules fit into but others are excluded from, and buried within these protein pockets

are arrangements of charges that, due to close proximity to the molecules held within the pockets, can have strong catalytic activity. A good example of a protein like this is the protease trypsin. Trypsin has pockets that only long-chain, positively charged residues can fit into; but when one binds to the pocket it triggers a catalytic reaction that results in cleavage of a peptide bond. The resulting effect is that trypsin will selectively cleave proteins at regions where arginine and lysine are present, but not elsewhere. This is referred to in proteomics as site-specific cleavage. Cleavage is a very important topic in proteomics, and will be discussed at length in a later section in this chapter.

Whilst there is a fairly good understanding of how different properties contribute to the structure; even if the structures of all proteins were known it is still impossible to know exactly what function a protein will have from either its sequence or structure alone. The most effective method for measuring what a protein does involves measuring that protein directly – referred to here as a *de novo* investigation. This can be done through purification of the protein and direct analysis; although while some proteins can be removed from the cellular environment and still retain their function (*in vitro* analysis - in glass), in other cases proteins produce different functions when removed from a living environment (*in vivo* analysis - in life).

In practice, beyond *de novo* investigation into the function of the protein, the accumulated knowledge of pre-existing proteins of a similar sequence is used to infer what newly discovered proteins do. Whilst this practice is not perfect, as similar protein sequences does not necessarily mean similar protein function, it provides a lot more utility to proteomic data generated in organisms that haven't been studied as extensively.

1.3.3 Nucleic acids

In the previous section, the importance of the primary sequence of proteins, and how it affects their catalytic activity – for the sake of the semi-permeable catalytic bag model – has been described. What hasn't been described so far is how that primary sequence is determined. To understand this requires an investigation of the main features of the 'genetic material' or 'nucleic acids'. It is worth briefly noting that the term 'nucleic acid' refers to 2 different acidic polymers found within a cell – deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). These are fundamental to the variability of the catalytic properties of the cell, and will be explained in more detail in this section. The polymers are referred to as 'nucleic acids' as they were originally extracted from a central cell structure called the nucleus; however whilst not all cells have a nucleus, all living things and even viruses contain nucleic acids.

In a living organism, the genetic material or DNA carries all the information for a cell

to produce proteins, which in turn facilitate life. DNA is a double-stranded, α-helical polymer made up of four unique base molecules – adenine (A), cytosine (C), guanine (G) and thymine (T) that are joined in a specific sequence. Like protein, DNA has a regular repeating backbone structure, where the bases are joined by negatively charged phospho-diester bonds, and the bases themselves are planar hydrophobic disks that stack on top of each other. The double-helix is an anti-parallel dimer, where one strand is the inverse of the other – in this case, where an A is present on the first strand, it will be hydrogen-bonded to a T on the other and vice-versa. The same is also true for C and G. The structure was first identified famously by Watson and Crick in 1953, work that lead to them being awarded the Nobel Prize in Physiology or Medicine along with Maurice Wilkins in 1962 (Watson and Crick, 1953). The sequence of all the DNA in a cell is referred to as the genome sequence. Within the genome sequence, the specific sequence – as with proteins – is vitally important.

From a conceptual point of view, DNA can be considered as the blueprints for the catalytic bag. The instructions for forming the complete list of every protein that can possibly be produced, under any condition, is found within the genome. The blueprints for an individual protein are referred to as a 'gene', and the collection of all protein-coding sequences within an organism is referred to as the 'genome'. As mentioned above, however, the exact numbers and even the specific presence of any of the catalytic components is variable. Not all of the potential catalysts that the genome has blueprints for will be produced in the same quantities, in fact not even all the potential catalysts that can be produced will be under all conditions. This can be considered using a facile analogy of a factory producing cars – there will be a blueprint for a wheel, an axle, a door; however a standard car will require 4 wheels, 2 axles, between 3 and 5 doors, etc. depending on the model. In a cell, whilst the genome exists as a central repository for all the proteins that can be produced, a second nucleic acid, RNA, is responsible for actually converting those blueprints into reality.

Before protein is produced, selected parts of the DNA sequence must be 'transcribed' into a substance called mRNA (mesenger RNA) by a protein called RNA polymerase. mRNA is a subset of RNA (ribonucleic acid) exclusively used for the transfer of information from DNA into protein (Brenner et al., 1961). RNA is a single-stranded polymer molecule with a highly variant structure depending on the sequence. Whilst RNA also has a wide variety of functions, including structural features and regulatory controls, these are diverse and fall outwith the scope of this review. The DNA bases are transcribed into equivalently named RNA bases – A, C and G; with the exception of thymine, which transcribes to uracil (U), carrying the sequence information from the DNA forward. This sequence of bases are read by a molecule called a ribosome, which translates the sequence into a protein. As described above, proteins act together to produce functions within a cell

based on the genome sequence, and so the movement of information flows from DNA to RNA to protein. This is referred to as the central dogma of molecular biology (Crick, 1970).

This direct link between the genome and the catalytic capabilities of a cell has important consequences for engineering. Conceptually, by removing a gene from the genome, the capability for a cell to produce the protein it coded for has been removed as well. The inverse is also true, it is possible to add a gene or collection of genes to a genome and add functionality. This forms the fundamental principles of genetic engineering. This first methods for targeted genetic modification of an organism were developed by Cohen and Boyer in 1973 (Cohen et al., 1973), and 2 years later in 1975 the historic Asilomar conference on recombinant DNA molecules was called due to concerns arising over the engineering of life (Berg et al., 1975). Whilst finding a gene that relates to a specific protein can be done relatively easily, the principles behind rationally engineering a system are more complicated.

Production of mRNA is controlled by a complex series of feedback reactions, which enable the production of proteins in response to the specific requirements at the time. One of the major difficulties in understanding this is that a snap-shot of cellular response is heavily influenced by the previous state the cell was in, which can make understanding and engineering cells difficult. A modern cell cannot function without a comprehensive collection of all of the materials mentioned above in a prior state, which has been provided in the form of an unbroken chain of living organisms since the first origins of life around 4 billion years ago (Haldane, 1929).

When physically 'translating' mRNA into a protein sequence, there is a clear issue of disparity between the number of nucleic acid bases (4) and the number of commonly used amino acids (21). To get around this issue the bases of mRNA are read in groups of 3 bases at a time, these triplets of bases are also known as 'codons'. There are unique codons for each amino acid, as well as specific codons indicating the beginning and end of a protein coding sequence (start and stop codons, respectively). As there are 64 possible codons (3 bases with 4 possible states per base, $4^3 = 64$) and only 23 codon 'states'; with 21 amino acids plus start and stop codons, the genetic code is degenerate (Crick, 1968). This degeneracy means that different codons code for the same amino acid. By measuring the occurence of codons within the translated genome a 'codon usage table' can be produced. This shows the rate of occurence of individual codons within specific organisms, which have been found to be biased towards certain codons in certain species; however comprehensive studies of this phenomenon only became possible in the post-genomic era (Duret, 2002).

The presence of an 'annotated genome' for an organism is vital for modern proteomics-

level investigations. This simply refers to genome sequence being known, and also being run through computational tools to identify putative genes and proteins.

1.3.4 Metabolites and membranes

The previous section described how the nucleic acids can be used to both monitor, and ultimately to control or engineer the catalyst content of the semi-permeable catalytic bag model. The conceptual origins for the information that makes up the catalysts (or proteins) has been given, but no clear path to their creation has been laid out. To understand the origins of proteins, we return to the glycolysis example from the beginning of this section. Pyruvate, which was an intermediate in the ethanol production pathway, is one of the building blocks for creating the amino acid alanine. The chemical molecules within the cell which are used to produce all other parts, are referred to as 'metabolites'.

Metabolite is a catch-all term that is given to cellular contents which are smaller than around 500 Da. This includes a variety of molecules with wide-ranging properties, and since the term is driven by size, rather than a function or process of creation as proteins and nucleic acids were, it is much more difficult to succinctly describe the role of metabolites within the cell. All intermediates are metabolites, all things crossing the outer cell membrane into the cell are metabolites, and all parts of the cell are built from metabolites during the cell cycle.

Clearly with such a diverse range of chemical molecules; maintaining a functioning cell requires either spacial separation of parts, or a series of strong kinetic chemical drivers, to ensure reactions go in the optimal route for life. In practice, both methods are used. Amphipathic metabolites called phospholipids, which have a charged phosphate group attached to a hydrophobic hydrocarbon chain, spontaneously self-assemble into membranes in aqueous environments – with the hydrophobic 'tails' aggregating together and the hydrophilic charged groups interacting with the water in the environment. Biological membranes are vital to life, and are interspersed with proteins that perform a range of functions, from structural roles, where they stabilise and strengthen the barriers, to creating gates and molecule-specific channels through them, to physically moving membrane-bound 'containers' around the cell.

Membranes are essential for life, and are used to drive a number of reactions that would otherwise be impossible. The ubiquitous example of this was mentioned above – the generation of ATP. ATP is a nucleotide-containing molecule that is fundamental to the creation of DNA within the cell (it makes up the A, of ACTG), but is also the fundamental energy carrier molecule used across all species. The molecule has 3 phosphate groups joined in tandem on one side, but in precisely controlled reactions, these phosphate

groups can be removed or transferred to other molecules (such as proteins, as mentioned above), releasing energy in the process.

Creation of ATP from ADP and P_i is a highly conserved and ubiquitous system of life, and is a good example for how cells utilise membrane separation to perform a function. The reaction is endothermic, meaning that it requires external energy to take place spontaneously under standard conditions. This energy in the cell is driven by a 'proton gradient'. A proton gradient requires a highly acidic reservoir, which spontaneously diffuses into a less acidic environment due to positive charge interactions between the atoms – this process is analogous to water being stored behind a dam against the force of gravity. A number of reactions within the cell release protons into the environment – one of these is glycolysis, which was mentioned previously. By spatially controlling where these reactions take place within the cell, energy can be expended to pump these output protons into a membrane against the electrical gradient. A small collection of specialised proteins are found sitting across these membranes called ATP-synthases. Collectively they form a carousel-like device, which works in a similar manner to a water wheel, and converts energy from the flow of protons in a controlled manner to generate rotational energy. The ATP-synthase has binding sites for both ADP and P_i, and as the molecule rotates it pushes the reactants together and excludes H_2O, shifting the reaction equilibrium to favour ATP production. ATP is then released from the protein and the reaction continues.

Membranes are also used to exclude molecules that might otherwise be damaging to the cellular constituents, and are used for a range of other reactions within the cell.

1.3.5 Integrating the system

Throughout the cell, there are a number of proteins which – when working in a concordant manner – take the fundamental molecules of life and convert them into self-replicating states. These states are commonly referred to as metabolic 'pathways' – or chemical routes by which a specific set of reactants can be converted into the products needed to facilitate life. One method to interpret this series of catalysed reactions is to order them into a matrix of stoichiometric equations, where relative amounts of reactants and intermediates are considered, rather the actual amounts.

These equations can then be assembled and simultaneously solved for a single, hypothetical equation of all the inputs required by a cell to undergo a single cellular division, or for a culture to produce a single mole of biomass (Saha et al., 2012). This method of analysis is called flux balance analysis (FBA) modelling. There are online repositories of these equations, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG), which

have been arranged into a general metabolic network. 'Omic level data can be overlaid on this to gain a general understanding of changing systems in terms of chemical outputs of a cell (an example of this is given is chapter 5).

As mentioned above, proteins are often assembled into 'pathways'. These often give the impression of linear flows or pipelines of materials around a cell, indeed this is a convenient model for understanding biological life, however it is not strictly true. Instead, the apparent flow of material around a cell is a naturally occurring phenomenon that is a combination of the following features:

- Close proximity of all catalysts and intermediates
- High reaction rates of catalytic conversions by proteins
- Better specificity for certain types of intermediates than others

This can be exemplified in the case of the proteolytic enzyme trypsin. It cleaves peptides with incredible accuracy at targeted locations in the primary chain, however if left in a reaction long after its preferred target is exhausted it begins to exhibit non-specific cleavage action. This effect is small to the point of being negligible whilst the substrate is present, however, generating a practical specific effect – especially within the time-frame of the presence of intermediates within a cell.

It may seem overly specific to consider this differentiation, since it doesn't appear to make much difference on a practical level, however this wide-reaching availability of intermediates to all catalysts and the presence of non-specific effects can be the causes of major impediments to rational genetic design of a construct. Likewise, they can be responsible for a leak of metabolites from one pathway into another in an unexpected manner. There are also literally billions of protein permutations, and slight modifications to the sequence or environment can entirely change their function.

Despite these limitations, there are a number of tools that can be used to predict the function of new proteins, based on their evolutionary relatedness to pre-existing measured proteins. The most commonly used tool for this purpose is the basic local alignment search tool (BLAST) (Altschul et al., 1997). This works by taking the 'target sequence', either from DNA or protein, and breaking it up into smaller sequences. The smaller sequences are then searched against a database of known protein or DNA sequences and aligned to specific points in the sequence. Initially matching sequences are then searched against again, with gradually extending lengths of the target sequence; this continues until only a series of sequences with strong similarity remain. This search method is robust to both individual base changes, as well as deletions and insertions into the sequence. There are more advanced algorithms that perform the search more rapidly, however there can be trade-offs in sensitivity for more distantly related sequences (Edgar, 2010).

When a new genome is sequenced, sections of DNA that appear to be gene-like – meaning they match a series of rules, such as beginning with a start codon and being a suitable length before encountering a stop codon – are identified in the genome. These are then searched using the BLAST-like tool to generate an informatic profile of the genes, where proteins that appear similar to pre-existing proteins are labelled as such and linked together. If information is known about the related protein, it can be used for the protein about which nothing is known – although sites are generally careful to label purely informatic information as such. It is important to note that with a new organism without an evolutionarily similar biological model organism, these techniques are of limited value – leaving large sections of the database labelled as 'putative' or 'hypothetical' proteins. These profiles are stored in online databases, such as UniProt (www.uniprot.org), and will gradually be updated over time as novel informatic tools and proteomic information become available. The large databases regularly exchange and update information between themselves, which helps maintain a good general standard of knowledge worldwide; however some industrially relevant or proprietory organisms – such as Chinese Hamster Ovary cells, have information held in databases that are not publicly available.

On method of assessing protein functions are Gene Ontology terms (GO terms) (Ashburner et al., 2000). GO terms are a notation system that describes a series of key functions the protein has been found to be related to. GO terms refer specifically to functions, rather than a specific species or protein, and so whilst there would be a term for oxidoreductase activity, there is not for a protein that triggers it, such as cytochrome C. GO terms are grouped into three major categories:

- cellular component, which dictates structural properties or localisation within the cell – this can include details such as 'membrane protein'
- molecular function, indicating how the protein physically operates – for example an Fe binding protein
- biological process, which describes any metabolic pathways the protein is associated with – such as glycolysis

As a result of this, a single protein can have multiple GO terms, and a single GO term can be attributed to many proteins within an organism. The documentation on the gene ontology website is useful for getting a better understanding of the full scope of the system (http://www.geneontology.org/page/documentation). It is important to note that like all current informatic conventions, GO terms are not currently a perfect system. The terms are fallible, as there is no way to differentiate researcher inputed terms from automatically assigned terms. In addition, not all functions have been identified, and so some terms may be missing from proteins where the association has not yet been made. Over time this will improve, as more information enters the system, however it is

important to be aware of potential errors or bias in function analysis on the 'omic level.

1.3.6 Investigating the biological system

Informatic *in silico* and *in vitro* techniques for assessing individual proteins in isolation have been described, but these are not as effective as a direct investigation into the effect of the protein (or gene) directly inside the living organism. These studies are referred to as *in vivo*. A common practice in genetics is to remove the genes that code for a protein from the cell, a technique referred to as creating a 'knock out' (KO or δ) mutant, and make observations on how its functionality changes. If the cell is no longer viable – ie. if it dies, then the gene is considered to be 'essential' for basic cellular survival. If the cell no longer produces a particular function then the gene can be linked to that function – for example removal of the alcohol dehydrogenase gene from yeast removes the capacity for yeast to make alcohol, then we could infer that the protein is related to ethanol production from glucose.

The typical final test for determining the function of a protein is to add either the removed gene back into the mutant, or to replace it with a well-understood gene from a different organism expected to perform the same function to the organism, and see if original functionality is restored. If function is restored then the gene is classified as performing the same function. If this does not restore the original state of the cell (or phenotype) then the genetic modification either caused off-target effects elsewhere in the genome, the attempted restoration failed, or the alternative gene for a different organism either did not perform all of the same functions as the gene that was removed, or produces other off-target effects.

The inverse can also be done, a gene that has been found to perform a specific function in an external organism can be added to the genome of a target organism to see how the outputs are affected. This is generally expected to produce a novel function – and is the fundamental principal in genetic modification of organisms used in biotechnology. This technique is used when an organism has properties that are valuable, but it grows poorly in an industrial setting; and so the functional parts of the organism are translated to an organism that is easier to grow on scale. An example of this was mentioned previously, where Amyris transferred the genes from the artemisinin pathway from *Artemisia annua* into a microbial host. Again, if the gene has been shown to responsible a specific phenotype and fails to do so after being moved to the new organism, it can fail for the same reasons mentioned above.

A genome can be altered by putting a piece of DNA into the cell, and then utilising the pre-existing mechanisms for repairing DNA damage or re-ordering DNA. Through

careful design, this can be used to either remove sections from the genome or add in external genes. The initial limiting factor to integrating DNA into the genome of an organism is physically bypassing the outer membranes. Some cells will naturally take up DNA from their environment – these are referred to as naturally competent cells. Other methods of penetrating the outer cell membrane include temporarily making them porous chemically, introducing breaks in them with sonic energy, physically forcing DNA through the membrane with kinetic energy, altering the membrane polarity with a technique called electroporation.

During the CyanoFactory, a number of knock-out and gene insertion studies were performed to understand the most effective way to produce H_2 in *Synechocystis*. The next section will focus on more specific details of cyanobacteria, hydrogenases, and specifically the organism *Synechocystis*.

1.4 *Synechocystis* and Hydrogen Production

1.4.1 Summary of the organism

Cyanobacteria, also referred to as 'blue-green algae', are a large group of widely diverse unicellular micro-organisms. They are believed to have been the first group to develop oxygenic photosynthesis more than 2 billion years ago, and are responsible for the aerobic conditions in the atmosphere on earth that support life as we know it (Olson, 2006). Species are found in both salt water and freshwater, with a number of species being halotolerant. Cyanobacteria tend to be intolerant to high light intensities, and some have been reported to change their buoyancy in response to different light regimes – notably through the use of structures called gas vacuoles (Walsby, 1972). Cyanobacteria are most well-known for their roles for over-taking algal growth during algal blooms of eutrified water, where the cells take advantage of reduced light levels to thrive (Mur et al., 1977). Several of the well-known species, such as *microcystis* and *trichodesmium* are more widely known because of their roles in harmful algal blooms, where they produce an array of toxins, which can have a detrimental economic effect on domesticated wildlife, such as cattle (Beasley et al., 1989). There area also a number of beneficial products that can be engineered (Ducat et al., 2011)

Synechocystis became a model for cyanobacterial species through serendipity to a certain extent. It is naturally transformable, although the transformation process can be accelerated through electroporation or sonication transformation methods(Tran et al., 2009); and it was the first cyanobacterial species to be sequenced - the third fully sequenced organism in history(Kaneko et al., 1996). It bears a striking resemblance to the chloro-

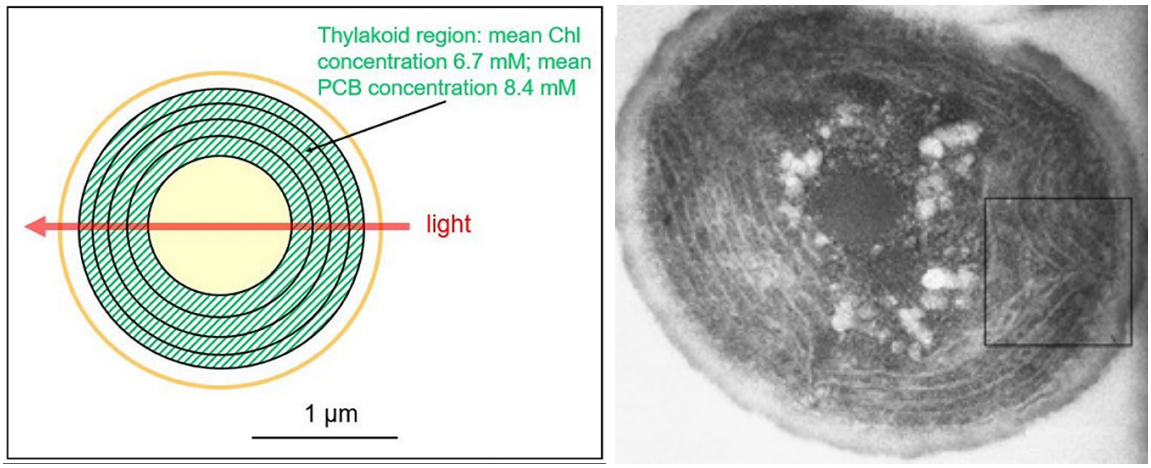


Figure 1.5: Left: A simplified diagram showing the internal membrane structure within *Synechocystis*, where moving in from the outside, the outer yellow circle is the outer membrane, the white circle is the periplasm, and the shaded region is the thylakoid membrane system. An arrow indicates the passage of light through the organism, image adapted from (Schuergers et al., 2016). Right: an electron micrograph of a *Synechocystis* cell. The thylakoid membrane structures can be clearly seen, image adapted from (Nickelsen et al., 2011).

plast and was believed to be an ancestral precursor to the structures found in many economically important C4 higher plants.

Synechocystis is a unicellular photosynthetic cyanobacteria that can live both autotrophically, generating energy from light by photosynthesis to survive, and heterotrophically by digesting glucose. Division occurs approximately once every 20 hours, although this can vary depending on conditions. Growth is slightly accelerated under heterotrophic conditions, to around 14 hours per division, and doesn't occur at all during dark phase in autotrophic conditions. The cells are spherical, approximately 3-4 um long and contain no gas vacuoles – a feature which enables them to be differentiated from physically similar, but bloom-forming species such as *microcystis* (Walsby, 1981). The general structure of the cells can be seen in figure 1.5 (p. 46) The cells consist of approximately 70 - 110 mg/g DW fatty acids, with over 70% of those fatty acids being palmitic acid (Tran et al., 2009). By mass, the largest cellular component is protein, which makes up approximately 700 mg/g DW (Touloupakis et al., 2016). The species contains a series of layered membranes, called thylakoids, which are embedded with photosynthetic machinery that generate energy enabling the cell to live solely on light energy – a state known as photoautotrophic growth (Heidorn et al., 2011).

It is the model organism for cyanobacterial research as a result of its fully sequenced and annotated genome, which was first completed in 1996 (Kaneko et al., 1996). The genome consists of a single large chromosome containing 3317 genes, and seven additional plasmids with a total of 408 additional genes, resulting in a total of 3725 genes and a

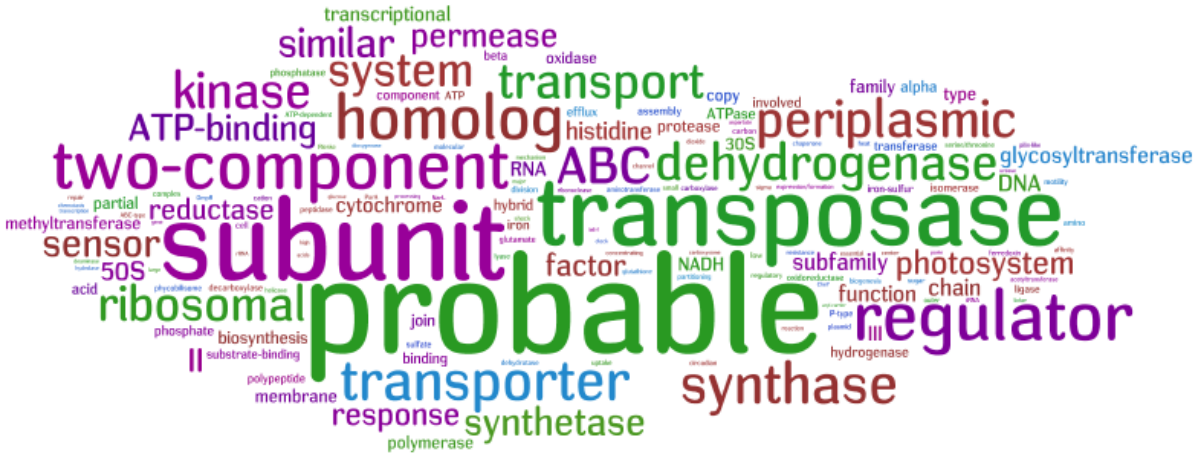


Figure 1.6: A word-cloud showing the gene names, with size related to the frequency within the gene list. Unknown function genes, or genes that are not closely related to any currently categorised proteins, are excluded (Kazusa, 2016))

total genome size of 3.95 mbps (mega-base pairs) (Kazusa, 2016), although of these genes only 1060 have been confidently identified, with the majority of the remaining genes coding for transposable elements, or being of unknown function. This is exemplified in the word-cloud in figure 1.6 (p. 47).

The organism is polyploid, meaning it has multiple copies of its single chromosome. Initially, up to 12 copies of the genomic chromosome were identified under normal growth conditions within *Synechocystis* (Labarre et al., 1989), although the separation of these chromosomes has been found to vary irregularly, as a result of uneven segregation of the genetic material during cell division (Schneider et al., 2007). Further study identified that this was an underestimation, with some strains containing over 200 copies of their genome – a value higher than that seen in the radiation-resistant organism *radiodurans*, although this value was seen to vary between 40 and 218 copies depending on strain and growth stage (Griese et al., 2011). A follow-up investigation also looked at stress effects, and found that phosphate levels were a key factor, when phosphate levels are abundant the cell accumulates multiple copies of its genome; however under prolonged phosphate starvation this copy number can drop until the cells ultimately became monoploid (only a single genome copy per cell) (Zerulla et al., 2016). Traditionally this polyploidy poses significant difficulties when performing genetic knockouts, as the high copy number can result in partial uptake of the genetic construct, resulting in 'heterozygous' cells containing the uptake biomarker. As a result of this and the slow rate of growth, creation of knock-out (Δ) mutants can take several months; and in cases where the mutant is significantly less viable than non-transformed variants it may be impossible (Heidorn et al., 2011). On a personal reflection, the recent research identified by Zerulla et al. (Zerulla et al., 2016) suggests that it might be possible to radically accelerate the rate of mutant generation in

Synechocystis by running mutations under phosphate-limited conditions; although this would have to be weighed against the reduction in growth rate caused by such conditions.

Synechocystis has a highly complex system of membranes, composed of outer, plasma and thylakoid membranes. On the outer-most layer is the mucilaginous sheath, a layer of polysacharides and glycolipids; beneath this rests the largely unknown surface layer (or S-layer). Underneath the surface lie the the outer and plasma membrane, which surround the peptidoglycan cell wall, and the thylakoid membranes are found within the cell (Heidorn et al., 2011). The majority of the protein content, around 60% by concentration, is made up of 4 phycobilisome pigment proteins, which are localised to the light harvesting machinery of the organism, known as the antenna (Colyer et al., 2005).

1.4.2 Hydrogen production in *Synechocystis*

In cyanobacteria, hydrogen can be produced by two enzymes, either directly by hydrogenases or indirectly by nitrogenases. Within the hydrogen-evolving species, H_2 is used as an electron sink. Hydrogen producing enzymes are typically activated when the environment is highly reducing, as this triggers the production of radicals that can cause oxidative stress, which is harmful to a cells continued existence through the destruction of structural and genetic components. Oxygen is by far the favoured terminal electron acceptor, as it is much more electronegative (3.41) than hydrogen (2.20). This means it will absorb electrons more readily than hydrogen, and can therefore be used to drive a much more efficient set of energy production and storage reactions in the cell. In the absence of O_2, or when the environment is excessively reducing, other mechanisms are required to prevent oxidative stress build-up.

Hydrogenases and nitrogenases are typically inactivated by the presence of O_2, due to the nature of how hydrogen is generated. The enzymes channel electrons into a fixed area containing protons in a controlled manner, which drives the reaction $2\,H^+ + 2\,e^- \rightleftharpoons H_2$ to the right hand side. If O_2 enters the active site of the hydrogenase enzyme, it is immediately converted into a free radical that begins to cause oxidative damage to the surrounding environment, which results in destruction of the enzyme. Oxygen tolerant hydrogenases exist, which exclude oxygen from the active site, however given the evolutionarily favourable (more efficient) reactions present in the cell for dealing with reductive stress using oxygen, they are much rarer. Hydrogenases are metalloenzymes that catalyse the combination of protons and electrons into hydrogen gas ($2\,H^+ + 2\,e^- \rightleftharpoons H_2$). Nitrogenases are the enzymes that fix atmospheric nitrogen in the form of ammonia, where hydrogen gas is produced as a side-product as seen below (Tamagnini et al., 2002).

$$N_2 + 8\,H^+ + 8\,e^- + 16\,ATP \longrightarrow 2\,NH_3 + H_2 + 16\,ADP + 16\,P_i$$

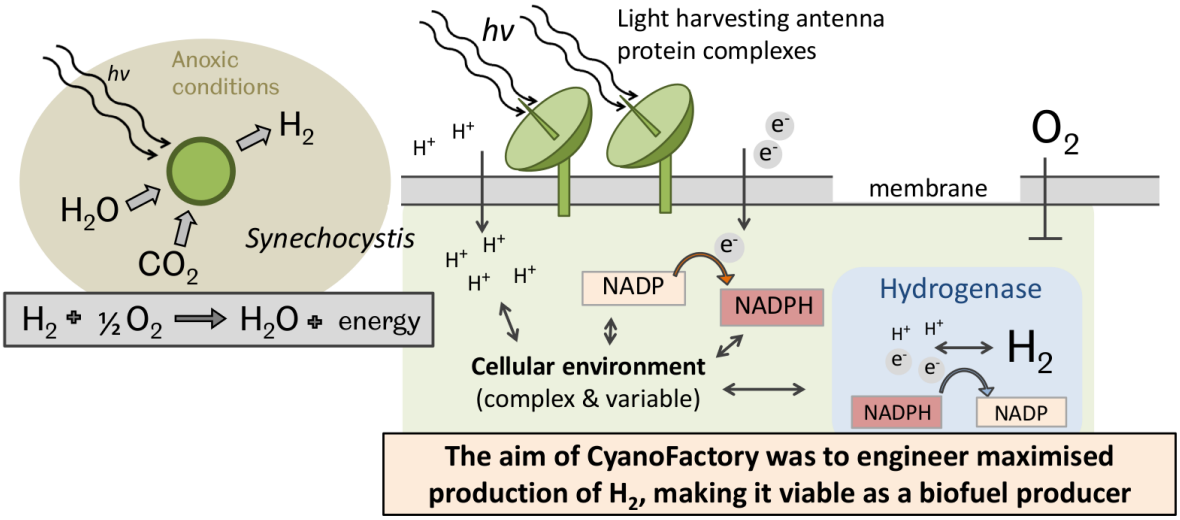


Figure 1.7: A figure indicating the general mechanism of hydrogen production within *Synechocystis*. On the left the general equation for H_2 production is given, along with the combustion reaction for H_2. Moving to the right, the antenna structures harvest light energy (*hv*) and in doing so drive an electron gradient across the membrane. These electrons are used to convert NADP to NADPH, which interacts with the hydrogenase to drive the equilibrium of the bidirectional hydrogen production equation to the right. This process is inhibited by the presence of O_2.

Within the active site of the bi-directional hydrogenase, there is a square-planar arrangement of moieties and coordinated metal ions. These act simultaneously as an electron sink and a reaction surface, where two protons are held in close proximity to the metal ions and preferentially draw electrons to form a single molecule of H_2. The electron supply to the hydrogenase in *Synechocystis* is provided by NADPH, and an overview of the system is shown in figure 1.7 (pg. 49).

Synechocystis contains a native bidirectional [NiFe]-hydrogenase, which is encoded by a cluster of genes known collectively as the hox cluster. It is believed that the primary function of this enzyme is to operate as an e^- valve during photosynthesis, allowing the organism to respond rapidly to changing levels of light in the environment without slowing e^- transport (Appel et al., 2000). Hydrogen is produced in *Synechocysitis* as a terminal electron acceptor in highly reducing environments, achievable as it pushes the reversible reaction towards the right due to an excess of e^- (McIntosh et al., 2011).

The the multi-subunit hydrogenase is formed from 5 proteins and contains two functional moieties, the hydrogenase moiety, responsible for the formation and degradation of hydrogen, and the diaphorase moiety, responsible for collecting and transferring e^- during the reaction (Eckert et al., 2012). The assembly of the hox cluster proteins is localised to the cell membrane where it can be effectively coupled to the photosystem proteins embedded within the membrane (Schultze et al., 2009). It is important to note that the

link is not as clear as implied here, based on the reported observations of a variety of mutant phenotypes in recent work (Eckert et al., 2012).

The functional portion of the hydrogenase is composed of the HoxYH subunits. This has been demonstrated with overexpression experiments that show increased activity, as well as with KO experiments that show a loss of function (Germer et al., 2009; Pinto et al., 2012a). Transcription of the hox genes is controlled by a number of factors, including light level, inhibition of the calvin cycle and anoxia (Kiss et al., 2009). This is thought to be coordinated by the transcriptional regulators AbrB-like protein, Sll0359 and LexA; although these are not exclusive coordinators of expression as they fail to account for oxygen and redox related response (Gutekunst et al., 2005; Oliveira and Lindblad, 2008; Kiss et al., 2009).

Measuring the internal systems that drive effects like hydrogen production, requires a number of focused techniques. In the remaining sections of this chapter, the core techniques used to analyse 'omic data are described and discussed.

1.5 DNA and RNA analysis

There are a number of different methods for determining the sequence of DNA, however the original method, dideoxy Sanger sequencing, was devised by Sanger et al and remained in popular use for almost 30 years (Sanger et al., 1977). This sequencing method involves the use of DNA polymerase to replicate a strand of DNA, but in addition to the normal mix of nucleotides, the sample is doped with a small amount (1%) dideoxynucleotides for one of the bases. When a dideoxynucleotide is incorporated into an elongating strand of DNA, it is blocked from further expansion. This occurs by random chance, as both the normal base and the dideoxy variant are present, resulting in a series of differently sized pieces of DNA that exactly correspond to a position of the base. Four separate reactions are run in tandem, with one dideoxy base variant included in each pot. These four reaction mixes are then run together on a gel and the bands can be read in sequence to give the sequence of the DNA. A variant of this technique exists where each of the dideoxy nucleotides are functionalised with different coloured tags. The reaction proceeds as above, however the tags can be added into a single pot and run on a capillary gel with a detector at the end. The detected colour sequence corresponds to the DNA sequence. The DNA sequencing for human genome project was conducted using this technique (Venter et al., 2001; Lander et al., 2001), however since then, there has been a rapid advancement in new sequencing technologies. There are a large number of 'next generation' DNA sequencing technologies - which have been recently comprehensively reviewed by Goodwin et al. (Goodwin et al., 2016).

Transcriptomics is the measurement of all the messenger RNA (mRNA) within the cell. When analysing RNA, there are a few key factors that need to be addressed. Beyond its role as a transitional step between DNA and protein, RNA plays a variety of non-messenger roles within the cell related to protein synthesis. These include providing structure for the ribosome – the construct that facilitates protein synthesis (rRNA) – and transferring amino acids to the ribosome for protein elongation (tRNA). Overall, mRNA makes up around 5% of the total RNA within the cell (Warner, 1999), although this can vary slightly with species (Kopf et al., 2014).

RNA is a single-stranded molecule that is generally considered to be unstable, and so when analysing the transcriptome it is typically converted back into DNA using enzymes called reverse transcriptases, producing cDNA (coding DNA) for analysis. Traditionally the measurement of cDNA was used to infer the amount of protein within a cell, as RNA was thought to translated into protein fairly directly. This was found to not always be the case, due to different rates of translation and different levels of turnover of mRNA and protein (Washburn et al., 2003), and so a more accurate model for the presence of RNA within a cell must also take into account the rates of protein production, RNA and protein turn-over, and the absolute levels in a given time.

The first full-genome scale combined proteomic/transcriptomic study was carried out by Schwanhausser et al. on mouse fibroblasts (Schwanhausser et al., 2011). Previous studies had shown a disconnect between the levels of mRNA and proteins; however this was suggested by Schwanhausser and colleagues to be caused by experimental differences, such as studies carried out in differing labs at different times or effects of reagents that were used during the experiment. They showed that the difference between protein and transcript levels measured in tandem were not as great as previously thought – there are clear correlations between protein copies and RNA copies per cell; this correlation is improved drastically by also considering the rate at which protein is translated from the RNA, which was found to be the major influencing factor for cellular protein levels (Schwanhausser et al., 2011). It was later found that additional levels of control may in fact play a larger role in protein translation than previously thought, due to a systems feedback effect called coupling (Dahan et al., 2011). This study was then repeated in *Mycoplasma pneumoniae*, this time also measuring the rate of protein turnover within the cell (Maier et al., 2011). They showed that gene expression is largely decoupled from protein levels and that translation efficiency is more prominent when considering actual protein levels. This same effect was also observed in *Sulfolobus* species, granting additional credence to the concept of codon usage being a key controlling point for cellular protein levels (Zou et al., 2012; Gingold et al., 2012). To date, a study of this nature has not been conducted for *Synechocystis*.

RNA-Seq is a method of sequencing the complete transcriptome of an organism at a

relatively low cost using next generation sequencing techniques (Wang et al., 2009). It does require the initial production of a cDNA library, however, and this can take a great deal of time and effort. As a result the method can be described as highly efficient, but with a high initial cost and low costs thereafter.

RNA-seq was first performed in *Synechocystis* for an ethanol production study by a group in China (Wang et al., 2012a). Considering the comprehensive quantitative information obtained from the transcriptome, it is easily the most time and energy efficient method of periodically observing the entire transcriptome under a variety of different conditions. This method has also already been employed to observe hydrogen production in the green algae, *Chlamdymonas reinhartii* (Toepel et al., 2013).

1.6 Proteomics

Proteomics is a discipline that analyses a collection of proteins — polypeptides which perform functions in vivo — extracted from a biological sample. Unlike other 'omics methods, such as genomics and transcriptomics, proteomics is complicated by additional factors. These include alternate splicing of the RNA before translation, which changes part of the primary sequence of the protein, or the addition of post translational modifications (PTMs) such as the addition of phosphate. These changes can radically change identifiable features of the protein such as function, weight or iso-electric point. Beyond simply identifying the presence of a protein within a specific proteome, this discipline can be expanded to identify a change in concentration of specific proteins within a sample; or to isolate specific PTMs which may alter functionality.

Whilst phenotypic variations can be measured based on how the organism interacts with its environment relatively simply, such as increased hydrogen production, the causes and limitations to these changes — and therefore the useful biological leads — are dictated by a plethora of small alterations at the metabolic level. The central dogma of molecular biology shows that steps taken by the organism to cope with its metabolic environmental conditions, at both a genetic and post-transcriptional level, will be reflected in the constitution of the proteome (Crick, 1970). As a result, by identifying and quantifying protein within an organism under specific conditions, a framework for a metabolic model to measure phenotypic response or adaptation can be formed.

Traditionally, proteins were identified using Edman degradation (Niall, 1973). This method sequenced a protein one amino acid at a time by cleaving the terminal amino acid and then identifying it; however it was prone to errors, can only sequence a single protein at a time, took hours or days to complete each run and required that the protein being sequenced was not blocked on the amino-terminus to Edman reagents. In the 1990s,

tandem mass spectrometry became the standard method of protein identification, as it bypassed these problems (Wilm et al., 1996). Technology has advanced over the past two decades and now it is possible to process a large number of proteins simultaneously, giving tandem mass spectrometry proteomics a much higher throughput rate than other methods (Steen and Mann, 2004).

1.6.1 The Proteomics Pipeline

The proteomics pipeline is the process that a sample goes through from experimental method, ending with a list of proteins and how they reflect the metabolic state within a cell. A summary of this process is available in Figure 1.8 (p. 54).

The first step is the extraction of the proteins. There are multiple methods of performing this task, which are discussed in detail in Chapter 3. The extracted sample is denatured then digested with a protease. Undigested protein detection, or top-down proteomics, can be performed with mass spectrometry but will not be discussed here. During the denaturation step, cystine disulphide bridges need to be reduced and then alkylated to prevent reformation. This improves the rate of proteolytic cleavage and also degrades tertiary structure based connections, which cannot be determined from genomic information alone. The digestion is commonly performed using trypsin, due to its high fidelity to specific cleavage sites, after either a lysine (K) or arginine (R) residue (Olsen et al., 2004).

As there are a great number of variables between different runs on a mass spectrometer, the simplest way of getting quantitative information of a difference between two samples is to label them separately and then combine the samples into a single work-flow. It is important to note that due to the limited nature of mass spectrometry proteomics, the technique cannot be used to prove the absence of a protein from a sample. This is because due to the nature of tandem mass spectrometry only a small subset of the sample is actually measured in a way which transforms it into useful information. This is further limited by features such as the dynamic range of a sample — the difference in concentration between the most abundant and least abundant peptides being measured. The spectrometer runs a large number of scans to attempt to capture the scope of the information, however relatively low abundance peptides will still fail to be detected. These peptides may be seen as a single m/z on the survey scan, but without further fragmentation data they cannot be identified.

As the sample is highly complex at this point, it is fractionated with liquid chromatography. This helps to reduce the dynamic range by separating out the peptides into multiple fractions, each containing a smaller number of peptides. This is done with a

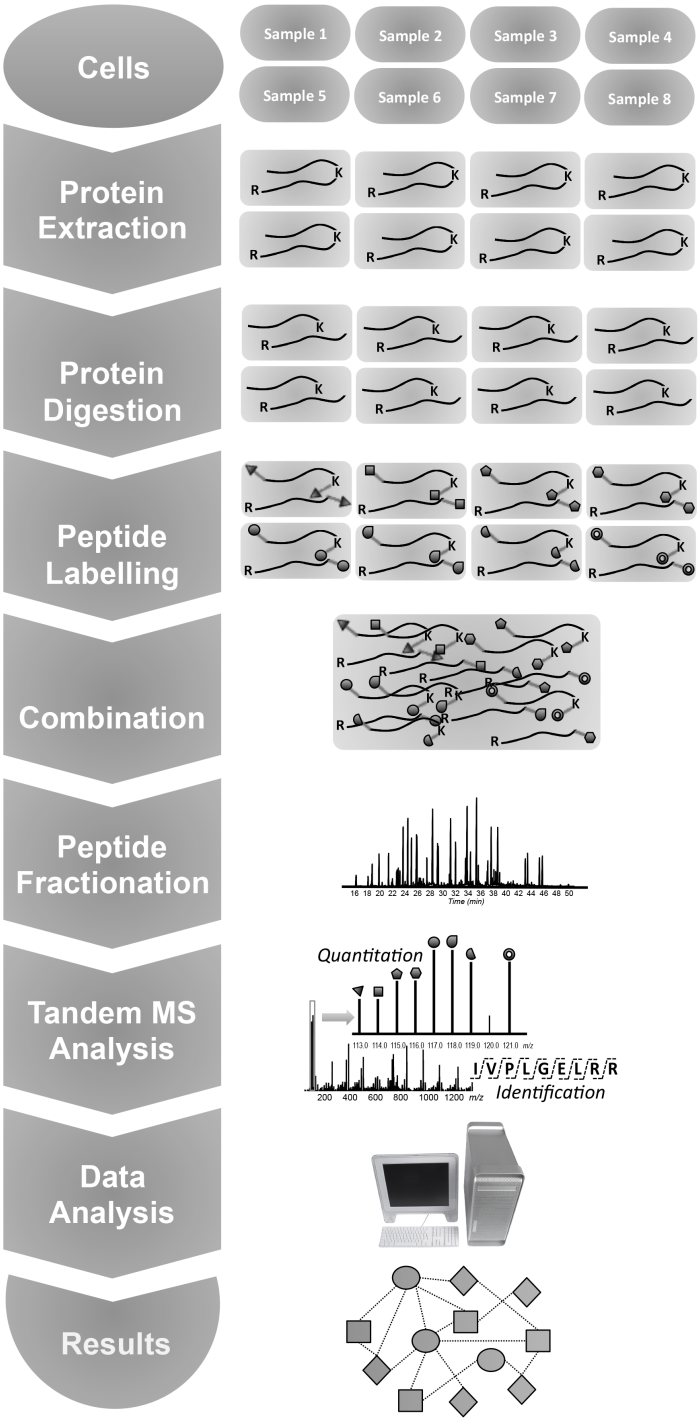


Figure 1.8: Proteomics Pipeline: Cells from different samples are collected, the proteins are extracted and cleaved before labelling. The labelled peptides from all the different samples are combined and then fractionated with liquid chromatography to reduce complexity. Fractions are then analysed with mass spectrometry and the data is analysed to produce results. For more details, please refer to the text. Taken from (Couto et al., 2013; Evans et al., 2013)

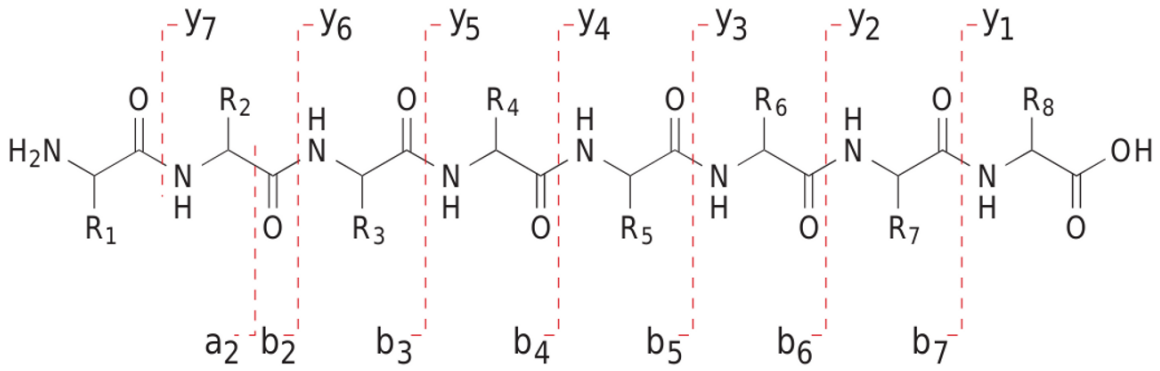


Figure 1.9: The ideal cleavage patterns of a peptide following collision. R_n are the functional groups on the amino acids and dotted lines indicate fragmentation - for example fragmentation between R_4 and R_5 would produce fragments b_4 (amino fragment) and y_4 (acid fragment). Observation of these fragments within a spectrum enables identification of the sequence of the amino acids within a peptide. Taken from (Steen and Mann, 2004)

method such as HILIC, where peptides are separated based on hydrophobicity. If the sample is not fractionated then peptides from high-abundance proteins will mask the presence of peptides from low-abundance proteins. Ideally during this stage peptides are grouped together within the fractions giving better intensity readings whilst hopefully reducing the dynamic range.

Tandem mass spectrometry (please see Section 2.4.4 for more details on Mass Spectrometry) was first proposed in 1978 (Yost and Enke, 1978). The actual detection method works by running two simultaneous mass scans on a sample, the first scan monitors the mass to charge ratios (m/z) of all material (including peptides) moving through the machine and hitting the detector. This is known as the MS or MS^1 scan. An on-board computer algorithm then identifies m/z values of interest from this scan.

Objects of the selected 'mass of interest' are filtered from the sample and draw through into a collision chamber, where they are fragmented. This level of fragmentation is fine-tuned to produce single cleavages within peptides, occurring at more or less specific points in the peptide bonds between the amino acids (Biemann, 1992). The output of this collision is fed to a second detector, where it produces an MSMS or MS^2 scan. In an ideal situation, due to a large number of molecules passing onto the detector, high-intensity peaks are seen for m/z values with a difference corresponding to the weight of each amino-acid in the peptide sequence, facilitating identification. See figure 1.9 (p. 55).

As the proteolytic cleavage (with trypsin, for example) occurs at specific sites, a theoretical digest can simultaneously be performed on a digital proteome, deduced from the genome of an organism. This provides a list of possible 'hits' for detected peptides to be validated against. The expected m/z values from the sample can be simulated, pro-

ducing a database of expected values. This can be used to identify peptides of interest within the data produced from the mass spectrometer. The observed and expected values are compared to assign possible identifications, with numerical methods being employed to attempt to counteract background noise and missing information. Alternatively, a method known as *de novo* sequencing can be used, where peaks within the spectrum are used to determine the amino acid sequence — as well as any modifications to those amino acids. These peptide sequences are then searched against the existing database of known proteins, leading to protein identification. When an MS^2 scan is confidently assigned to a peptide unique to a particular protein, the protein can be considered present within a sample. If the peptide is not unique to a particular protein, it is considered 'shared', shared peptides are often ignored as they cannot accurately be assigned to an individual protein.

There are several complicating factors to the database-search method of identification. Firstly, PTMs can't be detected from genomic sequence alone and therefore peptides with modifications will cause discrepancies between expected and observed m/z values. It is important to note that this issue can be avoided when using *de novo* sequencing if a comprehensive list of modifications is available — however this may exponentially increase the search time. Fortunately, not every peptide in a modified protein will be modified, and as a result a protein can still be identified in spite of PTMs. Additionally, there is the issue of isotopic variation of elements within a sample, for example the presence of ^{13}C will alter the mass of the peptide and therefore must be accounted for in the identification stage. The occurrence rate of this effect is known, and as the peptides are relatively small (around 6 - 20 amino acids long each) it can be accounted relatively easily using a method known as Isotopic Distribution Deconvolution. There is also a benefit to this effect - by identifying the distance between isotopic distribution peaks, it is possible to identify the overall charge (z) of the molecule being measured. These will occur at a rate of 1 in a singly charged molecule, 0.5 in a doubly charged molecule, 0.33 in a triply charged molecule and so on.

Mass spectrometry also suffers when identifying fragments of very similar mass, for example leucine and isoleucine have identical masses, making them indistinguishable on this basis. Finally, since not every peptide within the spectrometer can be measured high abundance proteins can completely mask the presence of low abundance proteins. The difference in abundance between the lowest and highest copy proteins is referred to as the dynamic range, and having a large dynamic range can result in reduced identification efficiency. Such as the case of phycobiliproteins in *Synechocystis*.

1.6.2 Mass Spectrometry

Mass spectrometry is essentially a highly accurate molecular weighing scales. Samples are loaded into the machine, ionised and then separated by their mass to charge ratio (m/z) before hitting a detector. There are multiple methods of ionising and separating molecules, resulting in a variety of different spectrometers — although they all follow the same basic principle described above. Preparation, and also the method of 'simplification' of the samples varies, depending on the type of ionisation method employed by the spectrometer. Two major methods of ionisation are the most common within proteomics: Electro-Spray ionisation (ESI) and Matrix Assisted Laser Desorption/Ionisation (MALDI). These are both 'soft' ionisation techniques, which cause very little to no fragmentation to the molecule being charged. Minimal fragmentation is essential for the survey scan in tandem mass spectrometry.

In ESI, the protein sample is digested (as described above) before fractionation. This is done gel-free, using High Performance Liquid Chromatography (HPLC). The sample is fractionated with a mass spectrometer competent buffer such as acetonitrile ($CH_3-C\equiv N$), which does not interfere with the operation of the spectrometer. The solution is run down a charged needle, situated next to the mass analyser collector. There is an electric current applied between the needle and the collector, which pushes charged ions down to the tip of the needle. On reaching the tip, the sample is repulsed into a Taylor cone with droplets escaping from the end of the cone and flying towards to spectrometer. These gradually reduce in size as the buffer dissociates, increasing the charge density of each droplet and spraying out ions as the droplet shrinks. On complete evaporation, typically only a single charged molecule remains. This is then processed through the mass analyser.

In MALDI, the sample is first separated using 2D gel electrophoresis. This is done by firstly running the protein sample on a column gel that is gradiated by pH, which separates proteins out by iso-electric point. The column gel is then loaded into another gel and standard gel electrophoresis is used to separate the proteins out by size. Different phenotypes are run on different gels, which produce a similar output as the sample is mostly identical. Relative abundances are measured by spot-size on the gel and proteins believed to be differentially expressed by a significant amount are picked from the gel and digested (as described above). This assessment is made by graphical analysis software, as the task is daunting for a researcher when attempting to measure a full cell proteome. Selected samples are then loaded onto plates within a MALDI matrix material. The machine ionises the sample by firing a wavelength of ultra-violet light at the matrix, causing its degradation. This simultaneously ionises the sample, whilst releasing it into the gas phase and launching it into the mass analyser.

There are four types of mass analyser: Time of Flight (ToF), Quadrupole, Quadrupole

Ion Trap (QIT) and Orbitrap.

The simplest form of mass spectrometer is the ToF separation technique. In this method, all the ions are accelerated to the same kinetic energy (E_k) and passed through a flight tube to a detector. As $E_k = \frac{1}{2}mv^2$, where m is mass and v is velocity, molecules with a smaller mass move more quickly and therefore hit the detector first (Mirsaleh-Kohan et al., 2008).

In the quadrupole, the sample is passed between 4 parallel rods before hitting the detector. These rods generate a tunable electrical and radio-frequency field that alters the flight path of ions moving through it, causing ions with a non-permissible m/z to collide with the rods and never reach the detector. The frequencies can be run through a range of steps that allow a spectrum of m/z values to be detected (Wolfgang and Steinwedel, 1956).

QIT works similarly to quadrupole, with the exception that it operates an exclusion detection method instead of a permissive detection method. Ions pass into the field and are then trapped by a dynamic alternating current. These collected ions are only ejected when the quadrupole is tuned to let those with a specific m/z pass through. Ejected ions then hit the detector and the m/z is determined by the frequencies of the field (Schwartz et al., 2002).

Finally, the orbitrap mass analyser works by trapping ions within an electrostatic field, without the application of radio frequency. The ions orbit around an axial electrode and harmonically oscillate with a frequency proportional to $(m/z) - \frac{1}{2}$. These oscillations produce an effect on the central electrode, which are detected and can be converted back into (m/z) values for high-accuracy mass detection (Makarov, 2000).

1.6.3 Quantification

So far, only the detection of the presence of a protein has been described. Whilst this is useful for confirming that predicted proteins exist or for identifying the presence of a particular organism or gene within a sample, it doesn't provide any useful information about the phenotype being investigated or provide an insight into the molecular function of the phenotype. The absence of a protein for a phenotype is non-informative as well, since as previously mentioned peptides can be missing from the data even if they are present within the sample. As a result, when comparing phenotypes some degree of quantification of protein is required.

There are numerous approaches for quantifying proteins within a proteome. These include measuring relative abundances on a gel prior to running through the spectrometer, making

an estimation of abundance based on a known standard, or by labelling samples prior to running them. There are two forms of quantification, absolute and ratio-based. In absolute quantification, the concentration of a protein is estimated based on the values measured, whereas in ratio-based quantification the amounts of protein in a sample are measured relative to another sample - such as an alternative phenotype or an experimental control.

In gel-based quantification, a gel is run as described above in either one or two dimensions. It is then labelled with a protein-specific stain, allowing two samples to be compared using image analysis techniques. For quantification in LC-MSMS there are a number of different techniques available, three common methods include label-free, metabolic labelling techniques — such as Stable Isotope Labelling in Amino acid Cell culture, (SILAC) — and isobaric tagging.

Briefly, label-free can be done by measuring intensities protein spiking involves inserting a known concentration of a purified protein into a sample during a run on the machine. Since the protein and its quantity are known, it can be used as a baseline for quantifying other proteins, however there are limitations with this method. In an ideal situation, all peptides from a protein will be passed from the sample to the detector, however peptides can be lost at different stages in the detection process. Some peptides don't ionise particularly efficiently, this can be due to the sequence of amino acids and the chemistry of the resulting molecule. There can also be an uneven separation of peptides into different fractions from the HILIC, which can reduce the signal intensity. As a result of both of these effects, a measured peptide spectra can be of a lower intensity than is expected. This leads to an under or over estimation of the amount of protein in a sample, depending on whether the lost peptides are from the sample or the spiked protein.

In SILAC quantification, one phenotype is fed/grown on media containing amino acids isotopically labelled with ^{13}C or ^{15}N. These non-radioactive 'heavy' labels are typically arginine or lysine, although this depends on the protease being used during analysis. They are stably incorporated into the proteins within the cells and over time become present within the entire proteome. When the peptides from these organisms are extracted, they can be pooled for detection in the mass spectrometer. As the peptides run through the MS^1 scan, the separate phenotypes can be observed due to a mass difference of 1 (or more, depending on the number of heavy amino acids present in the sample) (Ong et al., 2002).

An isobaric tag is a molecule that is made up of three parts, the binding group, the reporter group and the balancer group. The binding group specifically binds covalently to primary amines. In the case of peptide tagging, this ensures that the N-terminus of every peptide can be labelled (lysine residues will also bind the tag). The reporter groups

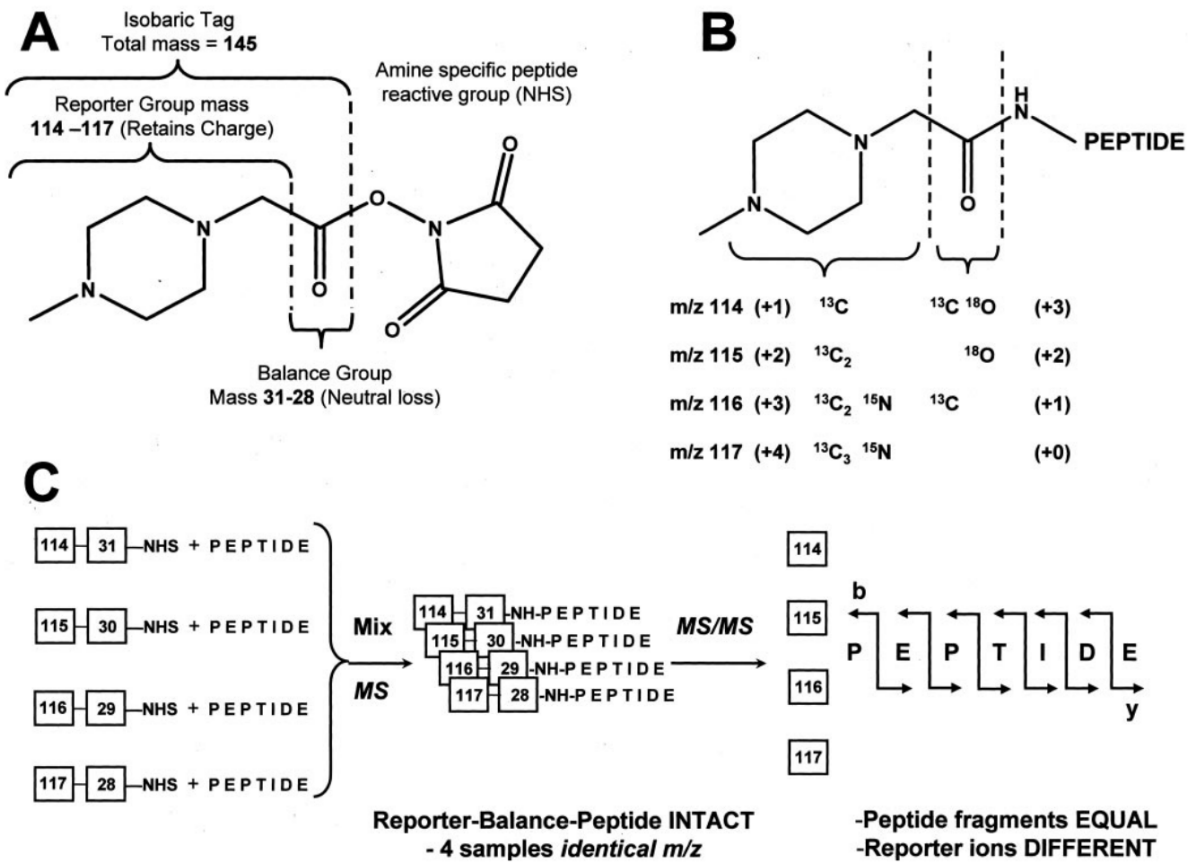


Figure 1.10: A summary of isobaric tags. **A**: The isobaric tag is made up of an amine specific peptide reactive group enables the tag to bind to the peptide, and a reporter and balancer group that weigh equal amounts cumulatively. **B**: This image shows the tag covalently bound to a peptide. In this example, the isobaric tag is an iTRAQ 4-plex. There are 4 tags with m/z values differing by 1, and each phenotype is labelled with a different tag. At this point the weight of all phenotypes has been increased by an equal amount, due to the balancer group. **C**: When the peptides move into the spectrometer, they remain isobaric within the survey scan, but during the collision phase they fragment. The reporters from the tags are then measured by spectrum intensity, which can be used to determine quantifications. Taken from (Ross et al., 2004b)

have the same chemical sequence, but contain ^{13}C and ^{15}N isotopes so that each label differs in molecular weight by approximately one. Finally, the balancer group is also made up with ^{13}C and ^{15}N isotopes complementary to those in the reporter group, causing the overall mass of all the tags to be equal when they first pass into the mass spectrometer. See Figure 1.10 (p. 60).

Each tag is covalently bound to a different replicate or phenotype within the experiment and after this step the samples are all pooled together for detection. At this point it is impossible to distinguish between different samples based on weight or molecular properties, resulting in no bias for particular tagged peptides moving through the LC or the survey scan. When the tagged peptide is picked for an MS^2 scan and undergoes

fragmentation, the reporter from the tag separates and can be measured by the detector. The weight of the reporter falls within a relatively clear part of the spectrum and as a result, relative intensities of the different reporters (and therefore peptides) can be measured from a single injection. There are two commercially available isobaric tags, Isobaric Tags for Relative and Absolute Quantification (iTRAQ) by Sciex, (Ross et al., 2004b) and Tandem Mass Tags (TMT) by Thermo Fisher (Thompson et al., 2003).

In principle these two methods work in the same way, differing only in the compositions and masses of the reporter tags. iTRAQ is currently available in 8-plex, allowing 8 different samples to be combined in a single run of the mass spectrometer, whilst TMT has recently released a 10-plex tag system, although this can only be resolved on very high accuracy machines such as an orbitrap spectrometer (Werner et al., 2012; McAlister et al., 2012). Once quantitative data has been collected from the mass spectrometer, it needs to be processed prior to enable understanding of what the results mean biologically.

1.7 In-silico models

1.7.1 What's in a model?

"All models are wrong, but some models are useful" (Box, 1976).

A model, in simplest terms, is a line drawn between the points on a scatter plot suggesting a trend. In doing this, we predict that the values that fall between measurements will follow this line – we make a guess about what the underlying causes of data would do in the spaces we do not measure. By relating this to some form of algebraic expression, we can take some controllable value and use it as a 'predictor' for an observable one. Consider the case in figure 1.11 (p. 62). The dotted grey line in this figure is a linear model with the equation $y = 0.1 \times x$, which in real terms tells us that for every 10 'units' we raise x by, we can expect y to raise by one 'unit', within a certain degree of error.

In the the first graph, the model appears to fit the data well, but as more data is gathered in the unmeasured space between points, additional effects are observed and could in turn be better described by a more accurate model. The important point to realise here is that whilst the first model did not capture the entire complexity of the underlying system, it enabled an observation of a trend. The 'noise' or error remaining between the data and the model shows us that there are features within the system that have not yet been captured, but the observation of $y = 0.1 \times x$ is still useful for predicting trends generated by this system – if the researcher can accept that their predictions will contain a margin of error.

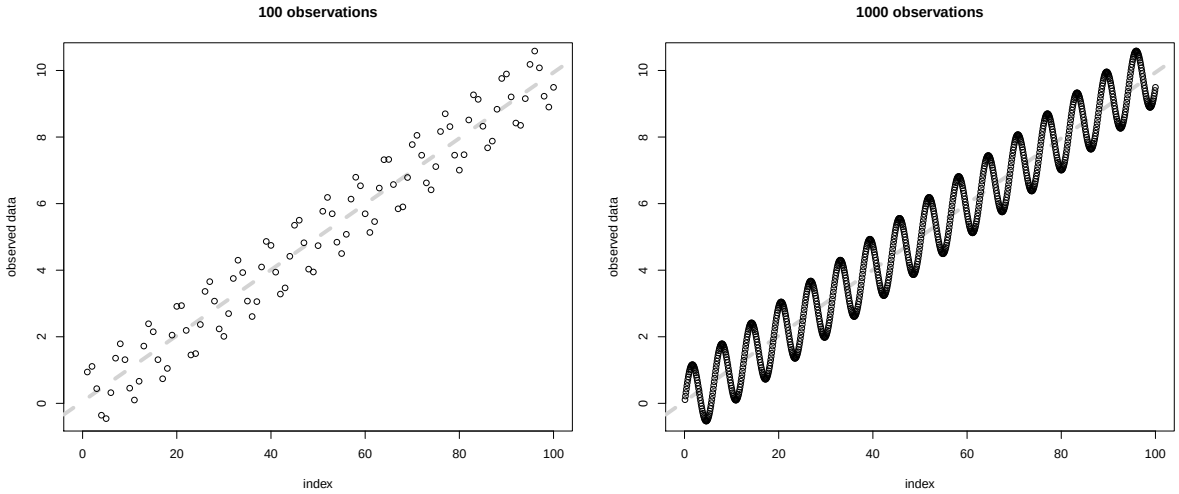


Figure 1.11: Two plots populated with data generated from the same formula, $y = 0.1 \times x + sin(x)$. In the first case the linear model appears to fit well (left), but further observations may reveal this to be an incomplete picture of the underlying trend (right).

Metabolic models of cells operate in a similar manner, they take recorded experimental values and attempt to describe what observable effects will take place within the cell. Metabolic models can be used for a variety of predictions for production. These range from relatively simple models, which consider enzyme kinetics to predict the amount of a product produced by a given protein (Johnson and Goody, 2011), to more complicated systems that predict entire pathways of enzymes and metabolite pools inside and outside the cell to produce a predicted production rate (Montagud et al., 2015).

More detailed models do not always produce better results, and it is worth noting that highly complex models run the risk of 'over-fitting' – or trying to predict the cause of noise in a system when the experimental data doesn't capture these observations. Consider figure 1.11 (p. 62) again; whilst it is possible to predict the $sin(x)$ from this data (as the underlying function is simple), if multiple underlying effects were present it is possible that a combination of unmeasured effects could appear to have a pattern that did not truly exist. The best example available of this is the line of best fit automatically applied in Microsoft excel graphs – the trend line is generated from few data points that have been highly interpolated. This generates a line that travels through every point, even if the system contains noise that cannot be explained by the independent variable.

The risk with over-interpreted models is that they suggest a relationship between the dependent and independent variables that does not exist, which can mislead the researcher. Models are, after all, only aids to help a researcher make associations between variables that are difficult to compute; they play a role in iterative research, enabling generation of new, testable hypotheses and appropriately designed experiments (Box, 1976). It is also important to point out that a model describes a relationship or correlation, not ne-

cessarily a causation; determining causation requires a reductive analysis – for example, consider the following observation: in a closed system, A is increased, and B also increases (A and B are correlated); however when B is increased and A does not increase, B is not causative to A.

1.7.2 The Monod model and FBA

Jacques Monod pioneered the metabolic model of cell division in *E. coli* (Monod, 1949). In his experiment, he equated maximum cell growth to the availability of nutrients in the media; making the first link between a stoichiometric equation and the terminal point of cell growth – halving the available carbon source resulted in half the maximum cell density, similarly; doubling the rate of production of a toxic metabolite or the rate of reaching a critical ion imbalance such as pH also halve the maximum cell density. Interestingly, in the same work he also mapped the various stages of bacterial cell growth as they are widely known in biology today: lag phase, acceleration of growth, exponential phase, retardation of growth, stationary phase, and cell death (although as he points out, in some cases one or more of these phases are so rapid that they are not observed).

Building on the work of Monod, Flux Balance Analysis (FBA) is one method of attempting to approximate the various metabolite pools within the cell. It works on the principal that by using stoichiometric principals described earlier in this chapter, a balanced equation can be calculated for the entire cell (Orth et al., 2010). To overcome the issue of the system being under-determined; where missing information leads to several possible 'correct' outcomes, or some/all of the equations could be set to 0 to ensure they balance, the model has a 'biomass equation' that it tries to solve the simultaneous equations for. This creates a bounded solution space – a set of absolute limits required to be satisfied when balancing the equations. The model then solves for the highest possible amount of biomass production in a given time – the same 'objective function' that holds true in evolution-driven systems (in most cases, the fastest growing organism dominates an environment).

In other cases, where adequate literature on the organism being studied is not available, missing values can be borrowed from similar organisms or 'fudged' with fake place-holder reactions. This is required in some cases, such as where a pathway has no bridging reactions and so the biomass equation cannot be solved, or when a series of known reactions form an island that does not participate in the biomass equation. Although an FBA model may not predict the same outcome for the same model every time, the solutions it generates are generally not wildly different from each other and they can be run through multiple iterations to generate a statistically favoured state or probabilities for various states.

As it is run on a series of metabolic reactions, which are mapped to enzymes and metabolite pools, an FBA model can be particularly useful for predicting how an organism will be affected by a genetic modification, or limiting/removing a substrate from the growth media. In these cases, a series of cheap *in silico* experiments can be run to determine that the cell will still be viable (able to grow) – potentially avoiding costly experimentation that would not work. It is important to note that scepticism of *in silico* observations is still high in the scientific community – for good reason.

Many existing models still require extensive testing under experimental conditions to identify their limitations; which arise from the model being drawn together from disparate literature sources, generated from different labs, under different experimental conditions, with different genetic backgrounds, and in the case of models of 'non-model species' – from different organisms entirely. The future is bright, however; as computational analysis of data improves, and more frequent, higher quality 'omics-level analyses are performed; our capacity to test and verify these models is improving.

1.7.3 Reconstruction of the *Synechocystis* model

Synechocystis has recently been the subject of computational modelling to analyse the organism for potential biofuel production by Montagud et al (Montagud et al., 2013). This reconstruction – iSyn811 – contained 956 reactions relating to 811 genes; representing over 75% of the confidently identified proteins known in the organism, and an improvement over the previous model iSyn669 (Montagud et al., 2010). The model was designed to investigate the metabolic potential of the organism to produce two key biofuels: ethanol and hydrogen; and utilised transcriptomic data to improve on the previous model.

In addition to testing stoichiometric balances with FBA against experimental values, the model also tested for 'coupling' or directionality and flux limits to the reactions (Burgard et al., 2004). This makes the model more accurate, by controlling not only which reactions are reversible, which are irreversible, and what direction matter flows in through the directional reactions; but also the rate limiting steps within these clustered groups, providing maximal and minimal flux limits. During the CyanoFactory project, the proteomic data generated at Sheffield was forwarded to the research group that generated these models, to put observations of system-level changes into context.

1.8 Thesis summary

1.8.1 Chapter 1

In this thesis, systematic improvements to the collection and understanding of proteomics data were generated, using analytical and experimental methods. This introduction has (hopefully!) served as a primer to the field of industrial biotechnology, biofuel and H_2 production, proteomics, and the cyanobacteria *Synechocystis*.

1.8.2 Chapter 2

The main body of the thesis begins with Chapter 2, which is made up of 2 separate literature reviews that were prepared originally as separate manuscripts. The first is an assessment of the historic development of the field of proteomics in *Synechocystis*, and the second an investigation of emerging proteomic technologies and how they will influence the development of production strain analysis in the future. The second was published in current opinions in biotechnology, however a publication covering a similar topic to the first was published by another group removing the niche in the literature that the study was intended to fill. Despite this, the work carried out in this chapter provided important ground work for a number of the investigations conducted in chapter 4 for improving proteomic methods in *Synechocystis*.

1.8.3 Chapter 3

Chapter 3 details the proteomic investigation of altered media hydrogen production, which were carried out as part of the CyanoFactory project. This chapter covers the lab-based practical aspects of the PhD, and the conclusions highlight key changes relating to the organism when grown under environmental conditions that enable hydrogen production. Two studies make up this chapter, one was carried out at the end of the first year of PhD study, whilst the second was a related analysis carried out at the end of the study utilising the improvements that had been designed over the course of the thesis.

1.8.4 Chapter 4

Chapter 4 covers investigations into optimal experimental proteomic methods is carried out. This is coupled with data analytical assessments of both the chassis, based on experimental data collected by both Sheffield and other partners during the CyanoFactory

project, and key analytical techniques for summarising the data to create a simple output for a biological specialist to interact with. Two of the studies presented in this chapter have also already been involved in publications, with a third being included as part of an industrial study of *E. coli* in a bioreactor.

1.8.5 Chapter 5

Chapter 5 details a comparative study between two popularly used proteomic tagging systems in the context of the target chassis – iTRAQ and TMT. This study highlights confounding effects of the *Synechocystis* background and details the accuracy, and therefore the practical working ranges, of these technologies. It also covers an informatics investigation into the proteomic background of *Synechocystis*. Work in this chapter is currently being prepared for submission as a publication.

1.8.6 Chapter 6

The thesis closes with chapter 6, which draws together the various conclusions from throughout the thesis and cohesively lists the work that is included in the body of the thesis. It then highlights the areas that have been developed over the course of this study, as well as summarising areas where further investigation is needed. The final section of the conclusions lists the core contributions to science that have taken place over the course of this PhD.

1.8.7 Methods and appendices

At the end of the thesis, a list of computational methods used in the thesis are included in a digital repository. The sections included in this chapter are discrete data bundles, containing the code, data and relevant figures; each bundle is referenced by a DOI. A description of each of the bundles is given in the computational methods chapter. The appendices contain each of the deliverable reports from the now complete CyanoFactory project, which cover a broader base of topics than those contained within the main body of the thesis. A full figure list and table list with caption details are also included.

Chapter 2

Literature Review

2.1 Introduction

This chapter is made up of two separate literature investigations, the first related to *Synechocystis*, looking specifically at the limits under which the organism has been investigated, the areas relating to potential biotechnological applications, and the inferred responsiveness of the organism to a variety of environmental changes. These investigations are vital for predicting and understanding the proteomic changes that occur under different growth conditions. In addition, within this section, the literature is assessed for its progression over time in a bibliometric analysis, suggesting future trends that may be applicable to other organisms being developed for biotechnological applications. This study was intended to be submitted for publication, however due to a similar study by another group it is no longer applicable at this time; and has been retained as in-house knowledge of the field.

The second section of this chapter is verbatim a paper that was submitted to *opinions in Biotechnology*. This section covers in detail the latest trends in proteomics for production strains – or strains that have biotechnological application for production. *Synechocystis* made up a case study in this paper, but the general focus was on the improvements that need to be made to industrial applications for proteomics compared with where the field currently is with applications for medical studies. There is a lot of scope for improvement in this field, and it is likely that looking at pioneering cases, such as *Synechocystis* – which has a relatively advanced collection of proteomic research associated with it compared to other microalgae – can inform and accelerate quality improvements in other strains.

2.2 Proteomics in *Synechocystis*

Authors: Andrew Landels, Jenifer Parker, Narciso Couto and Phillip C Wright.

2.2.1 Abstract

Cyanobacteria are photosynthetic micro-organisms that hold huge industrial potential for the biotechnology sector. Predictable engineering is required to realise this potential, however this is completely dependent on comprehensive systems-level information that in many cases has not yet been completed. Here we present a summary of recent mass spectrometry based gel free proteomic work, one of the key fields required for understanding and engineering a biological system. We focus on *Synechocystis sp.* PCC 6803, the first cyanobacteria to be sequenced and as a result backed by the largest body of supporting

proteomic literature, to highlight general research trends and the current limits of the literature.

2.2.2 Introduction

Cyanobacteria are a highly diverse group of photosynthetic microorganisms responsible for the fixation of 25% of all carbon dioxide on earth, making them the largest global source of carbon fixation. They accomplish this by also being the largest harvesters of sunlight energy. They are differentiated from micro-algae by being prokaryotic organisms which lack internal cellular compartments such as a nucleus. In this review we will focus on a single species of cyanobacteria, *Synechocystis sp* PCC 6803 (herein referred to as *Synechocystis*). *Synechocystis* is a small, naturally transformable cyanobacterium with a relatively small genome that has been colloquially dubbed 'The Green E. coli' in the literature (Branco dos Santos et al., 2014). *Synechocystis* was the first cyanobacteria to be fully sequenced, with a genome size of 3947 kilobase pairs, which encodes a predicted 3575 proteins (Kaneko et al., 1996). The photosynthetic machinery is located in the thylakoid membrane, hosting highly effective light-harvesting complexes.

Synechocystis has been used as a model for photosynthetic systems in eukaryotic organisms numerous times, which can be seen with a brief search of the literature producing hundreds of matches. It is also industrially relevant, with a rapidly increasing number of publications investigating *Synechocystis* as a bio-producer for a variety of fuels and specialist molecules (Table 2.1, pg. 70). To put this into context, prior to December 2009 there were no publications linking *Synechocystis* to biofuel production – there are now more than 80. These bio-molecules range from organic fuel molecules like ethanol (Song et al., 2014; Dienst et al., 2014; Qiao et al., 2012a)), to gaseous products like hydrogen (Pinto et al., 2012a) and even synthetic production precursors like isoprenoids and lactic acid (Kudoh et al., 2014).

Modern technologies, such as online tandem mass spectrometry and next generation sequencing, have revolutionised fields like proteomics and genetics by creating the 'big data' approach to biology. Despite the challenge, comprehensive proteomics is critical to bioengineering. The proteome is a highly dynamic system that responds on both the translational (presence/amount of each protein) and post-translational (activity/functionality) level. Addressing the challenges involved in generating high-quality data from this field is essential for realising the industrial bioengineering potential. For an introduction to shotgun proteomics in cyanobacteria, please refer to the review by Ow and Wright 2009 (Ow and Wright, 2009).

Here we focus on a number of key topics related to gel free proteomics in *Synechocystis*,

Table 2.1: A table of the different biotechnology products that have been investigated for production in *Synechocystis*.

Biomolecule	Publication
Isoprenoids	(Lindberg et al., 2010), (Bentley et al., 2014), (Kudoh et al., 2014), (Englund et al., 2014)
Feed-stock for fish	(Anemaet et al., 2010)
High-value organic molecules	(Reinsvold et al., 2011)
Lactic Acid	(Angermayr et al., 2014), (Varman et al., 2013)
Glycerol	(Savakis and Hellingwerf, 2015)
Hydrogen	(Pinto et al., 2012a), (Rögner, 2013), (Montagud et al., 2013)
Fatty Acids and Lipids	(Liu et al., 2010), (Gao et al., 2012), (Cai et al., 2013), (Chen et al., 2014a)
Hydrocarbons	(Kämäräinen et al., 2012), (Wang et al., 2013)
Ethanol	(Qiao et al., 2012a), (Wang et al., 2012b), (Dienst et al., 2014) (Sengupta et al., 2013), (Song et al., 2014)
Hexane	(Liu et al., 2012)
Butanol	(Tian et al., 2013a), (Varman et al., 2013), (Zhu et al., 2013), (Anfelt et al., 2013)
Sucrose	(Du et al., 2013)

In addition to identifying key *Synechocystis* proteomics papers we will discuss the current research limits in the field, standard procedures for generating proteomic data, a discussion on effective quantification methods in current use, and future directions for the field.

2.2.3 Standard procedures in use and protein identification challenges

There is currently no community-standard method for generating proteomic data from *Synechocystis*, however there are a number of techniques in the cell-pellet processing pipeline that follow broadly the same patterns. Cell disruption is the process of lysing the cells. There are two techniques that appear in the majority of publications reporting high protein identifications: sonication and bead-beating. These have become popular because they are automated, effective and highly reproducible. Other traditional techniques, such as liquid nitrogen grinding, have a great variation between different users and the methods are difficult to report accurately and reproduce. High-pressure methods, such as the French press or Yeda press, have been shown to be thoroughly ineffective at disrupting *Synechocystis* cells, as certain strains are pressure resistant. More accurate comparisons between extraction methods is impossible as no study has directly compared extraction methods whilst maintaining the same conditions downstream.

Cell debris processing is where the majority of divergence in publications occurs. This refers to the process of taking the disrupted cellular material and processing it into peptides for mass spectrometry analysis. Broadly there are two methods of doing this, separating the whole proteins out on a gel and using in-gel digestion to produce peptides or digesting individual samples in solution and fractionating the peptides out by their individual features. The method chosen is heavily dependent on the down-stream processing and the focus of the experiment being carried out – if the user is conducting an exploratory investigation that does not require merged quantification methods then the protein-level separation has been shown to have significant advantages over peptide-level separation. A more in-depth comparison between these methods is made later in this review.

Peptide post-processing is an optional step, included in studies where the user is interested in identifying post-translational modifications or applying a peptide-specific tag. Post-translational modification purification is a rapidly growing topic which requires a specialised approach for each different modification being concentrated. For further details on current trends in PTM research methods, please see this review by (Huang et al., 2014). With over 20% of all *Synechocystis* proteomics papers citing their use, the most commonly used peptide-specific tags are iTRAQ (isobaric tags for relative and absolute quantification). Alternative quantification tags have only rarely been used in *Synechocystis* with a single study reporting the use of TMT (tandem mass tags). To date no clear comparison between iTRAQ and TMT tags has been made.

Quantification of proteins in proteomics is generally done in one of two ways, either through a gel-based method using 2D gel electrophoresis stain intensity analysis, or else through a spectral intensity counting method on the mass spectrometer. The quantification method used is dependent on the mass spectrometer available; gel methods are commonly associated with MALDI spectrometers whilst spectral counts are used with the LC-MSMS setup. Almost all studies to date have processed spectra for identification with the software Mascot (www.matrixscience.com). Alternative programs have been used previously, but a flat comparison between different software has not been completed.

Proteomic techniques in *Synechocystis* have been rapidly evolving over the last 5 years. This can be seen broadly from the number of protein identifications that have been reported per publication (Fig 2.1 p. 72). A severe limitation on protein identification in *Synechocystis* is the presence of small, high-abundance phycobili-proteins (Gan et al., 2005). These proteins are integral to the light harvesting antennae, which account for around 40% of all proteins in the cell or 20% by weight due to their relatively small size – which can be observed on a standard protein poly acrylamide gel electrophoresis (PAGE) analysis as shown in figure 2.2 (p. 72). The focus of these studies has also been changing with time, ranging from initial investigations into the membrane eventually leading to

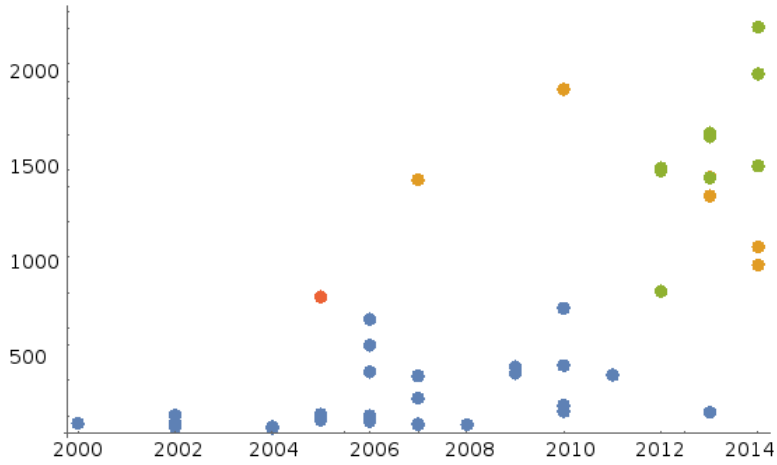


Figure 2.1: The number of proteins identified in each proteomic study of *Synechocystis* per year. All studies that confidently identified more that 1000 proteins are highlighted in yellow, all the green points were conducted by the same lab over the last 4 years, and the point in red was the first study that focused primarily on increasing the number of protein identifications.

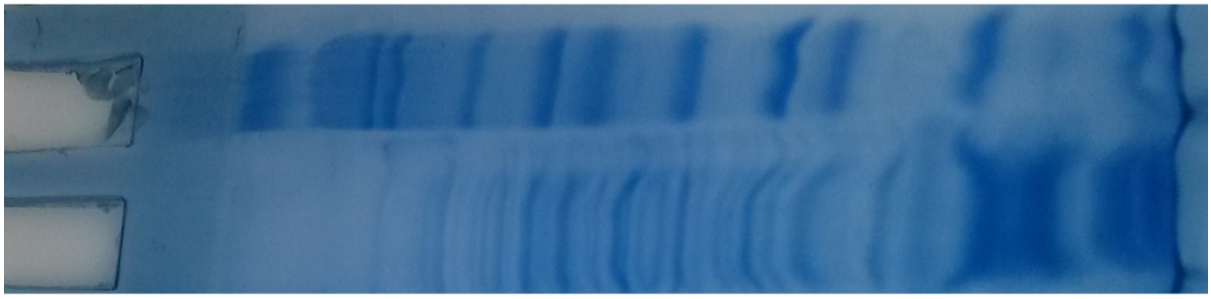


Figure 2.2: A slice of a gel image, showing a size ladder in the top row and *Synechocystis* proteins in the bottom row. The four strongest bands on the gel are the phycobiliproteins, which dominate the protein sample. The blue colour on the proteins here is as a result of dyeing with the Bradford reagent, however the phycobiliproteins also showed as a blue shift on the band whilst the gel was running.

much broader-reaching studies investigating technical improvements to techniques and systematic development of the organism for biofuel production (figure 2.3 p. 73). There have been several limits that have been explored in proteomic investigations, as detailed in table 2.2 (pg. 74).

2.2.4 Synechocystis proteomic studies

To date, only around 65% of the proteome of *Synechocystis* has been successfully identified through mass spectrometry, although this is higher than the number of annotated genes in the genome.

By focusing on publications that have reported the highest numbers of identifications,

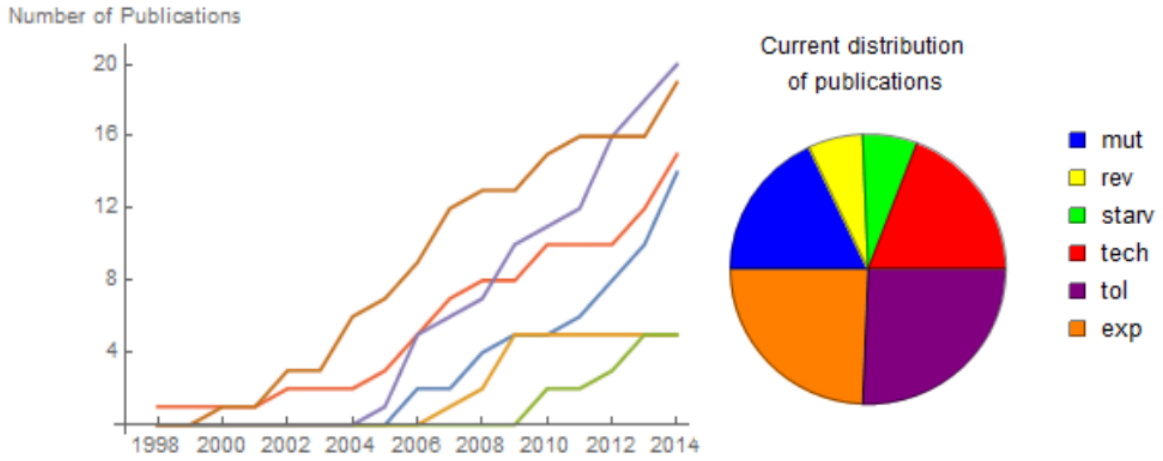


Figure 2.3: The topics being published in *Synechocystis* over time. Each publication was given a tag based on the topic, which are as follows: mut – Mutant study; rev – Review; starv – Stress, Starvation; tol – Stress, Tolerence; tech – Technical improvement study; exp – Exploritory Studies. In cases where a publication addresses multiple topics, it was assigned multiple tags, to reflect current knowledge and direction of interest in the field. No single publication was given multiple counts of the same tag, regardless of the size of the study.

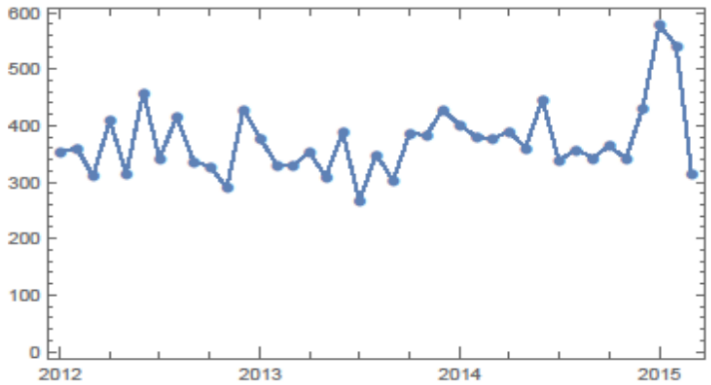


Figure 2.4: The overall number of publications per month since January 2012. The levels remained largely steady until a large spike in publications at the beginning of 2015 – the dip at the end of the graph results from an incomplete set of data, as this was collected over the first 16 days of March, 2015; suggesting a continuation in the publication trend. This figure is taken from the supplementary materials of (Landels et al., 2015).

Table 2.2: A table listing the current limits that have been investigated with a proteomic study in *Synechocystis*. Whilst these data cover a broad base of topics – many of which are discussed in more detail in the following sections of this section – as can be seen from figure 2.1 (pg. 72), a number of these studies conducted before 2010 may be of limited use compared with data that could be obtained with better analysis capabilities.

Condition	Limit
Temperature	Low: 20 degrees, High: 38 degrees Shock: 44 degrees
pH	Low: 5.5 High: 11
Complete starvation	Nitrogen – 6 days Phosphorus – 6 days Sulphur – 6 days Iron – 6 days
Longest study	Low phosphorus – 60 days
Light	Highest intensity light: 300 mmol photons/m2.s Highest intensity UV: 1 W/m2 Lowest intensity: Complete Darkness
Biofuels	0.25% Butanol 2% Ethanol 0.9% Hexane
Salt	High: 6% NaCl (9 days)
{CO2}	High: 3% CO_2 Low: air level
Metal Toxicity	Cadmium 40 μM Cobolt 40 μM Nickel 40 μM
Metal Tolerence	Cadmium 40 μM

we highlight techniques and features of proteome investigation that yielded the highest degree of success. The single point in red is the first investigation into improving protein identification in *Synechocystis*, performed by (Gan et al., 2005). The authors highlighted the importance of separating out the membrane and soluble fractions of the proteome prior to digestion, to enable solubilisation and digestion of the membrane fraction before recombining the two fractions together. Here a total of 776 proteins were positively identified with isoelectric focusing for protein and peptide fractionation prior to reverse phase giving the best overall number of positive proteins identifications (Gan et al., 2005). The technique is now a standard in this field.

In Figure 2.1 (p. 72), the points highlighted in green all reference the same protein extraction protocol from the Zhang group (Qiao et al., 2012a), and are produced in the same lab using the AB SCIEX Triple ToF 5600 spectrometer. These publications provide both qualitative and quantitative data through the use of the isobaric tagging agent iTRAQ. The most notable difference between these studies and others in the field with far fewer positive protein identifications appears to be the use of the high quality mass spectrometer. Whilst this is clearly an effective solution to the problem, it is unhelpful to labs that lack the resources required to purchase more expensive, powerful machines.

The orange points are all other published studies that have produced more than 900 unique protein identifications. The earliest *Synechocystis* publication to exceed this threshold was a bioinformatic-based protein identification investigation (Ishino et al 2007). This study was notable as instead of using offline LC separation, proteins were size selected using electrophoretic mobility and in-gel digestion before analysis with reverse-phase LC-MSMS. This qualitative method is affordable and highly effective – reporting 1442 unique proteins with good confidence but has yet to be used successfully with quantification methods, such as isobaric tagging.

In 2010 Wegener et al identified almost 2000 proteins ((Wegener et al., 2010). This study combines the output of 12 different conditions, each double-injected into the mass spectrometer. This dramatically increases the number of total identified proteins, suggesting that this comprehensive method of obtaining data from samples isolated from a variety of conditions is sufficient to increase the level of proteome coverage at the expense of time. The three remaining highlighted studies that exceed the selection criteria utilise depletion methods; one at the cell-pellet extraction level where phycobili-proteins are washed out of the membrane fraction (Zhang et al., 2013a) and the other two through post-digestion peptide purification – in this case a side-effect of preferentially selecting cys-containing peptides (Guo et al., 2014) or phospho-peptides (Talamantes et al., 2014).

There is evidence to suggest that the reduced protein identification is a stochastic feature. The study by Wegener et al doubled the number of unique protein IDs that had been made

at that time whilst identifying the majority of already identified proteins (Wegener et al., 2010). Notably, their work combined data from a number of different stress conditions with data from standard conditions. Low-abundance proteins in *Synechocystis* can be considered to fall below a 'stochastic selection threshold'. Whilst it is feasibly possible to identify a unique peptide from one of these proteins, it is highly unlikely that the second unique peptide required for confident identification would be found. In the Wegener study, stress proteins that would normally be of low abundance became much more prevalent, therefore enabling identification. By running multiple comparisons and producing much more data, they lowered the 'stochastic selection threshold' for the peptides generating large numbers of identifications.

2.2.5 Post-translational modification proteomics studies

Advancements in proteomics technology and high accuracy mass spectrometry can now lead to the identification of thousands of phosphorylation sites in a single experiment from a eukaryotic organism (Collins et al., 2007). The significance of Ser/Thr/tyr phosphorylation of bacterial proteins has been an advancing research area in recent years (Jers et al., 2008; Soufi et al., 2008) and bacterial phosphoproteomics is picking up momentum with global studies in a vast number of bacterial species. In the last 2 years 3 separate studies have been published centred on the phosphoproteome of *Synechocystis* 6803 (Spät et al., 2015; Lee et al., 2015; Mikkat et al., 2014a). Two of the studies were conducted on WT *Synechocystis* grown in standard conditions, and whilst whole-proteome analysis identified as much as 5% of the proteome containing putative phosphorylation sites, determined by both data analytics on the protein sequences and staining within a 2 dimensional gel experiment, less than 1% of the proteome was found to actually be phosphorylated under standard growth conditions (Lee et al., 2015; Mikkat et al., 2014a). The third study investigated how phosphorylation changed under nitrogen starvation and found a significant increase in the number of detected phosphoproteins – up to the complete 5% of predicted phosphorylation sites over the course of the investigation (Spät et al., 2015). Phosphorylation events were found to be much more frequent under nitrogen starvation, with the number and frequency of phosphorylation events increasing after 24 hours.

Phosphorylation is a much more rapid and efficient response method than protein turnover, a fact that is particularly prominent when nitrogen – a key building block of protein – is limited. As mentioned earlier, within *Synechocystis*, the antennae proteins represent the largest single usage of nitrogen within the cell: they have a fast rate of turnover due to reductive damage caused by photobleaching and the generation of reactive oxygen species, and they make up a huge proportion of not only the protein mass but the total mass of the cell (15% of each cell by mass). It is therefore unsurprising that the most sig-

nificant phosphorylation control event in *Synechocystis* under nitrogen starvation occurs on the photosystem II control protein(Spät et al., 2015). In addition, this same effect has also been observed in other another related cyanobacteria, *Synechococcus sp.* PCC 7942 (Schwarz and Forchhammer, 2005). Nitrogen limitation has been shown to increase glycogen formation (Joseph et al., 2014b), although this effect is also observed when the antennae structures are truncated genetically and nitrogen is not limited (Joseph et al., 2014a).

In addition to phosphorylation, both acetylation and glutathionylation have been identified as post translational modification systems operating in *Synechocystis* (Mo et al., 2015; Chardonnet et al., 2014).

2.2.6 Proteomics stress studies

A number of proteomics studies have been performed using *Synechocystis* cultures that are submitted to stress conditions (summarised in Table 1). Almost all of these use 2D-gel protein separation with specific image analysis software to identify changes in protein levels. Once excised protein spots are analysed by mass spectrometry analysis for protein identifications. Here were summarise the key studies centred on proteomics of *Synechocystis* under a number of different stress conditions.

Understanding the different cellular responses is integral to correctly interpreting proteomic data. A large number of studies conducted in omics tend to be 'exploratory', in the sense that the dataset is considered to be a measurement of responses in the same way that phenotype is used in genetic studies. Unfortunately, this approach is flawed, because where the phenotype is considering the cell as a complete unit performing a terminal function – as the fermentation example given in the introduction; the proteomic response is dynamic due to the internal cellular environment, where an over-abundance of something will drive forward a different reaction. If it were possible to measure the complete dataset, such as is the case with nucleic acids, it would be possible to interpret at least the complete change to the system; however even in this case the analysis is still limited because the functions of many of the proteins are either incompletely understood or completely absent. In other words, looking at one or two proteins with related functions being found to be less abundant and stating 'the cell is changed in this way' is incorrect, without making some other empirical observation.

That is not to say that proteomic observations of stress responses are without merit. Whilst they cannot be used to make direct statements about cellular function, they can be accumulated to give a comprehensive understanding of how the cell is physically functioning at a given point in time. In addition, they are the only omic measurement

form that directly investigates the functional aspects of the system; other methods make inferrences that are either prediction-based (transcriptomics) or response-based (metabolomics).

2.2.7 Temperature and light

As mentioned above, the photosynthetic machinery of *Synechocystis* is the largest protein constituent by a considerable margin, and as a result of photobleaching – where high intensity light regimes result reductive damage to this machinery and the generation of reactive oxygen species (ROS) – the turnover of these proteins is also high. The photosystems, particularly photosystem II, are particularly vulnerable to thermal and light stresses. *Synechocystis* has been shown to grow at 34 °C, albeit at a reduced rate, and to survive heat shock treatment of up to 44 °C.

In 2006, Slabas et al performed analysis on the soluble proteome of wild-type and a thermal tolerance mutant (δhik34) of *Synechocystis* PCC 6803 (Slabas et al., 2006). Although the study only identified 66 altered proteins, they found that the mutant demonstrated improved protein degradation and re-folding effects, both of which are important under thermal stress conditions that can cause errors in protein synthesis. Whilst proteomic technology in *Synechocystis* at the time of this study was limited, follow-up microarray studies verified these findings at the transcriptional level (Suzuki et al., 2006; Rowland et al., 2010). It found that the mutant was transcriptionally in a similar state to the heat-treated wild type, an effect that was amplified when exposing the organism to high temperatures for 60 minutes.

An additional response relating to cellular energy supply has also been found (Fu and Xu, 2006), where the upregulated proteins are also involved in making glycogen available as glucose to the cell for respiration. This switch is important in photosynthetic organisms, as a disruption to the availability of cellular energy can have consequences for survival and growth. It is important to note that this fundamental switch between energy supply methods is typically the most prominent effect observed in *Synechocystis* studies where the photosystem has been affected in any way, which is typically all stress studies.

In the absence of a heat shock, sustained higher temperatures have also been investigated in the context of thermo-tolerant mutants and WT under high light intensity in a study that measured over 1200 proteins in 2D gel electrophoresis analysis, and found approximately 30 significantly changed proteins (Miranda et al., 2013). This study highlighted another category of proteins, small chlorophyll A / B - binding proteins, played a role in thermal response when coupled with high light intensities (1000 $photon.\mu mol.m^{-2}.s^{-1}$). These proteins demonstrated a number of responses focused around the membrane pro-

teome. In addition, the metabolic response was also assessed, and interestingly whilst the previous study showed that stress triggered genes that led an increased availability of glucose, and verified this with an enzymatic assay, during high light-stress at high temperatures the entire cellular metabolism was found to be depressed.

Exposure to high intensity ultra violet B (UV-B) radiation over the short to long term (8 - 96 hrs) has also been measured with proteomic methods (Gao et al., 2009). Understanding prolonged UV-B response in *Synechocystis* is useful for outdoor applications, such as large scale PBRs. As this was a purely proteomic investigation, it generated a number of different leads for further investigation, but could only make broad statements about the general functions of those proteins. Interestingly, different proteomic expression profiles were identified for the short term and long term exposure regimes. Broadly, long term exposure suggested a depression of the photosystems and metabolism, and an increase in cellular defence and repair mechanisms.

2.2.8 pH

During photosynthesis, the media surrounding *Synechocystis* has been found to transiently raise in pH to levels as high as pH 10. This pH change is often attributed to the uptake of CO_2 from the media, however this belief is likely held because adding additional CO_2 will acidify the media and bring it back down to physiological levels. Conceptually, simply removing dissolved CO_2 from water alone should not be capable of bringing the pH to a level higher than that of distilled water, which is over 100 fold lower than the levels observed during photosynthesis. Instead, this pH rise is more likely to be attributed to the photosynthetic activity directly. During photosynthesis, protons are actively pumped into the plasma membrane until the point where the pH gradient is too high for the reaction to continue. At night, or in darkness, when this process stops, the pH drops back down to levels closer to pH 7. This natural variation in pH that the organism experiences gives it a good tolerance to a relatively wide-range of pH values, ranging from pH 7 to 10, although the growth rate is diminished with increasing pH due to a reduction in photosynthetic efficiency and significant cell damage is observed at pHs higher than this range (Touloupakis et al., 2016). A proteomic study has been conducted on the outer plasma membrane under pH 11, comparing it to pH 7 (Zhang et al., 2009). It found a significant upregulation in the amount of ATP-binding cassette transporters, which in a transcriptomic study would suggest replacement and repair, but in a proteomic study suggests reduced transport efficiency or reduced protein turnover. Distinctions like this are important when combining different levels of 'omic data and should be considered in any multi-omic study.

Low pHs are generally rare in the natural growth regime for *Synechocystis*, and so the

organism shows reduced growth rates below pH 7, and death of the culture within 120 hrs at pH6, and 24 hrs at pH 5.5 and below (Kurian et al., 2006). A study investigating non-transient proteome changes under pH6 found that whilst the periplasm experienced significant changes, internal stress responses to reduced pH were not detected within the cytoplasm. This suggests that the periplasm is robust enough to tolerate acid pH above 5.5, at least on a temporary basis. Significant changes were observed in the periplasm, however. Upregulation was found in proteins related to managing the byproducts of photosynthesis, such as carbonic anhydrase which converts CO_2 to carbonate. It appears from these studies that cell death in this range was due to a general lack of available energy, due to inhibited photosynthetic activity (Kurian et al., 2006).

2.2.9 Biofuel tolerance

As mentioned in the introduction, evaluating the capabilities of a specific organism is very important when attempting to produce a system as a production system. When considering *Synechocystis* as a suitable producer of biofuel, it is important to test whether the presence of the fuel will cause damage to the cells before engineering a mutant that is capable of producing them.

A number of studies into the proteomic responses of *Synechocystis* to the presence of biofuels have been conducted recently by the Zhang group, specifically growth in the presence of hexane (Liu et al., 2012), butanol (Zhu et al., 2013; Chen et al., 2014b) and ethanol (Qiao et al., 2012a; Wang et al., 2012b). All these studies used high quality iTRAQ LC-MS/MS technology to investigate proteomic changes along with a modern advanced mass spectrometer, focusing on the changes in the regulatory mechanisms. These datasets are therefore very high quality, and in some cases have been further verified with other 'omics datasets. The Zhang group has therefore provided an excellent resource for an informatics-based approach into the possible use of *Synechocystis* as a liquid biofuel production platform and systemic stress responses.

A major issue with using *Synechocystis* as a conventional biofuel producer is that the organism demonstrates remarkably low tolerence to organic biofuels. Raising the extracellular concentrations to as little as 1.5% ethanol, 0.8% hexane and 0.25% butanol have impeded cellular growth, which places serious doubts on how useful *Synechocystis* would be as a production chassis for their production. In addition to this, the cellular response in all cases has been found to be complex and multi-faceted; which blocks a simple 'single mutant solution'. This is likely as a result of there being no pre-existing metabolic pathway in place to either degrade or compartmentalise these small organic molecules. As *Synechocystis* lacks internal compartments, the most effective engineering solution would need to examine how to eject the biofuel into the media from within the cell, and how to

block it from re-entering.

When considering biofuel production, this evidence strongly supports the augmented production of naturally occuring fuel molecules in *Synechocystis*, such as palm oil or biohydrogen. Whilst additional engineered solutions may be possible, the complexity involved in such a solution would likely require too much initial investment and have too many potential points of failure to be economically viable.

2.2.10 Starvation studies

The effect of carbon dioxide limitation was evaluated interrogating *Synechocystis* proteome at 6, 24 and 72 hours using isobaric tags for relative and absolute quantification (iTRAQ) (Battchikova et al., 2010). 19% of *Synechocystis* proteome was identified but only 76 proteins were observed to be up-and down-regulated when cells were shifted from 3% carbon dioxide to air levels of carbon dioxide at the time points above mentioned. Under limiting carbon dioxide environment, major changes in inorganic carbon uptake, carbon dioxide fixation, nitrogen transport and assimilation and protection of photosynthetic machinery from excess of light were observed. Conversely, acclimation to low carbon dioxide down-regulated chaperones biosynthesis indicating that oxidative stress is not induced under this limiting condition. Small changes were observed for proteins belonging to phycobilisomes, photosynthesis complexes and translation machinery (Battchikova et al., 2010).

Label free proteomics was employed to investigate the metabolic response of *Synechocystis* to short and long term exposure to 33 different environmental conditions (Wegener et al., 2010). In total 1955 protein (53% of the predicted proteome) and 1198 proteins were identified and quantified respectively. From the quantified proteins, 306 proteins were found regulated upon carbon dioxide depletion, 548 were found regulated upon nitrogen depletion,349 proteins were found regulated upon phosphorous depletion, 390 for sulfur and 392 with iron depletion. Although different conditions were investigated, a general trend in protein expression was observed and particularly affected were proteins involved in amino acid biosynthesis, glucose metabolism, TCA cycle and cytochrome b6f complex which were found highly up-regulation. Moreover, in all conditions tested, cells adopt similar metabolic behaviour and a common stress response was the activation of atypical pathways for acquisition of carbon and nitrogen from urea and arginine. A comparison between transcript (RNA level) from an independent study and protein expression from this study revealed in general, poor agreement between transcript and protein metabolism. From all comparisons, this correlation was lower for sulfur depletion and highest for nitrogen depletion. Nevertheless, stress specific genes showed similar expression patterns in both transcriptomics and proteomics. Overall these results shows

that different perturbations generate a common response; nitrate uptake is reduced and arginine metabolic enzymes are 1.5-3-fold regulated compared with controls suggesting an active arginine deiminase pathway in cyanobacteria. In this study, key enzymes for amino acids biosynthesis were overall strongly up-regulated suggestion they cells try to maintain a metabolically active eventually to activate alternative nutrients for growth. Levels of proteins involved in photosynthesis were not significantly affected; only the efficiency of photosynthetic light reactions resulting in lower ATP production was reduced (Wegener et al, 2010).

2.2.11 Salt stress

The dynamics of the proteome during salt acclimation (NaCl concentration 684 mM) was investigated by two dimensional gels followed by protein identification by peptide mass fingerprinting (Fulda et al., 2006). From in-gel digestion, 500 proteins were identified and 55 were induced under salt shock and after long term salt acclimation. While protein synthesis was nearly blocked, photosynthesis was reduced by 60% of the value in the control cells. However, salt acclimation activates ABC-type of transporter for compatible solutes and the photosynthetic and respiration systems are tune by exchange electron transport activity through photosystem I and the cytochrome oxidase activity. Moreover, salt stress also activates heat shock proteins to protect and repair proteins under stress, DNA-binding stress protein and RNA-binding proteins and proteins involved in defence against reactive oxygen species which are known to be inducible at mRNA level in high light-stressed cells. In total, 45 proteins showed a greater staining intensity, on 2D-gels, and accumulate more than 2-fold; these proteins belong to the salt specific stress proteins i.e. proteins involved in the synthesis of compatible solutes, general stress, enzymes of the basic carbohydrate metabolism and hypothetical proteins (Fulda et al., 2006).

Li et al 2012 relate acclimation mechanisms regulation to histidine kinases as a sensor involved in the salt perception by using genetic mutations and DNA arrays. Hik33 is one of salt sensors its mutation revealed 26 and 28 differentially regulated proteins under normal and stress conditions. Regulated proteins were found to be related with plasma membrane rearrangements due to the Hik33 lost (Li et al., 2011).

2.2.12 Integrated ‘omics studies in *Synechocystis*

As described in the introduction, the next generation of ‘omics analysis will require the integration of multiple ‘omics datasets to generate a fuller set of data relating to all the internal cellular functions. Experiments where ‘omics measurements are made independently of each other have been demonstrated to be of limited to no use for understanding

interactions within a system (Schwanhausser et al., 2011); and so only experiments that combine multi-omic analysis in a single experiment will be detailed in this section – a summary of achievements in different systems-level 'omic assessments of *Synechocystis* has been recently reviewed by hernandez et al (Hernández-Prieto et al., 2014). Notably, a high-profile meta-analysis analysis of all transcriptomic work that has been performed in *Synechocystis* to date was also recently published, demonstrating that the majority of regulatory effects within the wild-type organism were strongly associated with photosynthetic responses and showed distinct patterns in response to (Hernández-Prieto et al., 2016). This is an new area of study, and so there are relatively few studies that integrate proteomics with other 'omics analysis.

A number of integrated proteomic and transcriptomic studies have been performed to date; all of which were conducted in the Zhang lab (Zhu et al., 2013; Song et al., 2014; Qiao et al., 2012a; Gao et al., 2015; Pei et al., 2017). Within these studies, the transcriptomic aspect of the analysis was performed using RNA-seq. The studies range from identifying general proteomic or small-RNA responses to biofuel production stress (Song et al., 2014; Pei et al., 2017), to an investigation of nitrogen stress (Huang et al., 2013) and even a network-level method for categorising hypothetical proteins (Gao et al., 2015). Despite these numerous studies, there is still a requirement within the field for a study to investigate the relative rates of protein synthesis and RNA stability (Schwanhausser et al., 2011) – until this is completed, creating an accurate link between protein and transcript data will remain a challenge for the field.

One of the key metabolic flux experiments first performed in *Synechocystis* was the transient $_{13}$C metabolic flux analysis; where Shastri spiked in $_{13}$C labelled bicarbonate as an auxotrophic carbon source and monitored the uptake in a number of monitored metabolites (Shastri and Morgan, 2007). This was then analysed by Young et al., who identified a short-circuit in the central carbon metabolism pathway that fundamentally changed future metabolic models of *Synechocystis* (Young et al., 2011). The earliest metabolomics-integrating proteomic study was performed by Miranda et al in 2013 – described above in the heat and light stress section of this chapter (Miranda et al., 2013). Whilst they did not directly integrate the metabolite and protein data, they did use gene cluster networks – determined from prior microarray experiments – to map their proteins into a network.

A more comprehensive combination of multi-omics data was performed by Ren et al in 2014 in the Zhang lab, where a comprehensive proteomic assessment was performed, backed by targeted metabolomic and transcriptomic analyses under acid stress (Ren et al., 2014). This was particularly effective, because the mutant being analysed showed only a small number of significant protein changes; and so targeted transcript measurements by RT-qPCR were feasible. The transcript measurements generally showed consistent magnitude and direction of change to the protein measurements; however not in every

case. This is not surprising, given the findings of Schwanhauser et al., that there is better correlation between cases when the protein expression rates are considered as well (Schwanhausser et al., 2011). Ultimately, the researchers successfully mapped a small network of responses to acid stress and improved understanding of how the cells react to reducing the pH.

Finally, the most recent metabolite integrating study was also performed in the Zhang lab, this time looking at 3 hydroxypropanoic acid (3-HP) – a key metabolite in the synthesis of a number of different high value products (Wang et al., 2016). The process of this study followed closely the previous study from the same lab, monitoring the levels a number of 'key' metabolites in a targeted metabolic analysis; then tracking the gene expression profiles of genes of interest with RT-qPCR. They highlighted a number of targets for future bioengineering, such as requiring a higher overall energy profile within the cell and highlighting stress responses that were activated in response to increased 3-HP production.

Including metabolite data in this manner, be it targeted or more broad-spectrum, will be key when attempting to engage industrial practitioners in proteomic analysis of *Synechocystis* – or indeed in production strains in general (Landels et al., 2015). This need is covered in the next section of this chapter, and so will not be discussed further here, but generating repeatable practical observations that are useful on an industrial level remains a challenge.

2.2.13 Concluding remarks

The work in this section demonstrates the broad range of investigations being made into the *Synechocystis* proteome. These range from stress responses, to physiological shifts between normal environmental conditions, to key studies related to industrial biotechnology. This background research is key to predicting the changes that should happen under different environmental cases; an essential step in understanding the true effects of a given change in the context of proteomics.

The number of papers described here show that whilst there is a substantial body of research related to proteomic studies within *Synechocystis*, the rapid rate of advancement in the field – particularly with regard to the number of proteins being identified in each study, suggests that there is scope for the repetition of classic proteomic studies that have already been conducted in *Synechocystis*: particularly in experiments where a whole-cell response to a stress or tolerance factor has been described, as was done by Wegener et al in 2010. The potential to increase both the quantity and quality of established data alone justifies the additional costs of the repetition. In addition, the majority of 'omic

level studies currently combine between two and three replicates on observations – which whilst conventional for scientific research, is realistically insufficient when working with such large datasets that are prone to non-independent variation that confounds classical statistical analysis.

If accurate bioinformatic models are to be completed in the future, the community will require repeated experimental verification of results at independent institutions. In addition, experimental designs including overlaps between existing data, for which there is a large amount of information, and new data, which need to be interpreted. Comparing everything to the Wild Type is insufficient when the normal operational features of the Wild Type are still elusive.

2.3 Advances in proteomics for production strain analysis

Authors: Andrew Landels, Caroline Evans, Josselin Noirel and Phillip C Wright.

Highlights

- Proteomics is widely used in production strain analysis
- The value of specific strategies is discussed with reference to case studies
- Methodologies often based on prior application to eukaryotic systems
- New developments target quantitative accuracy and proteome coverage

2.3.1 Abstract

Proteomics is the large-scale study and analysis of proteins, directed to analysing protein function in a cellular context. Since the vast majority of the processes occurring in a living cell rely on protein activity, proteomics offer a unique vantage point from which researchers can dissect, characterise, understand and manipulate biological systems. When developing a production strain, proteomics offers a versatile toolkit of analytical techniques. In this commentary, we highlight a number of recent developments in this field using three industrially relevant case studies: targeted proteomic analysis of heterologous pathways in Escherichia coli, biofuel production in *Synechocystis* PCC6803 and proteomic investigations of lignocellulose degradation. We conclude by discussing future developments in proteomics that will impact upon metabolic engineering and process monitoring of bio-producer strains.

2.3.2 Introduction

Chemical biotechnology is a field directed to harnessing living organisms as cellular factories, for bio-based production of small molecules and polymers (Fischer and Schaffer, 2014; Becker and Wittmann, 2015). These biological production systems are less well understood than traditional chemical engineering processes due to their inherent complexity. As a result, advanced molecular techniques like proteomics are required to engineer more efficient processes and develop new applications.

First defined in 1995 as a portmanteau of 'protein' and 'genomics', proteomics is the large-scale study of proteins within a cell, tissue or organism (Wasinger et al., 1995). It is a rapidly evolving field focused on identification and characterisation of these proteins and their proteoforms (isoforms and post translational modification (PTM) variants). Quantitative methods in proteomics have enabled comparative analysis of protein expression profiles, typically providing 'snapshots' of cells and proteins in different stages of bio-production. Recent studies have also measured protein turnover by determining rates of protein synthesis and degradation. These techniques offer a means to gain information on mechanisms of bio-production for purposes of optimisation and process monitoring. To date proteomics has found application to well characterised strains such as E. coli (Wisniewski and Rakus, 2014), emergent bio-producer strains like the cyanobacteria *Synechocystis* PCC6803 (herein referred to as *Synechocystis*) (Chen et al., 2014b); as well as metaproteomic analysis of mixed microbial communities (Abraham et al., 2014).

Bibliometric analysis (see supplementary material) of recent proteomics publications has highlighted a couple of key trends in producer strain studies: Proteomics in producer strain analysis tends to focus much more on understanding mechanisms and responses, or suggesting molecular pathways, indicating that in general production analysis is lagging behind the general trend toward targeted proteomics (fig. 2.5, p. 87). We cover the topic of targeted proteomics in more detail below and highlight a small number of cutting edge studies in our first case study.

In this commentary, we present a typical approach for conducting a proteomics experiment, highlighting key terms and concepts. We then outline novel proteomics approaches using post-2012 examples, focusing on three industrially relevant case studies: a method-specific approach, a strain-specific approach and a process-specific approach, concluding with a discussion of the impact of recent developments in the field.

Figure 2.5: A selection from a rank-plot of the 200 most frequently used words in abstracts of production-strain proteomics publications, ranked by frequency. Words in blue are higher-ranked in production-strain proteomics than in proteomics in general; words in red are lower ranked, words in black have not changed relative position. Faded words have changed rank by 5 places or fewer and words in bold-face are only present in the production-strain list. The solid line indicates the change in rank. This figure presents a snapshot of the full list, which is available in the supplementary material.

2.3.3 Proteomic analysis pipeline

High-throughput proteomic methods commonly used in biotechnology approaches utilise the 'shotgun' or bottom-up technique (Yates, 2013), where the proteome is digested into peptides that are typically 5 – 14 amino acids long. Whole proteins or larger polypeptides can also be analysed (top-down, middle down respectively), but this strategy has several technical issues detailed elsewhere (Zhang et al., 2013b). A digested proteome is complex, containing thousands of peptides with varying abundances. The mix requires fractionation, typically using offline high performance liquid chromatography (HPLC) or in solution isoelectric focusing, which splits the single sample into lower complexity fractions. Doing this collects together peptides with similar features – such as hydrophobicity, charge state or isoelectric point – and significantly improves the quality of the final data. Samples are then subject to nano-flow reverse phase HPLC, coupled directly to mass spectrometer. This process is referred to as MS-MS or MS2.

The mass spectrometer (MS) initially scans the masses and intensities of all eluting peptides from the HPLC, on a scale of seconds to milliseconds, this is an MS 'survey scan'. Eluting peptides are then selected for fragmentation from the survey scan, either in a data dependent (DDA) or data independent (DIA) acquisition mode. DDA targets a specific peak from the survey scan for further analysis, whilst DIA fragments all ions from the survey scan simultaneously. The data is then analysed computationally to identify and characterise the proteome. Two detailed reviews provide further information, Altelaar et al (Altelaar et al., 2013a)9 provide an overview, whilst Zhang et al (Zhang et al., 2013b) cover the topic more comprehensively.

2.3.4 Approaches in proteomics

Proteomic approaches can be subdivided into discovery and targeted modes, for characterisation of the proteome and analysis of an identified subset of proteins respectively. Their application and relevance are outlined in the case studies. Examples of workflows, gel and non-gel based, together with their major benefits and drawbacks and examples of their application to bio-producer strains are outlined (Table 2.3, p. 89). Classical proteomics employs two-dimensional electrophoresis (2DE) for analysis of expression profile, where protein identification requires MS as a second step. Gel free quantification methods achieve protein identification and relative quantification using MS with significant advantages over 2DE (Gygi et al., 2000). Protein and peptide labeling methods, metabolic (eg SILAC) or chemical labels (eg iTRAQ) have been widely employed in proteomics but 'label-free' methods are increasingly gaining in popularity (Evans et al., 2012; Christoforou and Lilley, 2012). To date, no direct comparison of all these techniques has been reported (Table 2.3, p. 89). Technique selection is dependent on the biological context, number of samples to be processed and compared.

Targeted proteomic approaches are directed to the detection and the precise quantification of specific subset of proteins of interest. This complements discovery proteomics and applications include verification of candidate proteins and process monitoring (George et al., 2015). Quantification is based on detection and measurement of proteotypic peptides that represent the protein, based on unique amino acid sequence. Specificity and sensitivity are both conferred via 'reaction monitoring' for the presence of (co-eluting) fragment ions, linking precursor and product transition information. Application of high-resolution mass measurement and acquisition of full fragment ion spectra have enabled developments, including higher throughput and specificity conferred by parallel reaction monitoring (PRM) as recently reviewed (Lesur et al., 2015).

Inclusion of stable isotope forms of reference proteotypic peptides, at known concentrations, enables absolute quantification. QconCATs (concatenated proteotypic peptide sequences), are custom designed recombinant proteins, which can be metabolically labelled, purified and tryptically digested, to provide a set of standards for absolute quantification of multiple proteins in parallel (Batth et al., 2014). Label free approaches are popular due to limited sample pre-processing requirements prior to analysis compared to label based methodologies. Examples include Intensity-Based Absolute Quantification (iBAQ) and Absolute Protein Expression (APEX), which have been compared for different sample types and MS platforms (Arike et al., 2012; Ahrne et al., 2013; Krey et al., 2014).

Technique	Mode	Example	Brief description	Benefits	Drawbacks	Recent examples
Two dimensional electrophoresis	Discovery	2DE, DIGE	Gel based separation of proteins employing immobilized pH gradients and polyacrylamide gels	A low cost approach to protein separation and sample analysis	Protein identification requires additional MS step	Mikkat et al. 2014 [19] Analysis of the *Synechocystis* phosphoproteome based on visualization of phosphoproteins with a phosphoprotein-specific dye Li et al. 2014 [64] Analysis of lipid-associated pathways in chlorella using DIGE
Metabolic labeling	Discovery	SILAC, 15N	*In vivo* protein labeling. Use of heavy variants allows discrimination from unlabeled (light) and thus relative quantification of peptides between samples	Labeled samples are mixed prior to tryptic digestion, minimizing variation introduced during further processing eg subcellular fractionation	Metabolic modification may be required to ensure stable isotope is sole source of specific nutrient eg amino acid, ammonium salt	Ciesielska et al. 2013 [20] Identification of *Starmerella bombicola* proteins associated with production and regulation of sophorolipid production (biosurfactant precursor)
Chemical label	Discovery	iTRAQ, TMT	*In vitro* Labeling at the peptide level, achieves simultaneous protein identification and quantification in multiplex format	Multiplex capability, facilitates inclusion of replicates Applicable to a range of protein samples such as cell pellets, subcellular fractions	Underestimation of fold change	Tang et al. 2013 [21] Metabolic engineering by gene knock out or overexpression of specific enzymes that lead to enhanced biofuel precursor production in *Saccharomyces cerevisiae*. Identification of targets for further optimization. Chen et al., 2014 [5•] Case study – identified Slr1037 as a regulon to provide novel target for transcriptional engineering of *Synechocystis*
Label free	Discovery Targeted	Compatible with both Data Dependent Acquisition and Data Independent (MSE, SWATH) Acquisition approaches	Peptide intensity based measurements, based on peak integration or spectra counting	Multiple samples can be compared	Quantification is achieved from independent sample runs. It is thus dependent on highly reproducible HPLC separations prior to MS	Yap et al. 2014 [22] Analyzing protein alterations associated with physiological changes occurring during different growth phases of *Lactococcus lactis*.

Table 2.3: **Proteomic workflows - Application, Benefits and Drawbacks** Commonly used Discovery and Targeted proteomic methods are outlined with reference to specific applications.

2.3.5 Case Study: Targeted proteomics for process optimisation

A key area of proteomic application is assessment and modelling of heterologous pathways. Whilst assessing how an inserted pathway is affecting the proteomic background provides useful information on how the organism is responding; for pathway engineering purposes it is often more informative to assess either the pathway proteins directly, or a specific subset of the proteome known to interact with it. Targeted proteomic methods like selective reaction monitoring mass spectrometry (SRM-MS) are useful for collecting highly repeatable, high-accuracy, quantitative data. These techniques are gaining popularity in bio-production pathway modelling and optimisation, as well as providing a means of assessment of standard parts and devices in synthetic biology. Proof of concept of this approach has been demonstrated for heterologous pathway expression in E coli as model/paradigm for optimization of heterologous pathways.

SRM-MS has been validated against analysis of red fluorescent protein expression levels in an expression plasmid and output of the tyrosine production pathway, controlled with a variety of different strength constitutive promoters (Singh et al., 2012). These methods can also be coupled with quantification methods that incorporate a standard, such as QconCAT, to generate absolute protein quantification levels (Batth et al., 2014). A major advantage of this is that absolute protein values can be incorporated into kinetic metabolic models alongside metabolite data; whilst relative quantification values – which are more commonly associated with global proteome assessment – cannot (Weaver et al., 2015). In practical production terms, it has also been used to optimise production of biofuels such as isopentanol (George et al., 2014) and biosynthesis precursors like terpenes (Redding-Johanson et al., 2011).

2.3.6 Case study: *Synechocystis* PCC6803

Cyanobacteria are a phylum of photosynthetic bacteria that offer promise in solar-powered bio-production. *Synechocystis* is a fully-sequenced, naturally transformable strain of cyanobacterium that is gaining popularity as a model production chassis (Machado and Atsumi, 2012). It is currently being investigated for a variety of different products including precursors such as isoprenoids and lactic acid from CO2 (Englund et al., 2014; Angermayr et al., 2014), as well as biofuels like ethanol, butanol and hydrogen (Chen et al., 2014b; Qiao et al., 2012b; Pinto et al., 2012b).

When engineering a strain for production, proteomic methods integrated with transcriptome data are used in about 20% of studies. These studies are frequently used to assess cellular response to production stresses (supplementary bibliometric analysis). Proteins and pathways found to be regulated in response to stresses are good candidates for

forward-engineering strategies; either through the more traditional method of managing metabolic flux in the organism, or by understanding and controlling responses. An initial proteomics analysis of butanol stress in *Synechocystis* highlighted multiple simultaneous pathways activating in response to the stress (Tian et al., 2013b), was followed up by a transcriptome study (Zhu et al., 2013). Joint analysis of these data identified slr1037 as part of a butanol-specific paired signal transduction system (Chen et al., 2014b). This was verified with a knock-out, which was more robust to butanol stress whilst maintaining wild-type growth rate under standard conditions.

PTMs are a highly conserved method of regulation in biological systems. Despite the enrichment strategies required for identifying these low abundance features, it is possible to assess PTMs on a systems level (Altelaar et al., 2013a). Three pioneering studies have been conducted in the last two years, cataloguing system-wide PTM responses in *Synechocystis* and demonstrating their role in regulation of photosynthesis and central metabolic pathways (Mikkat et al., 2014b; Chardonnet et al., 2014; Mo et al., 2015).

Despite the advantages of proteomics, it requires tuning to the organism being studied. In *Synechocystis*, photosynthetic antennae proteins make up 20% of the proteome by mass (Gan et al., 2005). This results in a large dynamic range relative to other producer-strain bacteria, such as E. coli, and so the advantage conferred by light harvesting capacity comes with the negative limit to proteomic coverage. This problem exists in any case where a small number of proteins are present at a very high abundance, relative to the rest of the sample such as the case with RuBisCO in plants (Gupta et al., 2015). Work has been carried out to reduce the abundance of these antenna proteins for better production and improved proteomic coverage (Kwon et al., 2013). The dynamic range problem can be alleviated to an extent through depletion strategies and use of high resolution, high throughput MS.

2.3.7 Case study: lignocellulose degradation

Lignocellulose is a complex polysaccharide constituent of plant cell wall. It is a promising substrate for production of molecules like ethanol and lactic acid; however, lignin inhibits the action of many common enzymes by sequestering the cellulose and xylose. This creates a bottleneck in efficient bio-production from this material [34]. Proteomics is at the forefront of deciphering solutions to this problem, either through fully integrated systems analysis or assessment of solutions utilising multiple organisms simultaneously.

Integration of metabolite, transcript and protein data, termed systems analysis, can be used to generate comprehensive models for how a cell is responding. The protein data offers an impression of the current state of the cell, whilst the transcript analysis iden-

tifies responses at high-coverage coverage and the metabolite data give an impression of flux (Huang et al., 2014; Kasavi et al., 2014). Proteomics data from several lignocellulose degradation investigations, including Clostridium and a variety of filamentous fungi, have been integrated with transcript and metabolite data to understand how activated pathways affect cellular dynamics (Schellenberg et al., 2014; Klaubauf et al., 2014). This approach has also been used to assess how xylose as a carbon source affects metabolism in yeast, to design better production strategies (Latimer et al., 2014).

The xylose catabolism study utilises pre-existing models for assembled pathways, and is typical of a proteomic work-flow for pathway analysis where identified and quantified proteins are overlaid on a metabolic model. Models are assembled using pathways found with literature analysis and new models are typically constructed by modifying an existing model from a similar organism. A detailed commentary of advancements in this process is provided by King et al in this issue (King et al., 2015). Where a pre-existing model is lacking, other general exploratory assessments, like principal component analysis, can be used instead to look for general trends in protein expression data (Alonso-Gutierrez et al., 2015). This technique can highlight proteins, or entire conditions, that cluster together depending on the focus of the study; with suitable experimental design this can provide information in lieu of a completed metabolic model. Due to the complexity of even seemingly simple strains, data interpretation still is a limiting factor in systems level analyses. This has led to the use of cutting-edge informatics, such as machine learning, being employed in analysis of the data (Kelchtermans et al., 2014).

Whilst individual organisms have shown effectiveness in degrading lignocellulose, natural systems utilise a combination organisms performing distinct roles to achieve the effect more efficiently (Boaro et al., 2014). The study of these communities through the analysis of the proteins is referred to as 'metaproteomics', where the community is profiled instead of focusing on specific organisms, or specific pathways (Vanwonterghem et al., 2014). A growing number of studies are being carried out using metaproteomics towards the ultimate aim of engineering these systems through 'Synthetic Ecology' (Pandhal and Noirel, 2014). This emerging field will likely be of importance to the bio-production community in the future.

2.3.8 Addressing the challenges and perspectives

Proteomics is a very dynamic area: the proteomic toolkit is constantly expanding with both the development of both novel technologies alongside new uses of existing technologies. In the latter category, development of quantitative proteomic technologies with higher multiplexing capability, neutron encoded TMT and SILAC reagents, improves the multiplexing capability of TMT (Werner et al., 2014; Merrill et al., 2014). This enables

both higher throughput and the additional benefits conferred by increased sample replication. Metabolic labels such as SILAC can be directed not only to expression profiling, but also to protein dynamics (Trotschel et al., 2012). The Super-SILAC approach, involves mixing samples from different conditions to generate an internal standard for cross sample comparison, and can be combined with iBAQ to give absolute copy-number level protein quantitation (Soufi et al., 2015a).

Improved sample preparation strategies improve protein identification rates and coverage. In general, reagents that are very effective at solubilising proteins are incompatible with MS. The technique of Filter Aided Sample Preparation (FASP), allows removal of detergents such as SDS and other contaminants. The FASP technique, first described in 2009 (Wisniewski et al., 2009) has been adapted for improved proteolytic digestion (eFASP) (Nel et al., 2015) and for compatibility with chemical labelling strategies for analysis of proteins (iFASP)(McDowell et al., 2013) or affinity purified protein complexes (abFASP) (Huber et al., 2014). Protein-protein interactions (PPI) is a growth area in proteomic analysis, particularly since proteins typically function in complexes which are temporally and spatially dynamic within the cell (Wright et al., 2013) and analysis of PPI has value in system based network modelling (Kasavi et al., 2014). Novel strategies include enhanced capability of protein-protein interaction analysis by use of affinity enrichment MS, as an alternative to classical 'pull down' approaches (Keilhauer et al., 2015).

In terms of MS instrumentation, performance is continually improving in terms of speed, sensitivity and resolution, such that complete proteome coverage is achievable (Mann et al., 2013). Unlike DDA, DIA methods are inherently not applicable to use of metabolic or chemical labels. This limitation is being overcome with approaches such as NeuCoDIA for multiplex analysis (Minogue et al., 2015). In general, whilst MS instrumentation was directed to operating in discovery or targeted modes, next generation instruments offer both capabilities. For example, DIA methods are enabling discovery and targeted modes of data analysis to be integrated, by retrospective 'MRM like' mining of discovery data for specific peptides. With MS developments there is a need for corresponding capabilities in processing software to fully mine the increasingly complex proteomic datasets.

2.3.9 Conclusions

In this commentary, we have highlighted using bibliometric analysis of the field and specific case studies that proteomics is widely used in production strain analysis. Due to the dynamic nature of the field, there are a number of cutting-edge developments that have improved quantitative proteomics over the review period and are continuing to emerge. These have generated step-changes in technical and conceptual approaches to process optimisation, in both single species as well as microbial communities (Abraham

et al., 2014; Pan and Banfield, 2014).

2.3.10 Acknowledgements

A.L and P.C.W are funded by the European Union Seventh Program for research, technical development and demonstration for funding under grant agreement No 308518, CyanoFactory. C.E. and P.C.W. thank the EPSRC for funding (EP/E036252/1). J.N. is funded by the Fondation Recherche Médicale (FRM: ING20140129444).

2.3.11 Recommended reading

Papers of particular interest, published within the period of review, have been highlighted as either of special interest, or of outstanding interest:

- Of special interest
  - Chen et al, 2014: This study provides a clear example of a follow-up investigation to a proteomics-led discovery, following identification of targets using a combination of proteomics and transcriptomics under butanol stress.
  - Yates et al, 2013: This review provides a background on the progression of shotgun proteomics. This is an interesting account of the historical progression of the field and nicely frames the subject.
  - Batth et al, 2014: This study highlights the creation of a tool for high-speed, high-accuracy quantification of proteins in E. coli. It demonstrates how synthetic biology approaches to proteomics are championing high-accuracy quantification data at the expense of producing a large number of identifications.
  - Latimer et al, 2014: This study uses proteomics and metabolomics to determine how engineered S. cerevisiae catabolises xylose – a sugar created in the breakdown of lignocellulose. It neatly frames how proteomics can be used to determine changes in pathway flux to determine the fate of metabolites in an engineered system.
- Of outstanding interest
  - Zhang Y et al, 2013: This comprehensive and detailed review covers a wide range of proteomics-related topics and discusses depths beyond the scope of this commentary. We highly recommend this review to readers who would like to learn more about a variety of topics in proteomics.

- Altelaar et al, 2013: This review provides an excellent introduction to the topic. We highly recommend it as background reading for readers who are interested by proteomics but are unfamiliar with the subject matter being discussed in this commentary.
- Soufi B et al, 2015: This demonstrates the use of super-SILAC, a cutting edge quantitative proteomics strategy in E. coli under ethanol biofuel production stress.

Chapter 3

Proteomics of hydrogen production

3.1 Chapter Background

This chapter follows a series of experiments performed to interpret the internal proteome-driven effects of a media change on *Synechocystis* under H_2 producing conditions. Growth rates and practical experimental features are given here, as these provide key background information for the proteomic investigation, followed by the proteomic data. The proteomic investigation described here was performed before the improvements listed in earlier chapters were carried out, and was the driving force for the method development work carried out. The same experiment has also been repeated, with the proteomic samples being used as the background for the isobaric tag comparison conducted in chapter 5. A comparison between the before and after states would have been a neat close to this thesis, however as there are so many changes it is not a fair comparison for anything other than recognising that the full body of changes collectively provides a huge improvement, with 345 confident protein identifications with quantification data increased to over 1000 when all changes are implemented.

Where other chapters mainly investigate method development, the work in this chapter represents an example case of the output-driven work conducted throughout the CyanoFactory project. There were more investigations conducted, covering the features described in Chapter 1 of this document; however due to the broad-reaching and varied nature of the different studies, it was necessary to exclude the majority of analyses to generate a defined scope for this thesis. The deliverable reports for the CyanoFactory project are included at the end of Chapter 6 as appendices, to give an overall summary of all the areas that were investigated.

Narciso Couto provided general tuition and assistance with the protein extraction, HPLC and mass spectrometry lab work carried out in this chapter.

3.2 Abstract

Synechocystis is naturally capable of producing H_2 under a defined subset of conditions – notably the absence of O_2, which causes destruction of the active site of the hydrogenase. The levels of H_2 production can be affected by a number of environmental factors, including available nitrogen, sulphur and intracellular oxygen. The Burrows media is a media that was designed through factorial design and optimised for H_2 production in *Synechocystis* by Elizabeth Burrows, a researcher in the US. In this study we investigate the effects of this optimised media on the growth rate and internal rearrangement of metabolic pathways within *Synechocystis*. The major findings show an increase in all carbon metabolism proteins in Burrows media, whilst a reduction in proteins associated

with photosynthesis.

3.3 Introduction

The organisms in the Cyanophyta phylum are considered to be some of the most ancient, and have been attributed to triggering the oxygenation of the atmosphere – enabling the development of modern life as we know it. Like all photosynthetic organisms, they have the capacity to survive both anaerobic and aerobic conditions – despite the fundamentally different energetic basis needed survival in these very different conditions. In many ways, it can be more convenient to think of the organisms within this phylum as not a single, variable organism, but two distinct organisms rolled into a single cell.

Here we attempt to describe changes that take place under isolated conditions, such as light and dark, aerobic vs anaerobic; but in the natural world many of these changes are not independent – darkness stops photosynthesis and so the production of O_2, generating an anaerobic environment – and attempting to isolate individual features when performing a systems-level analysis makes the investigation significantly more challenging. In this case, the question should not be – "how does factor x affect the internal system", but rather "what is the natural state of the system, and how does it respond to factor x". These two approaches may seem similar, but the fundamental difference is that the natural state is not necessarily the lab-grown state; but rather the state in which the organism evolved in and therefore optimised itself to.

3.3.1 Hydrogen production in Synechocystis

As described in chapter 1, under environmental growth conditions outside the laboratory environment, *Synechocystis* is believed to use H_2 as an electron sink under highly reducing conditions. In these cases, expression of the Hox cluster genes is increased, leading to the production of the *Synechocystis* native bi-directional Fe−Ni hydrogenase(Rögner, 2013). As the hydrogenase is bidirectional, it can convert excess electrons into H_2 as a relatively inert temporary storage. In addition, it can also run in reverse to degrade H_2 into electrons for use within the cell. As a result, H_2 measurements will reach an equilibrium after a certain concentration of H_2 is present in the headspace – previous studies have indicated that this state is reached between 6 and 24 hours (Pinto et al., 2012a). This experiment was focused on the initial changes that take place, rather than under equilibrium, as in an industrial production situation the excess H_2 would be need to be siphoned away from the system to ensure the harvest efficiency is as high as possible.

3.3.2 Burrows Media – background

Synechocystis alters metabolic state under a number of environmental factors, as shown in chapter 2. The key factors that affect H_2 production were investigated in a factorial experimental design (Burrows et al., 2008), which demonstrated that when nitrogen and sulphur were limited in the media the rate of hydrogen production was accelerated. In addition, carbon limitation under photoautotrophic growth was found to limit the rate of hydrogen production when the cells were switched to a fermentative state, and so carbon was added in excess.

The nitrogen levels in the media were optimised in a follow-up study (Burrows et al., 2009), as limiting nitrogen under photoautotrophic growth resulted in a reduced growth rate and therefore a reduction in overall nitrogen abundance. Interestingly, whilst the nitrogen levels were optimised in this experiment, they are likely responsive to the growth regime of the organism. Ultimately, there should be an amount of nitrogen present within the media to match the amount of desired biomass accumulation during the growth, so that the limitation effects are experienced by the cell just before they are exposed to H_2 producing conditions. In addition, previous investigation indicated that nitrate can increase the internal levels of cellular oxygen (Baebprasert et al., 2011, 2010), and so ammonium chloride was used as the nitrogen source in the media.

The Burrows media used in this experiment was tuned to this level, based on the experimental growth rates that had been observed in BG-11 media in our growth conditions. This was done by using the biomass equation from the flux balance analysis (FBA) model – generated by our collaborators in Valencia (Montagud et al., 2015) – to find the number of moles of ammonium needed to generate a mole of biomass. The molecular weight of the biomass equation was then used to calculate the number of moles of biomass present in the dry weight of 50 ml culture grown to OD 0.8, and the required concentration of ammonium per litre of media was calculated accordingly.

This media mix has also previously been tested under an enhanced reducing external state, where cyanide was added to the media to drive the production state by excluding oxygen from the cell (Burrows et al., 2009). These conditions were going to be replicated in this study, however due to hazards associated with working with cyanide, along with incompatible safety regulations within the department for the disposal of cyanide-contaminated media, these alterations were not included.

3.3.3 Summary of expectations

Due to the limited nitrogen availability in the Burrows media, the specific growth rate of the cells may hit a plateau if the cells undergo enough divisions to deplete the nitrogen levels in the media. Additionally, there are likely to be effects within the cell driven by detection of the limited levels of nitrogen and sulphur in the environment that will limit photosynthetic repair. There may be a higher rate of nitrogen turnover within the cell, as components need to be recycled into essential proteins for the cell to function; which may be visible as in increase in the overall levels of ribosomal and degradation proteins present. This may also trigger the production of cyanophycin as a storage response to longer-term nitrogen limitation, as is seen in phosphate depletion.

Nitrate assimilation is also a major route of increased oxygen levels within the cell; in the Burrows media the nitrate has been replaced with ammonia, and so it is expected that the ensuing reduced intracellular oxygen levels may lead to a more rapid rate of oxygen depletion and therefore hydrogen evolution (Baebprasert et al., 2011, 2010). If the overall energy levels in the cell are low, it is possible that carbon stores will either be more rapidly degraded compared with the BG11 media; however depending on the temporal state of growth the cell is in and starvation triggers, the metabolic pathways relating to carbon assimilation may be up-regulated.

3.4 Methods

3.4.1 Media comparison

BG11 Media

BG11 media (Stanier et al., 1971) was made up from 9 standard stock solutions. All solutions were made up to volume using de-ionised water unless otherwise stated. Solution 1 was sterilised by autoclaving, however to avoid alterations to the chemical composition of the media all other solutions were filter-sterilised. The stocks were prepared as follows:

1. 15.00 g $NaNO_3$ in 1000 ml.
2. 2.00 g K_2HPO_4 in 500 ml.
3. 3.75 g $MgSO_4 \cdot 7\,H_2O$ 500 ml.
4. 1.80 g $CaCl_2 \cdot 2\,H_2O$ in 500 ml.
5. 0.30 g Citric acid in 500 ml.

6. 0.30 g Ammonium ferric citrate green in 500 ml.
7. 0.05 g EDTA Na_2 dihydrate in 500 ml.
8. 1.00 g Na_2CO_3 in 500 ml.
9. Trace metal solution (per litre):
    - 2.86 g HBO_3
    - 1.81 g $MnCl_2 \cdot 4\,H_2O$
    - 0.22 g $ZnSO_4 \cdot 7\,H_2O$
    - 0.39 g $Na_2MoO_4 \cdot 2\,H_2O$
    - 0.08 g $CuSO_5 \cdot 5\,H_2O$
    - 0.05 g $Co(NO_3)_2 \cdot 6\,H_2O$

The final media composition per litre:

- 100.0 ml stock solution 1
- 10.0 ml stock solutions 2 - 8
- 1.0 ml stock solution 9
- 889.0 ml autoclave-sterilised, deionised H_2O

For $BG11_0$ media, solution 1 was excluded and sterile water was used instead.

Burrows media

The recipe for Burrows media was derived from a combination of several recommendations by a series of publications authoured by Elizabeth Burrows (Burrows et al., 2008, 2011, 2009). These publications detail a number of changes determined by factorial design that lead to increased observable H_2 production in *Synechocystis*. All stock solutions were filter sterilised - please note that autoclaving the final solution will cause it to become cloudy and filled with particulates, which causes problems for further analysis. The stocks were prepared as follows:

1. 19.32 g $NaHCO_3$ in 500 ml
2. 2.00 g K_2HPO_4 in 500 ml
3. 3.10 g $MgCl_2$ in 500 ml
4. 0.30 g Na_2SO_4 in 500 ml
5. 1.80 g $CaCl_2 \cdot 2\,H_2O$ in 500 ml

6. 0.30 g Citric acid in 500 ml
7. 0.30 g Ammonium ferric citrate green in 500 ml
8. 1.00 g NH_4Cl in 1000 ml
9. 0.05 g EDTA Na_2 dihydrate in 500 ml
10. Trace metal solution (As above)

The final media composition per litre:

- 100 ml solution 1
- 10 ml solutions 2 – 8
- 1 ml solutions 9 – 10

3.4.2 Synechocystis growth

Cells were grown in 250 ml shaking flasks each containing a maximum of 50 ml culture to ensure a good exchange at the air-liquid interface. The flasks were stoppered with cellulose bungs and rotated at 150 RPM in an illuminated shaking incubator at 25 degrees Celsius. The light was measured at 40 ± 10 $\mu mol.m^{-2}s^{-1}$ photons ($\approx$ 3000 $\pm$750 Lux) during operation.

For analysis of changes taking place after acclimation to a specific media, the cells were grown in the following regime. Initially, the cells were grown in BG11 for a growth cycle (approximately 2 weeks), to ensure they were in mid-log phase during transfer. The cells were then grown in the target media again until mid-log phase, to avoid transient acclimation changes on moving between media. This also facilitated degradation of nutrient stores within the cells in the case of being transferred to Burrows media. At mid-log, the cells were subcultured again into the target media; then grown to mid-log before harvest and analysis.

Cell growth was monitored daily by measuring 1 ml culture on spectrophotometer in polystyrene cuvettes at A_{730}. These measurements were done in triplicate, and diluted so the optical density (OD) was always between 0.05 and 0.5.

The cells were visually inspected under the microscope at for signs of stress or physical morphology alterations. Cell counts were conducted on OD_{730}1.0 cultures. These counts were performed using a haemocytometer, where 10 μl of culture was added to the haemocytometer, and left for 2 minutes to settle. The cells were counted by 0.05 mm^2 squares, until at least 200 individual counts had been observed. *Synechocystis* cells tended not to separate clearly after division, as can be seen in figure 3.1 (p. 104) and so circular cells

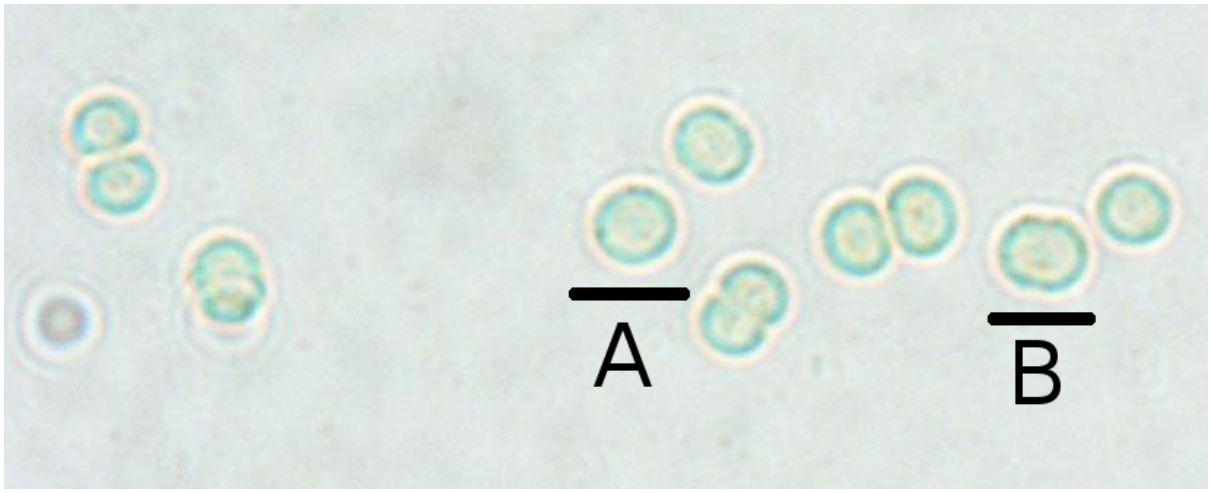


Figure 3.1: *Synechocystis* cells being visualised under a microscope. When counting with a haemocytometer, cells that appeared as type A were considered to be a single count, whilst cells of type B, where a septum had begun to form for cell division, were counted twice. This was done to provide some degree of consistency between cells counted near the start of the measurement, and those counted towards the end or on a recount. These images were taken with a light microscope under a 100-fold magnification, where the bars are approximately 2 microns across.

were counted as one cell, whilst cells that were either a pair or showed the initial signs of septum formation were counted as two. The count was then divided by the number of whole squares that had been counted, and multiplied by 4×10^6 to determine the number of cells per ml. All counts for these measurements were conducted by averaging technical triplicates and reporting experimental quadruplicates.

Dry mass was determined by taking $\approx$ 50 ml culture at A_{730} OD $\approx$ 1 – taking note of the exact volume and optical density for calculations later. The cells were collected by centrifugation and washed in lysis buffer – as described in Chapter 4 – and transferred to a pre-weighed 2 ml eppendorf tube. This was then centrifuged in a bench-top centrifuge for 20 minutes at 17000 RPM (at 4 degrees Celsius) and remaining supernatant was removed by pipetting. The pellet was then flash-frozen in liquid nitrogen and lyophilised in the freeze-drier under vacuum overnight to remove remaining liquid. The eppendorf was re-weighed and the difference between the prior and post states determined the mass of the cells.

Rough whole-cell protein values were measured by taking whole-cell lysate before protein concentration (see methods in Chapter 4) through use of the Bradford assay, although these were found to vary with repeated measurements. Values for % protein were also taken from the literature and other CyanoFactory collaborators – averaging 60%. This was conducted before the investigation into optimal protein quantification methods described in Chapter 4 – no accurate whole-cell protein measurement was determined in

this study, as the methods described were conducted on purified proteins rather than lysate.

3.4.3 Hydrogen production and measurements

For growth in anaerobic conditions, 100 ml serum vials were autoclaved, and filled with 95 ml buffer. These were then heated to 90 degrees Celsius and then cooled to room temperature, whilst being sparged with either Ar or air. The process was carried out using a bespoke set-up designed by Vi Nguyen at Sheffield University and took approximately 15 minutes. On completion of the heat-cool sparge cycle, each of the vials were rapidly covered with a rubber septum to prevent air contamination, and then capped and sealed using a crimping device.

Once the OD_{730} of the lowest concentration flask of cells reaches 1, the OD of each flask was recorded. The volume was evenly split between 2× 50 ml falcon tubes for direct replicate comparison.The culture was spun down (5000 RPM, room temperature, ramp up and ramp down set to 7), the supernatant removed by pipetting (20 ml stripette), and the pellet gently re-suspended in 5 ml degassed dd-H_2O. This concentrated culture was then transferred to a 5 ml syringe and leur-lock needle.

The concentrated culture was injected into the vials using the following technique, to maintain an anaerobic environment and to prevent back-pressure. A second, empty leur-lock syringe and needle was inserted into the vial above the liquid level; the syringe containing the culture was then inserted below the liquid level and the contents were injected. The empty syringe passively filled with the head-space gas during this injection. The syringe that contained the culture was brought above the liquid level but not out of the vial. The plunger on the syringe was kept depressed, and the passive syringe filled with head-space gas was depressed to pressurise the head-space slightly. Whilst this pressure was applied, the empty syringe was removed – preventing an influx of gas. The passive syringe was then depressed as fully as manually possible and rapidly drawn out of the vial to create a positive back-pressure to prevent passive air influx. Parafilm was wrapped over the top of the serum vials to further limit gas exchange. Timings of the hydrogen production experiment were started at the point of cellular injection. Figure 3.2 (p. 106) shows the serum bottles prior to hydrogen collection.

The head-spaces were sampled hourly for H_2 production using a passive collection method. The parafilm was peeled back and a balancer syringe, containing 1 ml carrier gas (Ar) was inserted. This was allowed to passively retract to relieve the head-space pressure. This was depressed slightly to maintain pressure and a collection syringe (a gas-tight GC injection syringe) was then inserted into the top of the sample – the plunger was

Figure 3.2: Samples prior to H_2 sampling. Each serum bottle has been capped with a rubber septum and sealed over with parafilm.

fully depressed to avoid a drop in pressure. Once the collection syringe was inserted, the pressure on both of the syringe plungers was relaxed. The balancer syringe was then depressed to enable passive collection of 1 ml head-space gas in the collection syringe, which was then sealed. The back-pressure was re-applied using the balancer syringe and the collection syringe was removed. The balancer syringe was depressed as fully as possible to replace the back-pressure and was removed as before. The parafilm was then re-applied to prevent gas exchange.

The samples were measured on a gas chromatography system using manual injection as described in (Maeda et al., 2007). H_2 was detected using a thermal conductivity detector maintained at 200 degrees. Ar was used as a carrier gas at a flow rate of 20 $ml.min^{-1}$, with the column temperature at 70 degrees, injector temperature at 100 degrees. Hydrogen retention was approximately 0.46 minutes. The method was used qualitatively, rather than quantitatively. This was done because the focus of the experiment was investigating the cells under a shift to a H_2 production state, rather than an assessment of volumes under different conditions. Notably, very low volumes of H_2 were produced by this method. Peaks associated with N_2 and O_2 were also observed in this method – indicating cases where gas leakage had occurred. Although time-consuming, the experiment was restarted in cases where external oxygen contamination was observed.

Once H_2 was clearly detected in all target samples, the experiment was stopped (this typically took between 2 – 4 hours). The samples were de-capped, decanted into pre-chilled 50 ml falcon tubes and rapidly centrifuged at 4 degrees Celsius for 10 minutes. The pellets were then treated as described in the protein extraction method in Chapter 4. The cold temperatures were intended to prevent the general proteome response within the cell from changing too far. This method failed to detect the hydrogenase; as it was likely destroyed upon contact with oxygen. Direct observation of the hydrogenase proteins would likely require that the experiment was conducted in an anaerobic chamber, the cells collected by rapid filtration to shorten processing time, before being flash-frozen in liquid nitrogen. This experiment was planned, but not carried out due to time and resource constraints.

3.4.4 Experimental design

To ensure statistical robustness in the experiment, all experiments were conducted in quadruplicate – this offers the largest single statistical improvement possible to the classic triplicate method, it is also the most cost-effective statistical improvement possible. The experimental set-up was designed as outlined in figure 3.3 (p. 109). This set-up ensured a close relationship between all samples and potentially clearer proteomic outcomes, although the design was susceptible to failure as the downstream measurements

were dependant on the robustness of the upstream growth and preparation.

As shown, the replicates were then paired up and pooled together. This was expected to improve the quality of observations, as purely stochastic variations between replicates would likely dampen or cancel each other out completely, whilst systematic variations would re-enforce their respective signals, producing a clearer overall set of data for statistical analysis. A side effect of this approach is that the stochastic variation can, in certain circumstances, be attributed to alternative organisation within the cell, and so this method would go further to mask the true state of the cell – however since a cellular population is being studied, rather than an isolated single cell, this issue persists within the populations regardless.

3.4.5 HPLC and Mass Spectrometry

The proteins were purified as described in Chapter 4. All pooling of samples was done at the cellular level before extraction, to ensure homogeneity of sample treatment. The proteins were reduced and alkylated with MMTS, to break and block cysteine disulphide bridges, before being digested with reagent grade trypsin. The peptides were labelled with iTRAQ reagents as per the manufacturers instructions, in the pattern shown in figure 3.3 (p. 109).

The samples were separated by high performance liquid chromatography using a Hyper-Carb column over a 60 minute gradient. The fractions were then pooled as described and separated by reverse phase on either the QStar, maXis or QExactive mass spectrometer; as described in Chapter 5.

3.5 Results

3.5.1 Synechocystis growth

The growth rate for *Synechocystis* was monitored in both BG11 and Burrows media over a period of 6 days, which was found to be approximately the amount of time needed for a 10-fold increase in biomass. A growth curve was prepared, suggesting a doubling time of 44.7 hours in BG11 media; however this was calculated when the cells were grown on a 12/12 light-dark cycle, and as *Synechocystis* do not divide without the presence of light (Heidorn et al., 2011) this rate can effectively be considered as a growth rate of 22.3 hours per division; or a rate of 1.032 cell divisions per hour. In the second growth cycle, 2 of the cultures were evenly split into both BG11 media and Burrows media at OD_{730} 0.066. BG11 showed a specific growth rate of 21.4 hours per division; whilst Burrows showed

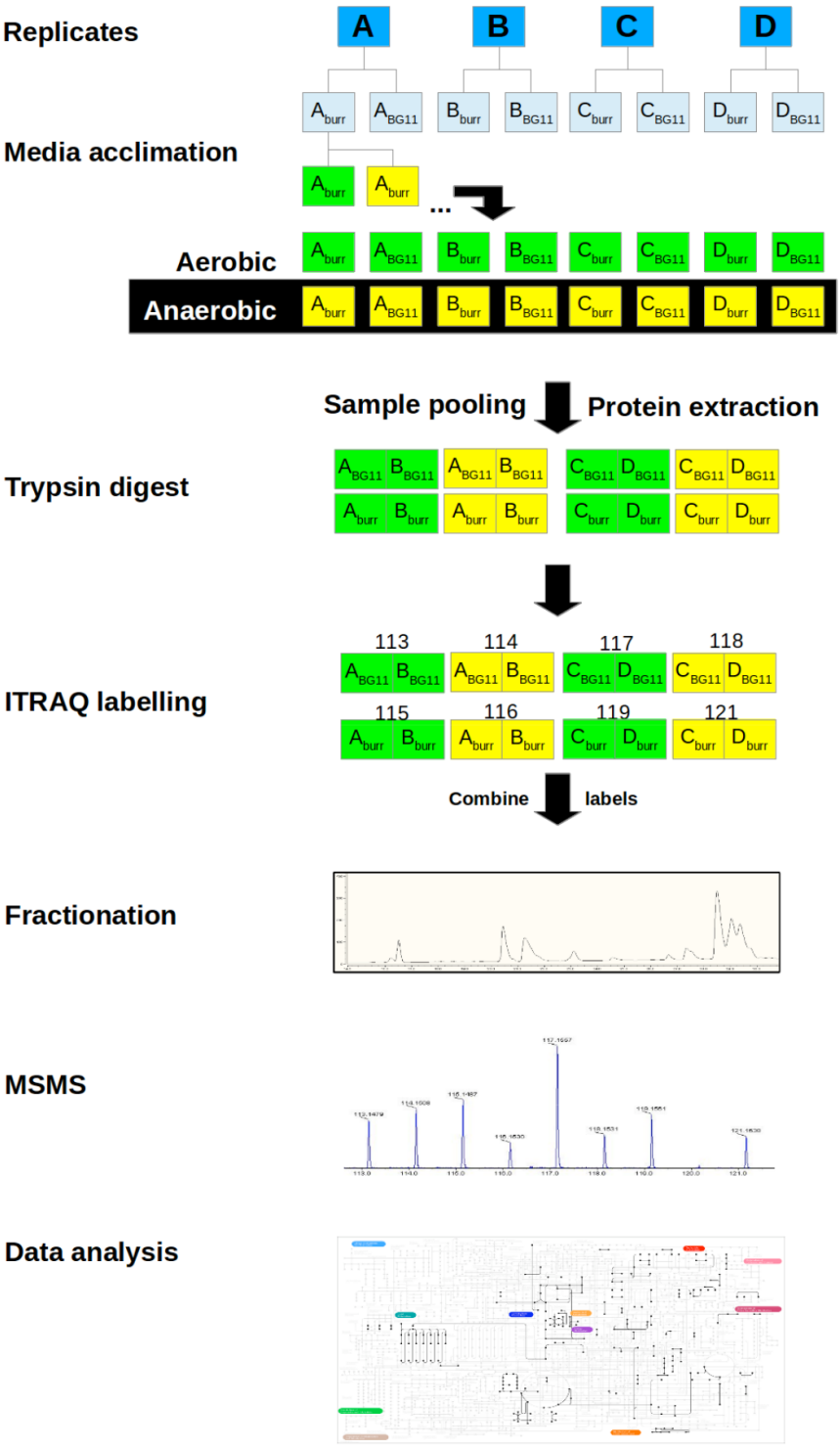


Figure 3.3: A flow-chart outlining the experimental design used in the BG11 vs Burrows proteomic experiment. The entire experiment starts from 4 separate flasks which produced paired replicates through the experiment; which are subsequently exposed to differing media and environmental conditions. This was done to keep the proteomic background as similar as possible between the replicates.

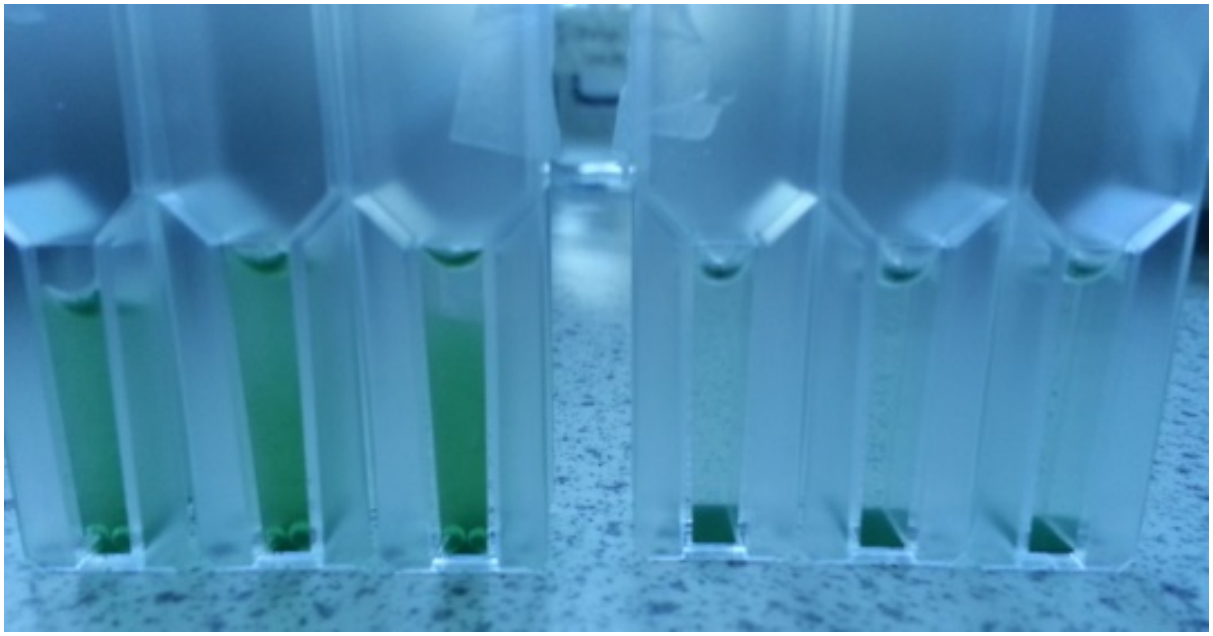


Figure 3.4: Cell culture was grown in BG11, transferred to either BG11 (left) or autoclaved Burrows media (right) and 100 μl was left overnight within a cuvette, which was topped with parafilm to prevent evaporation. Some settling was observed in BG11 media, but a much more substantial separation event was observed in the Burrows media. As a result, for all further experimentation the constituents of the Burrows media were prepared sterile and combined under sterile conditions to autoclaved dd-H_2O, which did not show the same settling effects (data not shown).

a rate of 22.9 hours per division. This suggests a slight reduction in growth rate under Burrows media, however not by an amount that fell outside the initial range, which was caused by uneven light distribution within the incubator.

When the Burrows media was autoclaved, it produced sediments that caused the cells within the media to flocculate together and drop out of solution over a single night. This was found to be a repeatable phenomenon, as demonstrated in figure 3.4 (p. 110), and so Burrows media was prepared using filter-sterilised constituents.

The cells were transferred to shaken serum bottles overnight, to detect potential changes occurring within samples as a result of being transferred to a sealed environment – as shown in figure 3.5 (p. 111). None of the conditions appeared to generate a qualitative change in the cells that would affect the larger-scale experiment, although the vibrant green colour seen suggested that the cells had not reached a point of extreme nitrogen depletion.

Figure 3.5: Cells were transferred at OD 1 into serum bottles under different growth conditions overnight, to determine if any clear physiological changes would take place, such as the settling observed in the autoclaved media. Transferring the cells to serum bottles did not appear to make a clear difference over a 24 hour period under either media condition when bubbled with either air or nitrogen.

3.5.2 Hydrogen production

The experiment was run until qualitative hydrogen measurements were made for all of the test cases that were grown in anaerobic conditions. Three of the four anaerobic Burrows media samples had begun to produce observable H_2 by the second hour of the experiment, whilst only one of the anaerobic BG11 samples produced observable H_2 in this same period. All anaerobic samples were found to produce H_2 within 4 hours of starting the experiment (Table 3.1, pg. 111).

Table 3.1: Each of the samples was checked for H_2 presence in the head-space each hour after the culture was transferred to the serum bottles. In this table, a positive detection of H_2 is denoted as a *o*, whilst a sample where H_2 was not detected was denoted as *x*. The samples are listed in sequence by replicate number, from left to right. Whilst both aerobic and anaerobic samples were measured for H_2 production, no H_2 was detected in the head-spaces of the aerobic serum bottles over the measurement time.

	Time (hrs)			
Media condition	1	2	3	4
Burrows, Aerobic	*xxxx*	*xxxx*	*xxxx*	*xxxx*
Burrows, Anaerobic	*xxoo*	*oxoo*	*oooo*	*oooo*
BG11, Aerobic	*xxxx*	*xxxx*	*xxxx*	*xxxx*
BG11, Anaerobic	*xxxx*	*xxox*	*oxox*	*oooo*

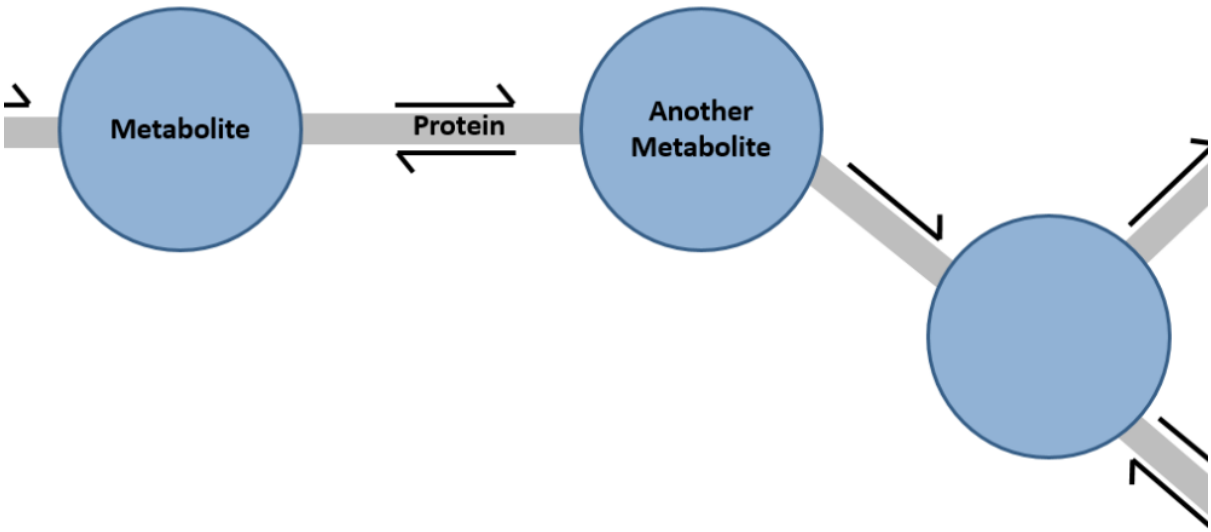


Figure 3.6: Within the KEGG structure, metabolites are nodes and proteins are edges. Proteins that are found to be statistically 'up-regulated' or 'down-regulated' in a condition will result in the colouring of any node that they point to. Conflicts aren't resolved in this with kinetics, and so the last colour over-laid onto the figure will dictate the apparent fold change; as a result these figures are guidelines rather than definitive informative graphics.

3.5.3 Proteomics – BG11 vs Burrows

The full detailed list of key protein changes identified here are available in a multi-sheet excel spreadsheet the digital appendix, DOI: 10.15131/shef.data.5327476.

The initial study of the proteomics data focused on looking at the pathway-level responses. This was done by mapping the statistically significant changes within the proteome to a metabolic pathway map – produced by KEGG. This was done using the KEGG-maper tool, where the observed proteins are overlaid onto the general KEGG metabolic pathway in black, and statistically significant changes are coloured in either red (lower protein abundance) or green (higher protein abundance). It is important to note that the KEGG map is a graph, where the nodes (circles) refer to metabolites and the edges (connecting lines) refer to proteins that inter-convert between different metabolites – as shown in figure 3.6 (p. 112).

Figure 3.7 (p. 113) acts as a primer for understanding the other KEGG maps in this chapter, where the general pathways have been highlighted to aid understanding for the reader at a glance – without requiring the more in-depth interactive visual tools available on the KEGG mapper web tool.

From the proteomic analysis, 345 proteins were confidently identified at a 1% false discovery rate (FDR) with at least 2 unique peptide matches. Of these, 335 proteins contained 2 or more unique iTRAQ labelled peptides for quantification, with over 206 unique pro-

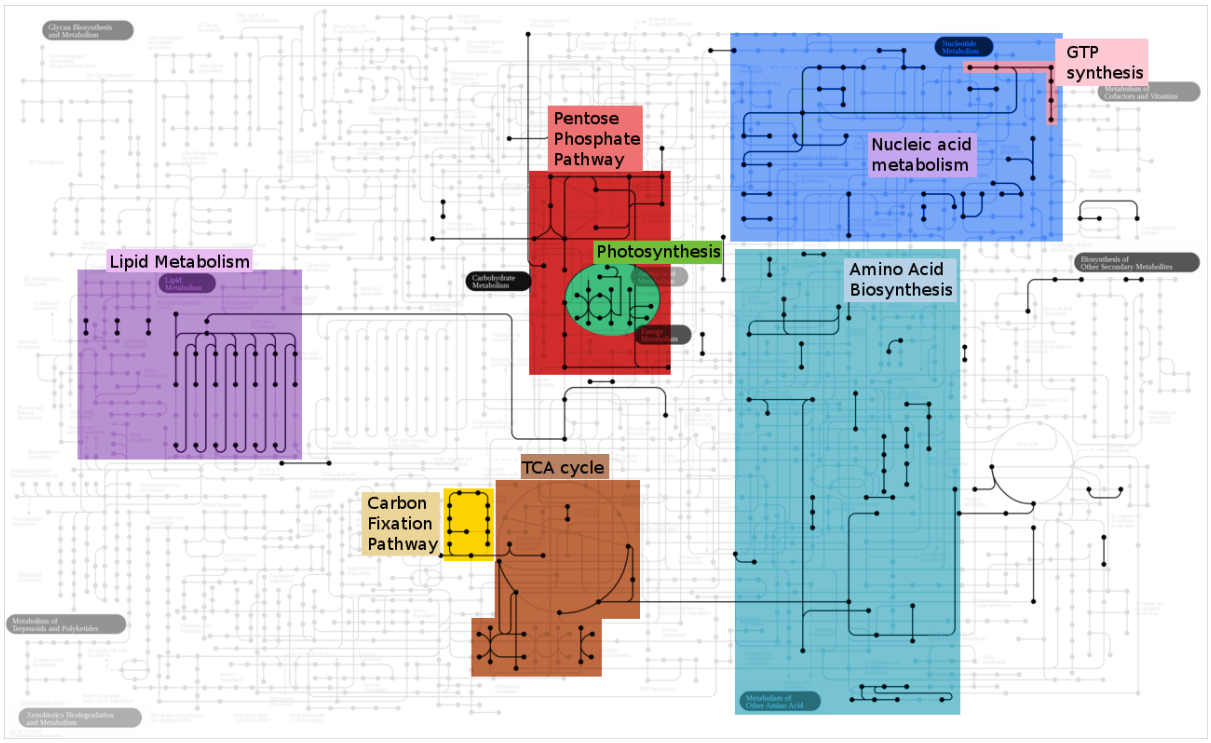


Figure 3.7: A coloured KEGG metabolic pathway map, the nodes are metabolites and the edges are proteins. The proteins that were identified in the study are highlighted in black – it is important to note that only the proteins were identified, and so the nodes are inferred by the identification of an edge. The different pathways have been approximately grouped and highlighted in a colour, to aid understanding of the major effects in the different comparisons in this chapter.

teins being quantified as significantly different over the entire study. These values were obtained with a single injection from 30 fractions.

In anaerobic dark conditions, 76 proteins were found to be differentially expressed. There are reductions in the pentose phosphate and carbon fixation pathways, with an increase in some parts of the TCA cycle. There is also a strong reduction in the phycobiliproteins, but an increase in the enzymes relating to NADP and NADPH to compensate for electron transport associated with hydrogen production (fig. 3.8, pg. 115).

The 'Burrows' condition media was compared to BG11, the standard media used by the consortium. Under standard conditions, 137 proteins were found in concentrations significantly different to BG11 when compared across all biological replicates (fig. 3.9 pg. 116).

Reduced protein quantification was observed across amino-acid biosynthesis pathways and nitrate-related pathways. This is concordant with the absence of nitrate in the media and general nitrogen starvation. A reduction of abundance in proteins involved in metal chelation or with metal ligand properties was also observed, along with other sulphur-rich proteins. This was an expected observation, as sulphur is a limiting factor in the 'Burrows' condition.

An increase was observed in phycobiliproteins and enzymes relating to electron transport. These are perhaps responsible for the increase hydrogen production observed in the Burrows media conditions. Large sequences of pathways in the central carbon metabolism, including carbon fixation and the pentose phosphate pathway were found in lower quantities, however individual proteins between these points in carbon metabolism were also significantly increased. Ribosomal proteins (non-network) were made up 39% (18/46) of the identified proteins with significantly increased levels with a fold-change greater than 1.5, indicating a high level of protein turnover.

In anaerobic dark conditions shown to produce hydrogen, 141 quantified proteins were observed to be differentially expressed between BG11 and 'Burrows' conditions. Large portions of differentially quantified proteome remain similar to the aerobic investigation. Of particular interest is the further increase in 'Burrows' conditions of phycobili-proteins during anaerobic-dark conditions, as the opposite effect is observed in BG11. There is also a further reduction in proteins in the central carbon metabolism pathways.

Directly comparing 'Burrows' in aerobic and anaerobic-dark conditions produced 53 differentially quantified proteins. The differences between these two conditions were less pronounced, although a further increase in phycobili-proteins and enzymes relating to NADP and NADPH, further validating the observed increase in antenna proteins for electron transport. There was also a further reduction in carbon fixation and the pentose

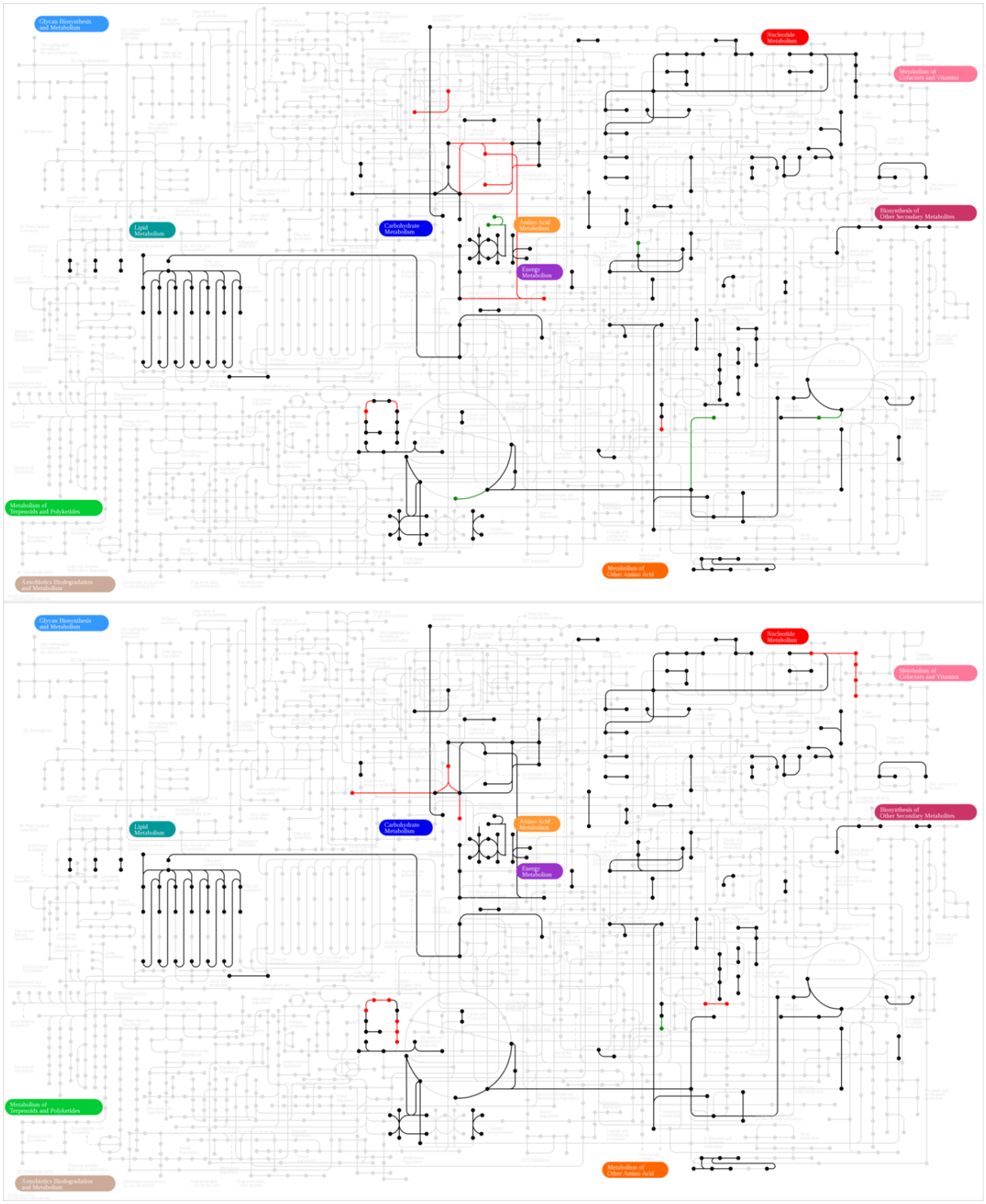


Figure 3.8: KEGG pathway maps highlighting the changes between aerobic and anaerobic states in BG11 (top) and Burrows media (bottom). In both cases there is a relative reduction in carbon fixation, however BG11 shows a large reduction in the pentose phosphate pathway, whilst Burrows shows a systematic switch off in the GTP synthesis pathway.

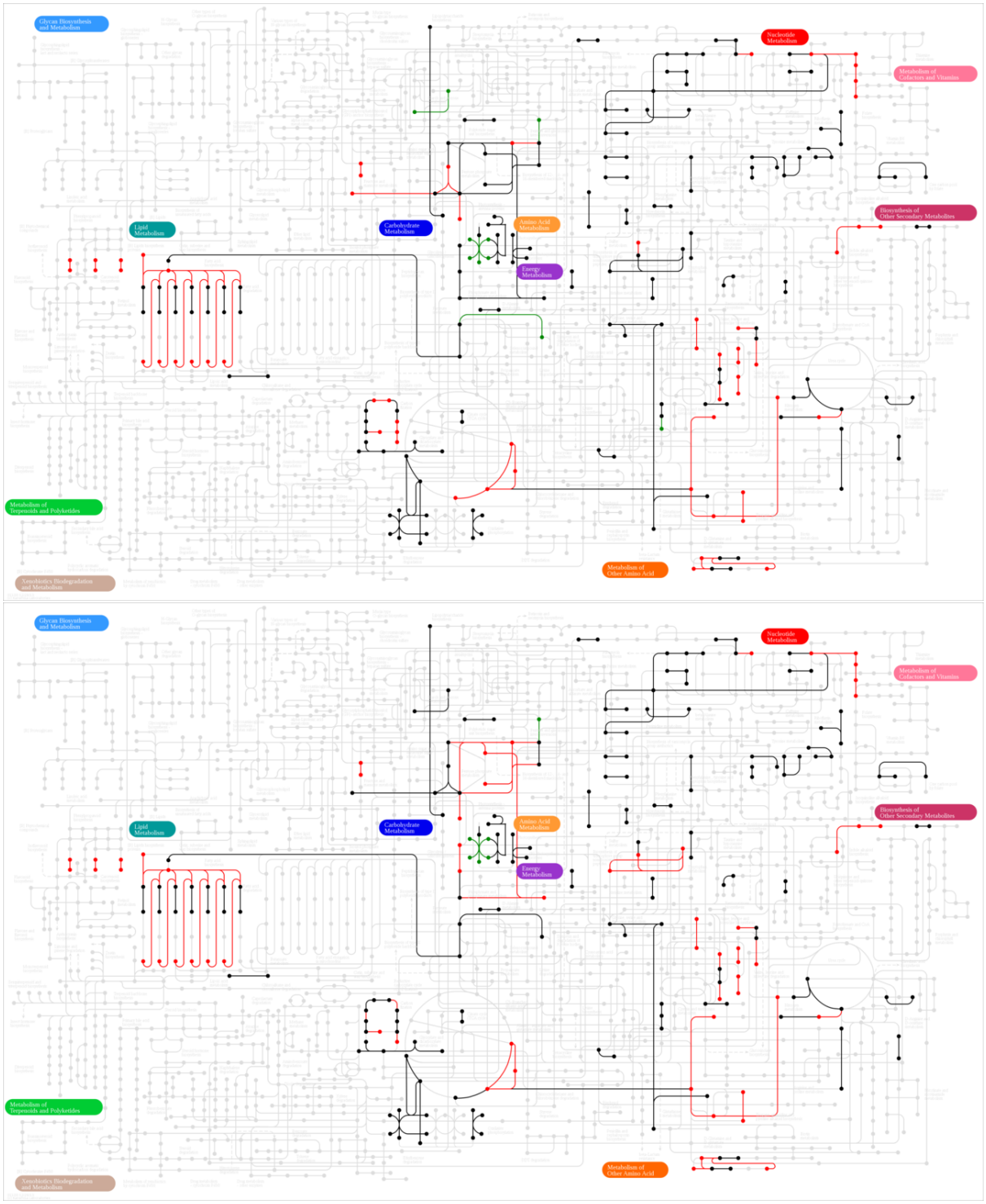


Figure 3.9: KEGG pathway maps highlighting the changes between BG11 and Burrows media under anaerobic (top) and aerobic (bottom) conditions. Across both states, the effects of the media change are uniform and highlight the dominant effects the media produce on the cells, suggesting a completely independent effect to oxygen availability. In both cases, proteins that heavily consum nitrogen – in this case the photosynthetic machinery – are less abundant in Burrows, whilst machinery that recycles nitrogen, such as amino acid biosynthesis, is much more active. In Burrows ATP and GTP production are both up, along with lipid metabolism, indicating a higher turnover of cellular energy and membrane breakdown.

phosphate pathways, similar to BG11 in the same conditions.

3.6 Discussion and Conclusions

The observations seen for altered carbon metabolism under the Burrows media conditions can largely be explained through the principal of nitrogen starvation – which was found to be the major contributing factor to increased H_2 production under the factorial designed analysis performed previously (Burrows et al., 2009). Under nitrogen starvation, previous study has demonstrated that the thylakoids and intracellular membranes are degraded and cellular inclusions of glycogen are increased (Allen, 1984). These effects have been visualised by electron microscopy, demonstrating glycogen accumulation in WT *Synechocystis* (Singh and Sherman, 2005).

Methods proposed by Lefteris (Touloupakis et al., 2016), a collaborator based in Firenze working on large-scale H_2 production within a bioreactor, demonstrated that simply letting the culture divide over time without replenishing the media resulted in a depletion of available nitrogen; and under this state placing the culture into darkness optimised H_2 production rates. On completion of H_2 production phase, the culture is then spiked with nitrogen or normal growth media, returned to light, where it returns to a state of oxygenic photoautotrophic growth, rather than the fermentative state required for H_2 production.

Research suggests that reduction in the prior levels of phosphate could accelerate the switching into this state, as phosphate depletion triggers the production of cyanophycin – a nitrogen and energy storage compound (Allen, 1984), which could present the cells with a faster rise out of lag phase when restoring the cells to a state of normal growth. This would require experimental and modelling design, to determine if the levels of phosphate could be controlled to deplete before the nitrogen levels without negatively impacting upon the cell growth.

Chapter 4

Improving proteomic methods

4.1 Chapter background

Proteomic data is an important but under-utilised tool for systems-level engineering of biological systems, particularly in biotech production strain field. In this chapter a number of core methods are developed for improving the quality of data generated by proteomic analysis, with a focus to improving the quality of data generated from *Synechocystis*. Whilst the methods are ultimately focused on *Synechocystis* applications, not all investigations were carried out on data generated from *Synechocystis* experiments.

The techniques presented here cover three main areas: Firstly, investigating laboratory techniques that can improve proteomic methods for *Synechocystis*, such as better extraction methodologies and looking at alternative protein quantification methods for determining the prior protein concentration. Secondly, investigating data-driven techniques for expanding or controlling effects arising from tag-based proteomic investigations, through balancing the effect of missing labels, understanding the lower-limit cut-off in low-abundance protein samples and merging of multiple investigation datasets together. Finally, investigating data-driven techniques that are more generally applicable to proteomic studies, such as improving the false discovery rate, automated identification of post translational modification frequencies, and cluster-based methods of interpreting proteomic data where a prior hypothesis is not determined through the use of Gene Ontology terms (GO terms).

These techniques were then applied in proteomic experiments utilised during the CyanoFactory project. The work carried out in this chapter has contributed to two publications, (Pinto et al., 2015) and (Chiverton et al., 2016); with other parts of this work currently in the process of being included in other publications.

4.2 Introduction

Given the multi-faceted collection of work carried out in the Sheffield University Chemical and Biological Engineering department, whilst the general research was performed with a view to generating tools that improve *Synechocystis* data quality, not all analyses within this chapter are carried out on the organism. Other focal organisms for the investigations in this chapter include *Chinese Hamster Ovary* (CHO) cells, and *Escherichia coli* (*E. coli*). Briefly, these organisms are extremely valuable for different biotechnology applications. CHO cells are the vehicle for the most high-profit pharmaceuticals currently in production, and are used for the creation of monoclonal antibody (MAb) proteins as they produce human-compatible glycosylation post-translational modifications on the protein chains, whilst *E. coli* is an fast-dividing and extremely well-characterised organism that

is widely used for developing and testing novel molecular biological systems; and was the first organism used for Synthetic Biology.

A number of small investigations were carried out to optimise the proteomics pipeline for *Synechocystis* – namely producing tools that would interact at points between the moment the experimental biomass was collected or received up to the point that the final data output from the investigation was completed. The first set of these was investigations into the experimental techniques used for protein harvesting, sample characterisation and preparation steps for processing on the mass spectrometer. These are referred to as 'experimental' improvements here, as they are physical modifications to, or measurements made of the sample to improve the downstream quality of the final data produced. All other investigations are data-driven and occur after the samples have been run on the mass spectrometer.

The second set of investigations looked at developing the data processing techniques during proteomic data analysis. These techniques included the following:

- detection of the presence/absence of proteins in a whole-proteome sample
- increasing the multi-plexing capabilities of tag-based quantifications
- utilising gene ontology terms to drive a cluster-based analysis of a protein system

There were a number of other investigations and automated scripts produced over the course of the study, including the inter-conversion between different proteomic data types, quantification of post translational modifications (PTMs) and assessment of the optimal approach for determining a false discovery rate (FDR) in proteomics. Whilst these are useful tools for the lab, they do not constitute novel research or a contribution to scientific knowledge, or in the case of FDR are hard-coded into analysis pipeline in most popular analysis software and are therefore not readily applicable to other studies; and so the code for this has been included as an appendix, but will not be discussed further.

As a curiosity, a comparison between different analytical software was conducted, however the major finding was largely one of personal preference driven by flexibility and convenience rather that production of an obvious scientific benefit. On analysis, MaxQuant produced the largest amount of individual files for conducting a variety of investigations. It also produced data in the most amenable format for down-stream data processing with both R and Mathematica, and so upon discovery the program was used for all further proteomic analysis, rather than being compared against other software available for accuracy or effectiveness of the identification algorithm.

4.3 Protein Extraction

The work in this section was carried out in collaboration with Narciso Couto, who assisted with lab work and ran the HPLC analysis.

4.3.1 Abstract

In proteomic analysis, the first step when analysing an organism is to lyse the cells and effectively extract the proteins. This can be challenging in hardy organisms like *Synechocystis*, where the cells have a tough periplasm and a complex internal system of membranes. In this section, methods were extracted from the pre-existing body of literature; and the most effective methods were combined together in an attempt to produce a more effective extraction technique. The technique combines a series of rounds of bead-beating with sonication, and when compared with another mechanical disruption protocol it was found to result in more protein identifications, as well as a more replicable extraction protocol. The technique did not appear to cause a biased reduction in observable proteins across a specific size range, and was also found to be effective for downstream metabolite extraction.

4.3.2 Introduction

When measuring proteins within a cell, they need to be separated from other cellular materials, such as lipids and DNA, for analysis. There are currently a variety of ways to extract protein from cellular samples, which all follow the basic pattern of lysing the cell and then separating non-desirable cell debris from the proteins whilst maintaining a good yield of protein content. The second step is usually done with a number of rounds of centrifugation, using buffer conditions that filter for proteins by solubility.

When considering cell lysis, there are large differences between all organisms. Some are particularly resistant to extraction techniques, and can be challenging to physically break apart for analysis; whilst others are much more fragile, and more robust techniques can lead to disruption of the protein sample due to shearing or rapid release of proteases. For every organism there exists a set of boundaries within which the optimal extraction of proteins, metabolites and nucleic acids can be performed, without causing significant damage to the extracted material. *Synechocystis* is a relatively hardy organism, due to the complex membrane structures, coupled with a periplasm which provides a range of resistances to damage from external sources, such as pressure and pH as described in Chapters 1 and 2 (Heidorn et al., 2011). From the global analysis of proteomic investigations in *Synechocystis*, it is evident that there are cases reported where limited protein

extraction took place. This is inferrable through the low level of reported proteins, where another reported for this low value – such as a low resolution mass spectrometer – was not given.

Limitations at the protein extraction level can have significant consequences for downstream processing. For example, if a subset of proteins, such as membrane-associated proteins, are not solublised during the extraction then a subset of the entire dataset will be missing from the analysis. There is no amount of correction or normalisation that can substitute for missing data, and so determining a reproducible technique that is broad-reaching enough to collect a representative sample from the proteome is essential for robust proteomic analysis in the organism.

There are a number of core extraction techniques that are widely used in *Synechocystis* proteomics. These are:

- Sonication

  In this method, the cells are subjected to ultrasonic energy at low temperatures. The energy causes pin-points of stress in the membranes and breaks up aggregations, causing the membrane to break down leading to lysis of the cellular components. Sonication as a method for lysing *Synechosystis* is relatively new, with all of the studies being conducted in the last 3 years. The recent emergence of this method is concordant with a recent improvement in the number of identifications in each study, and the method uses a defined set of parameters that enable repeatable investigations across different labs.

- Bead beating

  In this method, small glass beads are inserted in the sample. It is then vortexed resulting in the beads colliding with either each other or the walls of the tube, trapping the cells between two hard surfaces and crushing them open and resulting in mechanical lysis. First proposed by Norling et al. for analysing membrane fractions, this is the most established and popular cell lysis method observed in the literature – with 27 papers citing its use. There are potential issues with proteins and other materials adhering to the beads during processing, causing issues related to both recycling of the beats and reduced extraction efficiencies.

- Nitrogen cracking

  In this method, the cellular sample is frozen with liquid nitrogen and ground with a mortar and pestle. This is repeated several times until the researcher is satisfied that the cells have been lysed. Usually the number of rounds of grinding will be indicated within the methods, however due to the inherently human nature of the protocol, this is typically a judgement call based on experience (Axmann et al.,

2005). Similarly to bead beating, the cells are exposed to mechanical crushing forces – amplified by the use of very low temperatures that make the membranes brittle and more susceptible to shearing.

- Freeze-thaw cycling

  In this method, the cells are exposed to a series of freezing and thawing cycles. This causes the cells to rupture due to both osmotic stresses and physical ice-crystal disruption of the cell membranes. This method has the advantage of being passive, however it takes longer to complete than the active methods described above.

- High pressure cell disrupter and French Press

  These methods are grouped together as the both utilise high pressure to trigger cell lysis. The cells are run through machines that apply high pressures to the cells. Whilst widely applied successfully in other organisms, such as *E. coli*, *Synechocystis* has been shown to be extremely resistant to pressure changes (Heidorn et al., 2011), and as a result such methods are limited in application. This resistance is observed further observed in the meta-analysis, with a reduced number of proteins observed in studies that apply such techniques.

- Solubilisation with SDS and FASP

  Sodium dodecyl sulfate (SDS) is commonly used to extract proteins from a sample. SDS is a amphipathic molecule, meaning that it contains both hydrophilic and hydrophobic regions, which results in its detergent properties. SDS denatures proteins and covers them in uniform charge, making it ideal for gel-based electrophoretic separation, however it causes downstream problems with processing for proteomics and has therefore not been used in whole-proteome LC-MSMS studies. During liquid chromatography separation, SDS contaminates the column and is difficult to fully remove, resulting in destruction of the processing equipment. Additionally, within the mass spectrometer the detergents ionize well and are at a large excess relative to the proteins. This reduces the number of spectra collected on peptides, reducing the overall data quality.

  Filter Aided Sample Preparation (FASP) is an improvement on SDS extraction, enabling the removal of the SDS and avoiding downstream processing issues with LC-MSMS (Wiśniewski et al., 2009). This preparation removes the SDS from a sample, enabling it to be used as an extraction method in identification. The first step is to breakdown the entire cell through a combination of sonication and boiling in SDS; the proteins are then affixed to a membrane by filtration and further processing is done to the proteins in-situ on the membrane. This method has the advantage of retaining all the soluble and insoluble proteins for analysis, without

having to use 2D-gel extraction techniques. The proteins are then reduced and alkylated, the SDS is removed from the sample by buffer exchange with urea. Following this step, the urea is acidified and the sample is then desalted. At the time this study was conducted, the method had not been used previously in *Synechocystis* (based on a literature search – 2014), however the method has since been used in *Synechocystis* in 2015, resulting in the confident identification of over 2100 proteins for a phospho-proteomic study (Spät et al., 2015).

The FASP method was proposed as a general protein extraction method for *Synechocystis* during CyanoFactory in April 2014, however due to strong opposition by other members of the consortium, it was eschewed for a combination of other tried and tested methods from previous studies.

4.3.3 Methods

A comprehensive list of papers studying the proteomic response of *Synechocystis* were identified using the PubMed search engine (https://www.ncbi.nlm.nih.gov/pubmed/). The search terms used were "(Synechocystis) AND Proteom*", the search was performed in June 2013. The full list of papers was downloaded and processed manually, with key information relating to the topic of the paper, the method of protein extraction and the number of obtained proteins identified being collected. This data was then used to generate a meta-analysis to link the number of identified proteins to the extraction technique.

A new method combining sonication and bead-beating was tested. This involved first balancing the cells by OD_{730}, diluting all samples to that of the lowest concentration sample. To ensure enough material was collected for the analysis, 4 replicates were grown, paired and pooled together to ensure more balanced cell numbers – this meant that the slowest growing and fastest growing replicates were combined. As these are experimental replicates, these rates are largely determined by the initial cell concentration which varied slightly. To mitigate any effects of early stationary changes, all cells were harvested at mid-log phase – which equated to OD_{730} of between 0.6 – 0.9.

The pooled cells were centrifuged in 50 ml volumes for 10 minutes at 4 degrees Celsius at 10,000 RPM. The ramp-up rate on the centrifuge was kept at 9 (maximum), but the ramp-down rate was reduced to 7 to avoid disrupting the cell pellet. The supernatant was removed gently, taking particular care as the air-liquid interface passed over the pellet to avoid disrupting it – during supernatant removal, all supernatant was emptied into a beaker, to preserve the sample in case of pellet disruption. Once the liquid was removed, the tubes were inverted and stood on blue roll, being left for 2 minutes to let any residual

media fall from the tube and to let the pellet air-dry briefly. The cells were re-suspended in 1 ml of 4 degrees Celsius lysis buffer (recipe described in chapter 5) and transferred to a 2 ml protein lo-bind eppendorf tube. This also acted as a wash step for the cells. The tubes were centrifuged at 17000 RPM for 60 seconds and the lysis buffer removed by pipetting. If the proteins were not being extracted immediately, pellets were snap-frozen in liquid nitrogen and stored at -80 degrees Celsius.

At this point, the cell lysis-extraction methods diverged. In the nitrogen-grinding extraction, the cell pellet was re-suspended in 500 μl of lysis buffer and transferred to a mortar and pestle that had been pre-chilled with liquid nitrogen. The solution was pipetted into the mortar and a small volume of liquid nitrogen was added to freeze the sample. Manual grinding was performed for approximately 10 minutes per sample, with additional liquid nitrogen added as the sample thawed. When additional liquid nitrogen was added, care must be taken to ensure that the sample is not thrown out of the mortar due to rapid evaporation – this can be avoided with slow, careful addition of nitrogen. After this, the sample was scraped from the mortar and pestle whilst still frozen and transferred into a protein lo-bind tube for further processing. The mortal and pestle are then thoroughly cleaned with distilled water and ethanol between each sample, to avoid cross-contamination.

In the sonicating and bead beating method, the pellets were re-suspended in 500 μl of lysis buffer, and an equal volume of glass beads was added to the sample. The glass beads showed strong electrostatic interactions with the plastic weigh boat, causing issues for adding accurate, reproducible amounts of beads per sample. To avoid this, a 200 μl pipette tip was found to hold approximately the correct volume of beads when filled to the top and flattened off, and due to the altered geometry didn't demonstrate the same problematic electrostatic interactions, improving the accuracy of this part of the technique. At all points during protein extraction, the samples were kept on ice to limit protease activity and changes in cellular state.

Two main cycles were used in combination – a bead beating cycle and a sonicating water bath cycle. In the bead beating cycle, the samples were put into the bead beater on a 60 second cycle, followed by a 60 second rest period on ice. For the sonication cycle, the samples were placing into a chilled sonicating water bath, which was further cooled by the addition of ice, for 60 seconds, followed by a 60 second rest period on ice. Two bead beating cycles were run, followed by a sonicating cycle, followed by two bead beating cycles, followed by a sonicating cycle. After these 6 cycles were complete, the samples were transferred to a 4 degrees Celsius centrifuge and spun at 17000 rpm for 15 minutes. 200 μl of supernatant was transferred to a clean protein lo-bind tube stored on ice, and an additional 200 μl lysis buffer was added to each of the samples. These were then run through 2 further bead beating cycles and one sonication cycle before being centrifuged

again as described above. The supernatant was transferred to the same lo-bind tube as the first batch of supernatant, which was then centrifuged for 15 minutes at 17000 rpm at 4 degrees Celsius, to remove any particulates that had been transferred during protein extraction. The sample was then passed forward for further processing.

The processing methods re-align at this point. In both cases, the supernatant was transferred to a clean lo-bind tube, which had 5 volumes of ethanol added to precipitate the proteins. The samples were then transferred to a 4 degrees Celsius refrigerator overnight to facilitate protein precipitation. The samples were centrifuged for 15 minutes at 17000 rmp and the supernatant was removed. The pellets were then re-suspended in 200 μl storage buffer, completing the protein extraction. These extracts were either used for further processing straight away – digestion, alkylation, labelling, mass spectrometry analysis – or stored at -80.

4.3.4 Results

As can be seen from the data (Table 4.1 p. 128), studies that utilise bead beating and sonication appear to generate the highest number of identifications. It is important to note that, as can be seen from figure 2.1 (p. 72) in chapter 2, that the majority of the high-abundance protein extractions have taken place in the last 5 years in line with improved proteomic methods. This has the potential to skew the results in favour of such an analysis, and so chosen techniques in a full investigation of this new technique should have more of a meritorious basis for inclusion beyond being included in previous studies.

The protein gel in figure 4.1 (p. 129) shows that the extraction technique does not appear to limit the proteins identified by size, although to determine that there are no sections of the proteome being excluded by this method, both a side-by-side comparison on a single gel, as well as an informatic analysis on the identified proteins and their localisations within the cell to ensure unobserved biases were not taking place.

Cells disrupted using this same technique have been utilised in metabolomic studies during the CyanoFactory (extract shown in fig 4.2 p. 129, data shown in appendix).

4.3.5 Conclusions and Discussion

Overall, when comparing these two techniques there are 3 main features that should be considered: time, reliability, and proteome coverage. The nitrogen grinding technique is clearly much faster for a single sample, taking approximately 12 minutes per sample opposed to the 90 minutes for measuring up to 16 samples simultaneously required for the sonication-bead beating method. The bead-beating time scales linearly, as the operator

Table 4.1: Details from a bibliometric analysis of the effectiveness of different studies utilising a range of extraction techniques. Based on the overall extraction technique and focus of the study, papers were categorised into papers measuring just the soluble protein extract (Soluble), just the membrane protein extract (Membrane), and those combining both fractions together (Full).

Number of Protein Identifications			
	Membrane	**Soluble**	**Full**
Sonication			
Median	123	n/a	1472
Mean	123	n/a	1315
Max	123	n/a	1509
Min	123	n/a	807
Papers	1	0	4
Bead Beating			
Median	57	n/a	1200
Mean	224	n/a	1039
Max	1350	n/a	1955
Min	36	n/a	67
Papers	13	0	7
Mechanical Grinding			
Median	n/a	340	776
Mean	n/a	326.285714	776
Max	n/a	646	776
Min	n/a	120	776
Papers	0	7	1
Freeze-Thaw cycling			
Median	n/a	15	n/a
Mean	n/a	15	n/a
Max	n/a	15	n/a
Min	n/a	15	n/a
Papers	0	1	0
Pressure based lysis methods			
Median	155	102.5	n/a
Mean	155	102.5	n/a
Max	155	105	n/a
Min	155	100	n/a
Papers	1	2	0

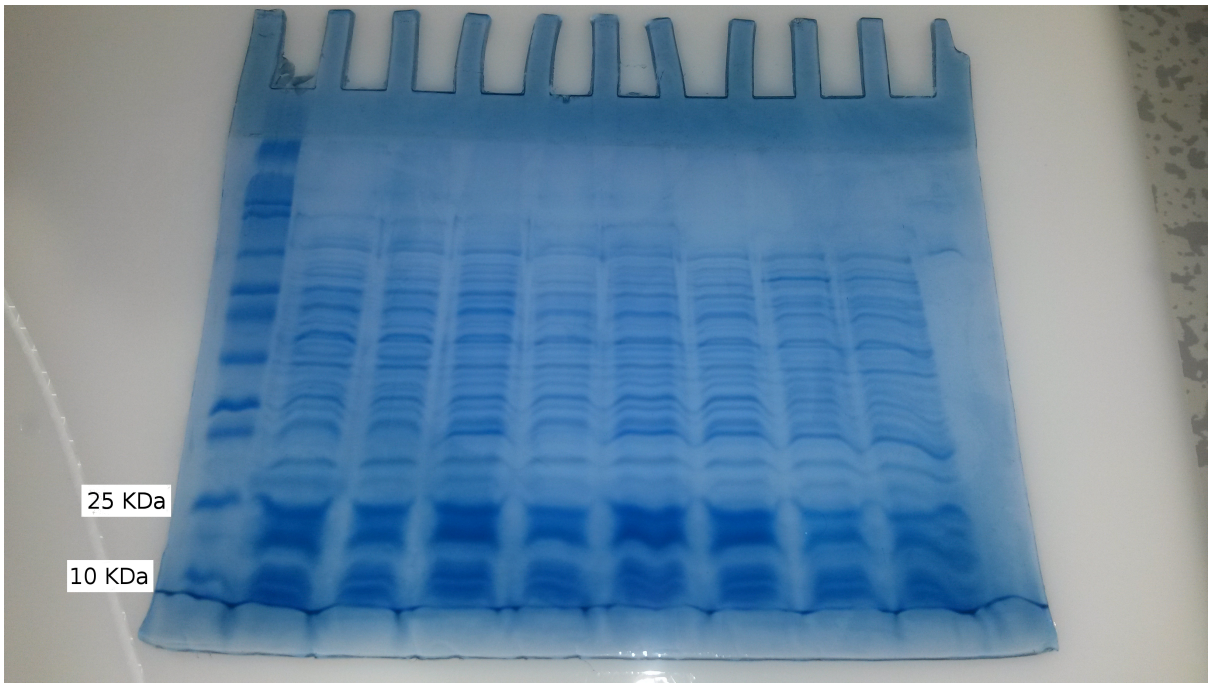


Figure 4.1: A protein gel showing a full set of samples extracted with the improved method, demonstrating a broad extraction range across the proteome. This extraction technique does not produce a bias against proteins based on size.

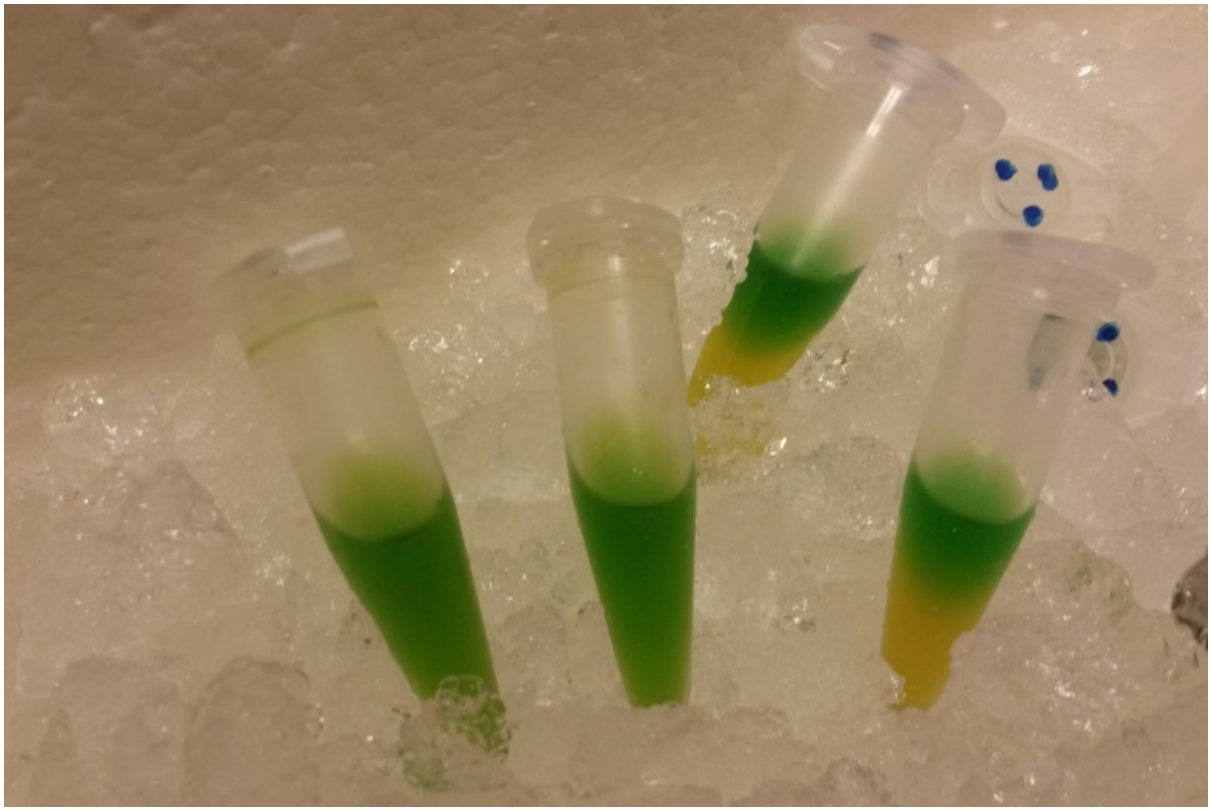


Figure 4.2: Whilst these metabolite samples were not analysed, this clearly shows that the same extraction technique is suitable for lysing cells for metabolomic analysis.

can only grind one sample at a time, and so when more than 6 samples are being processed simultaneously the grinding technique becomes slower. As a result, for an 8-plex iTRAQ experiment, the sonication-bead beating technique is faster and therefore more efficient on operator time.

In terms of reliability, the sonication–bead-beating method is a clear winner. The nitrogen grinding technique is heavily dependant on operator technique, and so the efficacy of extraction varies between different investigators. This same feature also makes it difficult to define the energy used to rupture the cells, and so determining if a different method is directly equivalent is also difficult. During initial attempts with the bead beating method, there was a large variation in the amount of glass beads added to each tube, which introduced an unreliable element to the extraction; however with the addition of the pipette tip method this was rectified.

In terms of proteome coverage, it is difficult to generate an accurate comparison. Further testing would need to be carried out comparing the same experimental replicate with the different methods; followed by testing on a mass spectrometer under identical conditions. Unfortunately, in this investigation the comparison was performed on different samples, taken at different times, measured with different mass spectrometers; so the best estimate that can be given for this is that on analysis of the dataset there did not appear to be a significant under-representation of proteins from a specific subset of the data. This finding needs to be experimentally verified, however.

No comparison between extraction methods for proteomics in *Synechocystis* has been done to date. As a result, there is the potential to generate a publication by building on the work conducted here, directly investigating a comparison of the methods identified here, whilst controlling upstream and downstream aspects of the work-flow to produce a fair comparison.

Whilst the novel technique generated a suitable range of proteins for investigation and appeared to produce a stable subset of proteins, further experimentation is required to assess whether the technique could be improved at any stage. This could be done by keeping the biomass from each round of extraction, solublising it and using gel-based techniques to determine the relative rates of protein loss at each stage. This technique could also to highlight any set-specific losses, such as fractions rich in membrane-associated parts of the proteome. Ultimately, due to the time-saving and increased reliability features of the novel method it was adopted for all further analyses carried out during the CyanoFactory project.

Beyond the tried and tested extraction techniques, there are also other avenues that could be investigated, such as the use of microwave radiation or supercritical fluids for aiding protein and metabolite extraction (Raynie, 2006). In addition, following on from the

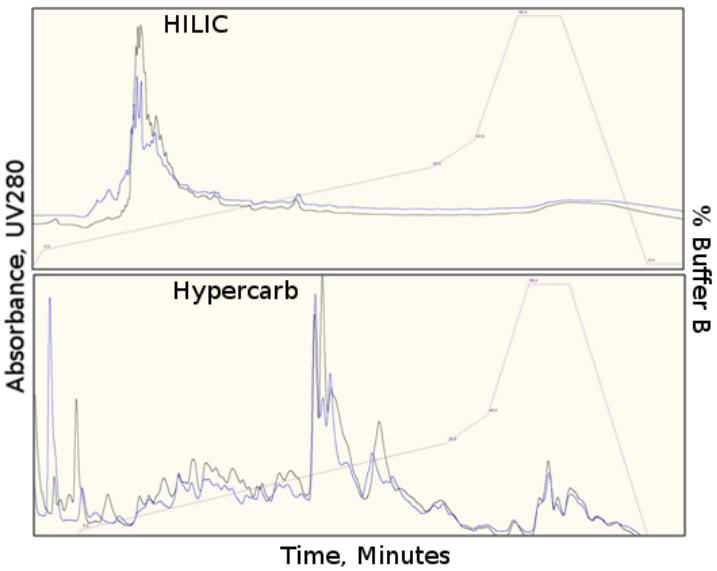


Figure 4.3: A comparison was run between *Synechocystis* peptides separated on both a HILIC column (top) and a Hypercarb column (bottom), using the same buffers and buffer ramp profile. The Hypercarb column showed a much more even distribution of peaks across the chromatography profile, suggesting a more even separation of peptides within the sample.

ground-breaking work by Gan et al (Gan et al., 2005), initial investigations suggest that there are further improvements that could be garnered from investigating other steps downstream in the proteome processing pipeline – as can be seen in the preliminary data shown in figure 4.3 (p. 131).

4.4 Protein Quantification

4.4.1 Abstract

Protein quantification is important when analysing samples by mass spectrometry. Different methods of protein quantification are typically subjected to errors. In this section, an analysis was performed to determine the most accurate method of protein quantification in *Synechocystis*. The Bradford assay – a protein dye colour-change assay, and Kalb assay – a direct UV spectrometry assay were compared using a BSA standard. The Kalb assay was found to produce lower quantifications than the Bradford assay, although they were more consistent with a densitometry analysis than the Bradford quantifications, which showed condition-specific effects. As a result, the Kalb assay was determined to be more effective for *Synechocystis* proteomic quantification.

4.4.2 Introduction

In proteomic analysis, it is important that the amount of protein being analysed by the mass spectrometer is equal between samples that are being compared. This is incredibly important for ensuring that the results observed from the analysis are meaningful and robust for a number of reasons.

As discussed in chapter 1, protein observations from a sample are genearlly considered to be incomplete when the sample is above a given level of complexity – as in the case in all full-proteome studies, meaning that not all proteins within a given sample will actually be observed during a study. This effect is stochastic and driven by the overall abundance of a given protein in a sample, running the same complex sample on two separate occasions will not produce the same set of proteins, but the variation will be emphasised more in the low-abundance proteins than in the high abundance proteins. Additionally, there is a general requirement for 2 or more unique peptides to be observed for a protein to be confidently considered to be present in a sample, as a check against a single contaminant resulting in erroneous identification of a protein, which further emphasises the bias against low abundance proteins within a sample. As with all things, proteins need to actually be measured to be compared. If proteins are unevenly weighted, then one sample will have a much greater range of observations that will not be present in the other, negating the benfit achieved by increasing the sample range.

Tag-based proteomic quantifications appear mitigate this problem by merging all peptides together, resulting in an even baseline set of proteins for observation. The resulting quantification measurements can then be normalised, either by balancing the sum of all the extracted ion counts for each tag through scalar normalisation, or by using median correction to balance the central range of observations. On face value these techniques appear to negate the need for careful balancing of initial protein concentrations, however in fact they only mask the problem. As discussed in chapter 4, a number of factors including ratio compression and peptide co-isolation, mean that the observed values do not represent the true proportions of a protein within a sample – particularly in low abundance spectra. This means that correction based on these data will compound bias effects and can ultimately result in either type I (false positive) or type II (false negative) errors from the analysis of tag-based data, where the observations would simply be missed in an un-multiplexed sample.

Having highlighted the importance of prior protein quantification for proteomic analysis, an investigation into methods that might introduce bias into this result is therefore necessary. The standard method widely employed for quantification of protein is the Bradford Assay (Bradford, 1976). This assay uses a dye called Coomassie Blue to determine protein quantification accurately within the 0.2 – 20 μg range, and is measured by a colour shift

Reagent	Absorbance wavelength (nm)
Coomassie	590
Folin	750
Bicinchonic Acid	562

Table 4.2: The colourometric protein quantification reagents and their respective absorbance wavelengths for calculating protein concentration.

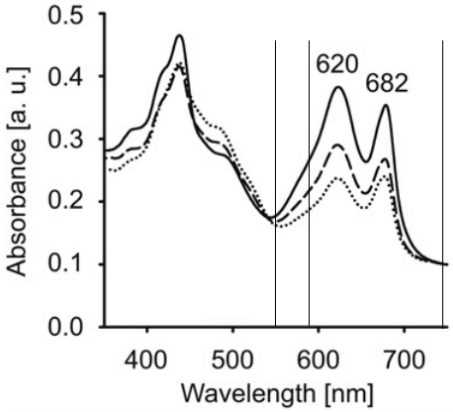


Figure 4.4: A wave-scan of whole-cell *Synechocystis* under increasing light intensity (solid, to dashed, to dotted lines respectively), adapted from (Kopečná et al., 2012). Three verticle lines have been added to the plot, highlighting the different absorbances for the bicinchonic acid assay, Bradford assay, and Folin's phenol assay (running from left to right). The peaks for phycocyanin and chlorophyll are indicated with the maximal absorbance values (620 and 682 respectively).

generated through interactions between the dye and basic amino acid residues (Compton and Jones, 1985). Whilst the test is robust and has a wide applicability to a number of different samples, different proteins do present minor variations in detection accuracy based on amino acid composition (Stoscheck, 1990). The Bradford Assay is an example of a colourometric protein detection assay, where the presence of protein causes a shift in absorbance of either a dye or a chemical complex. Other examples of this type of assay include the Lowry protein assay, which measures a shift in Folin phenol reagent under a reduction of Cu^{2+} to Cu^{+} (Lowry et al., 1951) and the bicinchonic acid assay – which also makes a measurement based on copper reduction, but uses bicinchonic acid, instead of Folin phenol, at the capture agent (Smith et al., 1985). These three methods generate absorbance measurements in the wavelengths given in table 4.2 (p. 133).

Synechocystis has 2 major regions of absorbance interference in a spectrophotometer, one at 620 nm – corresponding to phycocyanin, and the other at 680 nm – corresponding to chlorophyll (Kopečná et al., 2012). As can be seen in figure 4.4 (p. 133), which shows measurements made of cells at an equal density under different lighting regimes, these absorbances can change for *Synechocystis* where the concentrations and proportions of light-harvesting proteins are altered. Since this error is variable between samples, and since there is a clear overlap with the measurement of phycocyanin and the Bradford reagent, this demonstrates that the Bradford assay is unsuitable for measuring protein samples with high amounts of phycocyanin present.

Whilst bicinchonic acid and lowry assays may be more suitable for conducting these meas-

urements, as they measure in regions outside this interference, there is a more general issue for using colourimetric analyses for quantification in cyanobacteria: the abundance of brightly coloured compounds that absorb light in the visible spectrum. *Synechocystis* depends on dynamic control of a variety of light-harvesting compounds during a given light-dark cycle, and so measurements made within the visible spectrum are always susceptible to wider variations due to contamination with a range of metabolites that are either co-extracted with proteins or covalantly bonded, and as such persist through clean-up stages. These can be observed during the extraction process, as supernatants and pellets at each stage of extraction are brightly coloured, ranging from blues to yellows to reds. The relative abundance of these light-absorbing compounds varies depending on environmental conditions, and so analysis methods focused on the ultra-violet (UV) region of the light spectrum (100 – 400 nm) was considered to potentially produce more reproducable results. UV methods have the added advantage of being technically simpler to achieve, as they don't require the adition of reagents or incubation times. Analysis in the UV region of the spectrum requires the use of a quartz cuvette, since polystyrene and glass cuvettes have interferring absorbances at wavelengths below 340 and 320 nm, respectively (Stoscheck, 1990).

There are 2 UV absorbance regions in protein molecules, the first is the energy corresponding to the peptide bond, which absorbs at 205 nm, and the second is energy stabilised by aromatic residues, which absorbs at 280 nm (Warburg and Christian, 1941; Stoscheck, 1990). Whilst the 205 nm region is much more sensitive to changes in protein concentration, a large number of other molecules also absorb in this region, especially molecules containing c=c bonds(Stoscheck, 1990), which are present in all coloured organic compounds; and so measurements of *Synechocystis* proteome samples in this region were excluded for the same reason as measurements in the visible spectrum were excluded. On the other hand, measurements in the 280 nm region are less accurate between samples with varying protein levels, due to the changes in the relative number of aromatic residues in the proteins which could again introduce sampling bias. DNA and RNA both generate interference in the UV region of the spectrum, and so these must be corrected for to give an accurate protein quantification. This is typically done in the 260 nm region, where proteins do not produce a strong absorbance signal, but nucleic acids do (Stoscheck, 1990). An alternative method to measuring the 280 or 205 nm regions directly is to measure the 210 nm absorbance off-target at 230 nm (Kalb and Bernlohr, 1977). This technique, which produces similar accuracy to levels attained by the Lowry assay, has the advantage of not being strongly affected by varying proteins in the background. Whilst combining the best aspects of all quantification techniques, the actual mathod itself is relatively unheard of, with the paper only stating 591 citations over the last 40 years, compared with over 18,000 for the bicinchoninic acid method, 196,000 for the Lowry method, 230,000 for

the Bradford assay. The formula for calculating protein concentration from this method is as follows:

$$\mu g/ml.protein = 183A_{230} - 75.8A_{260}$$

Where A_{230} and A_{260} are absorbances at 230 nm, and 260 nm respectively (Kalb and Bernlohr, 1977). In this study we compare the quantifications produced by the Bradford assay with those produced by the Kalb UV calculation.

4.4.3 Methods

The spectrophotometer assays were performed as described by Stoscheck (Stoscheck, 1990) and Kalb (Kalb and Bernlohr, 1977). For the Bradford assay, a standard curve for BSA was created by diluting a 1 mg ml^{-1} protein solution. Protein concentration for *Synechocystis* samples was estimated initially by eye on a protein gel, then informed estimations by trial-and-error were performed by dilution, until the concentration was in the correct magnitude – in this case an initial 1 in 50 dilution of the protein samples was made to bring the absorbance readings to around 1 – then a series of 2-fold dilutions was made to determine the absorbtion coefficients. Quantification was performed on the 1 in 100 dilutions of the samples as these absorbance values were determinable on the BSA curve. The same 1 in 50 initial dilution was used for all samples, resulting in variations between protein concentrations in different samples. The protein gel was analysed with densitometry to determine approximate relative protein concentrations in a numerical fashion other than by eye, where the coomassie-stained gel was photographed and analysed optically by ImageJ in the UV-plate reader. All data was analysed in the software R and graphs were made with the ggplot2 package (Wickham, 2009).

4.4.4 Results

Initially the protein concentration levels were checked with SDS-PAGE and densitometry (fig 4.5 p. 136) to determine the relative quantifications. Prior to extraction, samples had been pooled and analysed with a spectrophotometer to determine relative concentrations. The different pooled samples were then diluted to produce samples with approximately equal numbers of cells.

The protein samples were then analysed with both the Bradford assay and the Kalb UV spectrometer assay, the quantifications were found to be the values in table 4.3 (p. 136). The Bradford assay gives consistently higher quantifications than the Kalb assay, although it also gives higher values for cells grown in BG 11 media than Burrows. The samples generally appear to change in the same way seen in the densitometry analysis.

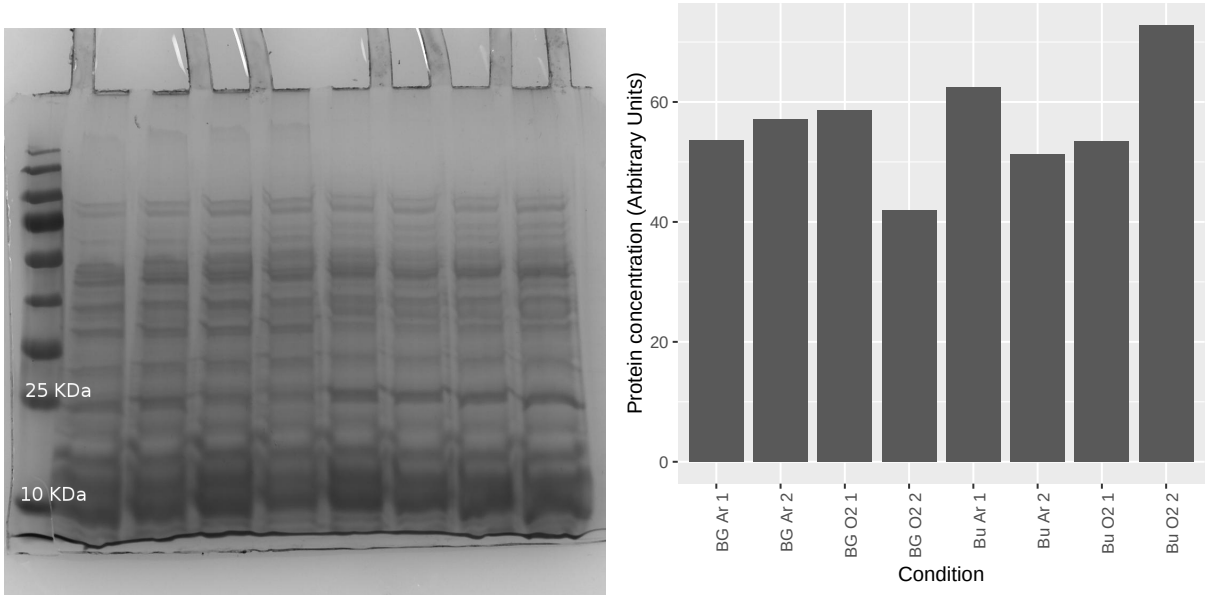


Figure 4.5: A SDS-PAGE gel stained with Coomassie blue and a densitometry analysis of the image. This shows the relative quantifications of protein between the samples.

Condition	Densitometry (Relative)	Bradford ($mg.ml^{-1}$)	Kalb ($mg.ml^{-1}$)
BG Ar	0.74	88	39
BG Ar	0.79	84	41
BG Air	0.81	94	42
BG Air	0.58	60	33
Bu Ar	0.86	66	38
Bu Ar	0.70	52	33
Bu Air	0.73	72	36
Bu Air	1	81	46

Table 4.3: A table showing the different quantifications obtained from the protein samples for each of the different measurement methods. The Bradford assay has consistently higher quantifications than the Kalb assay, and also shows higher quantifications for all the proteins from samples grown in the BG 11 media.

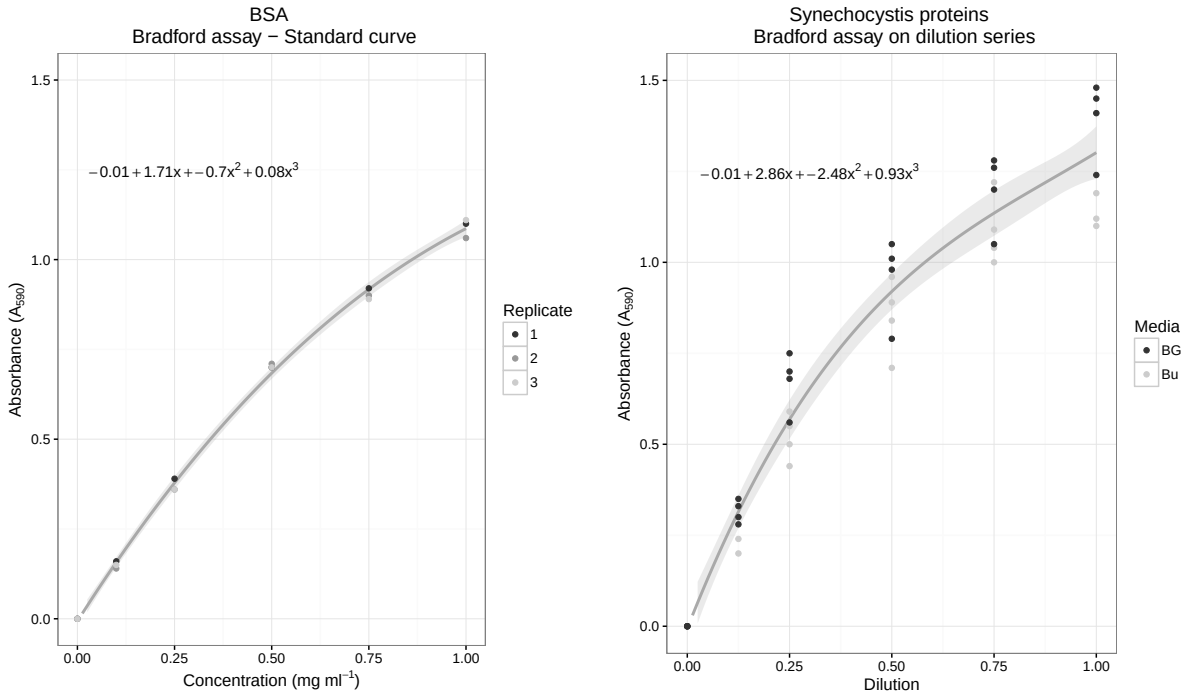


Figure 4.6: A comparison between serial dilutions of a known BSA standard and *Synechocystis* proteins from a proteomics experiment on H_2 production. Whilst the two curves are not supposed to match, the ratio of the coefficients in the general linear model should be consistant between the two. The *Synechocystis* proteins show a realtively higher contribution from high-order polynomial terms, suggesting non-linear interferrence.

Bradford analysis was performed on serial dilutions of BSA and *Synechocystis* samples to determine protein concentrations, shown in fig 4.6 (p. 137). The dilution series comparison shows a higher-order polynomial fit when comparing BSA to the whole proteome sample, suggesting an unsuitability for the the standard curve as a quantification tool in this case. Interestingly, there appeared to be a systematic difference between the concentrations observed for cells grown in BG-11 media as opposed to those grown in Burrows media. The dataset was split by media condition and investigated in more detail. A comparative dilution model was generated for each of the different conditions (fig 4.7 p. 138).

4.4.5 Conclusions and Discussion

The further analysis on the different conditions suggested that a Bradford assay could be affected by differing environmental conditions more than the UV Kalb assay. This was based on the better consistency observed between all measurements in the Kalb assay when compared with the Bradford. Ultimately, the findings of this chapter are interesting but are badly in need of further repetition of the experiment, under a variety of different protein-measurement conditions, to verify if the findings are actually legitimate or just a

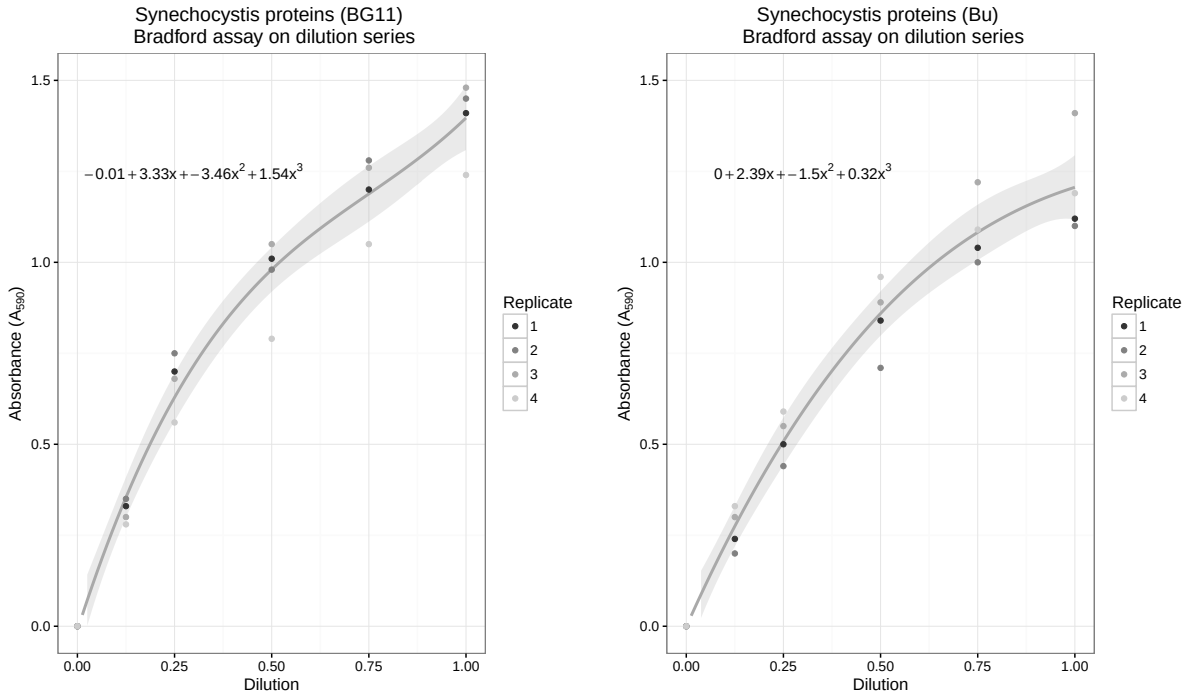


Figure 4.7: A comparison between serial dilutions of BG11 and Burrows media. The ratio of the coefficients in the general linear model should be consistant between the two; but as in figure 4.6, this is not the case. The dilution series are coloured by replicate. Cells grown in Burrows media appear to show the heteroscedastisity expected in a hierarchically-linked dilution series, whilst this is not as evident in BG11.

one-off effect.

From a practical point of view, the Kalb assay is simpler to perform than the Bradford. The Bradford assay requires more complicated reagents and a more involved investigation technique, with the proteins being reacted with a reagent and left for 5 minutes; whilst the Kalb method is a direct measurement in the UV spectrum. The Kalb assay requires the use of a quartz cuvette, as other commonly used materials like polystyrene or glass will absorb radiation in the UV spectrum – this adds expense to the technique. It is a one off cost, opposed to the reagents needed for the Bradford, and is quicker; so the over time the Kalb method would be more cost effective.

There was a disparity between the two measurements suggesting that either the Bradford method was over-estimating protein concentration or the Kalb method was underestimating protein concentration. Due to the research discussed in the introduction, and evidence of other cases where the Bradford assay over-estimated protein concentration in the presence of contaminating substances, it seems more likely that the Bradford is over-estimating, but is still possible that the Kalb is under-estimating protein concentration. Designing an experiment to verify one or the other being accurate is challenging, as all experiments for protein quantification discussed here have bias when measuring a

complex protein mixture.

A possible experiment to investigate this effect would be to run quantifications with both techniques at various stages of clean-up. This would highlight if a particular step resulted in a significant loss of signal, which could then be further investigated to determine whether a large amount of protein was being lost at that stage or if a major contaminant was being removed. It would also be interesting to comparatively quantify proteins that had been size-separated, to see if particular bias existed in a detectable size fraction of the proteome. This may enable the detection of a representative, unchanging part of the proteome to determine the overall protein quantification – much the same way such comparisons are performed with ribosomal RNA or caretaker genes in nucleic acid studies. Determining a simple, replicatable technique for such a process would likely be challenging.

This experiment was limited because all samples were measured during a single proteomics experiment, and as a result are susceptible to systematic contamination. To verify these results, additional repeats on different conditions would need to be carried out to ensure robust analysis. In addition, only the Bradford assay was assessed with a dilution study, and so the Kalb technique may also display the same level of non-linearity of absorbance. The Kalb method has the advantage of not being compared to a standard, however some observations suggested minor variations between different protein standards being studied. As a result, it would be beneficial to confirm this effect by performing a comprehensive, systematic study into the effect.

Finally, the curve-fitting model used here that looked at a cubic model was suggested by the Bradford test supplier to better fit the observed data, but a cubic model is mathematically a worse model to fit the data than a linear one in many ways – notably as if continued there could be up to 3 separate measurements that would result in any given concentration (polynomial functions with a cubic component typically cross the x axis 3 times). It is likely that that this model is suggested because a more accurate logarithmic function is more challenging for most operators to calculate. This limitation may also contribute to the limitations of the Bradford assay, where the calculation for the Kalb assay is linear.

As another possible improvement to proteome detection – if the phycobiliproteins are responsible for the visual spectrum contamination, it may be possible to remove the effect directly. The chromophore in biliproteins is covalently attached during protein formation to a cysteine through the formation of a disulphide bridge (Scheer and Zhao, 2008). Reducing and alkylating the protein mixture prior to measurement may remove the chromophores. When not covalently bonded to the biliprotins, the chromophores are much less stable and energy is given out as heat rather than absorbed in a specific area

of the visual spectrum (Scheer and Zhao, 2008). Testing this would be relatively simple, and would require the same steps normally taken during protein sample preparation – reducing the disulphide bridges with dithiothreitol (DTT) and alkylating them with MMTS to prevent reformation. The additional secondary and tertiary structural damage done to the proteins as a side effect of breaking these connections may also further contribute to the break down of colour in the sample.

4.5 Studies in a low abundance proteomic background

The work in this section has been included as part of a publication, however this description covers a more detailed explanation on the values and cut-off used, as well as the background challenges from a study of this kind and the parallels that can be drawn with other studies (Chiverton et al., 2016). The dataset used in this section was generated by Lesley Chiverton and Caroline Evans. Joselin Noirel provided key insights into the data analysis.

4.5.1 Abstract

Presence-absence studies are of great biological relevance; although because proteomic datasets are sparse (not all proteins are measured) and subject to type I errors (false positives), determining such states is increasingly challenging. In this study, presence-absence is determined through the use of an α cut-off technique. A cut-off is determined by modelling background noise within the sample, then all values that fell below this are set to an α value – in this case equivalent to the lowest possible measurement. This was conducted by designing an iTRAQ study with 2 missing labels, to determine the overall level of interferrence of other factors during the study (co-isolation, isotopic contamination). The experiment shows that whilst determining a cut-off and α level is possible in a given experiment, creating a general model for this is challenging, however the data do suggest that a general model should be based on the Poisson distribution.

4.5.2 Introduction

In *Synechocystis*, a major issue for proteomic analysis is the high-abundance subset of antenna proteins causing an expansion of the dynamic range(Gan et al., 2005). High abundance proteins within a sample is a significant challenge for a range of biological samples, as exemplified with the case of human blood plasma. Blood plasma is a key target for identifying disease markers (Anderson et al., 2004), since it interacts with all

cells in the body and can be readily extracted for analysis; however as much as 90% of the protein present in human serum is albumin (Chan et al., 2004). To facilitate analysis of these protein samples, numerous techniques have been used to remove these high abundance but unwanted proteins from the sample – as reviewed by Bellei et al (Bellei et al., 2011).

Whilst in the case mentioned above – where the high abundance proteins responsible for expanding the dynamic range beyond measurable levels – selective depletion is an option as the 'contaminating' proteins are not of interest to the investigators, this case is not true for *Synechocystis*, where the high abundance proteins are also of interest to the researcher. Unfortunately, as they make up such a significant part of the proteome, changes in the overall levels of the light-harvesting machinery can have profound consequences for relative quantification of the other proteins in the sample – where a 50% reduction in the antennae structures can result in a 10% reduction in the total protein content of the cell.

Analysing samples which are contaminated with a dominant protein, where simultaneously observing the levels of the contaminating protein and the proteomic background are important, is a real challenge. Statistically speaking, observing relatively low-abundance proteins within a sample becomes an exponentially challenging endeavour as the difference between the high-abundance proteins and the background stretches further apart. This is problem is discussed in fine detail in Chapter 4 of this thesis. The difference between the highest-abundance protein and lowest-abundance protein in a sample is called the 'dynamic range'.

In this section, we discuss an analysis that was performed on an industrially relevant clean-up step of a biosimilar, produced in CHO cells. The product in question was a monoclonal antibody. The two chains of this protein were parsed as 'X00001' and 'X00002' during analysis, due to the industrially sensitive nature of their precise sequences. CHO cells are favourable to work with for the production of drugs, as they have human-compatible glycosylation patterning and they will naturally secrete target proteins into the media. This ideally results in a constant production cycle of the protein, which can be purified from the media through the use of a protein-A column. During collection, some of the host cells lyse and release cellular constituents into the media. This investigation looked at persistence and possible co-concentration of certain proteins within the media, which could ultimately result in catastrophic outcomes for the patient. The researchers were interested in both the overall levels of the target drug throughout the study to detect cases of major product loss through the process, and also the relative levels of background contamination in the process.

In this work, firstly an observation was made on the intensities of missing labels on an

iTRAQ 8-plex study, to determine the background level of channel contamination for any given iTRAQ label. These observations were than used to differentiate between very low-abundance proteins, and proteins that are completely absent within a multiplex sample. A minimum threshold – referred to here as the cut-off level – was modelled from the labels that were unused in the dataset. Values that fall below this cut-off level are then replaced with the α value, a non-zero value that falls at the minimum observable level for a dataset. Determining presence-absence in a ratio analysis should not technically be possible ($\frac{n}{0} = \infty$), but can be conducted with this technique.

The paper describing these techniques was recently published (Chiverton et al., 2016). This study was directly investigating the efficacy of a depletion methodology, so whilst in a typical high-abundance background study, such as the ones described at the beginning of this section, would employ a depletion technique – such as protein-A purification to increase their sensitivity; it could be argued that such a methodology may generate artificial results in this study, or mask proteins that co-purify with the host protein and would, as a result, have been depleted in any such downstream analysis. As a result, even though our analysis required a number of informatic stretches of the data, it was experimentally the optimal choice available at the time.

4.5.3 Methods

For the proteomic experiment, 3 states of purification in the same sample were being compared: the extracellular media that had been protein-concentrated – referred to here as the culture harvest (CH), the flow-through material collected from a protein-A column purification (FT), and the eluate from the protein-A column purification (EL). Two replicates of each state were taken during this experiment, resulting in 6 samples, which were labelled with iTRAQ 8-plex tags. The full experimental methodology of this experiment is available as supplementary material in the publication (Chiverton et al., 2016).

iTRAQ labels 113 and 121 were intentionally left blank – these two labels should both have the lowest susceptibility to noise contamination. 113 because there are no lower mass labels that are susceptible to isotopic contamination through C^13 distribution, and 121 as there is no 120 mass label – as a result the isotopic contamination on this label should also be minimal. Whilst this feature is useful for determining high stringency noise models, it does mean that the modelled cut-off used will err on the side of presence, rather than absence, as it comes from a cleaner background than would be expected from other labels. The median of these background samples was used to determine high-level differences between the two cases. The cut-off value was calculated by taking the non-zero values in the dataset, finding the mean value, and adding four standard deviations to it. This was done to ensure that around 1 in 15,000 cases of noise fell above this cut-off,

which was chosen as the dataset consisted of 12,879 peptide observations in total. All values that fell below this level were converted to α, which was set at 0.1 – the minimum observable value in the dataset after the cut-off was applied. The α value was chosen to avoid skewing the quantifications away from the other measured values. During this study, the minimum observed value was also noted to calculate the full dynamic range observable in the mass spectrometer – where the highest-intensity values result in detector saturation.

An alternative data-processing technique was used during analysis of this dataset. iTRAQ-labelled peptides with missing labels were included in the analysis, out of necessity, as all values in the dataset that fell below the cut-off threshold were raised to it. Within the experiment, as a result of the ratio-derived quantification in the dataset, the values for α were not all equal. This was an important feature, but meant that the differences between presence and absence in different proteins were of different significances. In example, a high intensity peptide signal will generate a much greater ratio difference between the observed values and α, although peptides quantified at α should be considered 0 for all practical intents and purposes and so the ratio based on these values is largely meaningless. Using such quantifications, relative to an arbitrary level, could result in biases for low-intensity peptides to appear either retained or enriched incorrectly. If the initial value was very low, with some missing intensities in the prior conditions, then the range between α and observations will be small as well. As a result of this limited pool of available data, a non-statistical methodology was instead used to determine if a signal was changed in a meaningful way from a previous sample – this was described as follows:

$$0.9 \times (max.A) < (min.B) = A.Depleted$$

$$0.9 \times (min.A) > (max.B) = A.Enriched$$

Where $max.A$ is the highest intensity reading for a protein, and $min.A$ the lowest. An increase in relative concentration relates to a co-enrichment of the protein during the process, whilst a reduction suggests a depletion – with a 10% boundary to provide additional stringency on the technique. The advantage of this method is that it is harder to breach the enriched threshold, which in terms of the aims of the experiment was the best possible outcome to avoid false positives as far as possible.

4.5.4 Results

The protein quantification data showed a few interesting features that had implications for industrial preparation of biosimilar drugs, which are discussed at length in the publication

and will not be repeated here (Chiverton et al., 2016). Of interest in this chapter is looking at the findings for each of the analytical techniques that were trialled, and determining whether they were effective or suitable for the task they were trying to achieve.

The full potentially detectable dynamic range of the mass spectrometer was determined as 8.2×10^5, although practical limitations within the peptide dataset reduce this during experimental observations, due to factors like isotopic contamination.

In the noise models, the data show a Poisson distribution (fig. 4.8 p. 145). This is expected as a result of how the mass spectrometer works – it makes physical count events, and so whilst the numbers in the scale appear to be continuous when the counts become very large, when working at the very low intensity level it is clear that they are actually discrete. The mean intensity values observed for each of the excluded labels were 0.064 for tag 113, and 0.045 for tag 121, although the highest intensity measurement observed for 121 (138.6) was much higher than the highest value observed for 113 (17.98). These data show that peptide-specific effects may be contributing to the observed high-intensity peptides in 121, as highlighted previously (Ow et al., 2009); however 113 appears to generally experience a higher level of isotopic contamination at a persistent low level.

Within a Poisson distribution $\mu = \sigma^2$, so the intensities were converted to discrete values

$$\frac{\mu}{min}$$

where $min = 0.0047$, $\mu_{113} = 0.064$, $\mu_{121} = 0.045$ and $\mu_{model} = 0.55$

$$\frac{\mu_{model}}{min} = \frac{0.055}{0.0047} = 11.6$$

and

$$\sigma^2 = 11.6, \sigma = 3.4$$

$(\mu_{model} + 4\sigma) \times min =$ cut-off intensity ≈ 0.1

4.5.5 Conclusions and Discussion

A simple mean-value Poisson model was used to determine the background contamination level. It may have been possible to create a more detailed model derived from the other signals in the data, such as the peptide sequence. This could account for changes more specifically, however such an approach seemed overly complex based on limited requirements for this particular study.

For developing a higher-accuracy model of the background noise in future, it would make more sense to repeat the experiment with blank labels placed differently within the data.

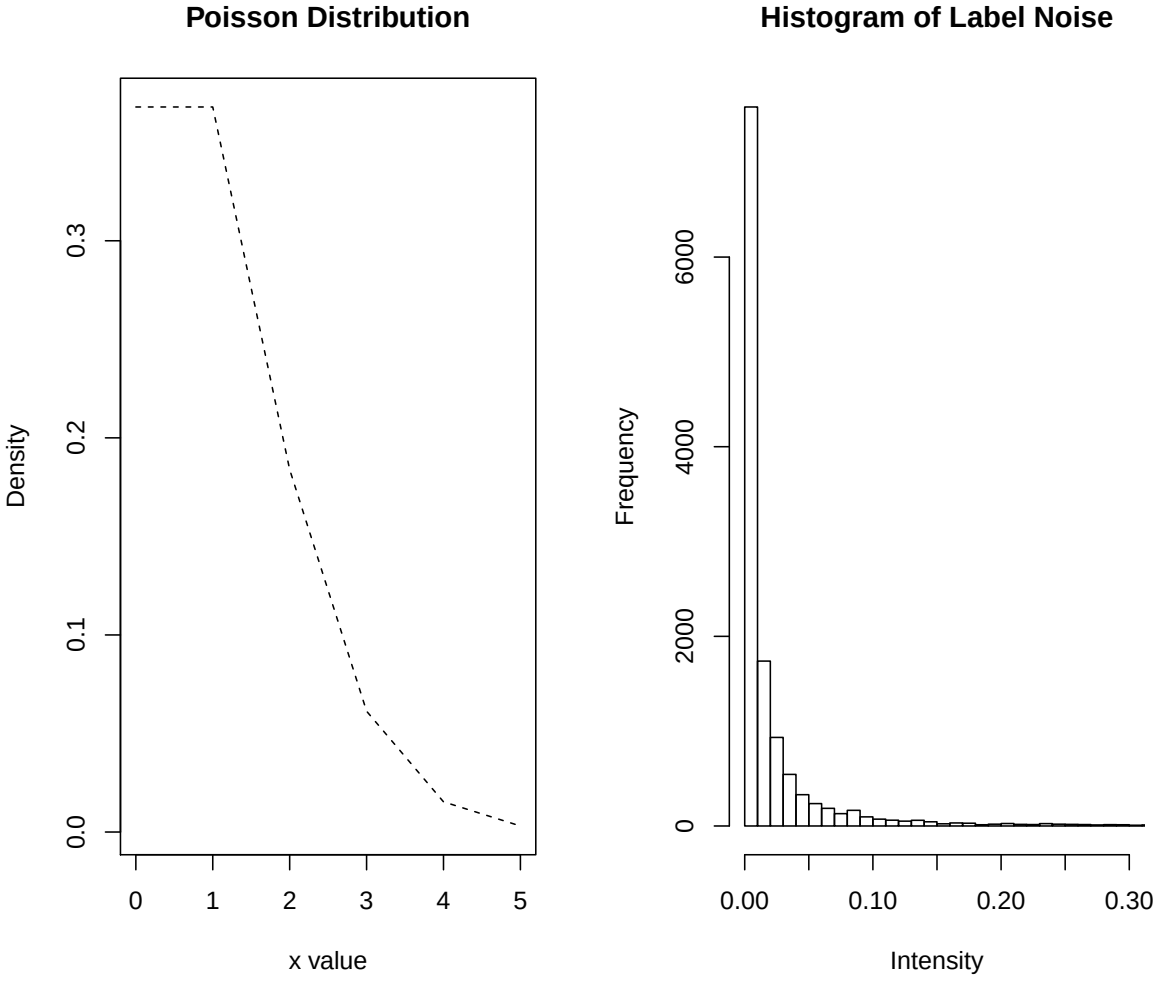


Figure 4.8: A Poisson distribution (left) and the histogram of the label intensities measured in the empty iTRAQ channels (right). Due to the discrete nature of the mass spectrometer measurements at low intensities, the data observed approximates a Poisson distribution, which was therefore used for the background noise model.

In this particular experiment, the ideal candidates for determining background contributions should be in tags 117 and 119, with EL samples in 118 and 121. This way, the samples act as a buffer between each of the samples – ensuring the maximum sensitivity for detecting low-abundance signal. They also give a direct read-out of the isotopic contamination coming from other proteomic labels.

Whilst this design was proposed, there were concerns that such an experimental design would prove too stringent to achieve the aims of the researchers – to detect very low abundance protein signals – and so the labels with less interference were chosen instead. Ultimately, the experiment would have been improved if both experimental designs had been implemented, as it would have enabled a much more accurate model for the data and would enable comparison between labels considered to be 'noisier' – such as 117, with labels considered to be 'cleaner' – like 121; however due to associated costs and limited value of return such an experiment is not currently economically viable.

The α cut-off technique is, in and of itself, an interesting method of tackling the 'missed-label' ratio quantification problem – which is usually approached with exclusion. An advantage of this is that there are potentially many more data points available to an investigator, which is always valuable when analysing data. The two major weaknesses of this technique are driven by how the model is calculated and the data-specific effects. If the model is too stringent, it can cloak the observations of low-abundance proteins, which could collectively be assigned to non-existence when they are truly present; however if it is too lenient peptides that are actually in the noise level would appear to be present when they are truly absent, but given more prominence. Besides this, the current model is highly susceptible to peptide-specific effects that could spoof the cut-off threshold.

Ultimately, if the researcher can accept that there is a practical limit to detection, and is willing to sacrifice a potentially meaningful ratio for one that is certainly artificial, then this methodology would enable clean and balanced observations of datasets where there are significant numbers of missing labels. Whilst a non-statistical approach was utilised in this analysis, which was required with this dataset due to the dominant MAb proteins in the sample, it does not preclude statistical analysis of such datasets in a more normal situation. Arguably, applying an α cut-off to more datasets could improve the statistical significance of iTRAQ or TMT quantifications in a variety of cases – such as presence-absence studies within the proteome.

4.6 Merging tag-based experiments

Joselin Noirel provided key insights into the tag-merging techniques and conceived of the proteomic data analysis method described in this section. The datasets used in this section

were generated by Filipe Pinto (IBMC, Portugal) and Narciso Couto (Sheffield, UK); and Bagmi Patternak and Pia Lindberg (Uppsala, Sweden).

4.6.1 Abstract

Within tag-based quantitative proteomics, the experimenter is typically chained to the number of samples they are able to compare within a single experiment. With some experimental design consideration, it is possible to chain multiple experiments together in a manner where the results being compared are not only meaningful on a qualitative level – proteins that are in higher or lower abundance – but also on a quantitative level – the levels of fold-change also being proportional between experimental studies. In this section, a number of these methods are explored and compared, utilising median correction and mean-ratio scaling. Experimental datasets exemplifying each different case are shown, demonstrating the improvements gained by each step of processing.

4.6.2 Introduction

There is currently something of a progressive arms race taking place between the companies that produce isobaric tag reagents. Originally, tandem mass tags (TMT) enabled simultaneous investigation of two samples within a single experiment, and then improved upon these reagents to enable the analysis of 6 samples simultaneously (Thompson et al., 2003). Shortly after this tagging technology was released, the rival label producers iTRAQ generated a 4-plex set of reagents which came coupled with a bioinformatic processing tool from Applied Biosciences (now ABSciex) (Ross et al., 2004a). This led to the iTRAQ reagents becoming far more popular that the TMT reagents, despite coming later to the market.

On the back of its relative success, iTRAQ released the 8-plex tag. In the mean-time, TMT has been buoyed by becoming the platform-standard isobaric tag for Thermo instruments, and have recently announced that they have reagents capable of up to 10-plex analysis. Whilst this ever-increasing sample multiplexing may be something that the research-consumer is interested in, there are significant limitations with this ever-expanding multi-plex capacity, as pooling more samples comes at a cost of direct signal identifications – for a more in-depth discussion of this phenomenon, please see chapter 4.

Beyond physically creating new chemistry for generating these isobaric tags, it is mathematically feasible to increase the multiplexing ability of these tags through the combination of multiple experiments into a single analysis. Although technically simple to perform, such a combination can be challenging, as it depends on experimental design with this

combination in mind before starting the experiment. In some cases, experimental datasets can be tied together from separate experiments, however to produce meaningful results, these datasets should be generated from the same mass spectrometer with the same protein extraction procedures and the same experimental conditions during growth. Failure to adhere to these requirements – whilst philosophically feasible, since cells should respond to the same conditions in the same manner – is practically very challenging.

The difficulty in this type of analysis is that it makes finding a cohesive set of changes within the proteome challenging, particularly at the finer level of protein regulation where the interesting findings are observed. The major issue is that there are multiple 'ground states' that a cell can exist in, which are controlled by kinetic features rather than fixed 'pathways'. This means that whilst broad-spectrum changes – such as the up-regulation of carbon uptake machinery or down-regulation of antennae proteins in *Synechocystis* – are conserved across different analyses as they are kinetically the most favourable inputs to the system; smaller-scale changes such as responses in metal-binding proteins might not be observed systematically. This is not to say that the finer-scale changes are not informative for a researcher – indeed they give valuable insights into the control processes present within the organism being studied; however it can be more challenging to observe identical responses when looking at the system coming with minor variations in starting conditions. These features are usually grouped together into a pseudo-random variable referred to as 'biological variation', unexplained changes that occur in a systematic manner at a level that can't be reliably observed. Such issues can lead some researchers to state that 'biology can't be modelled, as biological systems are unpredictable' – a statement that is clearly incorrect or else scientific endeavours into biology would be consistently fruitless!

When merging different methods together, assuming that the data is consistently produced and analysed in the same experiment, there are a number of approaches that can be taken. Regardless of the method used to merge multiple experiments, there are still a number of weaknesses with such an approach compared with an increased multiplex capacity that is chemically generated.

Firstly, and most prominently, is the shared protein requirement for analysis. 'You can't compare what isn't there' is a strong mantra in proteomics, and whilst techniques can be attempted to account for missing labels in multiplex experiments (discussed in the previous section of this chapter), if the protein to be compared is not present in both analyses then that usually signifies the end of the analysis. This also presents a chain-weakness in the analysis, where if one dataset presents with far lower quality data – ie. fewer protein identifications – than the others, then it limits the maximum quality of the overall data. Arguably, this effect is a suitable compromise to enable comparisons that could otherwise not normally be made; and it can be mitigated by the increased rate of observation that takes place in lower multiplex observations – notably the difference in

identified peptides between iTRAQ 4-plex and iTRAQ 8-plex can be as high as a 50% reduction in the 8-plex observations.

Secondly, whilst proteins might show the same direction of change, the magnitude of such a change is not always consistent. For example, a protein in two independent analyses shows a statistically significant up-regulation in 2 different experimental conditions, however in one the fold-change is 2, whilst in the second it is 10. This can be caused by a number of features that are not actually derived from the proteomic state of the organisms being investigated, including fewer peptide identifications in one dataset compared to another, additional noise contamination, or stronger ratio compression effects. The effects of this can have a number of consequences for the data, since the fold-change is inherently linked to the statistical significance.

Finally, since shared samples are needed across experiments in a number of designs, the full range of the multiplex comparison is not additive. For two 4-plex iTRAQ experiments, a and b, a completely independent pair of experiments would be $\{a_1, a_2, a_3, a_4\}$ and $\{b_1, b_2, b_3, b_4\}$, where a and b refer to the iTRAQ experiments and the subscript refers to the sample number. The same experiments with a shared control would be $\{a_1, a_2, a_3, a_4\}$ and $\{a_1, b_1, b_2, b_3\}$, where a_1 is the shared control and a total of 7 unique samples are compared. This is further reduced if an additional control is shared between the samples. Adding additional multiplex experiments into the design continues this trend, due to the consistent requirement for the shared sample, and so the number of experimental comparisons available in a given number of experiments would be:

$$(lab - con)exp + con$$

Where lab is the number of labels in the multiplex, con is the number of controls shared between the experiments, and exp is the number of separate multiplex experiments being merged.

In this section, three methods for merging labels are described – all with 8-plex iTRAQ analyses. The first utilises two shared samples quantified against a single label to determine ratios, the second uses a similar comparison but instead of being quantified against a single label the samples are quantified against the mean quantification from both of the controls. The final method takes the second method even further, removing identical controls entirely through a direct repetition of the experiment and normalising the experiments using the mean value of all the labels in the experiment. This method essentially sacrifices the maximum number of conditions to be compared for a greater increase in the number of experimental replicates. To close this section, an experimental design is proposed where 12 conditions, each with 2 replicates, are compared over 3 iTRAQs with an external reference comparison being conducted in a TMT 6-plex.

4.6.3 Methods

The simplest method to attain this type of comparison is to perform two independent analyses which share a control condition, where no further data analytics is performed.

$$\frac{a_n}{a_c}, \frac{b_n}{a_c}$$

Where a_n refers to the set of all labels in the first iTRAQ, b_n refers to the set of all labels in the second iTRAQ, and a_c is the control sample which is identical in both iTRAQ experiments. The protein lists are then filtered to only contain $a_{prot} \cap b_{prot}$, or proteins that appear in both iTRAQ a and b.

Protein Quantification

The datasets are then analysed in a standard manner for consistent statistically up and down regulated proteins between the experimental conditions from both analyses. This is the method used to detect significantly changing proteins and is used in other studies within this thesis in chapter 5 and the appendix. Initially, all intensities are converted to ratios through division by a control sample. This is done to convert the values obtained for each peptide to a relative change, rather than an absolute quantification – since the intensity of a peptide signal in the MS^2 scan is not necessarily proportional to the amount of protein present and can be affected by peptide-specific effects that distort the values.

The standard method for quantifying proteins is to generate a relative ratio to a single control sample. The major issue with this in a typical ratio-derived investigation, is that the noise or variance in a single label is completely collapsed. This method masks cases where protein measurements in the control sample are erroneous or missing. Alternative methods to this are described below.

These ratios are then log-transformed. Without this transformation, half of the observed values lie in the range of 0 – 1, whilst the other half lie between 1 – ∞; whereas afterwards the data is split evenly around 0.

For example, take the values 1 and 100. If 1 is the control sample then the ratio between the two is 100, whereas if 100 is the control sample then the ratio is 0.01; demonstrating that these values are unevenly distributed in linear space. Following a log transform, these ratios become:

$$log_{10}[100] = 2$$

$$log_{10}[0.01] = -2$$

Making the data much easier to work with, although it does require an exponential

transformation to obtain the original ratios again.

Identifying statistically significant changes

To determine an up-regulated protein, $a_{up} \subset a_{prot}$, for a given protein *prot*, under experimental conditions x and y,

$$a_{up} = \frac{\overline{\text{prot}_\text{x}} - \overline{\text{prot}_\text{y}}}{\sigma_p\sqrt{2/n}} \leq test$$

Where σ is standard deviation, n is the number of quantifications, *test* is the statistical cut-off decided by the experimenter for significance (traditionally 0.05 for arbitrary reasons), and

$$\sigma_p = \sqrt{\frac{\sigma_x^2 + \sigma_y^2}{2}}$$

The order of $prot_x$ and $prot_y$ is reversed to identify a statistically significant reduction in protein level between two samples.

In the majority of cases in proteomics there are multiple replicates included within the same experiment. There are two methods for resolving the multiple comparisons generated in this case, the first is to merge the peptide quantifications from the replicates together when performing the statistical test, although this isn't recommended due to collapsing the hierarchical error associated with both the physical experiment and the extraction protocol.

An alternative method is to consider each replicate independently, running multiple comparisons – for example with 2 replicates this generates 4 separate T-test P-values. Depending on how much stringency is wanted in the down-stream analysis, either the highest P-value can be taken of the four (highest stringency), or the median of the 4 values for a more relaxed stringency. If the latter method is being used, then P-values should be converted into Z-scores before being combined, to remove sample-specific bias (Pascovici et al., 2015).

A z-score is simply a P-value measured by how many standard deviations (σ) it falls away from the mean of the data. This conversion ensures that the averaging is performed on a linear scale, and avoids further bias from being introduced to the data. The z-score is calculated as

$$z = \frac{x_i - \overline{\text{prot}_\text{x}}}{\sigma_x}$$

Where for a given protein x, x_i is the label intensity, $\overline{\text{prot}_\text{x}}$ is the mean intensity of the protein and σ is the standard deviation.

Merging by statistically significant proteins

Using one of the methods mentioned above, a complete list of proteins that are found to by significantly 'up' or 'down' between a test case and some control sample – usually the WT – are compiled for each iTRAQ. The control sample between the two iTRAQs are considered to be identical – indeed these samples are often the same samples in both studies, although they do not necessarily have to be. As a result, samples from different iTRAQs can be compared not directly to each other, but to their statistical differences to the control sample, as exemplified in (Mota et al., 2015).

Scaling by control samples for quantification

Cases where the protein changes move in the same direction for each of the samples ($a_{up} \wedge b_{up}$) are considered to be unchanging, however where the signs are different and the proteins are present in both can be considered different. Cases where no significant change was observed are more difficult to assign, as they lack conclusive evidence either way.

A more advanced version of this method was published as part of (Pinto et al., 2015), where two control samples were added to each of iTRAQs being combined, as per the standard method. Briefly, the paper was investigating the effects of stable integration of protein production constructs into 5 putatively neutral sites within the *Synechocystis* proteome. The aim of this study was to ensure that there were no significant background effects occurring as a result of this integration across all of the sites, but since replication was required there were more test cases than iTRAQ labels available within an 8-plex. The full investigation report that was produced for this work is available as an appendix to this thesis.

To maintain the variation present within every label, instead of being divided by one of the control samples, each label was divided by the mean of the two control values:

$$\frac{ms_{xi}}{\mu(ms_{xc1} + ms_{xc2})}$$

Where in a given MS^2 scan x, ms_{xi} is the vector of the label intensities and $\mu(ms_{xc1} + ms_{xc2})$ is the mean of the control intensities.

Between the two iTRAQ experiments, the values were normalised against each other to enable inclusion of the quantitative data as follows. Following log transformation of the values:

$$iTRAQ_{b,x,i} \times \frac{\mu(iTRAQ_{a.Control,x,i})}{\mu(iTRAQ_{b.Control,x,i})}$$

Where for each protein x, $iTRAQ_b$ is the full set of labels for each protein in the second iTRAQ experiment, $\mu(iTRAQ_{a.Control})$ is the mean of peptide quantifications for each protein in the 'control sample' labels in the first iTRAQ, and $\mu(iTRAQ_{b.Control})$ is the mean of the peptide quantifications for each protein in the 'control sample' labels in the second iTRAQ.

This simple scaling vector, when applied to the second iTRAQ, is sufficient to cluster the proteins – a full break-down of this process in code form can be found in the appendices.

Scaling by mean protein intensity

Initially, iTRAQ datasets were sub-setted to only contain proteins present in both experiments. The data within each of the labels was then median-corrected as follows:

$$\frac{iTRAQ_{y,i} - i\widetilde{TRAQ}_y}{iTRAQ_{y.0.6} - iTRAQ_{y.0.4}}$$

Where $iTRAQ_{yi}$ is the vector of relative protein quantifications for label y in a given iTRAQ, $i\widetilde{TRAQ}_y$ is the median value in that vector, and $iTRAQ_{y.0.6}$ and $iTRAQ_{y.0.4}$ are the 60% and 40% quantiles, respectively.

As described above, the protein data is ratio converted prior to log-transformation; however instead of using a single control sample or the mean of control samples, the data is divided by the mean of the entire set of labels measured in the iTRAQ for each peptide, $pept_i \times \mu(pept_i)$, where $pept_i$ is the list of label ratios for a given peptide and μ is the mean. The two iTRAQ experiments were then scaled against each other with a slightly modified scalar to the one used for control samples above:

$$iTRAQ_{b,x,i} \times \frac{\mu(iTRAQ_{a,x,i})}{\mu(iTRAQ_{b,x,i})}$$

Where $iTRAQ_{x,i}$ is the vector of the label intensities for each protein x, in iTRAQ a and b, and μ is the mean.

4.6.4 Results

The control scaling and mean protein intensity scaling methods are shown here. For the control scaling methods, the data was clustered after being scaled by control sample. Since the method used identical samples across the two iTRAQs as the control, and the datasets were scaled by mean of these two samples, the control samples from the first experiment show very close clustering with their counterparts in the second experiment – as can be seen in figure 4.9 p. 155). This finding shows that once corrected, the

experimental variation for an identical sample is negligible. The other interesting finding with this analysis was the consistent clustering between the biological replicates across the two experiments, with the test cases clustering against each other more closely than they did against the control samples.

This observation led to the question of whether the requirement for identical samples across multiple experiments was truly necessary, or whether replicates were sufficient. A follow-up experiment, therefore, was conducted on two experimental iTRAQs that had no identical shared samples, but were full experimental replicates of the same experiment. These samples were treated slightly differently to the case used in Pinto et al, as no individual labels stood out as controls – since all the experimental replicates theoretically had an equal similarity between their experimental counterparts. For this experiment, the samples were therefore controlled against the mean of all intensities after median correction (fig. 4.10 p. 156) – as described in the methods.

Following median correction, the data were scaled to balance the measured intensities – as described in the methods. These were checked visually with a box-whisker chart to ensure that the correction process had not disrupted the median correction applied in the previous step, as shown in fig 4.11 (p. 157). As can be seen from the data, the median for each label remained stable, however in certain cases the tails of the data moved. This was to be expected, as the majority of proteins within the dataset should be present around the centre of the data and should therefore not change significantly, whilst the proteins changing by a significant amount move a much greater distance when a scalar is applied.

Whilst figures 4.10 (p. 156) and4.11 (p. 157) show an overview of the data, they do not show the internal quantification relationships between the individual proteins. For all datasets, the proteins quantifications were plotted against each other, as shown in fig 4.12 (p. 158). In this graphics grid, the 4 test conditions are plotted against each other, the first two rows showing internal comparisons. Note the final scatter-plot on the second row shows more variation than the other seven comparisons on the first two rows, reflecting the increased variation in the data seen prior to median correction in figure 4.10 (p. 156), where the dataset in question is the third box and whisker chart from the right.

The third row of this figure shows the comparisons made between the two iTRAQs prior to scalar correction, but after median correction has been applied. This shows the large amount of variation seen, particularly in the lower-abundance proteins within the sample. In this case, both control set labels from one iTRAQ were merged together and plotted against the merged labels from the second iTRAQ, so these plots contain twice as many data points as the graphs above. This break-down in relationship between the samples shows the common problem when comparing quantitative data across separate

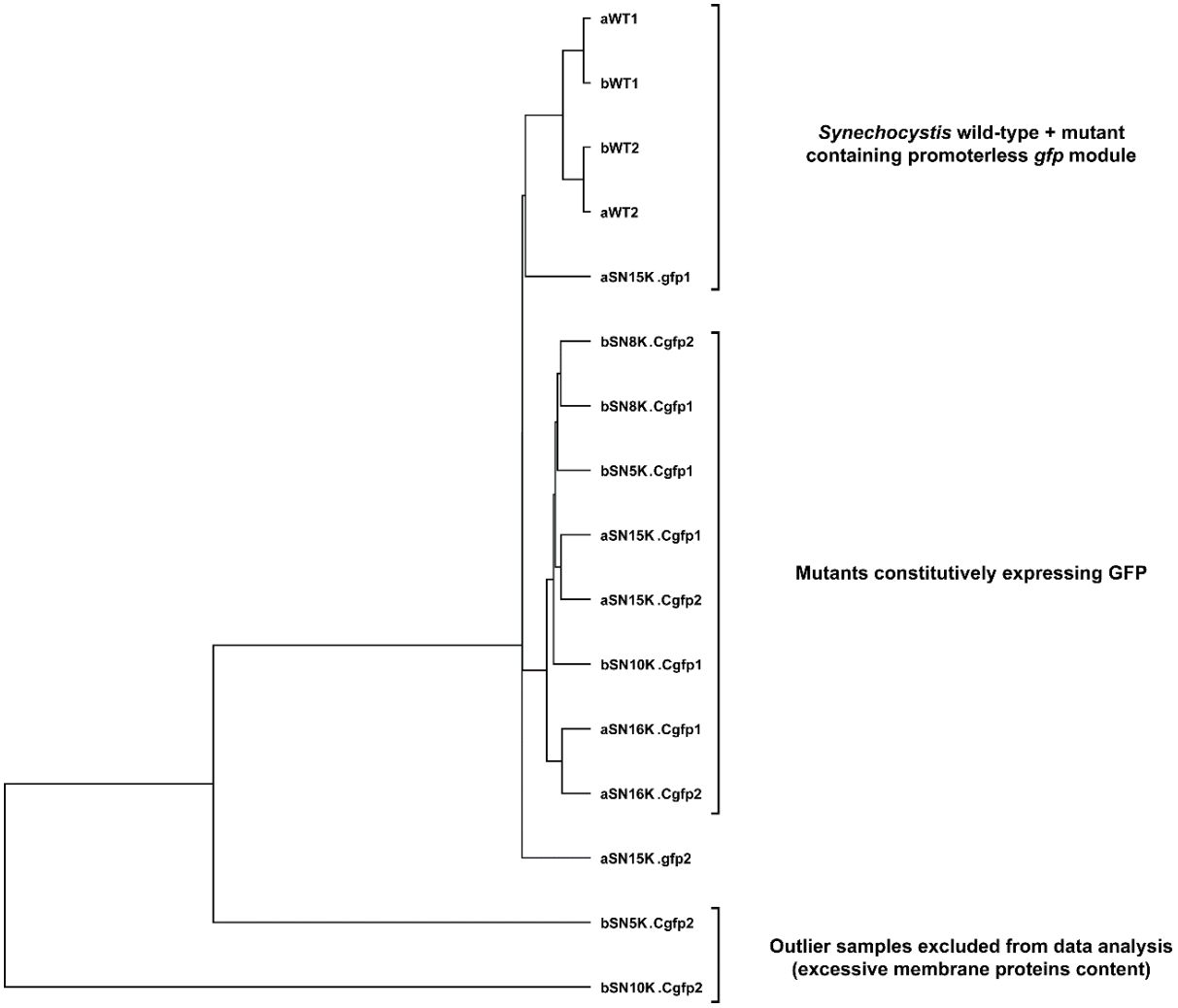


Figure 4.9: Cluster plot taken from supplementary material in (Pinto et al., 2015). This cluster plot was built through 2 replicates of a shared wild type (WT) samples across the two separate iTRAQ experiments. These are labelled as WT1 and WT2, with a and b denoting the iTRAQ experiment across all of the samples. After normalisation, the samples clustered very closely together across the two iTRAQ experiments. Two of the samples stood out during the analysis as containing a substantially different set of proteins, mainly related to the cell membrane.

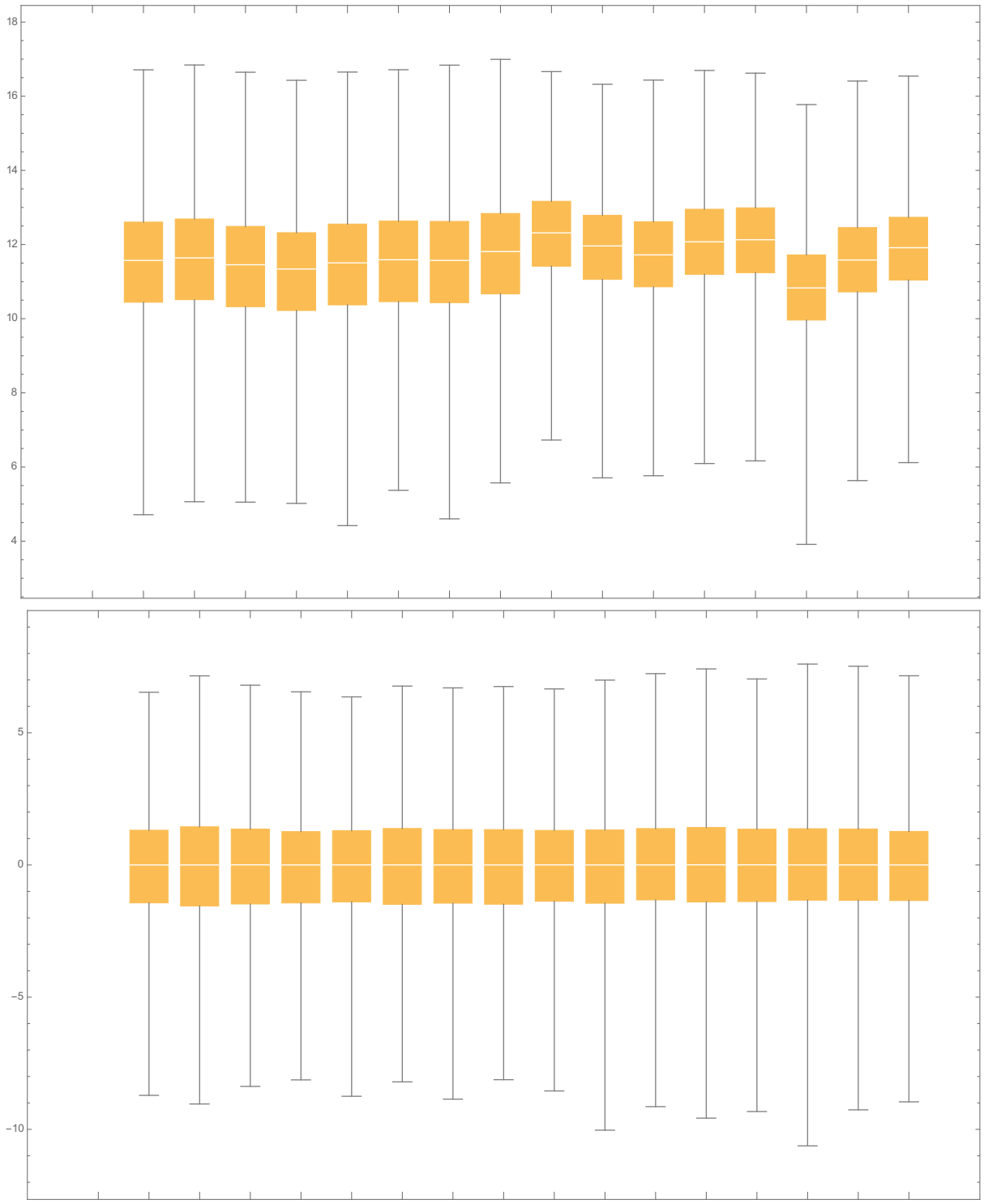


Figure 4.10: A box-whisker plot showing the range of peptide intensities (measured in direct counts) before (top) and after (bottom) median correction. *Post-median correction values are in log space.* Two iTRAQ 8-plex experiments were plotted side by side, the first 8 from one experiment and the second 8 from the second. All values in the bottom graph were normalised so that the median values were all equal, and so that the spread of the centre 10% of the data fell within the same range. This transformation improves the quality of the data in each experiment independently, but by itself doesn't improve the overall quality of comparison – please see fig 4.12 (p. 158).

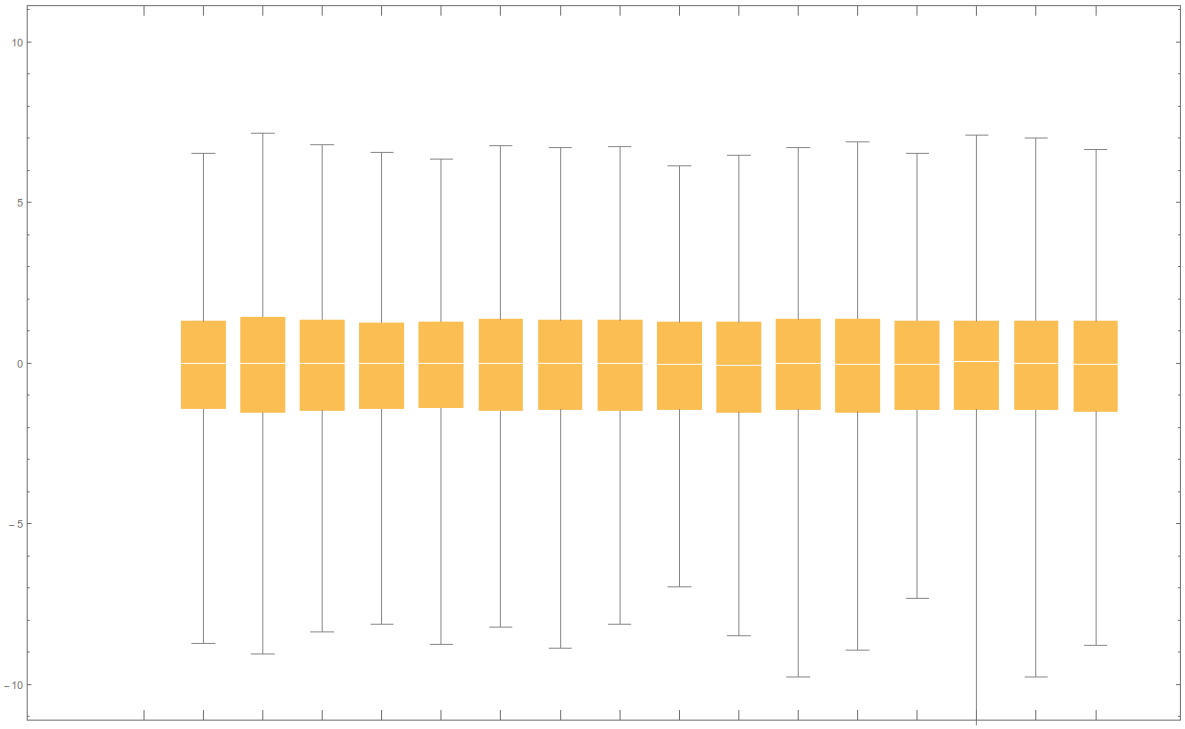


Figure 4.11: The same dataset from figure 4.10 (p. 156), scaled as described in the text. The scalar transformation doesn't affect the overall distribution of the data.

iTRAQ experiments from the literature. Even when the experimental conditions are highly replicable, the lower-intensity proteins show a much wider range of variation than the higher-abundance proteins.

In the fourth row, the same comparison was made as in the third row; however the correction scalar described in the methods had been applied. This showed a complete reduction in the variation observed previously, enabling a much neater relationship between the samples. It is important to note that whilst the variation seen in row two is diminished in this dataset, it has been diluted relative to the other measurements as there are twice as many data-points in these plots.

The different experimental samples were clustered using principal component analysis, and overlaid onto the same figure. As can be seen in figure 4.13 (p. 158), the sample that had a lower initial mean intensity prior to correction (D4*) still showed large differences compared to the other proteins in the sample despite the mathematical corrections that had been applied. This suggests that a reduction in the initial extracted protein quality generates a lasting and significant alteration to a proteomic data-set that cannot be corrected in data analysis, demonstrating the importance for the reliable, repeatable methods described earlier in this chapter.

Since clustering techniques like PCA determine relatedness between samples by looking at differences between samples – taken as 'components', 'outlier' samples with many

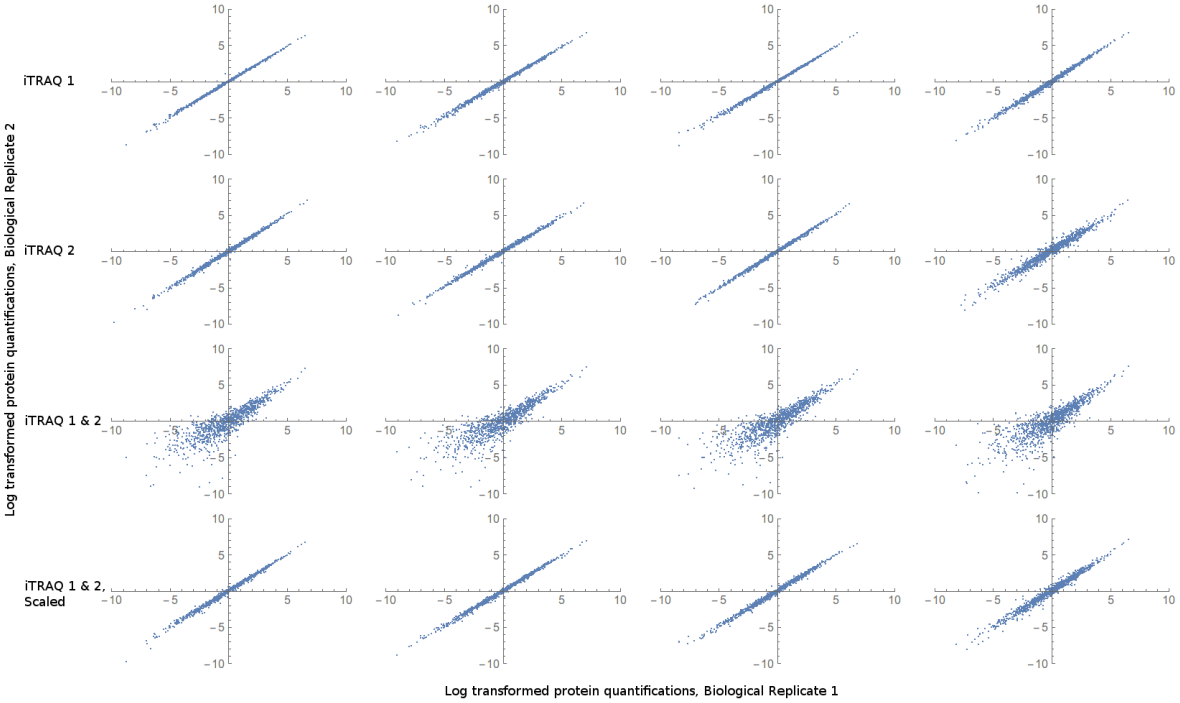


Figure 4.12: A graphs made up of 2 separate iTRAQ experiments. The first row and second row are each of the 4 different test conditions plotted against each other, **within** experiment 1 and 2, respectively. Row 3 is biological replicates plotted against each other **between** experiment 1 and 2 **before** scaling; and row 4 is biological replicates plotted against each other **between** experiment 1 and 2 **after** scaling.

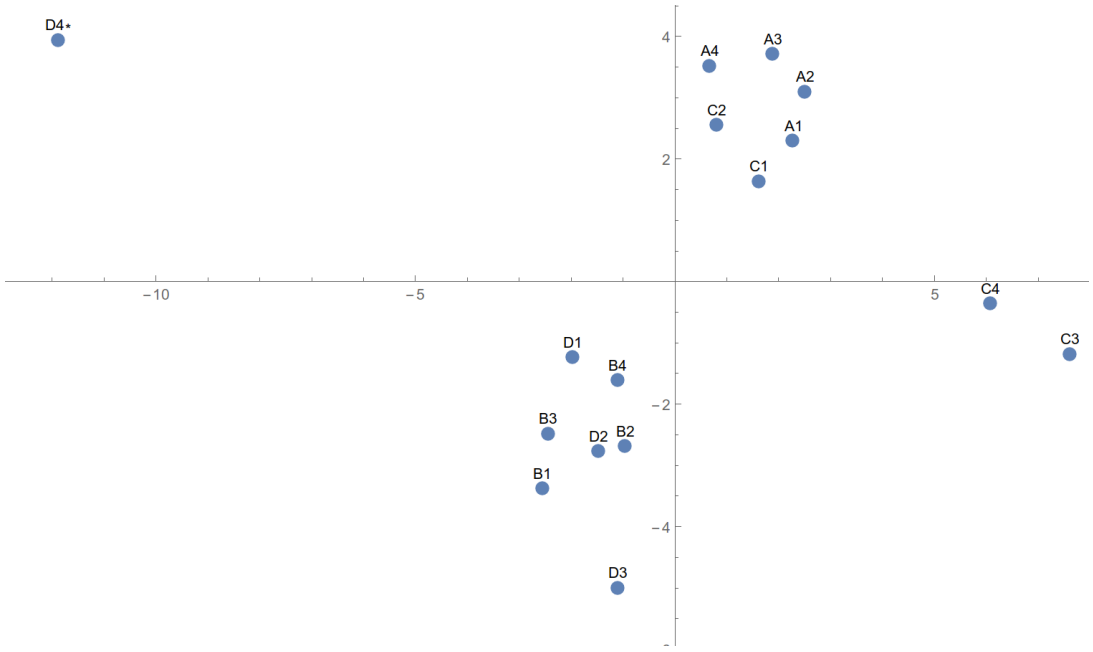


Figure 4.13: A principal component analysis (PCA) on the dataset, where the letters refer to experimental conditions and the numbers refer to replicates. 1 and 2 are replicates from the first experiment, and 3 and 4 are replicates from the second experiment. In this PCA analysis, sample D4 has been highlighted as an outlier.

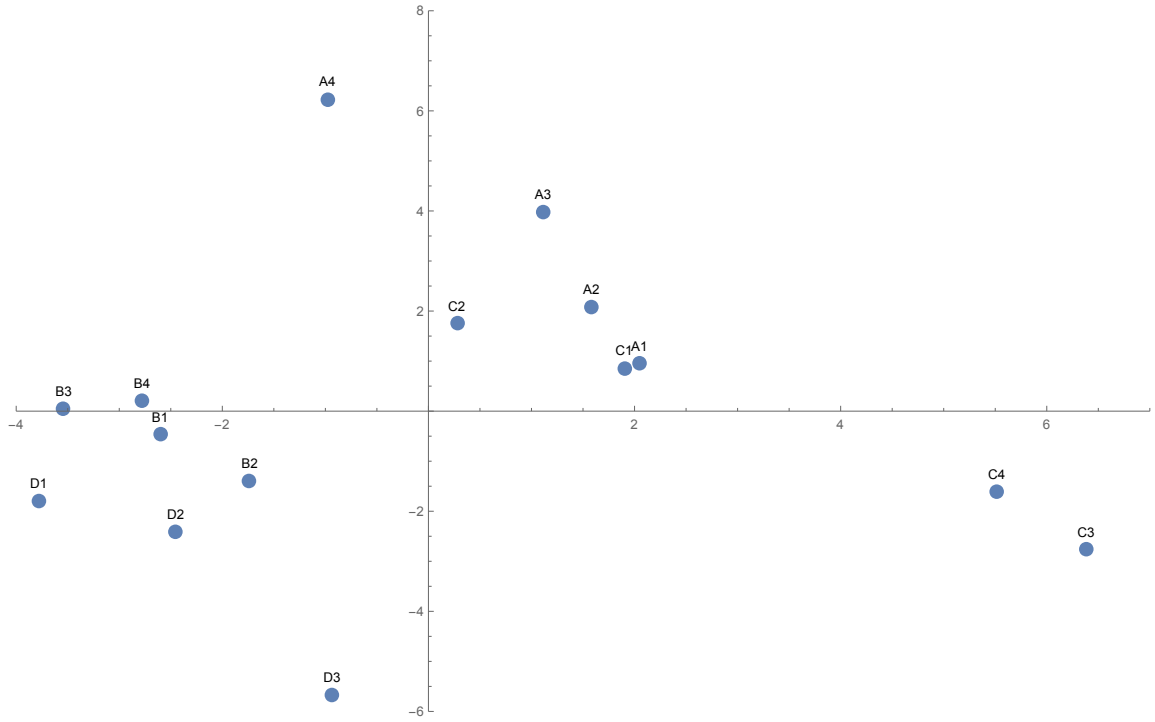


Figure 4.14: This PCA uses the same data as above, but with the proposed outlier D4 removed. As PCA is vulnerable to outliers, since they compress other effects in the data, this re-analysis was important to ensure the close clustering observed was not an artefact of the outlier.

more differences to the other experimental samples will therefore be responsible for a much larger proportion of the variation observed. Since the x and y axis in this analysis represent terms of variation, including outliers in an analysis of this type fundamentally changes the axes, and therefore the metrics determining how the different samples cluster together.

As can be seen, the first principal component – which is plotted on the x axis and explains the largest amount of variation between the samples – shows a mild separation between the A-C cluster and the B-D cluster, but shows a much clearer separation between sample D4 and all other experimental data across the entire study. As a result, the D4 data point was excluded and the clustering was re-calculated as shown in figure 4.14 (159).

After this correction was applied, the data showed a much more general separation across the entire variation space rather than being much more strongly dictated by a single value. The clusters in this case are therefore more diffuse, but enable an objective consideration of how effectively the merging method brought the two iTRAQ datasets together.

In this case, the letters on each point describe the experimental condition, whilst the numbers describe the replicate. All replicates denoted as 1 and 2 were taken from the first iTRAQ and those denoted 3 and 4 were taken from the second iTRAQ. There are cases of replicates taken from the same iTRAQ still show a closer relationship to each other than

to the second iTRAQ, such as C1&2 compared to C3&4 – due to inherent measurement differences in the two spectrometer runs; however generally the clusters appear to form independently of the iTRAQ, indicating a successful merging of the separate datasets.

4.6.5 Conclusions and Discussion

The methods described in this section show a clear ability to join multiple quantitative mass spectrometry experiments together, using methods where there is a shared identical sample across the experiments, as well as flat experimental repetitions. The potential for these techniques to enable much broader analysis of multi-plexing tag experiments is large, and the methods described here are relatively simple.

Generally, other studies into combining data from multiple quantitative proteomic tend to either ignore the quantitative variation (Mota et al., 2015), or attempt to use the data to create a general model for iTRAQ experiments (Hill et al., 2008). The issue with either of these approaches is that they ignore the hierarchical variations introduced at numerous different levels of the experiment, or try to generalise it – which can in turn mask findings from lower-abundance proteins in a given experiment.

Whilst the quantitative data has been mathematically collapsed in this analysis, and therefore cannot be used in a meaningful way across the experiments; the general directional changes – arguably the more useful feature of large datasets like proteomic analysis – are maintained.

In real terms, this method produces a set of proteins showing a consistent change in protein regulation; but escapes the pit-falls of targeted statistical analysis on a non-independent dataset. As a result, data from multiple quantitative proteomic analyses can therefore be used much more reliably for the commonly-used clustering methods, such as PCA, KEGG pathway mapping, and heatmap analysis.

The statistically significant changes method is by far the most common method for detecting changes within a proteome dataset (Pascovici et al., 2015), however it has a number of issues related to it. Statistical tests in biological samples are faced with a significant challenge: the relationships between different proteins are not completely independent; and whilst there are a large number of individual aspects of the system to measure, but there are generally a relatively small number of experimental replicates compared to the number of tests being conducted. So on the one had, the researcher cannot really apply parametric tests like the T-test – which assume an underlying independent, normal distribution – but many ‘omics-level experiments lack the sheer number of replicates needed to conduct non-parametric tests, where generating the number of replicates needed to increase the statistical power of these tests to produce meaningful results is prohibitively

expensive.

Ultimately, many opt to use the parametric tests, but recognise that there may be additional factors that confound the findings of the study. This is why generally the findings from proteomic experiments are verified by an additional experimental method, such as an enzymatic or observational study.

Additionally, as a large number of proteins are detected during a proteomic experiment, a multiple-test correction, such as the Holm-Bonferroni or Šidàk method is required (Abdi, 2007). These are good for reducing the number of false-positives observed in a sample, however they assume independence between the tests, which is not the case in a biological sample, and as a result generate a disproportionately high number of false-negatives.

Furthermore, even though the quantitative information is not compared across the studies, proteins that show a larger fold change in one experiment also show a stronger test statistic, such as a P-value or Z-score. Since this can vary between experimental repeats if it isn't scaled out, as described here, it can introduce hidden bias into experimental studies.

As a result of the issues associated with statistical significance; whilst having two separate lists of differences, this scaled merging of the data ultimately provides a neater way to merge experimental findings than the pre-existing methods. To address concerns of scaling the numerical difference to match that of a single experiment, it is recommended that the scalar method is run twice, with the scalar being calculated first for the second experiment, then for the first experiment by reversing the datasets in the algorithm.

The effect of this would be to generate two separate states for the data, which could then be compared to find differences, selecting only cases where the changes are consistent between the two samples. The advantage of this method over just comparing statistical tests and taking consistent changes, is that firstly a greater number of replicates are available for analysis in each case, as the scaled datasets can be used for both comparisons. In addition, truly inconsistent measurements remain inconsistent with linear scaling, whilst small scale effects are amplified (or compressed).

This should result in a less stringent set of parameters for measuring changes, enhancing the statistical power of the test and producing fewer false negatives whilst not disproportionately enhancing the number of false positives.

4.7 Cluster analysis – using GO terms

The dataset used in this section was produced by Caroline Evans, Jen Parker, and Graham Stafford as part of a CBMnet project with Fuji.

4.7.1 Abstract

Analysing an entire proteomic dataset can be challenging, particularly when limited or conflicting information is available from the published genome. Gene ontology terms are a collection of tags that give basic information about the properties a protein, or proteins related to it through shared evolutionary history. In this section, a method for clustering proteins together, determining grouping sizes, and assigning Gene Ontology labels to each cluster is described. The features of the clusters can be combined with the tags associated with them to produce a number of high-level statements to be made about the data, similar to the way statements about the metabolism can be made with KEGG pathway maps. Unlike KEGG, because of the broader reaching features listed in GO terms it is possible to include structural and organisational information not normally available in a summary analysis.

4.7.2 Introduction

As described in Chapter 1, GO terms are a useful convention that have been developed to enable a better understanding of non-model genomes. Many of the emerging genomes of interest within the field of industrial biotechnology, and particularly within the phycological branches of industrial biotechnology, are not model and are highly dependent on carry-over studies on better understood plants, such as *A. thaliana*. Whilst GO terms are a useful metric for determining the features of a proteome, it can be difficult to determine what the changes within them mean on a case by case basis – especially when a whole-cell system, or other similar large datasets tend to change in a systematic manner, with one part affecting others.

A number of methods for analysing large datasets tend to look at methods for clustering the data together into smaller groups. As described earlier in this chapter, these include heatmap analysis, principal component analysis and KEGG map overlay for protein data. Data clustering can be conducted in a number of unique methods, although ultimately each of these follows the same basic philosophy. Two points within the dataset are determined to be the closest together and are grouped, along with a metric indicating how closely related they are. These are then considered as a single data point and the process is repeated, until all data points and grouped data clusters in the set are

paired together. Clustering can either be performed bottom-up, where all data points are arranged into a hierarchy of relatedness, or top-down, where a number of clusters are assigned to the data and the points are grouped together.

Bottom-up clustering

There are two key variations that can occur in this method, due to challenges associated with clustering, the first is determining the most accurate final clustering, since pairing the two closest samples at the first level may not produce the shortest links between all data sets. Calculating the closest pattern of clustering is a problem that becomes exponentially more difficult when either the number of points to be grouped increases, and is often referred to as the 'travelling salesman problem', where a salesman needs to visit all the cities in a country once, but also wants to take the shortest possible path between all cities.

There are multiple proposed methods that address this problem, including brute-force approaches – where every combination is tried, and the best outcome is determined at the end by the closest clustering – however this is an exponential scaling problem that is impractical for large datasets. Algorithmically faster methods have also been proposed, such as the greedy algorithm – where every combination is tried at the first pairing and the two closest clusters determined at this point are grouped. This is then repeated at each subsequent pairing level, with a reducing number of comparisons to make each level. The complexity involved with this problem increases as a triangular number, instead of an exponential.

The second challenge is determining the 'distance' between two points in high-dimensional space. In cases where there are many dimensions, but a point is only changing in a systematic way in few of those dimensions, stochastic error can result in similar points being clustered far apart. In the case of proteomics, dimensions are considered to be different iTRAQ labels; so this problem can be framed as follows: If two related proteins are changing in a fixed manner in 2 labels (one experimental condition), but vary randomly in the other 6 labels, then the two proteins may end up clustered far away from each other, even though there is a systematic effect on those two labels, due to the contributions from the non-systematically changing labels.

The metric that determines how closely related two points are ultimately determines the order and magnitude of clusters, and so using different clustering methods can produce profoundly different clustering outcomes. In this thesis, all clustering was performed using the Ward distance (Batagelj, 1988), where the distance between any two points is considered in Euclidian space. This method considers a point – or the centre point of a

cluster – in all dimensions, and mathematically connects it to its closest partner. In 2 dimensional space this could be considered as drawing a straight line on a graph between 2 points, then measuring the distance as the square drawn from that line (since the points are varying in 2 dimensions).

This method was chosen as the author found it the easiest to relate to, and since linear measurements of variation across scaled proteomic labels in n-dimensional space scaled linearly with the increasing number of points in the dataset – which was a useful feature when analysing proteomic data in fixed-label experiments such as iTRAQ 8-plex. Alternative algorithms that scale more efficiently with progressively increasing dimensionality (ie. for a 10-plex, or indeed even a 40-plex experiment) (Aggarwal et al., 1999). Whilst these were not necessary for the work conducted within this thesis, they may be more useful in cases where large numbers of proteomic experiments are merged together, using the methods described in the previous section. This is important, because as the number of dimensions increases, all points become further apart from each other, due to stochastic variation.

Top-down clustering

Whilst bottom-up clustering methods are useful for creating a complete hierarchy, this is not always useful in larger datasets. In this chapter, a machine-learning technique called K-means clustering is also employed; where a fixed number of clusters (K) are requested from a system. These are calculated by assigning K arbitrary values randomly into the system and determining the distance between them and the data points – in this chapter using the Ward method. The points that are closest to the values are grouped together as K individual clusters, and the arbitrary value is changed to be the mean of all the values in its cluster. The points closest to this new value are then re-assigned into the cluster. The process repeats until the means no longer change, and then all points within a given cluster are assigned and the calculation ends.

Heatmaps

Whilst clusters can be assigned as described above, visualising the data for a researcher to gain meaningful understanding of the dataset on the basis of their scientific training and accumulated knowledge is essential for generating understanding from the data. As a result, a number of simple tools can be used to make the process simpler. The most commonly-used method for interpreting patterns in large datasets is the heatmap.

A heatmap is essentially a large table of numerical results, however the values are converted to colours to facilitate a fast-scan approach for understanding a dataset. In a

randomised dataset, this is generally impossible for humans, as without blocks of colour the trends are not readily visible – and so a heatmap is only a useful tool for visualising data after a clustering transformation, not just a standard table of data. In addition, the data must be appropriately scaled, otherwise a single high value may distort the heatmap into a solid block of colour that cannot be interpreted by the human eye. As a result, data in heatmaps is typically scaled by either the row or column, rather than being the raw values from the entire dataset. Despite its limitations, by organising and scaling the data appropriately, a heatmap can be a powerful tool for a researcher. They are particularly useful for interpreting groupings in a complex dataset with multiple dimensions – such as 'omics data.

Outline of the GO clustering method

In this section, a novel method for visualising data features is described. In the method, the data is clustered in a bottom-up hierarchical method and overlaid onto a heatmap. The data is then limited to a fixed number of clusters, determined with a K-means algorithm, to simplify the investigation. The most frequently appearing GO terms within the dataset are then assigned, by relative frequency, to each of the clusters. The overall output is a heatmap, with a series of GO tags associated with it. This enables a researcher to pick out general effects in a heatmap cluster – such as a systematic change in proteins under certain experimental conditions – and make statements about it on the basis of gene ontology.

4.7.3 Methods

To begin the analysis, all peptide data are collapsed into protein quantifications as follows. All code for this transformation is available on the accompanying digital code repository (DOI: 10.15131/shef.data.5327524).

Each of the labels are median-corrected, as described in the previous section. The geometric mean is taken (normal mean calculation on log transformed data) from the peptide quantifications as the protein quantification. These data are then arranged into a numeric matrix, with proteins listed in rows, and label quantification intensity in columns, using the program R.

To determine the number of clusters to break the data into, an iterative K-means analysis was conducted; where the data was clustered into a range of K clusters from 1 – 20. The sum of squares error was calculated between the data points in a given cluster and the cluster mean, summing multiple clusters together. These were plotted onto a graph,

and a manual decision was made as to the 'optimal number of clusters', where the point of diminishing return from an increasing number of clusters was reached. The optimal number of clusters was then used as a cut-off point on the bottom-up hierarchical cluster data, generating a number of subsets within the proteomic data.

This is the weakest step in this process, since despite numerous attempts, I have been unable to automate this step of the analysis. As a result, it remains subjective and requires tweaking by the user to determine the 'right' number of clusters.

The GO terms for the proteomic data being analysed were downloaded from the uniprot website. Using the program Mathematica, these were linked to each of the proteins inside the clusters. For simplicity in this analysis, only GO terms with more than 20 or more unique references within the dataset were extracted, and the remaining terms were discarded. The set of all the remaining GO terms within each cluster – assigned by protein – was tallied, resulting in a matrix indicating the number of proteins associated with a given GO term in each cluster, with clusters being listed as rows and GO terms as columns. Each of the GO tallies were divided by the sum total for each term, generating a value between 0 and 1 for each term in each cluster. The clusters with the highest proportion of counts for a given term was assigned with the GO tag.

This is biased towards larger clusters – for example, proportionally smaller clusters will typically contain a lower GO tally count, by virtue of having fewer proteins overall. This could have been normalised by dividing the number of counts by the number of proteins in a given list, but this in turn causes a bias towards assignments in smaller clusters.

The data was then plotted in a heatmap, showing both the full level bottom-up clustering and coloured groupings, and a bar chart highlighting the relative levels of GO terms in each of the different clusters.

4.7.4 Results

A number of simulations for determining the optimum number of clusters were run. Determining a useful cut-off is challenging, as it varies greatly from dataset to dataset. In this case 8 clusters were chosen, based on accounting for an $\approx 90\%$ reduction of the variance observed in the initial analysis. In addition, it also translated as the last 'linear' point in error reduction, after the initial large drops attained in the first 4 clusters (Fig. 4.15, 167).

This was overlaid onto the the bottom-up cluster pairs to determine the distance between the different branches that were being assigned, showing a similar pattern of diminishing distance to that observed in the K-means clustering error (Fig. 4.15, 167). This indicated

Figure 4.15: The error outputs from each of the K-means top-down cluster simulations (left), and the hierarchical graph of the bottom-up paired clusters, resulting from the selected K-pairs cut-off in the bottom-up clustering (right). The error within the dataset shows a diminishing reduction in error as the number of clusters is increased. The selected clusters in the hierarchical graph show this diminishing error is present in both bottom-up and top-down calculations, as can be seen from the length of the edges in the graph.

that K-means is an appropriate proxy for determining the distance between the clusters, however an algorithm for determining a suitable within-cluster distance for the optimum user level is not forthcoming from this data.

This analysis was applied to the data from a study in *E. coli*, where the proteome had been subsetted into investigations on a membrane protein fraction, a soluble protein fraction, and an excreted protein fraction. The membrane fraction is shown here as an example of how the GO terms were assigned to different clusters present in the heatmap (Fig. 4.16, 168).

4.7.5 Conclusions and Discussion

The method in this section was devised as a completely novel method for interpreting complex proteomic data using GO terms within a heatmap-based clustering tool. The problem was not trivial, as proteins have multiple GO terms associated with them, and unique GO terms can be assigned to multiple different proteins that could be changing in different ways within a dataset. Whilst this method was applied to an industrial investigation report with moderate success, it still requires further refinement and development, particularly in the 'optimal clusters' processing.

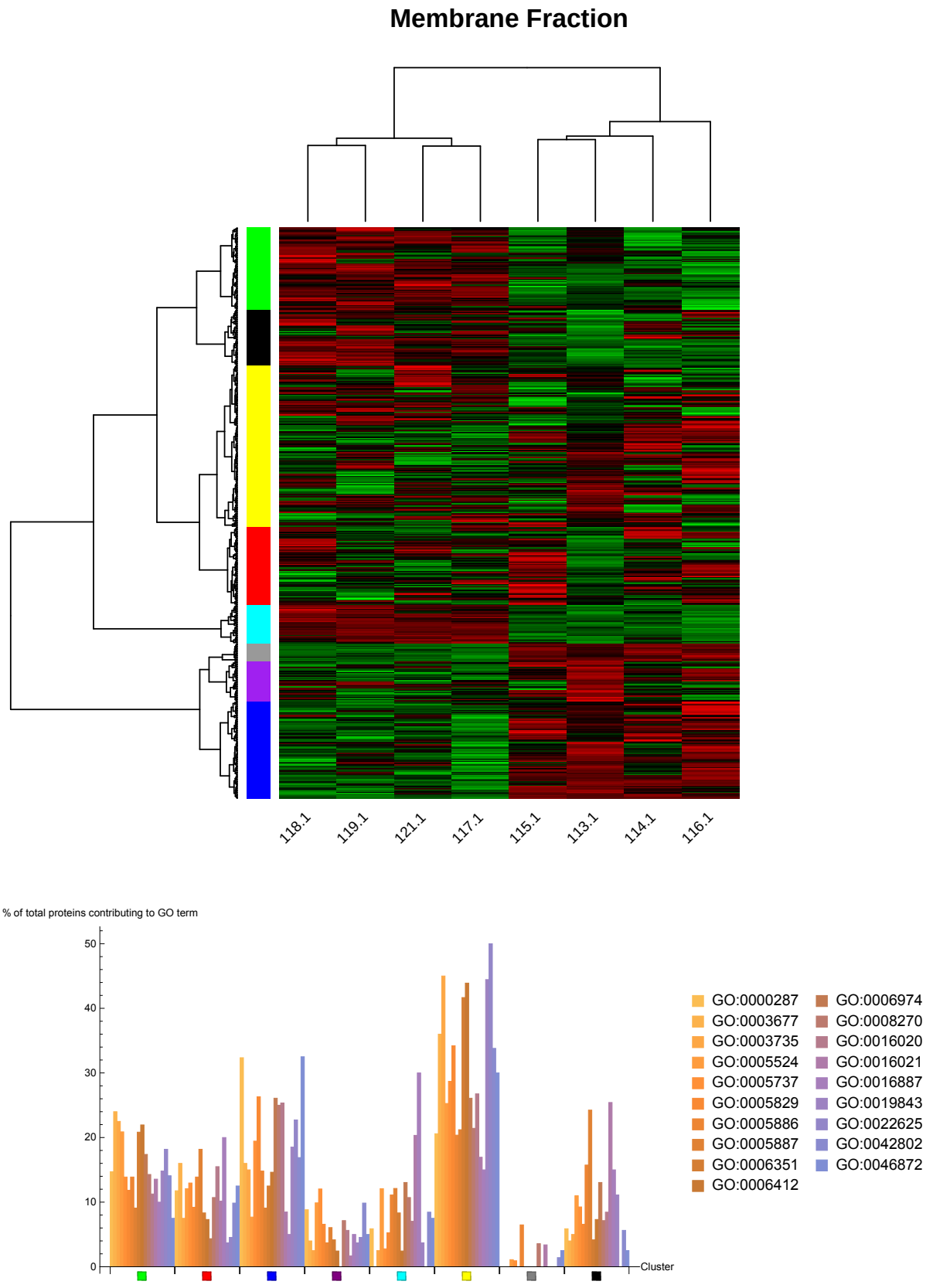


Figure 4.16: An example of how the output of the GO cluster tool would look when applied to a dataset. The heatmap shows the different grouped clusters, each assigned to a different colour (top). The GO analysis linked to the different clusters is displayed in a bar chart, where the colours indicate the cluster the analysis is linked to (bottom).

A major issue with the analysis in its current form is determining the cut-off point for a given cluster. In this case a justification was determined through a statement about the amount of error that had been explained and an arbitrary limit for diminishing returns; however a similar justification could be applied for picking four clusters as the optimum value, notably because they show the biggest reductions in overall data error.

In addition, from a visual point of view when overlaid on the heatmap they produce an intuitive set of groupings – for example, if the cut-off was applied at 4, the groupings would appear as (green, black), (yellow, red), (cyan), and (grey, purple, blue). These groupings 'appear' to be simpler choices, however that assumption is made from a human level observation. This is an interesting problem, as by using the clustering method it is possible to pull out links occurring within the data that aren't obvious to a researcher; although since the decision is arbitrary it is also possible that the computation is being over-complicated.

Determining how to assign the GO terms to the different clusters is a remaining challenge that is not trivial to solve. The aim of the method used was to highlight clusters that are relatively concentrated within a particular subset of the data, however different proteins have a different number of GO terms associated with them – so better studied proteins will tend to have a higher number of GO term tags. One option is to look at the proportion of tags within a specific cluster, by dividing the GO term tags by total number of proteins in a given cluster – normalising the relative levels by cluster size. This would work well in the example put forward in this chapter – where the most abundant terms were the only ones included – however one of the aims of this work was to make a generalisable assignment that worked for all tags, not just the most frequently occurring. Ultimately, with more time a more detailed investigation into developing the simplest and most accurate assignment could have been conducted.

Another problem with this tool is making meaningful statements about the changes taking place within a cluster from the protein-heatmap level. In this investigation there were 4 replicates of 2 test conditions, and so standard statistical methods would produce 3 clusters – increased protein levels in condition 1, decreased protein levels in condition 1, or undetermined. An advantage of this tool is that it can investigate more complex relationships between samples within clusters, however in its current state the tool has no capacity to simplify this process for the user. Ultimately, in its current form it provides a simpler way of running a large number of tests as the number of comparisons increases, where whilst 2 test conditions produces 3 different statistical states, even increasing this to 3 test conditions could raise the number of states as high as 13 – when considering direction of change and combinatorial effects.

The complexities in GO terms mean that whilst this tool has advantages, it is susceptible

to the same errors that arise during a typical GO analysis, including human error in assignment and the lack of directionality involved with some terms. For example, two proteins may both be related to nitrogen metabolism; however individual investigation into the publications for those proteins are needed to determine if they are involved with increasing or decreasing the available levels of nitrogen (Ashburner et al., 2000).

Ultimately, despite its shortcomings and the need for further development, the GO cluster performs the role it was originally intended for, which is to give the researcher a brief overview of what general changes are taking place within a proteomic dataset.

Chapter 5

Isobaric tag comparison

5.1 Chapter Background

Isobaric tags are an important tool in the proteomics toolkit for multiple protein comparisons in a multiplex format, which allows comparison between technical and biological replicates. In this chapter the two most popular tagging systems, isobaric tags for relative and absolute quantification (iTRAQ) and tandem mass tags (TMT), are compared in a *Synechocystis* background against a controlled pseudo-complex model mixture. This is important for making a rational choice about which tag set to use for a given experiment, and highlights some of the issues that surround the choice of tag. This chapter also covers an investigation into the proteomic background of *Synechocystis*, and looks at a label-free emPAI model of a blind dataset to determine the median protein concentration within the proteome. This background proteome investigation was used to determine the best region to run a tag comparison, so that the findings were optimised for a typical candidate protein within a proteomic investigation.

A number of individuals contributed to the work presented in this chapter. Andrew Landels wrote the chapter, calculated background concentrations, performed in-silico analysis, designed the experiment, created the proteomic background mix, assisted with the production of the spike-in mix and analysed the data. Bagmi Patternak and Pia Lindberg provided the dataset that was used to generate the in-silico analysis. Narciso Couto produced the spike-in mix from proteomic standards, combined the spike-in and background mixes, fractionated the samples by HPLC, and ran samples on the Maxis mass spectrometer. Caroline Evans ran the samples on the Orbitrap mass spectrometer.

5.2 Abstract

Quantification in proteomics is important for generating understanding of the underlying biology. Relative quantification with isobaric labels such as iTRAQ or TMT greatly simplifies this, however a number of external factors can affect the values returned. Using a controlled mix of four proteomic standards – ranging from 25 to 70 kDa – and a unique experimental design covering a large range of relative quantifications – from $1:40$ – we performed a direct comparison between the two labelling systems. These comparisons were performed over a range of different background conditions, from a small range ($1:4$) in a protein-sparse 'simple' background, to an expanded range ($1:40$) in a protein-rich, 'complex' background.

An investigation of the proteomic background was performed using emPAI and a secondary dataset, to determine the appropriate quantity of protein to spike into the background. This was done to ensure that it mimicked a typical protein in an experiment, but gener-

ated enough identified peptides to enable more accurate quantification. In *Synechocystis*, this was determined to be 0.5% of the total protein background by mass.

Both tags experienced compression in an increasingly complex background. The compression effect was found to be stronger in lower-concentration proteins, suggesting thatcombining label-free quantification methods, such as exponentially modified protein abundance index (emPAI), with quantitative proteomic methods could greatly improve the confidence in observed results.

The key findings from this investigation suggest that TMT produces a more and precise (variance) and accurate (mean) in a complex background, but the individual labels are less precise and accurate in a simple background. This conflicting finding is driven by an ≈50% increase in the number of confident peptide IDs in TMT experiments compared to iTRAQ experiments; and so whilst the tags are less effective on an individual level the additional data makes up for this.

This raises questions around the trend towards increasing the number of labels within a single experiment (TMT 2-plex, iTRAQ 4-plex, TMT 6-plex, iTRAQ 8-plex, TMT 10-plex), where using multiple experiments and knitting the datasets together may provide a more accurate set of data (See chapter 3).

5.3 Introduction

5.3.1 Background to isobaric tagging in proteomics

In bottom-up quantitative – or 'shotgun' – proteomics, tag-based stable isotope labelling is widely employed. As evidence for this, a search on google scholar with the terms *proteom* iTRAQ OR TMT* returns almost 18,000 publications citing the use of either iTRAQ or TMT labelling (*search date 18/10/2016*), the two main covalent tagging technologies employed within the field (Noirel et al., 2011; Evans et al., 2012; Altelaar et al., 2013b; Zhang et al., 2013b). This technique involves covalently linking a series of tags – with identical chemistry, but varying internal mass distributions through the use of isotope integration – to peptides generated from a number of different proteomic samples. These different samples are then pooled together and run collectively. When the peptide spectra are observed in the mass spectrometer, each tag produces a unique 'reporter ions' present within the spectrum that can be directly compared to give a ratio abundance between the different samples.

For an introductory description to the field of quantitative proteomics, please see the introductory review by Altelaar et al. (Altelaar et al., 2013b), and for a highly detailled

review of this process, please see one of the following comprehensive assessments of the field, by either Zhang (Zhang et al., 2013b) or Rauniyar and Yates III (Rauniyar and Yates III, 2014).

iTRAQ and TMT are two different commercially available reagents used to label peptides in proteomic experiments. The reagents bind via N-hydroxy-succinimide chemistry (Thompson et al., 2003), and so the tag is attached to peptides at all free amine sites – namely the N-termini of peptides and the epsilon amino group of lysine. Both tags generate low mass reporter ions under tandem mass spectrometry fragmentation, which produce a signal in a region of the mass spectra that is usually clear for peptides. Before undergoing this fragmentation step, the tags are all of equal mass (isobaric), where the reporter group is balanced through distribution of isotopes in a separate balancer or normaliser group. Isobaric labelling, in multiplex format, provides an additive effect on precursor intensities as the ms^1 signal is made up of the combined signals of 6 (in the case of TMT 6-plex) or either 4 or 8 (in the case of iTRAQ 4-plex and 8-plex) combined protein pools – (details shown in chapter 1), which theoretically increases the sensitivity of detection.

Another similar tagging method is N,N-Dimethyl Leucine (DiLeu) tags, which like iTRAQ and TMT are isobaric when measured at the MS^1 level, but produce a range of 4 identifiable tags at the MS^2 level (Xiang et al., 2010). Rather than being purchased from a supplier, these tags offer an economic alternative, as they can be synthesised in the laboratory at a greatly reduced price.

Issues relating to quantification accuracy with iTRAQ and TMT labels have led them to mainly being used as lead generation or 'discovery' technologies in medical-based studies, rather than confirming actual quantifications of proteins within a sample (Noirel et al., 2011). Indeed in some cases no statistical repetition is performed within the experiment until after potential leads have been identified, with 8 different samples being examined on 8 different iTRAQ labels (Evans et al., 2012). Similar issues arise for TMT. They are also employed for industrial production strain analyses (Landels et al., 2015). When investigating the effectiveness of proteomic quantification methods, two main aspects are typically studied – accuracy, or how close to a true value the measurements that are taken are (Keshamouni et al., 2006; Glen et al., 2008; DeSouza et al., 2008; Bantscheff et al., 2008; Kuzyk et al., 2009), and precision, or how closely grouped the measurements are (Chong et al., 2006; Gan et al., 2007).

5.3.2 Advantages of tag-based approaches

There are a number of advantages to using tag based quantification, it mitigates the technical challenges associated with label-free methods for comparisons between separate samples. Individual mass spectrometer 'runs' can be affected by a wide variety of factors, including stochastic variation arising from the physical measurements of the sample; which means that although an identical sample may be injected into the machine one two consecutive occasions, the chances of observing an identical data output are vanishingly small. This can cause problems for direct comparisons between samples, as it is challenging to determine where a sample difference is actually occurring between samples and when it results from stochastic variation between sample measurements. As a result, cases where label-free quantitative measurements tend to be more accurate are typically 'targeted' studies, which focus on a smaller number of peptides and infer information about the sample from them (Blein-Nicolas and Zivy, 2016).

Label-free analyses require a minimum of 3 repeated runs on samples (Cox et al., 2014) to generate the confidence required for these comparisons, which also has an operational cost associated with each experimental measurement, for both the physical operation of the machine and also the operator time requirements. As mentioned above, in label-based proteomics the samples are pooled, and up to the point where they generate a spectrum on the machine, are considered to be identical both chemically and by mass. As a result, the different samples experience an identical set of analysis conditions, and so each measured peptide can be compared directly to determine the relative abundances between the level of the protein between the two samples. This translates to a physical reduction in the required number of spectrometer runs and therefore the overall costs associated with the experiment.

5.3.3 Ratio compression in tag-based approaches

Whilst tag-based proteomic studies generate large amounts of quantitative data more efficiently and accurately, compared with spectral counting techniques, there are a number of limitations that have been highlighted with the technique that need to be addressed. One of these is the effect of ratio compression. In ratio compression, the ion count ratio within the spectrum, which should be representative of the absolute abundance of the peptide in the samples being compared, is underestimated in complex samples. The phenomenon was initially documented by Bantscheff et al (Bantscheff et al., 2008) in an attempt to generate a more robust and sensitive set of iTRAQ quantifications, and was attributed to an issue called isotopic contamination – or contamination with molecules that have a similar or identical mass. There are two main forms of isotopic contamination,

the first is contamination with near isobaric masses – such as the case where the 115 mass signal in the iTRAQ labels is inflated in arginine-containing peptides (Casado-Vela et al., 2010) through the inclusion of a arginine-derived mass (Gehrig et al., 2004) that has a mass to charge ratio of 115.08, whilst the iTRAQ 8-plex tag has a mass of 115.1 ± 0.01. A second case of this is contamination with an isotope of the phenylalanine immonium ion, which typically has a mass of 120.08, but due to the presence of naturally occurring C^{13} within the peptide results in an isotopic contamination with the iTRAQ 8-plex 121.1 ± 0.01 label (Ow et al., 2009). (Details of the full labelling regime are given in Chapter 1)

These examples of contamination presented here result in inflated values for an individual label, which can skew findings especially when internal biological repeats (multiple labels for the same experimental condition) or repeated experimental runs with randomised labels (multiple iTRAQ kits used on the same experimental proteomic samples) are not performed (Hill et al., 2008). These near-isobaric isotopes can be separated by high resolution mass spectrometers. The second form of isotopic contamination is generated from isobaric masses. This occurs in cases such as attempting to differentiate leucine and isoleucine, where the two molecules have an identical mass but are different atomic arrangement, or from attempting to perform quantifications on shared peptides – peptide sequences that are identical in two different proteins (Jin et al., 2007; Zhang et al., 2010; Dost et al., 2012). This type of contamination cannot be directly corrected for by improving the machine resolution, but can sometimes be inferred from other information in the dataset.

A number of studies demonstrated that where the physical ratios grew larger, the observed ratios failed to keep pace, showing a reduced difference or 'compression effect' between the samples (Pierce et al., 2008; Keshamouni et al., 2006; Glen et al., 2008; Bantscheff et al., 2007, 2012). This effect was amplified by increasing the ratio difference between the two proteins (Ow et al., 2009) and was shown to be a linear effect (Karp et al., 2010). The majority of experiments investigating this effect use either a control mix of proteins at known concentrations within a protein sample (Ow et al., 2009), or the same entire protein sample compared at different concentrations; and showed that the more complex the background proteome, and the higher the dynamic range of the sample (Bandhakavi et al., 2009); the greater the rate of sample compression. The effect was found to be the result of a phenomenon called co-isolation, where two peptides of similar mass were both analysed at the same time (Ting et al., 2011; Christoforou and Lilley, 2011; Wenger et al., 2011). This confounds the values that are measured, and because the measurement is ratio driven any addition to the signal will result in a compression of values.

Since the majority of proteins within a proteome are expected to be unchanging (a 1:1 ratio), the majority of contaminants are expected to push towards this state resulting

in compression. From a practical sense it helps to think about the phenomenon from a mathematical point of view, where if you have two values, 1 and 9, and add 1 to each, it reduces the ratio between the two values from 1:9 to 1:5; and if you add a value greater than 1 to both then the compression effect is increased. This problem can be solved through the use of MS^3 – a technique where the MS^2 scan fragmentation energy is tuned to fragment the peptide to produce y and b ions, and then these fragments are then further fragmented at a higher energy to release the isobaric tags in a separate scan for quantification (Ting et al., 2011; Christoforou and Lilley, 2011). (b and y ions are described in chapter 1)

Alternatively, a technique referred to as QuantMode employs proton-transfer gas phase reactions to reduce the charge state of the target species and separate it from the interfering m/z to improve quantification accuracy (Wenger et al., 2011). Ion mobility (IM) separations coupled to Q-TOF instruments have the potential to mitigate MS/MS spectra chimeracy, since the IM-MS has the ability to separate ions based on the collision cross section, in addition to the m/z (Shliaha et al., 2014). Unfortunately, machines capable of this level of analysis are still prohibitively expensive for many labs around the world and as a result this solution is not currently widely applicable; as a result a number of groups still work on alternative solutions to problems related with co-isolation, improving quantification quality (Martinez-Val et al., 2016; Dowle et al., 2016; Brodbelt, 2015; He et al., 2016; Cologna et al., 2015).

5.3.4 Reduced data return from tag-based approaches

Beyond ratio compression, a major issue with label-based techniques is the reduction in the physical number of identified spectra, compared with label-free techniques. During fragmentation of a tagged peptide, the collision energy in the mass spectrometer must be tuned to liberate the isotopic label from the peptide, which is different to the optimal fragmentation energy for fragmenting a peptide at the peptide bond to produce b- and y-ions – which are used for peptide identification; this reduces the efficiency of the machine and also the quality of the data produced. In addition, there is ideally a single fragmentation event at the MS^2 level, which facilitates the peptide identification stage of data analysis, however there is a finite amount of sample available at any given time within the mass spectrometer and so if the sample must be divided between more m/z signal outputs, then the signal in each of these outputs is reduced. This effect is related to the effect seen when proline-containing peptides are fragmented; the major signal observed is fragmentation at the proline residue, due to its capacity to stabilise a charge more readily than a standard amino terminus, resulting in a diminished signal from all other fragmentation events and an overall reduction in peptide identification

quality (Hunt et al., 1986).

The reduction in the number of identified spectra from different labelling techniques has been reported in the literature, where a label-free analysis showed a 40% higher incidence of identified spectra when compared with an iTRAQ 4-plex experiment (Wang et al., 2011) – although the comparison was performed on different spectrometers, with a higher-end device being used for the label-free analysis which precludes a direct comparison (Evans et al., 2012). Another study found that a metabolic labelling approach using SILAC (stable isotope labelling by amino acids in cell culture) – a tag-free labelling technique – resulted in 14% more protein identifications, but the tag-labelled samples demonstrated both a higher level of precision and accuracy (Li et al., 2012).

The number of observed peptides also diminished as the '-plex' number increased, with iTRAQ 4-plex showing the highest rate of identification, followed by TMT 6-plex, and then iTRAQ 8-plex (Pichler et al., 2010); whilst the study only identified a relatively small number of proteins (70 proteins from 250 unique peptides in the highest case), the data showed a successive reduction in the number of spectrum matches by 15% then a further 50% as the 'plex' level increased from 4, to 6, to 8 – although the exact reduction appears to vary from study to study, where a more recent and statistically robust study found that the level of reduction was 40%, not 60% (Mahoney et al., 2011). This effect was attributed to both the effect described above and to other fragmentation events occurring within the larger tags, as a result of higher-charged species resulting from the larger tag molecules attached to the peptides. The study claimed that precision and accuracy were equivalent between all cases, however due to the small sample size, finer details that are observable statistically were not available within the scope of this study, suggesting that the study we conducted here provides new information to the scientific community. Investigation into the higher-charged species effect resulted in a dynamic or 'on the fly' correction program to modify spectrometer operation, for parallel quantification and identification of spectra (Mischerikow et al., 2010). This was ultimately was the same approach that tackled the quantification compression effect mentioned above (albeit released earlier), but is again limited by the availability of a high quality mass spectrometer.

Ultimately, these reductions in data quantity and quality raised the question of limits of detection. A comprehensive study performed by Mahoney et al (Mahoney et al., 2011) highlighted that a minimum fold-change limit of significance – a threshold very widely used within the literature for a number of different studies – held little to no bearing on whether the protein levels were changing between samples; and that the only way to determine significance is statistically, however detecting changes below a 2-fold difference were challenging, even with the most advanced statistical techniques of the time at the researchers disposal (Oberg et al., 2008; Schwacke et al., 2009; Hill et al., 2008).

Whilst it's clear that labelled techniques reduce the number of identifications, a more important question is whether it has a direct bearing on the outcome of a proteomic investigation. A direct comparison between labelled and unlabelled techniques demonstrated that whilst an unlabelled experiment confidently identified more peptides, and therefore had more robust statistics, enabling the identification of 3-fold more statistically significant changes; these two approaches ultimately provided the same biological story and the additional quantifications gave no additional insights into the system being analysed (Neilson et al., 2011). This study can, in the context of the other research discussed here, be considered to make a statement about the data analytical methods commonly employed for proteomics, as discussed in chapter 3. Despite this, it is clear that having higher quantities of higher quality data will always be beneficial, as long as the additional cost is not too great.

5.3.5 This study

In this chapter, a direct comparison between the two most popular quantitative proteomic labelling systems, iTRAQ 8-plex (Ross et al., 2004a) and TMT 6-plex (See chapter 1 - introduction), was carried out to identify three main points: initially to give an impression of the relative benefits of quantification values between the two systems, given the inherent differences in the systems. Beyond that, it was important to identify potential systematic effects present in *Synechocystis* samples, so that fairer comparisons could be made between experiments using a different labelling system. Finally, if systematic differences were identified it was important to see if they could be accounted for mathematically or computationally to generate more accurate comparisons in post-experimental data processing.

Unlike other experiments of this type that have been carried out previously in the literature, our approach had a combination of four unique selling points. Firstly, independently varying protein quantifications were used for the spike in, rather than a simple dilution series. This enabled us to control the concentrations of the different proteins on different labels independently – more accurately emulating the situation present in a real sample. The previous methods in the literature investigated quantification bias by making a standard mix of different proteins at the same concentration, which were added in equal proportions to each of the labels being investigated – generating a 1:1 ratio across all labels. The labelled protein mixes were then combined in different ratios to control the levels of the spiked-in proteins. An advantage of this approach is that it enabling verification of proposed label-specific effects in the sample – for example, if the labels appear to be showing a systematic bias it can be studied to ensure it isn't just because it has a lower concentration of protein present in the mix. This study is a natural progression of the

field, and before publication of this work another group has independently approached the same conclusions that we did and reported this novel methodology (Ahrné et al., 2016).

Secondly, the protein levels were evenly balanced across all labels in the sample; so that when quantifying the mix without the presence of background contamination standard corrections, such as median correction or channel sum correction, could be applied to the dataset without negatively affecting the results. Without this, using a normal analytic technique such as standard software would re-balance a typical dilution series back to a 1:1 ratio, to try and approximate even protein loading on each of the samples. This is one of the fundamental assumptions in label-based relatively quantitative proteomics: that the total quantity of protein present on each label is equal and what is interesting to the researcher is the changing balance of the individual proteins within the sample, rather than a physical change in loading levels. Incidentally, whilst this assumption is generally held to be true, it is very rare for a 'total protein quantification per cell' value to be given in proteomic experiments.

Thirdly, as the study was targeted to a single organism, *Synechocystis*, the background spread of protein concentrations in a typical sample could be modelled to enable accurate control of the spike-in concentration to give specific information about the precision and accuracy of quantifications observed in proteins at different concentrations within the sample. In this case, the label-free quantification method emPAI (Ishihama et al., 2005) was used to calculate the distribution of proteins measured within a number of *Synechocystis* proteomic samples. From this data, a distribution histogram was generated. It showed a normal-like distribution that rapidly tailed off at the lower detection limit. This leaves an open question about what the distribution of the low-abundance proteome looks like in *Synechocystis*, however identifying the observable distribution provides a key advantage. The sample proteins can be spiked in at a level typical of proteins within the sample, reducing the risk of either swamping the sample by setting the spike-in concentration too high, or failing to observe the proteins at all because they were spiked-in at a concentration below detectable limits.

Ultimately, the main feature of this investigation was that for the first time the iTRAQ and TMT labelling systems were measured in the same experimental framework, using the same MS platform, enabling a direct comparison between the two systems in a controlled manner. Whilst there have independently been studies into a variety of iTRAQ and TMT labelling systems individually, this aspect of the experimental work is completely novel in the literature. Due to fundamental differences in the labels – the most obvious being the different number of labels in each of the systems, with iTRAQ having 8 and TMT having 6, some comparisons in this framework are obfuscated slightly; however the design has been adapted to account for this as far as possible. These steps are highlighted in the

iTRAQ	Myo	Cas	CytC	BSA	Sum
113	3	0.8	1	2	6.8
114	3	0.8	2	1	6.8
115	2	1	3	0.8	6.8
116	2	1	0.8	3	6.8
117	1	2	0.8	3	6.8
118	1	2	3	0.8	6.8
119	0.8	3	2	1	6.8
121	0.8	3	1	2	6.8
Sum	13.6	13.6	13.6	13.6	
TMT	Myo	Cas	CytC	BSA	Sum
126	3	2	0.8	2	7.8
127	3	2	0.8	2	7.8
128	0.8	2	2	3	7.8
129	0.8	2	2	3	7.8
130	2	2	3	0.8	7.8
131	2	2	3	0.8	7.8
Sum	11.6	12	11.6	11.6	

Table 5.1: Experimental design table showing the relative concentrations of the different proteins in the master mix. Efforts were made to balance the mass of protein on each tag and also the total mass of each protein in the sample to avoid bias so measured effects could be generally applicable to other proteins.

methodology for this experiment.

5.4 Methods

5.4.1 Experimental design

4 common proteomic standards (BSA, Cyt C, Myo, B-Cas) were prepared at a concentration of 1 mg ml-1. These samples were then digested with trypsin, the disulphide cysteine bridges were reduced and alkylated, and the peptides from each protein were combined in a single pot for each label at the concentrations given in table 5.1 (p. 181).

The quantities of peptides were balanced for each label, so that an even mass of all proteins was added to prevent concentration bias in the study. The base ratios covered a range of 1:3.75; this range was chosen to look at the effects of differentiating iTRAQ values that were close together. The labelling was performed as per the manufacturer's instructions. The set of labelled peptides was referred to as the 'master mix' and was stored at -80 when not in use. Three experiments were performed, the first analysed the mix itself, and so the labels were combined at a 1:1 ratio and run without a background

iTRAQ	Scalar	Myo	Cas	CytC	BSA	Sum
113	10	30	8	10	20	68
114	10	30	8	20	10	68
115	5	10	5	15	4	34
116	5	10	5	4	15	34
117	2	2	4	1.6	6	13.6
118	2	2	4	6	1.6	13.6
119	1	0.8	3	2	1	6.8
121	1	0.8	3	1	2	6.8
Sum		85.6	40	59.6	59.6	
TMT	Scalar	Myo	Cas	CytC	BSA	Sum
126	1	3	2	0.8	2	7.8
127	5	15	10	4	10	39
128	10	8	20	20	30	78
129	1	0.8	2	2	3	7.8
130	5	10	10	15	4	39
131	10	20	20	30	8	78
Sum		56.8	64	71.8	57	

Table 5.2: Experimental design table showing the relative concentrations of the different proteins after being applied to a scalar dilution. Where the scalar is stated, the concentration was achieved through relative dilution of the other labels, so 10 refers to a $\frac{10}{10}$ dilution, or undiluted label mix, whilst 1 refers to a $\frac{1}{10}$ dilution.

sample; this was done to verify the experimental accuracy of the operator, to ensure that having a pseudo-complex mix wouldn't provide interactions in an uncomplicated background, and to generate correction factors to account for any effects seen as a result of these two effects.

In the second experiment, the labels were combined at a dilution range of 1:10 without a background, as shown in table 5.2 (p. 182). As with experiment 1, this enabled verification of the experimental accuracy of the operator, and highlighted any further compression observed by increasing the concentration range within the pseudo-complex sample. Due to the experimental design, this step expanded the range significantly and uniquely for each protein due to the varying concentrations of protein on each label, as can be seen in figure 5.1 (p. 184), increasing it from $1 : 3.75$ up to $1 : 37.5$. Due to the experimental design, a range of values and concentrations were measured in each of the individual proteins, providing a broader set of values for assessing if the compression effect was linear. These ratios were applied to the spike in proteins by mass in mg, the full list of final mass of protein added in all experiments is available in table **??** (p. 183).

Finally, in the third experiment the labels were combined with the same dilution range as experiment 2, but in reverse-order, both due to a limitation of available sample and

			Ideal Factor	14.70588235	
	Protein – mg	100	Factor	14.5	
TMT	Cyt	BSA	Myo	Lys	Total (mgs)
126	11.6	29	43.5	14.5	98.6
127	11.6	29	43.5	14.5	98.6
128	29	43.5	11.6	14.5	98.6
129	29	43.5	11.6	14.5	98.6
130	43.5	11.6	29	14.5	98.6
131	43.5	11.6	29	14.5	98.6
Total (mgs)	168.2	168.2	168.2	87	591.6
After Dilution					
126	1.16	2.9	4.35	1.45	9.86
127	1.16	2.9	4.35	1.45	9.86
128	14.5	21.75	5.8	7.25	49.3
129	14.5	21.75	5.8	7.25	49.3
130	43.5	11.6	29	14.5	98.6
131	43.5	11.6	29	14.5	98.6
	118.32	72.5	78.3	46.4	315.52
	Protein – mg	50mg per tag	Factor	14.5	
iTRAQ	Cyt	BSA	Myo	Lys	Total (mgs)
113	14.5	29	43.5	11.6	98.6
114	29	14.5	43.5	11.6	98.6
115	43.5	11.6	29	14.5	98.6
116	11.6	43.5	29	14.5	98.6
117	11.6	43.5	14.5	29	98.6
118	43.5	11.6	14.5	29	98.6
119	29	14.5	11.6	43.5	98.6
121	14.5	29	11.6	43.5	98.6
	197.2	197.2	197.2	197.2	788.8
After Dilution					
113	1.45	2.9	4.35	1.16	9.86
114	2.9	1.45	4.35	1.16	9.86
115	8.7	2.32	5.8	2.9	19.72
116	2.32	8.7	5.8	2.9	19.72
117	5.8	21.75	7.25	14.5	49.3
118	21.75	5.8	7.25	14.5	49.3
119	29	14.5	11.6	43.5	98.6
121	14.5	29	11.6	43.5	98.6
Total (mgs)	86.42	86.42	58	124.12	354.96

Table 5.3: Table showing the amounts of protein, by mass, added to each experiment. The values were calculated from the ratio-design tables shown in this section.

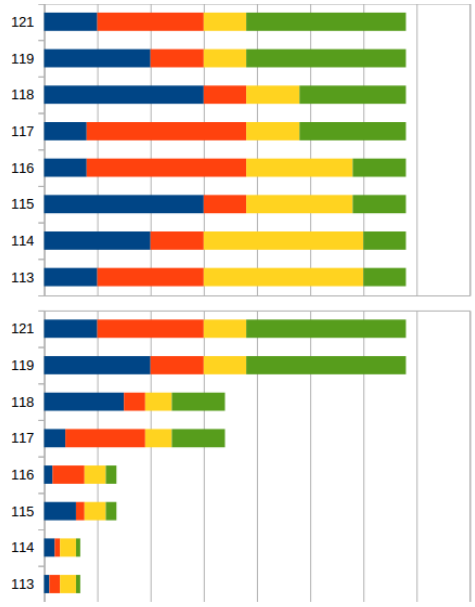


Figure 5.1: These stacked bar charts show how the internal protein concentrations within the experiment are initially balanced in the iTRAQ experiment, but are then expanded to a much larger range through the use of a simple dilution step. In this case, the different colours represent the amounts of different protein labelled with each iTRAQ tag. The top bar chart shows the master mix, whilst the chart below shows the mix after it has had a dilution step applied to it. This dilution step is the same as the one used in the 3^{rd} experiment described here, and the chart indicates the relative amounts of iTRAQ tag, and the corresponding protein levels, spiked into the complex background.

also to account for protein-specific effects, as shown in table 5.4 (p. 185). This sample was spiked into a complex Synechocystis background at a concentration that was typical of the majority of proteins within the background sample.

5.4.2 Calculating the background

To determine a suitable amount of protein mix to spike into the *Synechocystis* background, a statistical assessment was carried out on a sample dataset kindly provided by B. Patternak and P. Lindberg from Uppsala university, Sweden.

An in-depth proteomic dataset, comprised of 2 8-plex iTRAQ experiments investigating a mutant against WT *Synechocystis* under two different conditions, was generated on a Q-Exactive HF mass spectrometer. Part of this dataset was analysed to generate a distribution of protein concentrations within the proteome. The dataset was provided blind, so the mutant, the conditions and the iTRAQ tag assignments are all unknown to the operators to remove bias during analysis. To calculate the emPAI scores, the 'observable' peptide values were calculated as follows.

The complete proteome for Synechocystis PCC6803 – Kazusa strain, was downloaded as a fasta file from uniprot (taxonomy:1111708 – accessed August 2015, 3517 protein entries). This was then merged with the spike-in proteins to make a singular database for analysing the data, by doing this, effects on statistical methods such as false discovery were equal between all analyses. The fasta file was processed in Wolfram Mathematica (version 10.1) to generate an in-silico digest of each of the proteins, excluding any peptides that fell outside a 1000 – 7500 dalton window to replicate the presence of 2+ or 3+ ions observable in the 500 – 2500 m/z window used during the mass spec experimental scan.

iTRAQ	Scalar	Myo	Cas	CytC	BSA	Sum
113	1	3	0.8	1	2	6.8
114	1	3	0.8	2	1	6.8
115	2	4	2	6	1.6	13.6
116	2	4	2	1.6	6	13.6
117	5	5	10	4	15	34
118	5	5	10	15	4	34
119	10	8	30	20	10	68
121	10	8	30	10	20	68
Sum		40	85.6	59.6	59.6	
TMT	Scalar	Myo	Cas	CytC	BSA	Sum
126	10	30	20	8	20	78
127	5	15	10	4	10	39
128	1	0.8	2	2	3	7.8
129	10	8	20	20	30	78
130	5	10	10	15	4	39
131	1	2	2	3	0.8	7.8
Sum		65.8	64	52	67.8	

Table 5.4: Experimental design table, an inverse of table 5.2 (p. 182). In the iTRAQ experimental labels, paired proteins have flipped concentration (magic square effect), however the relative concentrations between the TMT labels have changed between the diluted test mix and the complex background. This was a limitation of the 6-plex:4-protein mix.

The emPAI scores for all identified proteins were calculated using the following formula.

$$emPAI = 10^{(\frac{N_{observed}}{N_{observable}})} - 1$$

Where $N_{observed}$ is the number of unique peptides observed for a given protein, and $N_{observable}$ is the total number of unique peptides that could be observed for a given protein.

This data was then graphed as a histogram to identify the protein concentration distribution and dynamic range. Dynamic range was calculated by taking the exponential of the difference between the maximal and minimal emPAI values. The dynamic range limit gave a practical minimum level for the spike-in data; for example if the observable dynamic range was reported as 10^3 then the lowest concentration of any spike-in protein should be greater than $\frac{1}{10^3}$ of the highest abundance protein in the sample. In *Synechocystis*, it has been reported that the 4 phycobilisome proteins make up approximately 20% of the proteome (Gan et al., 2005). This can be seen clearly from any data involving wild type cells grown in light conditions, with each of the proteins being approximately equivalent in concentration, and so the highest protein concentration here was approximated as being 5% by mass of the total protein sample. As the total mass of the protein sample injected into the mass spectrometer was known, the total amount to spike in was calculated by:

$$max = 0.05 \times total$$
$$min = max \times \frac{1}{1000}$$

Where max is the highest estimated concentration for any individual protein in the sample and min is the lowest detectable concentration of protein in the sample.

$$target = \frac{max}{50}$$

$target$ is the region of concentration the spike proteins should be in, as observed from the data distribution in the histogram 5.4 (p. 192). The code from this analysis, along with the code from all other analyses in this thesis, is available in the data repository <DATA REPOSITORY ADDRESS AND DETAILS>.

HPLC Buffers A + B

The two HPLC buffers contain a high (A) and low (B) conentration of acetonitrile (CH_3CN). These are mixed within the instrument to produce an elution gradient for evenly separating peptides into fractions, reducing the complexity of the mixture for increased measurements in mass spectrometric analysis.

1. Buffer A - 80% CH_3CN
    - 800 ml CH_3CN
    - 190 ml ms-grade H_2O
    - 10 ml 1M ammonium formate (NH_4HCO_2) at pH3
2. Buffer B - 5% CH_3CN
    - 50 ml CH_3CN
    - 940 ml ms-grade H_2O
    - 10 ml 1M ammonium formate (NH_4HCO_2) at pH4

As a result, 1 ug of the mix was added as a spike-in to each experiment. After being combined, the samples containing the spiked-in labels were fractionated by HPLC (high performance liquid chromatography), using a HyperCarb column over a 70 minute fractionation program that had been standardised for *Synechocystis* samples. In total, 48 fractions were collected, one per minute at a flow rate of 0.2 $ml.min^{-1}$ per minute, with the first 16 minutes and final 6 minutes discarded.

For the samples without a complex background, the samples were pooled together in groups of 12, producing a total of 4 fractions. This was done to reduce the overall mass spectrometry data acquisition time, as from the simplicity of the samples they were already fractionated sufficiently. In the spike-in experiment, 2 runs were performed – one with offline pre-fractionation and the other without. The effects of compression exacerbation have been shown in the literature to be enhanced by a complex background (Ow et al., 2009); however no data is available on whether iTRAQ and TMT labels experience this compression effect to the same extent. To limit the differences between the samples, the 'unfractionated' mix was in fact fractionated as described above, but an aliquot containing all the collected fractions merged together was spiked into the machine. In the fractionated mix, the samples were pooled together in groups of 8, producing a total of 6 fractions for analysis. The samples were measured with both the maXis UHR ToF and QExactive HF mass spectrometers (Bruker, Bremen, Germany; Thermo, Bremen, Germany).

5.4.3 Data analysis

The data processing route was kept as close to a typical data processing run as possible. As a result, the .raw data files were directly analysed with the program MaxQuant (version 1.5.3.30), with the recommended settings for each analysis except where stated. In addition to the standard post translational modifications, phosphorylation of serine was

added to improve the quality of the beta casein protein identifications. The false discovery rate (fdr) was set at 1%. The same fasta database from Uniprot used to produce the emPAI background was used for analysing the proteome (taxonomy:1111708 – accessed October 2015).

The evidence.txt files were analysed using the open source program R (version 3.2.2) to interrogate the data. Figures were generated using the ggplot2 package (Wickham, 2009) and the code used is available as an appendix. Briefly, msms scans from the spiked in proteins were isolated from the dataset. Each MS/MS scan was converted to ratio values by dividing the values for each label by the sum of all the label values. MS/MS scans with missing labels were excluded from the analysis to prevent skewing of the results away from the expected values.

5.5 Results

5.5.1 Background proteome distribution in *Synechocystis*

An in-silico investigation was carried out on the proteome initially to determine its features. The main purpose of this was to identify an appropriate concentration to spike the background protein mix for the tag comparison experiment, where it would be at a level typical of an identified protein within the *Synechocystis* proteome. This was done to ensure that the spike would be indicative of proteins normally present in the proteome and to give a realistic comparison of the two tagging systems in physiological conditions.

In *Synechocystis*, 96.5% of the proteome is distributed over the 5 – 100 kDa range, with a smaller number of very high mass proteins (fig. 5.2 (p. 189)). As expected, the proteins show a linear mass to length relationship, however the variation between these values increases at higher values, expressing minor heteroscadasticity. In our model we consider this effect to be negligible within the range where the majority of the proteome is expressed.

The linear relationship breaks down somewhat when considering the emPAI model, as only peptides within the mass spectrometer range of 500 – 2500 m/z can be observed during analysis. During data processing, spectra with a 2+ or 3+ charge are selected for, as these peptides produce the best spectra for analysis; as a result only peptides with a mass with the range $1000 \leq x \leq 7500$ were considered 'observable'. When considering this mass distribution, the mass to unique peptide distribution contains more variation fig. 5.3 (p. 191). It still shows a linear trend, with a practically observable maximum number of unique peptides for a protein of given mass, however the relationship between

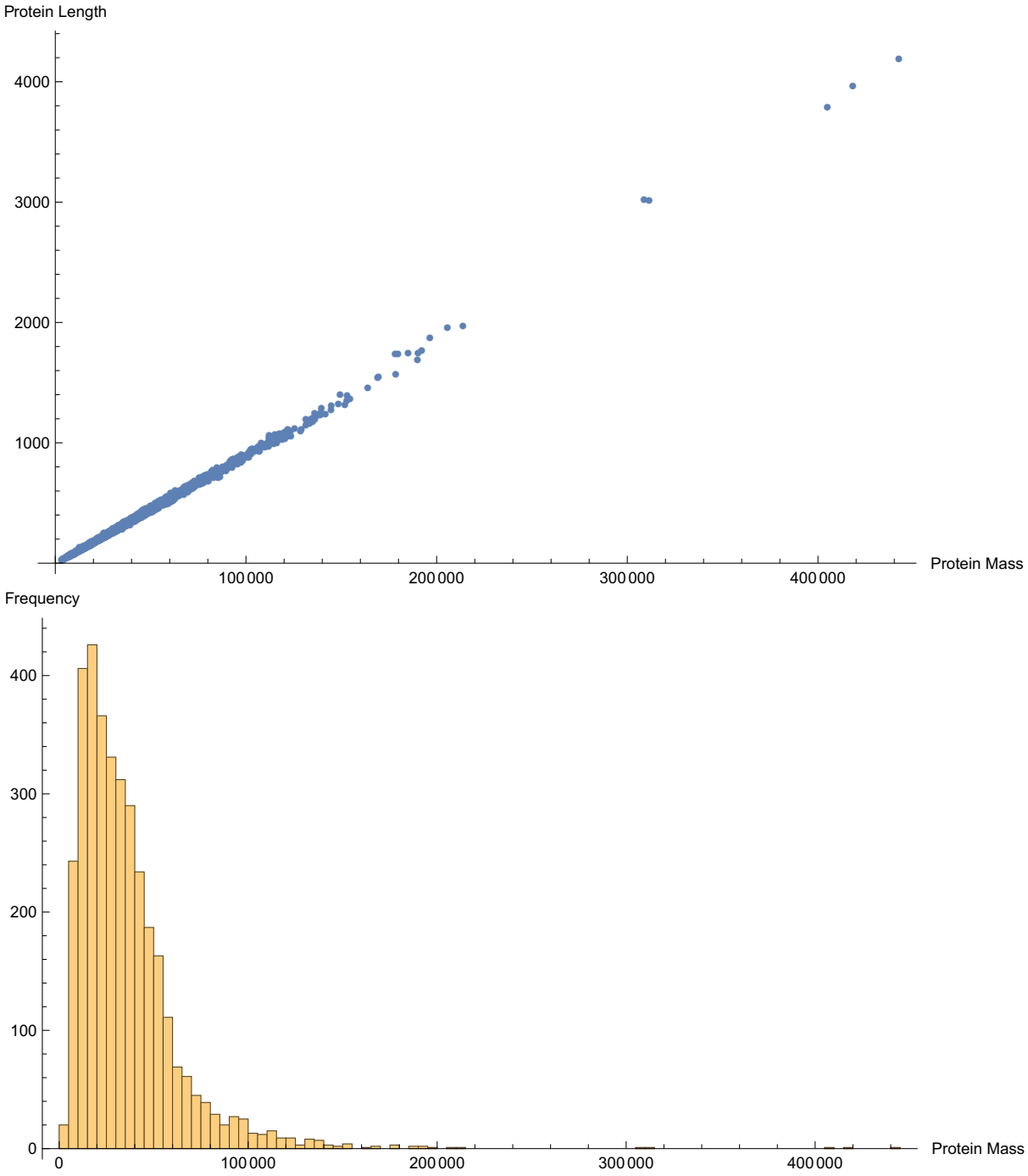


Figure 5.2: A scatter-graph showing the linear relationship between protein mass and length in *Synechocystis*, and the corresponding histogram of protein masses. 96.5% of proteins are present in the 5 – 100 kDa range. (Image created in Wolfram Mathematica, data obtained from uniprot)

protein mass and observable peptides is not as linear as the oversimplified version of the algorithm used in the program Mascot would indicate.

The 'blind Sweden' dataset used for generating the background consisted of 20 high resolution fractions from two iTRAQ 8-plex experiments of *Synechocystis* was analysed with MaxQuant software, which identified 4332 unique peptides, mapping to 1182 proteins present within the proteome downloaded from uniprot. In this data approximately 3% of all possible observable peptides have been identified assuming every unique peptide was present within the sample (123,860 possible observations), but equates to 9.24% of all observable peptides when considering only proteins where direct observed proof is available. A histogram of the protein concentrations as calculated by emPAI in fig. 5.4 (p. 192)) shows that dynamic range of the measured sample observed here is 10^3, calculated by converting the range of the histogram into linear values – this is described below.

$$\ln(2.2 - (-4.4)) = \ln(6.6) = 10^3$$

where -4.4 was the lowest observed value, and 2.2 the highest observed value, as seen in fig. 5.4 (p. 192

The observations of lower-abundance peptides drop off sharply at the left side of the histogram, which may be either due to a limitation of the proteome itself – where the physical limitation of the emPAI calculation has been reached due to the average number of unique peptides per protein – or could be due to a stochastic feature relating to the limits of the mass spectrometer used during measurement. As observable in fig. 5.3 (p. 191), with the exception of a single outlier with 335 unique observable peptides, the upper limit of detection falls in the range of 175 – 195 unique peptides. When plugged into the emPAI formula, this produces the range with an upper limit of 2.2 and a lower limit of -4.4, suggesting a saturation of the values in the formula taking place. For this proteome, using emPAI, the maximum practically observable dynamic range is 7.8×10^2.

On the assumption that the true concentration distribution is symmetrical, the upper half of the distribution can be used to estimate the true distribution more accurately than the lower half of the distribution. The median of -1.64 suggests that this range could be more accurately approximated as $(2.2 - (-1.64)) \times 2 = 7.68$, suggesting that the full dynamic range of the sample could be approximated to 2.2×10^3, around 10-fold higher than the measured value. Whilst this is lower than dynamic ranges reported in other organisms including *S. Cerevisiae* (4.5×10^4) (Picotti et al., 2009) and *E. coli* (3×10^5) (Soufi et al., 2015a), it is closer to the expected range of values and the reduced rate may result from upper-limit saturation, which cannot be corrected for in this model.

When the emPAI quantifications shown in fig. 5.4 (p. 192) are overlaid onto the histogram

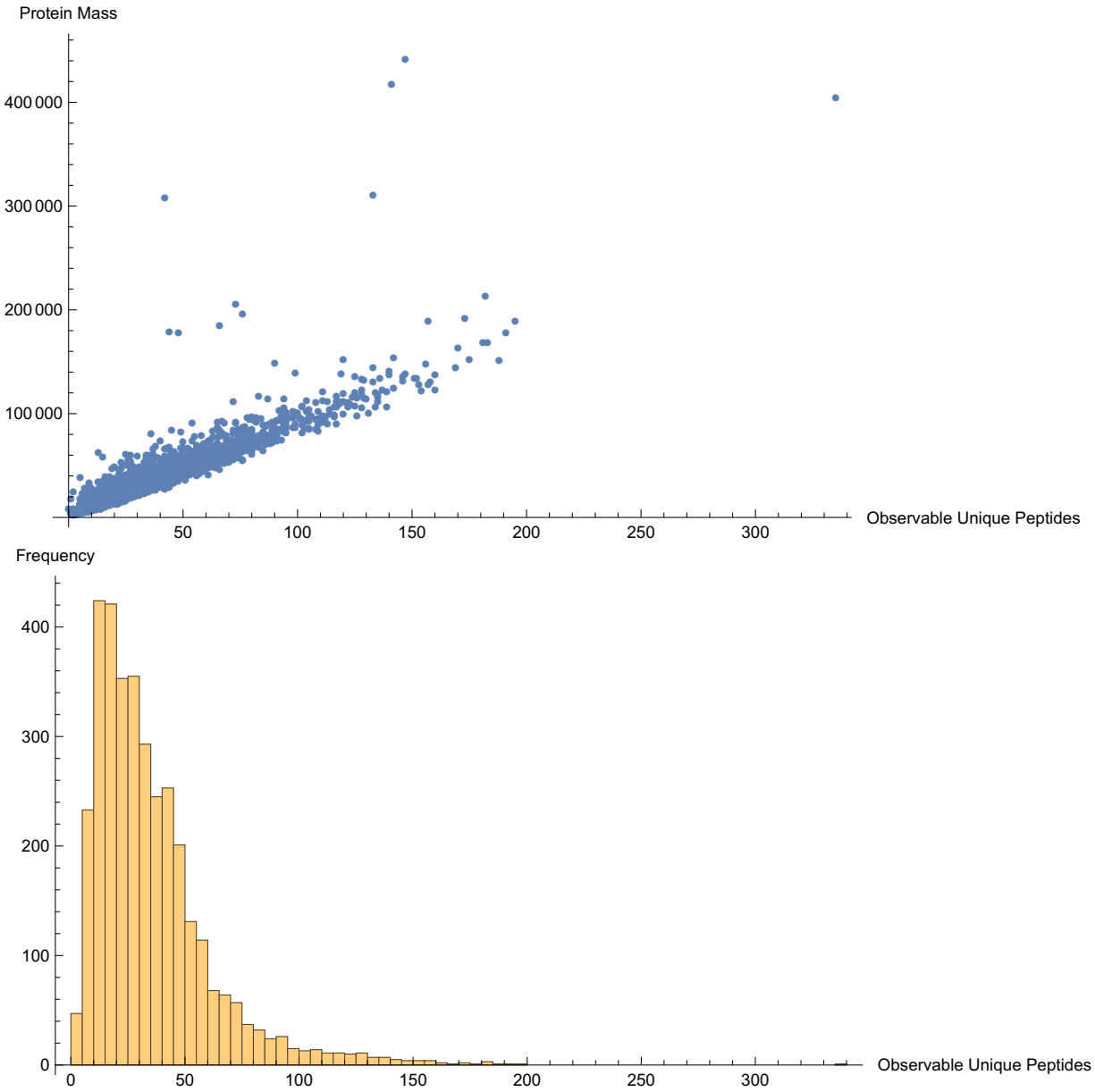


Figure 5.3: A scatter-graph showing the relationship between protein mass and the number of observable peptides within *Synechocystis*, and the corresponding histogram of the number of observable unique peptides. The scatter-graph shows far more variation between these values, compared with the length-mass relationship, despite this the relationship between the two values is largely linear with more stochasticity present. The majority of proteins are present in the 5 – 100 observable unique peptide range, and so in figure 5.5 (p. 193), the higher value proteins have been excluded due to sparsity. (Image created in Wolfram Mathematica, data obtained from uniprot)

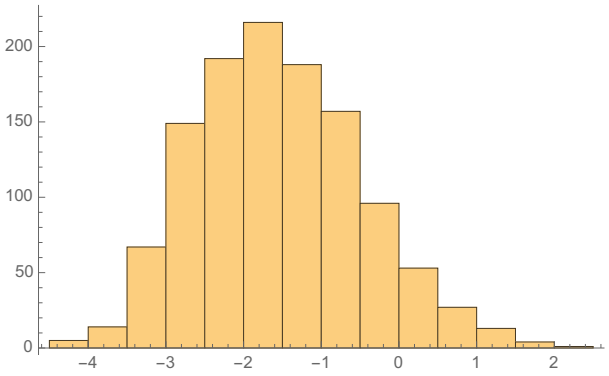


Figure 5.4: Protein concentration distribution, measured in natural log, generated by the emPAI formula. The distribution is almost Gaussian, however left tail is cut down and falls off abruptly. This is likely due to the fact that the proteins in this region are approaching the lower detectable limits of the machine. (Image created in Wolfram Mathematica.)

in fig. 5.3 (p. 191) and broken down into flat percentages within each histogram bin, we can see that as expected, a lower percentage of smaller proteins, or proteins which generate fewer tryptic peptides, are observed compared with larger proteins or proteins that generate more tryptic peptides fig. 5.5 (p. 193). The numbers become more stochastic as higher values are reached, due to the smaller number of discoverable proteins of that size, however the observed trend is fairly stable in proteins generating between 0 and 90 tryptic peptides and this range covers just over 95% of the complete observable proteome.

As mentioned previously, within *Synechocystis* there are 4 highly abundant proteins present – the phycobilisome proteins – which make up 20% of the total proteome. In the emPAI model of the background proteome presented here these proteins are assumed to be at saturation level, but provide a reference for the rest of the proteome in relation to the emPAI model. As the 4 antennae proteins are at an equivalent concentration to each other, each of the proteins can be estimated as making up approximately 5% of the total protein present by mass, so in the sample 5% of the total protein mass is the highest point on the distribution. Given the values stated above, the highest and lowest measurable protein concentrations in the dataset are as follows:

$$max = 0.05 \times total.protein$$

$$min = max \times \frac{1}{1000}$$

Given that the maximum value attainable in emPAI is:

$$emPAI = 10^{(\frac{N_{observed}}{N_{observable}})} - 1$$

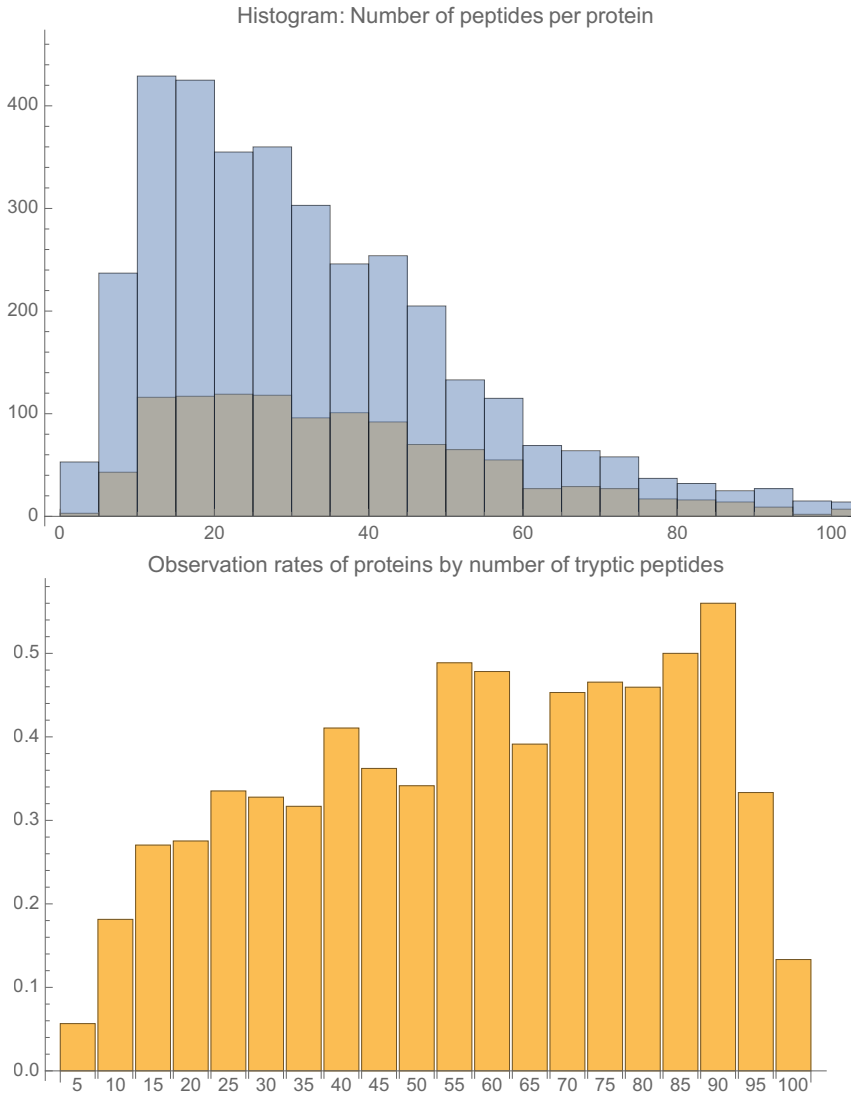


Figure 5.5: At the top, the histogram shows observed proteins from the dataset against all observable proteins in *Synechocystis* proteome, binned in groups of 5 by the number of unique peptides per protein. The blue bars show an approximation to the protein size distribution within the genome (see figure 5.2 (p. 189)), and the orange bars show the number of identified proteins from each bin that were observed, demonstrating the sampling distribution for the *Synechocystis* background. The observation rates are given in the bar chart below. This figure shows that there is a bias against the identification of very small proteins, with a general upward trend in the rate of identifications, until the statistics become unstable in the sparser 'higher-mass' region of the proteome. (Image created in Wolfram Mathematica.)

$$10^{(\frac{1}{1})} - 1 = 9$$

Where all observable peptides are seen in a dataset. The median emPAI value from the data is 0.193, and so to match the median protein concentration the spiked protein should be added at $\frac{0.193}{9}$, or approximately $\frac{1}{50}$ of max.

The spike-in mixture consisted of 4 proteins, however the internal variation on these proteins means that for any given tag the protein concentration will range up to $\frac{1}{40}$ below the original value. The aim of this experiment was to set up the proteins at a concentration surrounding the median value to give the data values for a typical protein in the experiment. The spike-in value was therefore added at a 20-fold higher concentration than the median protein level after dilution, ensuring that the range of the concentration would fall around the median value within the dataset. As a result of these observations, the most practical concentration for spiking in the background proteins in the full-range final mix was found to be:

$$target \times range.scale \times protein.proportion \times max \times total.protein$$

$$\frac{1}{50} \times 20 \times \frac{1}{4} \times \frac{1}{20} = \frac{1}{200} \times total.protein$$

The total protein added in each case is 25 mg per label, so 1 mg spike mix in iTRAQ 8-plex (200 mg total) and 0.75 mg in TMT 6-plex (150 mg total).

5.5.2 Direct peptide and protein counts

All proteins from the spike-in data were observed with at least 2 significant unique peptides, and in the complex fractionated background collectively made up 0.77% (iTRAQ) and 1.13% (TMT) of the overall quantified spectra. These values were expected to be the same, and the difference between the two can directly be attributed to an experimental error which occurred during the spike-in stage, as 1 mg of spike mix was added to both the iTRAQ and TMT backgrounds, where 0.75 mg of spike mix should have been added to the TMT mix. The spike-in concentration was high enough to ensure observation that all the target proteins within the experiment were observed in every experiment, including the complex unfractionated experiment; but were not the highest abundance proteins in the sample according to an emPAI estimation. This findings suggests that the estimations within the model were useful, however it seems unlikely that the spike-concentration used in the experiment reflected the true median level of the sample – this is discussed in more detail in the discussion section below.

As mentioned above, different tagging systems result in different numbers of peptide-spectra matches. Table 5.5.2 (p. 195) shows the summary information for all of the

Experiment	ID'd spectra	Quant spectra	ID'd spike	Quant spike	Prots	Frac
$iTRAQ_1$	1218	845	1060	707	28	4
$iTRAQ_2$	938	408	832	342	24	4
$iTRAQ_3$	13717	11786	106	91	1073	6
$iTRAQ_4$	3416	3025	39	35	583	1
TMT_1	1874	1211	1661	1030	44	4
TMT_2	1222	305	1119	256	27	4
TMT_3	17791	15491	206	175	1229	6
TMT_4	4539	4067	65	61	624	1

Table 5.5: Summary data of the number of peptide-spectral matches identified in each experimental run. The experiments are grouped by 4, with the first 4 using iTRAQ tags and the second 4 using TMT tags. The experiments in sequence are: (1) the spike in mix without a complex background and without dilution (range 1:3.75), (2) the spike-in mix without a complex background but with dilution (range 1:37.5), (3) the spike-in mix in a complex background and with dilution, and (4) the spike-in mix in a complex background with dilution, but without LC fractionation to simulate an even more complex background. ID'd spectra indicates the total number of spectra that were confidently identified, Quant spectra indicates the number of spectra that retained intact quantification data for all label channels, ID'd spike and Quant spike are similar to ID'd and Quant spectra, but relate directly to spectra that match the spiked in peptides from the control mix. Prots is the total number of proteins identified in the final mix, and Frac shows the total number of fractions that were injected overall.

proteomic experiments carried out in this investigation. iTRAQ experiments resulted in 75.9% of the spectra identified in TMT experiments on average – considering data from 44,715 confident peptide spectral matches, which falls within the range of 12% - 50% reductions in identified spectra reported in studies previously. In the complex background experiments, 77.1% and 75.3% of the peptides were identified in the fractionated and unfractionated experiments, respectively. This translated to the iTRAQ experiment identifying 87.3% and 93.4% of the proteins identified in the fractionated and unfractionated TMT experiments, demonstrating that the difference between the two tagging systems is exacerbated as the complexity in the sample reduces.

Interestingly, whilst both tagging methods show similar rates of quantification over the entire set of experiments, iTRAQ shows a pronounced (18.6%) increase in spectra with complete quantification information over TMT in the simple background experiments. Whilst this difference vanishes completely in the complex background mixes, both systems show an increase in the rates of complete quantification as the complexity of the background increases, suggesting that this increase resulted from an increased rate of background co-elution. TMT also showed a lower level of tolerance to increasing ratios between the individual labels; for example in the undiluted mix, 64.6% of the spectra that were identified had a complete set of quantifications associated with them, whilst in the

diluted mix this dropped to 25.0%. The same comparisons for iTRAQ were 69.4% and 43.5% respectively. In the diluted mixes, this actually translated to a larger number of quantified spectra available in the iTRAQ data, despite the overall reduction in identified spectra. As a result, it is likely that the extra quantifications in the TMT experiments are actually a result of background contamination, and so the iTRAQ tags appear to have a relatively better fidelity in quantification than the TMT tags in the experiment.

5.5.3 Quantification and compression

As mentioned above, a mixture of 4 proteins was spiked into different backgrounds for iTRAQ and TMT tags. The proteins used in the mixtures below are abbreviated in this section as follows: bovine serum albumin (BSA), bovine β casein (cas), equine cytochrome C (cytC), and equine myoglobin (myo). In the undiluted, simple background experiment, both tags showed an uncompressed linear relationship between the observed and expected ratios, across the full range of the data. This uncompressed linear observation was maintained in the diluted, simple background experiment for the TMT tags, but not in the iTRAQ tags experiment (fig. 5.6 p. 197).

In the iTRAQ experiment, both the BSA and, to a greater extent, the myo proteins show an over-estimation of the true fold-change. As a result, this skewed the linear fit for the whole dataset to suggest a general overestimation in iTRAQ labelling, but the effect appears to be protein-specific. This overestimation may be attributed to minor errors in dilution for the individual labels, and a dilution correction factor could be calculated to make the samples more linear, however such a calculation is of limited general value. This observation is interesting in the context of the data in table 5.5.2 (p. 195), which indicated that the larger number of quantifications present was a beneficial feature of the iTRAQ tags, as the increased observations seem to have increased the overall noise level in the dataset and so the additional value they bring may be limited.

These data suggest that while there is a large visible spread of the individual datapoints, statistically the precision of the linear model is high, due to the concentration of measurements around the expected ratio of the data. The data demonstrates a degree of heteroscedasticity, with the higher ratio values having more spread than the lower values. This would make sense as there is more likely to be deviation at the tails of the data, and since the data are ratios relative to 1 the greatest variation would be visible in the higher ratios. A log transformation of both the axes enables closer investigation of this effect (fig. 5.7 p. 198).

Under log-transformed axes, a very clear skew is observable in the TMT diluted mix. Whilst this does not appear as a significant deviation in linear space, since the variations

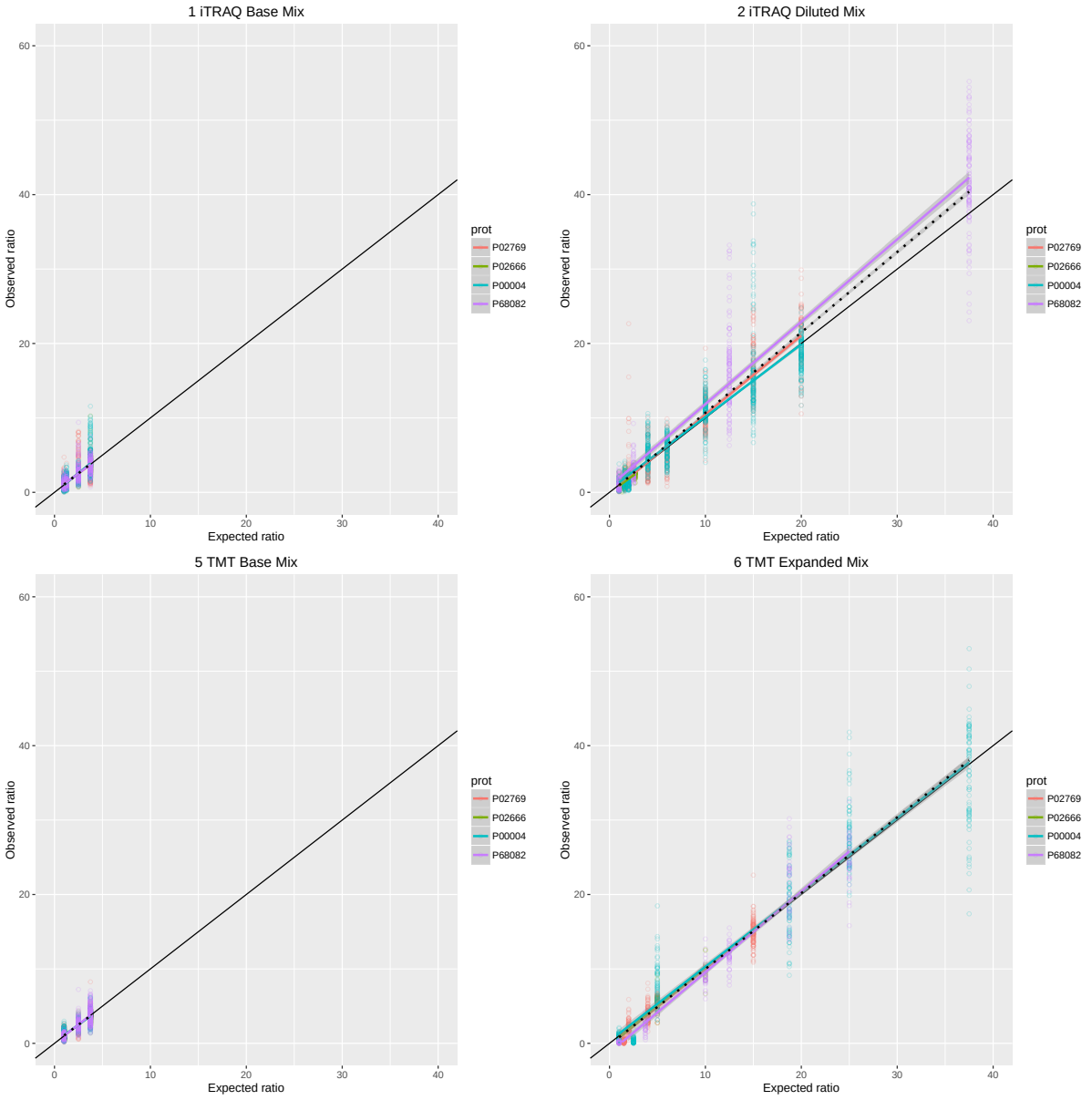


Figure 5.6: A 2x2 grid showing the simple mixtures and their diluted expansions. Individual proteins from the spike-in mix are highlighted in the corresponding colours in the legend, these are bovine serum albumin (P02769, red), bovine β casein (P02666, green), equine cytochrome C (P00004, blue), and equine myoglobin (P68082, magenta). The solid black line shows the expected relationship between the observed and expected ratios. The dotted line shows the best linear fit for the data when considering the entire dataset. iTRAQ data are shown on the top row and TMT data are shown on the bottom row. The shaded grey area around the lines indicates the variance in the linear models applied to the data, the broader the shaded area, the lower the precision. The hollow circles are individual data measurements and show the abundance and spread of the data measured at each point for each protein. (Images created with the ggplot2 package in R.)

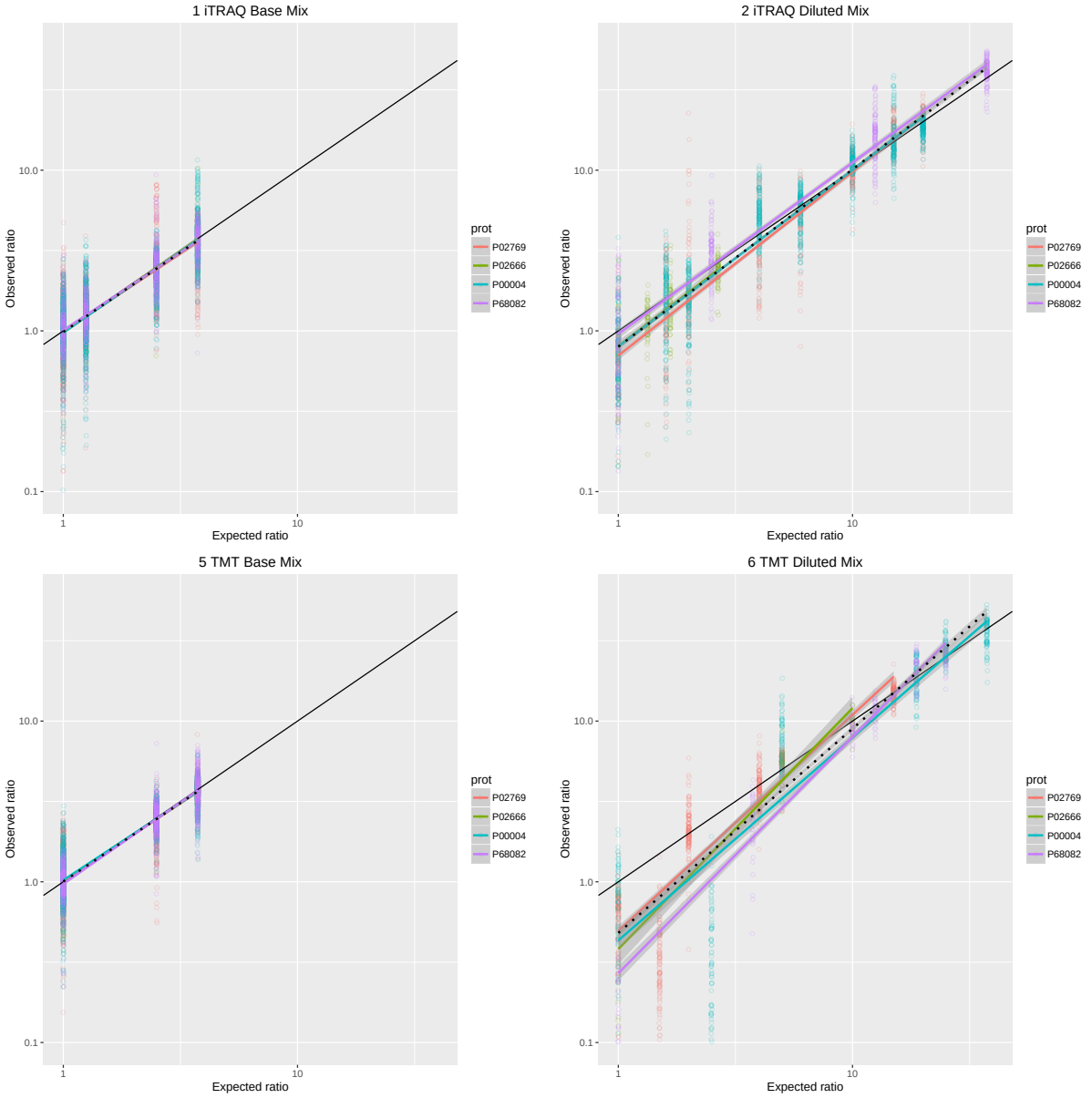


Figure 5.7: A 2x2 grid showing the simple mixtures and their diluted expansions from figure 5.6 p. 197 under log-transformed axes. Proteins are bovine serum albumin (P02769, red), bovine β casein (P02666, green), equine cytochrome C (P00004, blue), and equine myoglobin (P68082, magenta). The solid black line shows the expected relationship between the observed and expected ratios. The dotted line shows the best linear fit for the data when considering the entire dataset. iTRAQ data are shown on the top row and TMT data are shown on the bottom row. The shaded grey area around the lines indicates the variance in the linear models applied to the data, the broader the shaded area, the lower the precision. The hollow circles are individual data measurements and show the abundance and spread of the data measured at each point for each protein. (Images created with the ggplot2 package in R.)

are tightly clustered together in the lower-ratio region of the data, it is actually a more pronounced overall effect when considering the values purely from a ratio perspective. The reduced abundance of tags with complete quantification information is likely to be related to the low-abundance skew resulting in a general over-estimation of the ratio, although this effect is the result of a tag-specific effect. On investigation of the effect, the fault lies with quantifications from TMT tag 129, which makes up – due to a weakness in the dilution design – the majority of the low-abundance quantifications in the simple background diluted mix TMT experiment. Unfortunately, a flat scalar modification to the values would, whilst improving the median value of the tag intensity, not correct for the increased variance observed with the low-abundance tag. Fortunately, due to the experimental design, in the complex background mixes this tag is present at high abundances and the effect seen here is no longer present.

It is not clear from either figure 5.6 (p. 197) or 5.7 (p. 198), but there is a much higher density of measurements in the base mix experiments compared with the expanded mix experiments, as can be seen in table 5.5.2 (p. 195). This results in a very precise linear model for both of the base mixes compared with their diluted variants, despite showing what appears to be a larger amount of spread on the data. In the simple mixtures, the iTRAQ data appear to be both more accurate, whilst the TMT tags are slightly more precise, however this can be traced back to an issue with the TMT tag 129. On examination of the data, this same tag appears to be responsible for a large proportion of the missing values, with tag 129 missing 72% of the quantifications for all confidently identified peptides labels; and as a result has directly contributed to the observed reduction in full quantification data quality observed for the TMT tags seen in table 5.5.2 (p. 195).

When the mixtures are spiked into a complex background, the linear ratios seen exhibit clear signs of compression, as shown in figure 5.8 (p. 200). This effect can be more closely interpreted using a log-axes graph, as was used in the previous experiment (fig 5.9 (p. 201). There are a few different comparisons being conducted simultaneously, so the figures can be challenging to interpret, however the core comparisons are as follows.

The iTRAQ data shows 3 different measurements regimes, the first is the full [1 : 2.5 : 12.5 : 37.5] range, highlighted by cas. Next is the [1 : 1.6 : 2 : 4 : 6 : 10 : 15 : 20] fold-change range, with a high and low concentration protein generating the outputs, these are cytC – which is a smaller protein and therefore has higher concentration peptides, (as equal masses of protein were spiked in, not equal concentrations) and BSA – a larger protein with therefore lower concentration peptides. Finally, is a low concentration compressed range measuring from [1 : 1.3 : 1.6 : 2.6], at a concentration 3 fold lower than the median range, highlighted by the myo protein. It is worth noting that as the same master mix was used for all 4 experiments, there was a limited amount of material available. As a result, the dilutions in the simple background and the complex background run anti-

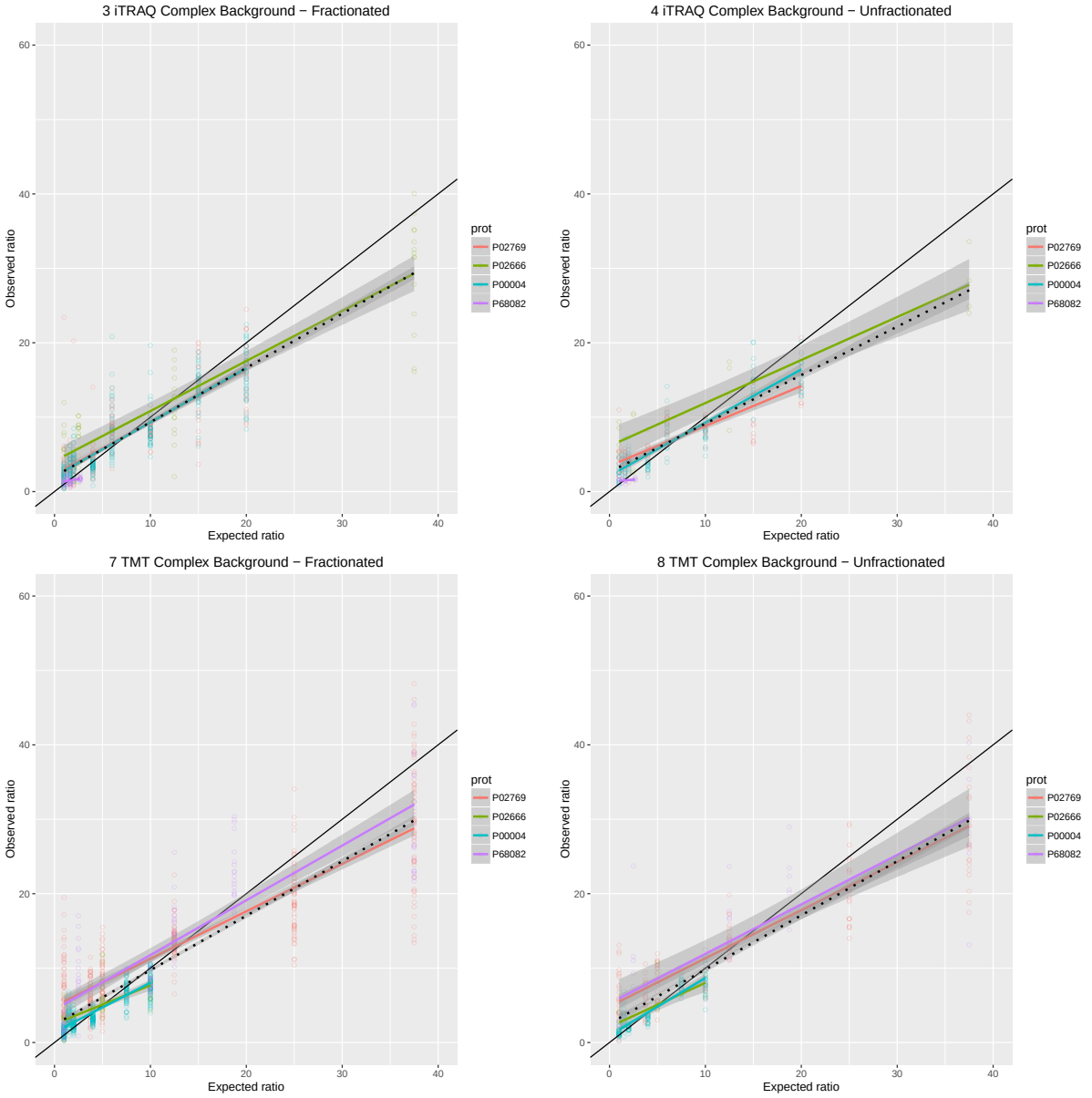


Figure 5.8: A 2x2 grid showing the complex bg mixtures. Individual proteins from the spike-in mix are highlighted in the corresponding colours in the legend, these are bovine serum albumin (P02769, red), bovine β casein (P02666, green), equine cytochrome C (P00004, blue), and equine myoglobin (P68082, magenta). The solid black line shows the expected relationship between the observed and expected ratios. The dotted line shows the best linear fit for the data when considering the entire dataset. iTRAQ data are shown on the top row and TMT data are shown on the bottom row. The shaded grey area around the lines indicates the variance in the linear models applied to the data, the broader the shaded area, the lower the precision. The hollow circles are individual data measurements and show the abundance and spread of the data measured at each point for each protein. (Image created with the ggplot2 package in R.)

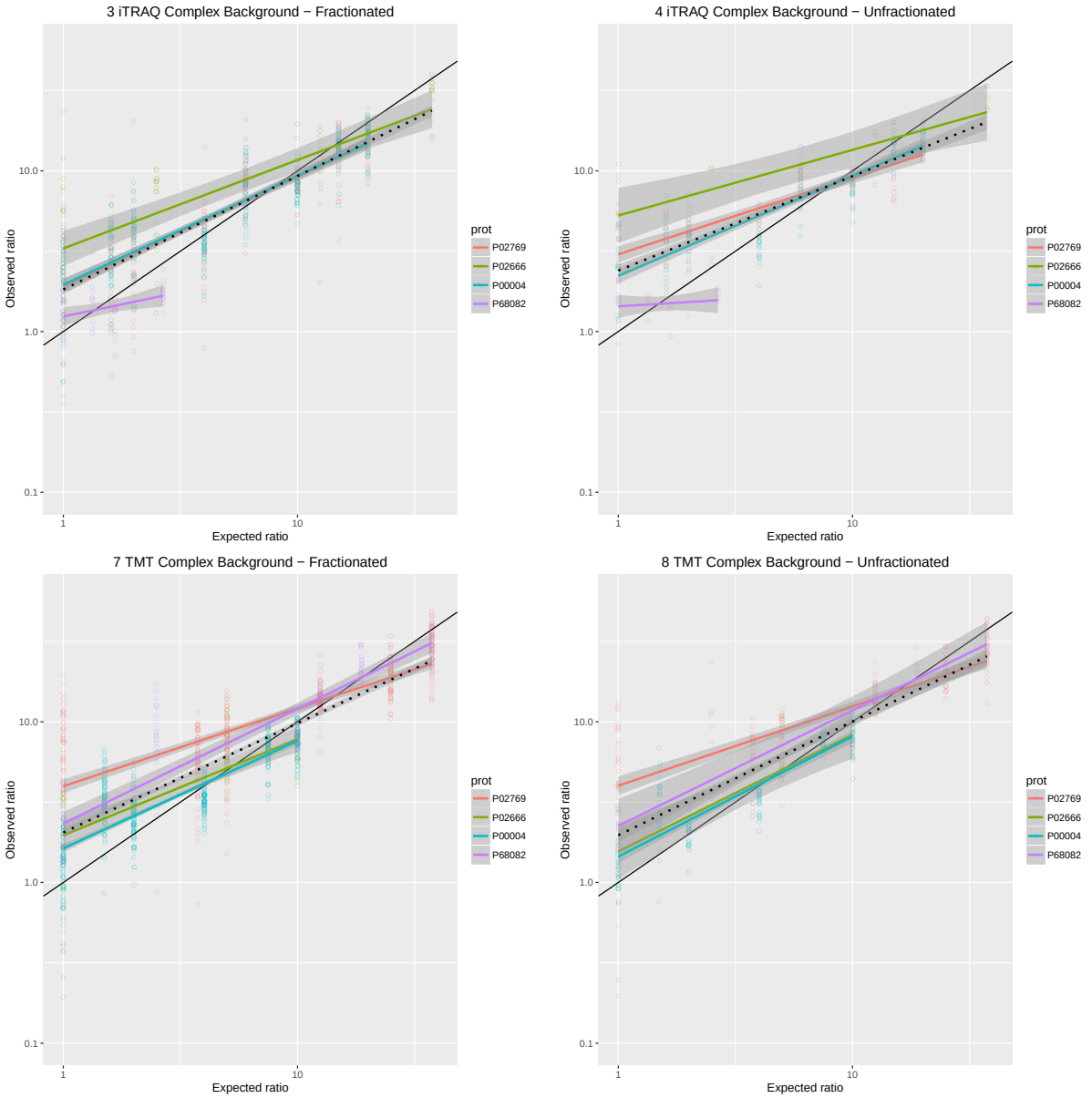


Figure 5.9: A 2x2 grid showing the complex bg mixtures from figure 5.8 p. 200 under log-transformed axes. Individual proteins from the spike-in mix are highlighted in the corresponding colours in the legend, these are bovine serum albumin (P02769, red), bovine β casein (P02666, green), equine cytochrome C (P00004, blue), and equine myoglobin (P68082, magenta). The solid black line shows the expected relationship between the observed and expected ratios. The dotted line shows the best linear fit for the data when considering the entire dataset. iTRAQ data are shown on the top row and TMT data are shown on the bottom row. The shaded grey area around the lines indicates the variance in the linear models applied to the data, the broader the shaded area, the lower the precision. The hollow circles are individual data measurements and show the abundance and spread of the data measured at each point for each protein. (Image created with the ggplot2 package in R.)

parallel to each other – this means that whilst cas shows the 1 : 37.5 range in the complex background and myo covers the 1 : 2.6 range, these proteins were switched in the undiluted experiment. It has no effect on the bsa or cytC, with the exception of switching the concentrations of iTRAQ reporter labels between the two proteins.

Looking at the protein-specific linear models, the three different regimes form three separate gradients on the figure. This occurs because to maintain the variance in all the labels, each dataset was normalised by the median reporter value rather than division by any given reporter in the dataset. As a result, the point at which each of the linear models crosses the solid black line indicates the median reported ratio within the sample. The gradient of the line indicates the level of compression, and so it can be seen that all proteins with the exception of cytC undergo increased compression in the more complex background mixture. It is interesting to note the the larger-ranged cas and shorter ranged myo experience higher rates of compression than either bsa or cytC in the fractionated iTRAQ background, as can be seen from the steeper gradient on the linear model. Cas shows a linear concentration gradient across all samples, suggesting that label-specific effects did not appear to contribute significantly to the results, however they may become apparent if more data points were available.

The shaded area indicates the precision of the linear model. This becomes broader for all proteins in the unfractionated background, showing an overall reduction in precision. The gradient on the bsa line drops in the unfractionated complex background compared with the fractionated complex background, however the same effect is not seen for cytC. This suggests that bsa was more affected by increasing isotopic contamination in a higher complexity background than cytC. Whilst a compression effect is visible on the low-abundance myo samples between the fractionated and unfractionated backgrounds, the resolution between the different measurements is not high enough to classify the different labels as significantly different from each other in either case. So whilst this observation gives some impression of how the additional complexity blocks the observation of small scale changes, this information makes little practical difference to analysis of a real biological sample. Across the iTRAQ samples,there appears to be a small overall fold change between the 1 : 37.5 range (5 : 30 observed min-max ratio) and the 1 : 20 range (3 : 15 observed min-max ratio). Within the ranges, the median 1 : 10 ratios were 3 : 9.7 and 5 : 11 respectively. These values suggested that the wider ranges were more susceptible to compression, and so proteins with a wider spread of values recorded across all the iTRAQ tags will have more internal compression as well.

The TMT data shows 2 different measurement regimes, the first spans the full 1 : 37.5 range, and is covered by myo [1 : 2.5 : 10 : 12.5 : 18.75 : 37.5] and bsa [1 : 3.75 : 5 : 12.5 : 25 : 37.5], with each protein taking different interval steps to highlight resolution along the range. The second spans the smaller 1 : 10 range at regular intervals with 2

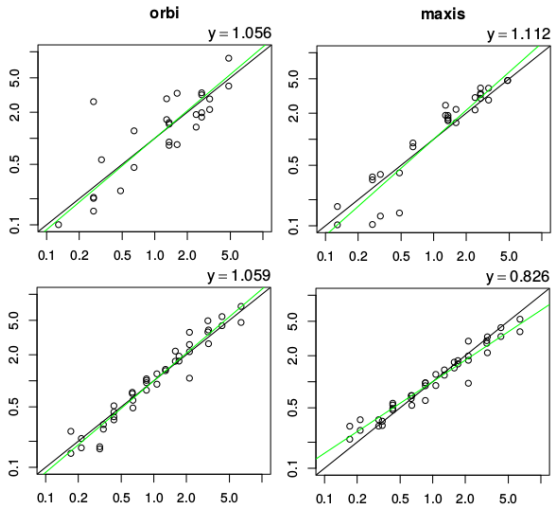


Figure 5.10: A comparison between the extended-range mix without the addition of a complex background, with TMT on the top row and iTRAQ on the bottom. The left column are the data collected from the QExactive HF, whilst the right is data collected from the maXis. In this figure, each clear circle is a protein quantification for a single label. All labels have been normalised to the mean and put into a log scale, so the points at the top and bottom of the image are from proteins with the furthest spread in ratio (1 : 37.5) and the central points the smallest. (Image created in R.)

protein replicates, ctyC [1 : 1.5 : 2 : 4 : 7.5 : 10] and cas [1 : 5 : 10]. Looking at the protein-specific linear models, the 2 separate regimes form largely parallel gradients in TMT. The compression affect appears to be much more uniform across all proteins, whereas in iTRAQ this was not found to be the case. Specifically, the measured ratios between 1 : 37.5 were 6 : 29 and 6 : 32, and between 1 : 10 were 3.5 : 7.5, showing far more linearity between measurements.

Interestingly, TMT tags do not appear to experience an increase in average compression with increasing complexity as iTRAQ tags do (dotted line comparison between fractionated and unfractionated), although this may be related to a number of factors, including the experimental design (2 proteins spanning the full range instead of 1), the more regular intervals between the low and high ratios in the TMT experiment, and the overall higher number of matching spectra contributing to the models. Individually, myo appears to experience higher levels of compression, bringing it in line with bsa, and the 1 : 10 ratio appears to actually experiences a slight reduction in compression; however this comes with a dramatic reduction in precision. It is interesting to note that TMT tags in the fractionated mix appear to show more heteroscedasticity, with higher precision being achieved at the low-ratio end, whilst iTRAQ variance is more homoscedastic in the fractionated mix. Both samples are more homoscedastic in the unfractionated mix, whilst displaying a reduction in precision.

A comparison was made between two mass spectrometers, the maXis UHR ToF and a QExactive HF, to observe how changing the model and type of mass spectrometer affected the different tags. This assessment was made on the extended mixture to determine limitations on quantification in a relatively simple background – although it could be argued that the background was pseudo-complex due to the protein concentration range used. As can be seen in figure 5.10 (p. 203), even at this concentration gradient the

maXis began to experience compression, however this is more noticeable in the iTRAQ study, which covered the full range more comprehensively than the TMT study when the data is investigated in this way.

This comparison was the first performed on the data, and demonstrates the difficulty in designing a truly comparative experiment on two systems with different numbers of replicates. The other comparisons in this section subdivided the data into protein groupings and made all values relative to the respective fold-changes around 1 in log space, as would be more typical in a traditional proteomics quantification experiment. This figure shows that when ignoring the protein concentration and performing a purely ratio-driven experiment, compression affects both the upper and lower measurements. This follows, as if the low measurements are abnormally high, and the sample is compared from the middle, then the upper measurements will be relatively lower also.

5.6 Discussion

5.6.1 emPAI vs tag-based quantifications in *Synechocystis*

The maximal range measurable in *Synechocystis* is around 8×10^2, however due to the rapid drop-off in measurements at the lower tail of the histogram, this suggests that the lower tail observations are mostly driven by the maximal number of unique peptides generated from the largest protein in the proteome. This in turn suggests that for accurate determination of fold-change by the emPAI spectral counting method in *Synechocystis*, we should either consider the median to maximum value, where this effect is not present – which would result in a measurable range of 45-fold; or else consider the space before the rapid drop-off occurred whilst the histogram is still symmetrical, suggesting a measurable range of 90-fold.

Whilst this is still greater than the range accurate quantifications can be made in iTRAQ or TMT systems, which is closer to 10-fold and suffers significantly from compression effects, it still represents a limitation in the technology. These numbers give an indication of practical measurement ranges within full-proteome studies, and can therefore be used to better understand the limitations of pre-existing proteomic data. This will have important ramifications for systems-level analyses, as there are previously reported incompatibilities between proteome observations and transcriptomic/metabolomic observations.

5.6.2 Features of the *Synechocystis* proteomic background

A dynamic range of 2.2×10^3 was estimated for *Synechocystis*, which is lower than both *S. Cerevisiae* (4.5×10^4) and *E. coli* (3×10^5). Whilst the observed range is only half that of the yeast proteome (Picotti et al., 2009), it is much lower than a similarly comprehensive study conducted in *E. coli* (Soufi et al., 2015b). In the *E. coli* study, the largest copy-number value they reported as 300,000, with no lowest value given. As a result of this, it is difficult to determine if the lower cut-off should truly be assumed as 1 count (or possibly even a fraction of this value), or if it should be considered to be higher as no definitive cut-off value was given in the methods. A literature search did not produce evidence for observations of single copy proteins when measuring the global cellular environment of a population of cells.

Fractional values could theoretically be reported when considering a protein that is only expressed by a subset of the cellular population, which would be a useful feature to assess in a metaproteomic study. It does raise in interesting question of whether there are stable single-copy protein systems present in nature. The difficulty of looking at protein abundances in such cases, is that protein abundances within the cell are dynamic by their very nature. As protein expression can be switched off completely, it follows that at some points in time there is a single copy of that protein within the cell; however if the single copy-number state is only a step to a different value, rather than a stably expressed value, then its value is of limited practical use. There are also studies which suggest that true genomic repression control is impossible, as all parts of the genome of a cell are transcribed into RNA at some point, including the telomeres in linear genome organisms. This is likely to be the reason why studies into on/off switches in Synthetic biology focus on probabilistic threshold switches with the aim to replicate a binary effect, rather than designing an absolute binary control system.

During the investigation, the high-abundance proteins were at the saturation level – where all observable peptides were seen for the four most abundant proteins. This caused the range between the median point and the upper limit of the model to be compressed. From a practical sense, this resulted in the spike-in volume being higher than the true median point of the dataset. Whilst this was a concern before starting the experiment, it was impossible to correct for based on the data available during the method development stage, however a suggested true value became apparent during data analysis. In the iTRAQ dataset, the spiked-in peptide mix made up 0.77% of the dataset, or $\frac{1}{25}$ of max. Since the proteins were targeted to be spiked in at $\frac{1}{50}^{th}$ the max concentration, this suggests that the saturated proteins are 2-fold higher than expected. Given the earlier estimation using the median, this which would change the overall dynamic range estimation for the *Synechocystis* proteome to 1×10^4 (4-fold increase as the right tail

was used to approximate the left tail). This value is still lower than the dynamic range observed for other proteomes, however, and *Synechocystis* would benefit from a dedicated investigation to determine a better approximation of dynamic range of the sample under different processing conditions. The Sweden dataset contains a large amount of replicated mass spectrometer runs of the experimental data, and so could be used to get a true impression of the range of the experiment – fractions of the dataset could be sequentially added to the model until the high-abundance proteins first reach saturation; although the data from this particular dataset may be too sparse for such an approach to succeed, as despite the large amount of technical replication only 30% of the proteins in the proteome were identified.

Within the proteome, as apparent in fig. 5.2 (p. 189) there are the 5 very high mass proteins (> 300 kDa), which are still considered 'predicted' and have not been experimentally verified. The two largest, slr0408 and slr1028, appear to be related only to each other phylogenetically and so have no functions computationally assigned to them, whilst the other three appear to be involved with cellular adhesion based on gene ontology matches. Whilst the chances of observing these proteins should be high, within the Sweden dataset only the two largest proteins, with unknown functions, were observed; suggesting a bias in the protein extraction technique used against membrane-fraction proteins. In fig.5.3 (p. 191), the largest outliers in this plot fall in the high mass range of the proteome. This suggests either a relative reduction in negatively charged amino acids in large proteins (larger tryptic peptides), a concentration of such residues in close proximity in the primary sequence (smaller tryptic peptides), or repeating/non-unique sequences within these larger proteins. This would make an interesting further analysis on the proteins at the upper limits of the *synechocystis* proteome, indeed of large proteins present in other organisms as well, however such an investigation is beyond the scope of this study.

The number of proteins identified in the fractionated samples where the mix was spiked into the background (1073 in iTRAQ and 1229 in TMT), are consistent with the dataset used to generate the estimations (1182 in iTRAQ). This suggests that the spike-in did not cause a significant reduction in the quality or quantity of the observations of the background proteome, and suggests that the observations made of the spiked in proteins are therefore typical of proteins within the *Synechocystis* proteome.

5.6.3 Minimum detectable limits

It is clear from the data presented in log scale that as the overall concentration of a protein reduced in the mix, it hits a critical threshold below which it was impossible to changes apart. From a ratio measurement perspective in this data the threshold appears to be changes below 4-fold in the unfractionated sample, and between 2 and 3 in the

fractionated sample.

In the iTRAQ diluted experiments, the small range protein ratio running from 1:2.6 had a maximal concentration of ≈ 47% of the median protein concentration, and a minimum of ≈ 17%. These values are calculated as follows:

$$conc = \frac{spiked}{\mu(spiked.base)} \times \mu.conc$$

$$max.conc = \frac{0.8}{1.7} \times \mu.conc \approx 0.47 \times \mu.conc$$

$$min.conc = \frac{0.3}{1.7} \times \mu.conc \approx 0.17 \times \mu.conc$$

Where *spiked* is the ratio of the protein being looked at, $\mu(spiked.base)$ is the mean concentration in the undiluted mix, which was made up at $\mu.conc$, the mean protein concentration within the sample (5.5.1, p. 194).

When comparing the fractionated and unfractionated samples, it is possible to determine a slight ratio change between the lowest and highest concentrations – seen by the gradient on the model, however the same is not true in the unfractionated experiment. This demonstrates a possible concentration cut-off for complexity measurements and would benefit from further investigation. By spiking in a median-concentration-balanced protein standard with ratios in a ladder form, it may be possible to determine a cut-off point for the concentration of the minimum detectable concentration within the sample. This phenomenon is interesting, as the experiments that showed that MS3 analysis had the capacity to remove ion interference worked on whole-proteome samples (Christoforou and Lilley, 2011; Ting et al., 2011; Wenger et al., 2011), and so are likely biased as a result by the presence high-abundance proteins where the whole proteome dataset is changing concordantly, rather than in an individual protein study such as the one presented here. It would be interesting to look at these advanced techniques in finer detail and see if they do indeed have the capacity to differentiate between low concentration proteins as effectively as high concentration proteins, or if they demonstrate the same levelling off effect observed here.

The quality of our study could have been improved significantly by including more proteins within the sample; increasing the number in the mix from 4 standards up to 12 or even 24 standards would have enabled a balancing effect between the iTRAQ and TMT designs and allowed for internal protein replicates. The experiment would also have benefited significantly by including more repeated injections, which were not possible due to machine-time limitations. Ideally, between 5 and 10 repeat injections of the samples would have added another dimension to the experiment, where assessing how increasing sampling rates affected all factors within the spectrometer, along with rates of detection.

A further improvement to this design would be to compare repeat injections of an unfractionated sample to one which had been fractionated offline, as fractionating a sample innately adds repeat injections by the nature of the technique (each fraction must be injected, therefore larger datasets are compared). In this experiment, the unfractionated experiment was a simulation for increased complexity, but for a true comparison an even number of injections for each sample, ie. 12 repeat injections of unfractionated sample against 2 repeats from each of the 6 fractions used in the experiment.

5.6.4 iTRAQ vs TMT, which is better?

There doesn't appear to be a clear singular winner in this contest. On one hand, the TMT tags result in more overall quantifications, which in turn means larger amounts of data for the researcher to work with; however on the other the results indicate that the iTRAQ tags actually produce better quality data for larger ratio differences between proteins in different conditions. As there is a hierarchy in the data, where the peptides need to be identified before they can be quantified, it is generally favourable to have a higher level of identification; even if this results in a lower proportion of quantifiable spectra. In addition, as sample complexity reduces, the benefit of having a higher number of identifications translates to a growing increase in the number of identified and quantified proteins. It is possible that this skewed the observed results beyond the normal levels, as there were more around data-points for the TMT values, and because the increased number of spectral matches is non-linear with increasing concentration the $\frac{1}{3}$ increase in protein concentration translated to a 50% increase in the number of observed peptides in the TMT experiments.

When comparing the base mix experiments to the expanded mix experiments, the proportion of spectra that have quantifications for every label drop by more than 50%. This shows that expanding the range of the concentration reduces the chances that the lowest-intensity labels will be observed. This effect is more pronounced in the TMT samples than the iTRAQ samples; however more MSMS spectra are confidently identified. It is possible that this effect demonstrates the division of collision energy, between fragmentation of the peptide to generate fingerprint spectra and fragmentation of the label. The same proportion of unquantified spectra is not observed in the samples which also contain a complex background proteome sample. The quantification rates for the background samples increase from between 25 – 60% in the simple background up to 85 – 90% in the complex background. Similar values are observed when looking only at the proteins from the spike mix (Table 5.5.2 p. 195). This is strong evidence of co-isolation of tags (Karp et al., 2010), where in cases that the intensity of the tag would be too low to detect in an uncomplicated sample, the background provides a baseline value. The

method used to identify spectra with missing values excluded any msms spectra with a 0 intensity for any tag, so background interference in the label region of the spectra results in fewer spectra excluded from the count. Since the same sample was spiked in to the background and therefore it had the same labelling efficiency, it is possible to observe rate of co-isolation in this dataset from the proportional increase in quantified spectra. In this case an increase from 41% to 85% (fractionated) and 90% (unfractionated) labelling efficiency was observed for iTRAQ, and an increase of 23% to 85% (fractionated) and 93% (unfractionated) in TMT.

The proportion of the spectra attributed to the target proteins from the spike-in mix was extracted from the data. In the experiments without background, the proportions of msms spectra attributed to the target proteins was similar between the two labels – around 90% of identified and 85% quantified spectra. In the experiments where a complex background was present, the target peptides made up between 0.7 and 1.5% of all observed and quantified spectra.

5.6.5 Proportionally more of the spike in proteins present

As shown in the results, there were proportionally more of the spiked in protein peptides present in the unfractionated background than the fractionated background. The increased proportion of spiked in signal is a bit counter-intuitive at first glance, however it is a function of the noise within the raw experimental data. By fractionating the samples, the overall level of background noise is reduced. This enables lower-intensity peaks to be identified, resulting in a larger number of peptide identifications from the sample and therefore proportionally fewer identified spectra linked to higher-abundance peptides. It has been stated that overall signal is consistently proportional to the amount of protein present in the sample. From these data, this does not appear to consistently be the case, as lower-abundance peptides are under-represented in the sampling process. This indicates the level of the spike-in samples relative, to the median concentration of proteins in the *Synechocystis* proteome. If the spike was at the median protein concentration within the proteome, then the number of identifications should have reduced when the complexity increased, as higher concentration proteins would be favoured.

As described in the methods section, the concentration of the spike proteins was balanced at 0.5% of the total protein mix, to ensure observation of the proteins within the sample; however once the dilution factors are applied to the simple mixture calculations, these values reduce to 0.27% and 0.23% of the overall samples for TMT and iTRAQ respectively. The observed proportions in the sample for each protein were 0.29% and 0.36% for TMT and 0.19% and 0.29% for iTRAQ, fractionated and unfractionated respectively. The iTRAQ values fall around the expected level, whilst the TMT values are abnormally

high, due to the accidental over-dosing of the TMT experimental data, however once this is corrected for the experimental observations match the expected values.

5.6.6 Balancing more replicates against fewer observations

Overall, whilst TMT generates more observations of the data per technical repeat, iTRAQ has the advantage of 2 additional replicates within the sample. This equates to a 33% increase in available data, which whilst not quite as high as the benefit observed from the extra spectra has the advantage of being able to capture a wider range of experimental conditions. As highlighted in chapter 3, methods which merge multiple mass spectrometer experiments together result in a reduction in the number of identified proteins, and this effect is more pronounced than the reduction in spectra. In addition, there is the question of cost. Running a repeat injection, whilst still generating additional costs due to additional mass spectrometer operation time, is much cheaper than running an additional experiment and purchasing additional reagents, running additional fractionations, etc. On balance, this suggests that if more experimental samples are available then an iTRAQ experiment should be run with additional technical repeat injections to improve the coverage lost due to reduced observations of the tagging system.

TMT tags show a reduction in complete quantification compared with iTRAQ tags. Whilst it could be argued that this effect could actually be beneficial in reducing the effect of compression, as the difference between the highest and smallest values would be greater, it comes at a significant cost to accuracy when improvements to the spectrometer operation are made, as isotopic contamination plays a more significant role in determining the quantification that is observed.

A future expansion on this experiment would take the same experimental methodology and do a systematic assessment on the systematically increasing 'plex' values to determine if the effect could be predictably modelled. The literature suggests that this effect is related to the label chemistry rather than the number of labels in the sample, as a previous study carried out comparing iTRAQ 4-plex and iTRAQ 8-plex labels found that the 8-plex had improved quantification accuracy for the same labels (Pottiez et al., 2012), however an objective assessment on whether the number of labels present within a sample affects the precision and accuracy of those quantifications has not yet been carried out. A key experiment to demonstrate this would be to compare a series of labelling techniques systematically – as described above, or to run a series of experiments in a single label system – such as TMT – but incorporating different numbers of labels in each run (ie. 126 & 127; against 128, 129, 130 & 131 for example).

General applicability of findings

This study has been focused primarily on *Synechocystis*, as that is the focus of this thesis, however previous studies on the tag-based effects observed here have been carried out in *S. cereviseae*, *H. sapiens* and *E. coli*. Whilst *Synechocystis* has the complicating factor of high-abundance photo-system proteins generating an artificially high dynamic range, the key findings are related to the data in general and how non-specific background noise reduces the observed data quality. As the compression effect is observed across a number of different species, and scales with the complexity of the sample, it seems fair to assume that this is not a *Synechocystis*-specific effect. Inversely, whilst it was an expected finding, this data definitively shows that *Synechocystis* is also subjected to the same compressive effects that other tag-based analyses have shown.

5.7 Chapter conclusions

In this chapter, the following main findings were identified:

- The *Synechocystis* proteome has an estimated dynamic range of 1×10^4, although this is likely to be higher due to saturation effects within the model.
- The median concentration of protein within the *Synehocystis* proteome was estimated to be $\frac{1}{1000}^{th}$ of the total protein in a sample by mass.
- Higher mass proteins (>120 kDa) in *Synechocystis* are incompatible with spectral-counting methods of label-free quantification, due to large amounts of stochastic variation in observable peptides.
  - This finding should be generally applicable to higher-mass proteins in other organisms, and also to proteins with a large number of splice variants in eukaryotes.
- In *Synechocystis*, the largest label-free fold change measurable with the emPAI method of quantification is 8×10^2, however due to unreliable quantifications at the lower tail, a practical, accurate fold-change measurement range is closer to 90, around 10 fold lower than the max estimation.
- TMT tags produce larger amounts of noisier data in a proteomics experiment due to an increased number of peptide spectral matches, this offers a larger amount of data for the experimenter to work with which mitigates most of the issues seen between the two tagging systems. This can be as high as 50% more available spectra in a TMT experiment.

  - This effect results in a higher number of protein identifications from a dataset, a result which is emphasised the larger the number of detected peptides is.

- TMT tags showed a higher rate of incomplete tags present at the 1:37.5 range in the uncomplicated background, with the lower concentration tag not being observed in 75% of cases compared with 56.5% cases for iTRAQ.

  - This suggests that iTRAQ tags has the potential to maintain better precision and accuracy at a wider range of values, and in the case of cleaner signal experiments (such as ms3 mass spec filtering) will likely produce higher quality datasets due to more complete sets of quantification data.

  - This also means that iTRAQ 8-plex as a tagging system has a higher absolute range than TMT 6-plex.

- In the experiment, neither iTRAQ nor TMT methods offered suitable resolution to detect small fold changes (difference of <2-fold) in a complex background similar to a typical proteomic sample.

  - This is emphasised by a combination of a limited number of data points and a low level of relative precision in both tagging systems, however both systems demonstrated the ability to differentiate between such cases without the presence of a complex background, and further investigation should be conducted into the low fold-change range under ms3 filtering.

- To balance the difference between TMT and iTRAQ tags, more technical repeats of injections of iTRAQ-labelled samples should be performed.

  - An increase from 2 injections to 3 (or a 50% relative increase in injection number for iTRAQ experiments) should be sufficient to balance the observations between the two systems, based on the number of observed peptide counts and assuming accumulation of observations in subsequent experiments are linear.

- Both tagging systems showed a dramatic increase in the number of complete quantifications (25% - 45% observations increasing to 85% - 90%) under a complex background, demonstrating how pervasive the effect of isotopic contamination is across both systems.

  - This suggests that the majority of tag-based observations previously will only have observed the very largest changes occurring in the proteome; and so there is a strong case for re-analysing previously measured systems and conditions

- The compression effect observed under typical experimental conditions for both sets of tags was similar, but TMT experienced a more pronounced compression effect under a more complex background.

Chapter 6

Conclusions

6.1 Key findings from this study

Throughout this thesis, a number of novel methods were proposed for improving proteomic investigations – mainly with the organism *Synechocystis*, but some of which were generally applicable to proteomics for biotechnological production applications. In this section, the key conclusions from this thesis will be summarised.

6.1.1 Energy

The key focus of the studies in this thesis were directed towards H_2 production. Due to limitations stated within chapter 1, including a lack of existing infrastructure, exceedingly high costs, and the large amount of development still needed to make the technology viable; it appears unlikely that biological H_2 production in *Synechocystis* is a viable method for producing energy – at least until the price of energy increases substantially to make it cost effective.

Based on the literature review in chapter two, it seems that despite a number of investigations taking place, it is unlikely that *Synechocystis* would be viable for organic biofuel production either – due to a naturally low tolerance to alkane-based biofuels such as butane or hexane. There is scope for engineering the required resistances into the organism, but given that other organisms show much more promise in this area it does not seem like an effective line of investigation. It is possible that there is scope to produce fatty acids for fuel, however it is more likely that these will be capitalised in another industrial pipeline, such as animal feed, due to the relative costs involved in their production, and slightly higher value when applied elsewhere.

Despite this, the natural hydrogenase properties in *Synechocystis* provide the possibility for regenerating some energy costs, when considered as part of a holistic bio-production strategy for animal feed, cosmetics additive production, or synthetic biology engineered fine chemical production. Due to the large amount of investigation that has been carried out on its background, *Synechocystis* has a lot of potential as an engineerable production host, however as it currently lacks a current high-value market it is unlikely to be an industrial production leader any time soon. It is, however, likely to provide the key background findings that will drive forward investigations with other microalgae.

6.1.2 Proteomics

Synechocystis has traditionally posed a number of significant problems for proteomic investigations, however in the last 5 years many of these appear to have been overcome

with the advent of advanced mass spectrometers. There are a number of solutions that have been proposed in chapter 3 to improve the accuracy, repeatability, and quality of proteomic data gathered from *Synechocystis* – these include a less biased protein quantification method (although this needs more robust experimental verification), an improved extraction method (although this needs to be compared to other novel extraction methods being utilised in the medical field such as FASP), and methodological tools for assessing relatively low abundance peptides. When utilising these recommended changes, over a 3-fold increase was observed in the number of protein identifications; however this was largely biased due to a change in the mass spectrometers over the course of the PhD.

In addition, a novel comparison of quantitative tagging technologies – iTRAQ 8-plex tags and TMT 6-plex tags – was carried out, to determine which provided better quality data. It was found that the iTRAQ tags could measure accurately over a broader range of values, but produced a substantially lower number of peptide identifications, resulting in an overall reduction of data quality. As a result, the take-home message from chapter 4 was that whilst iTRAQ tags have greater potential for precision, they ultimately lose out to the TMT tags for statistical reasons on a like-for-like comparison. Performing multiple injections may be suitable to recover this difference.

A novel informatic study of experimental data, to determine features of the proteomic background of *Synechocystis*, was also conducted in chapter 4. This showed that whilst the phycobiliproteins were by far in abundance, as had been determined in a number of previous studies of the organism, the remaining proteins generally had a normal distribution. The study could not determine any features about the range of lower-abundance proteins within the cell, due to a limitation on detection within the proteome, however a future study could correlate proteomic and transcriptomic findings to determine the distribution of non-expressed proteins, proteins expressed at low levels, and the protein distribution identified in this study.

This investigation also highlighted that the very largest proteins in the proteome were in generally low abundance, however as these were mainly membrane-bound proteins, it is possible that the experimental dataset – which was generated externally – may have had some bias in extraction. A further investigation of the data may reveal this bias, and a tool for detecting this form of bias as standard practice during proteome investigation would likely be a valuable addition to the proteomic community.

In chapter 5, a novel investigation was carried out on the proteomic changes experienced by cells under fermentative H_2 producing conditions. The findings of this study generally followed what was expected, based on background investigation into the literature. This provides valuable information for the proteomic community, as it further verifies the previous findings and adds to them slightly by considering *Synechocystis* in a state it

may have held prior to the event leading to oxygenation of the atmosphere.

6.2 Contributions to science

A number of literature and bibliometric analyses were performed, synthesising new information from the pre-existing literature.

- One of these has been published, providing an overview to expand future proteomic studies in biotechnological production. This will contribute to biologically engineered solutions for industrial settings in the future, by linking the bleeding edge studies in proteomics to industrial applications and synthetic biology.
- The remaining studies provided an in-house description of the current state of the proteomic field in *Synechocystis* – whilst an alternative version has been made available in the literature, it does not contain the same data tables, such as a summary of all conditions that have been studied with proteomics to date, or a list of all the methods used and the number of proteins that they generated, nor does it contain a study on the improvement of proteomic techniques in *Synechocystis* over time.
- The data tables described in the point above were used to develop a number of novel methods described in chapter 3 of this thesis – including the protein extraction analysis. They also provided the evidence that whilst the field appears to be advancing in general, the majority of the high-level studies that have been conducted in *Synechocystis* have all been conducted by a single lab in China with a high quality triple-tof mass spectrometer.

Novel processing tools and methods have also been devised, laying the groundwork for improved proteomic processing through:

- Production of a number of in-house processing scripts, facilitating analysis with Principal Component Analysis, Heatmaps and clustering, statistical identification of high abundance proteins, conversion between data formats during downstream proteomic data analysis
- 3 tools which merge proteomic datasets from separate analyses (as described in chapter 3, one of which has been included in a publication, and a second that was emulated in a publication)
- A tool which assigns Gene Ontology terms in a cluster-based analysis, which after further optimisation will be suitable for publication (chapter 3)

The work contained within this thesis has been put towards:

- 3 published manuscripts (within chapters 2 + 3), 2 accepted book chapters (within chapter 1), and 2 manuscripts currently being written (chapters 4 + 5)
- work conducted by 4 partner groups in an EU FP7 grant (Chapter 5 and appendices)
- An industrial summary report for production in *E. coli*
- Poster presentations at 5 national conferences
- Progress presentations at 5 consortium meetings

Over the course of this PhD, work carried out by the author has led to the propagation and training of the scientists of tomorrow through:

- Organising, obtaining funding for, supervising, and teaching a team in the international Genetically Engineered Machines (iGEM) competition, ultimately resulting in being awarded a gold medal at the finals in Boston, USA
- Supervision of 4 masters student projects
- Provision of 3 MSc-level university lectures on the topic of computational biology, synthetic biology, and data processing in proteomics
- Facilitation in the 'Engineering – You're hired' event
- Over 100 hours of facilitator work in the chemical engineering department, including teaching, introducing key concepts of the biological engineering module, devising a marking scheme, and marking of lab reports and individual assessments. This was done for first year, second year and masters level students.

6.2.1 Future work

If work on the topics described in this thesis were to continue and additional funding and time were made available for experimental verification and further data processing, the next steps would be:

- Finalisation of the practical methods identified in Chapter 3 – including verification of the findings for the extraction methods through a robust experimental analysis and of the protein quantification experiment through a broad-spectrum assessment of the different methods available beyond the Bradford analysis.
- Completion of the bioinformatic tool for utilising GO terms in a cluster analysis also described in Chapter 3. This will likely need a robust machine-learning aspect to solve the optimal-cluster problem, which dictates how the solutions to other problems would be determined

- Verification of the background proteome analysis conducted in Chapter 4 with transcript-level data from the same dataset. This would solve the low-abundance cut-off issues encountered during the analysis and could also provide an interesting assessment of the relationship between the RNA and protein levels in *Synechocystis*
- Inclusion of the other 'omics work described in the appendices within the body of the thesis. The broad-reaching nature of this study meant that only a single facet of the overall investigation could be included in the final thesis, however with additional time to complete the analyses that are currently underway, it would be helpful to generate a multi-level systems analysis of the many aspects of *Synechocystis* in an industrial setting. These would include single-day variations occurring to the background in a large-scale photo-bioreactor, the effects of increased salinity and temperature, a kinetic metabolic assessment of the carbon flux under light-harvesting mutants, the effects on metabolism of the organism in the event of environmental release due to an industrial-level contamination event.
  - *The experimental work for many these studies has already been completed, but a complete understanding of the body of data would probably require another PhD project in its entirety or dedicated post-doctoral study.*

Chapter 7

Computational Methods

7.1 Pre-amble

This chapter contains a summary of the code and datasets that this thesis is composed of. The code and data for all methods described in this chapter are available from the Sheffield University Online Research Data repository – managed through figshare, under either the creative commons licence for data, or the MIT licence for code. If you need any further information or data, please contact me by email on *andrewlandels [at] gmail [dot] com*.

7.2 Synechocystis growth rates

Growth rates were taken from 50 ml culture grown in 250 ml shaking flasks in multiple replicates over a period of 2 weeks. The cells were initially grown in BG11 media, then subbed into BG11 media and Burrows media.

This data shows that switching to the Burrows media did not significantly impact the rate of growth. Data DOI: `http://10.15131/shef.data.5327482`

7.3 Synechocystis proteomic data (H2 production)

Synechocystis PCC6803 was grown in two media conditions, standard BG11 and Burrows media; in two different head-space gas mixtures, air and 100% nitrogen (anaerobic). The experiment was continued until hydrogen gas was detected in the head-space, then the samples were collected by centrifugation and flash-frozen in liquid nitrogen. Proteomic analysis with iTRAQ labels was performed, the data was analysed with the EasyProt software (Gluck et al., 2013).

The first sheet of this dataset contains the full list of all proteins identified in this experiment, and the proceeding pages contain lists of proteins that were differentially regulated with statistical significance under the different environmental conditions (media and headspace). Data DOI: `http://10.15131/shef.data.5327476`

7.4 Kalb protein quantification, data

A protein quantification method, devised by Kalb and Bernlohr, was utilised to remove the influence of phycobilisome pigments on protein quantification prior to analysis. Control samples of bovine serum albumin (BSA) and bovine cytochrome c (cyt) were used

as standards of known concentration to investigate this method and how it scaled for individual proteins and mixtures of proteins.

This dataset shows that there was an over-estimation of cyt that scaled linearly, and a slight underestimation of BSA that did not appear to scale. When combined together, the effects of the linear scaling were still present, however were reduced by half. Data DOI: `http://10.15131/shef.data.5327485`

7.5 Protein Quantification in Synechocystis

In this code, a series of plots (pdf files) were generated from the attached data. This data was generated from a series of Bradford assays, which was performed as described previously by Stoscheck.

The code is written in R, and uses the ggplot2 package. The points are plotted and the lines of fit are generated using a linear model (polynomial, degree 3). The error region is a dark-grey ribbon, and is generated using the default settings on ggplot. DOI: `http://10.15131/shef.data.5327488`

7.6 Densitometry analysis of Synechocystis proteins

Proteins from BG11 and Burrows media were analysed by densitometry, using the software imageJ to calculate values for the different protein lanes - briefly, areas were drawn around regions on the plot and the density of the shading of the pixels was calculated in these regions. The gel image used to generate this data is included in the online data repository. The outputted densitometry values are reported in the csv file.

The code in the online repository is written in R, and uses the package ggplot2; initially the code imports the data, finds the sums the regional densities from each of lanes on the gel, then generates a bar chart from these values. DOI: `http://10.15131/shef.data.5327500`

7.7 Poisson noise model for low-abundance labels in iTRAQ

Proteomic iTRAQ analyses generate background noise, which can generate false positive results in very low abundance analyses. In this analysis, a sample dataset (published in

Chiverton et al) is used to generate a model for noise - this is possible because within the experiment, two iTRAQ labels were strategically left blank.

This code, written in R, produces two figures. The first is a histogram of the empty labels in the dataset; and the second is a histogram of a series of values produced using a Poisson distribution of random noise. This comparison shows that due to the discrete nature of mass spectrometer data at low intensities (on the scale of individual counts, as opposed to hundreds or thousands in typical measurements), a Poisson model would need to be used to accurately model the noise. DOI: `http://10.15131/shef.data.5327503`

7.8 Merging tag-based proteomic experiments

The code described in this section is split into two separate scripts, both written in Mathematica. The first (MaxQuant_to_SignifiQuant) converts the data format of files generated by the program MaxQuant and re-orders them into a format that can be input to SignifiQuant - a program in the in-house proteomics pipeline available at the Sheffield University Biological and Chemical Engineering Department. This code reads one or more files within a relevant directory, collects all peptide information, and writes a new file containing all required data. As such, it is both a conversion script and also a data-collecting script.

The second script investigates methods for merging together two biologically replicated datasets - specifically, one dataset represents a complete experimental replicate of the other. The theory behind this methodology is described in chapter 4.6. Briefly, this code examines the label intensity distributions, log-transforms the data, then utilises the median correction method to generate a fixed median value (0) and scales the data to generate an equal gradient between the 40th and 60th percentile.

The protein data in the repeat experiment are then scaled by the protein data in the initial experiment. This slightly disrupts the balancing by median correction, however not significantly. The data are then plotted against each other in a scatter plot, demonstrating systematic improvement of the quality of the between-experiment repeatability. A principal component analysis was then performed, showing a much closer clustering by experimental condition (principal component 1) than of experimental replication deviations (principal component 2), demonstrating success of the method.

This method shows effective combination of two proteomic datasets that are completely independent experimental repeats, demonstrating for the first time that this methodology is feasible in tag-based proteomic investigations. DOI: `http://10.15131/shef.data.5327506`

7.9 Cluster Analysis - Using GO terms

This code attempts to cluster proteins by relative intensity under different conditions, then assign frequencies of Gene Ontology (GO) terms to each of the clusters. The theory behind the code in this section is described in detail in the aforementioned thesis in chapter 4.7. The scripts in this section are written in R and Mathematica, and require the use of the uniprot website to generate the GO terms.

The first part of this pipeline is written in R. The input to this code is tag-based proteomic data, where the first column lists uniprot IDs for the identified proteins, and the subsequent columns contain protein quantifications (these can be absolute or relative). Initially, a list of unique proteins are output from the data. This list of uniprot IDs is then uploaded into a uniprot search. The uniprot table is updated to include all GO terms, by clicking on the 'columns' button, selecting the GO Terms drop-down, and checking each of the boxes. These settings are applied by clicking 'save' in the top right-hand corner. Once this is done, the data is downloaded for use further along in the pipeline.

The proteomic data read by the R script is clustered, and using a K-means analysis a critical cut-off point for the number of clusters is selected. This value is chosen manually, and is selected based on a "within-groups sum of squares" graph. This graph calculates the sum of squares distance between all points to a central mean, then applies two means, creates two clusters, and calculates the sum of squares again. This process is iterated until 20 means have been applied to the data. This is plotted as the aforementioned graph, where the analyst is aiming to have the minimum possible number of clusters, but also the lowest sum of squares. In the worked example provided, 8 clusters were selected (as highlighted by a verticle line on the plot).

The proteins were grouped into 8 clusters and assigned a side-colour. These clusters were exported into a csv file for use later. Finally, a heatmap was generated, using the gplots package, was used from the data. This heatmap had 2 dendrograms - one showing relatedness of the labels, and the second for the proteins. The selected clusters highlighted with side colours.

The next part of the analysis was performed in Mathematica. The GO Terms downloaded from UniProt were linked to each of the proteins on the list. To de-clutter the data and simplify the analysis, only GO terms with 20 or more unique references from the dataset were extracted, and the remaining terms were discarded. The set of remaining GO terms within each cluster were tallied, producing a matrix of GO terms and a count for each cluster. The values for each cluster were divided by the sum of all observations, producing values between 0 and 1 for each term in each cluster.

This list was then plotted to show the GO distribution across each of the selected clusters,

enabling analysis of GO term concentration within clusters of the dataset. DOI: `http://10.15131/shef.data.5327524`

7.10 Proteomic background in Synechocystis with emPAI

The code for the methodology described below was written in Wolfram Mathematica (10.1) and the notebook file is "iTRAQ_TMT-complexity_emPAI.nb"

An in-depth proteomic dataset, comprised of 2 8-plex iTRAQ experiments investigating a mutant against WT *Synechocystis* under two different conditions, was generated on a Q-Exactive HF mass spectrometer (data not included in this repository due to size constraints). To calculate the emPAI scores, the 'observable' peptide values were calculated as follows. The complete proteome for Synechocystis PCC6803 – Kazusa strain, was downloaded as a fasta file from uniprot (taxonomy:1111708 – accessed August 2015, 3517 protein entries), which is available in this respository.

This was then merged with the spike-in proteins to make a singular database for analysing the data, by doing this, effects on statistical methods such as false discovery were equal between all analyses. The fasta file was processed in Wolfram Mathematica (version 10.1) to generate an in-silico digest of each of the proteins, excluding any peptides that fell outside a 1000 – 7500 dalton window to replicate the presence of 2+ or 3+ ions observable in the 500 – 2500 m/z window used during the mass spec experimental scan. The emPAI scores for all identified proteins were calculated using the following formula.

$$emPAI = 10^{(\frac{N_{observed}}{N_{observable}})} - 1$$

Where $N_{observed}$ is the number of unique peptides observed for a given protein, and $N_{observable}$ is the total number of unique peptides that could be observed for a given protein.

This data was then graphed as a histogram to identify the protein concentration distribution and dynamic range. Dynamic range was calculated by taking the exponential of the difference between the maximal and minimal emPAI values. DOI: `http://10.15131/shef.data.5327539`

7.11 Comparing iTRAQ and TMT isobaric tags

The code presented here was written in R, and uses the ggplot2 package for generating graphs. The data was generated a QExactive mass spectrometer, and was analysed in MaxQuant software using the standard data processing pipeline. Included with the code for this section are the output from the MaxQuant pipeline, namely the 'Evidence' files, which contain peptide information and raw/processed quantifications.

The code in this section is a little detailed, due to the style of the experimental design: the experimental methodology is described in detail in chapter 5 of the aforementioned thesis. The files have been named in such a manner that they contain a series of switch identifiers (X_X_X, where X is either 1 or 0 and _ is an identifier for splitting the file-name string with a regular expression). These were used to trigger switches in the code, and determined if the dataset being analysed used iTRAQ or TMT tags, was a flat or extended concentration range of the protein mix, and whether the extended concentration mix was run in the forwards or reverse direction (the test proteins were flipped during the experiment, to avoid protein-specific skew).

At the beginning of the code, a number of functions for applying different transformation to the data are defined:

Tags() generates a matrix of expected tag values for the spike-in proteins.

Corrected.new() turns all the values into ratio values between 0 and 1, relating to the sum of a given row.

Trim() removes the file extensions from the filenames, to enable correctly labelled graphs.

FlattenData() collects all peptides relating to the spike-in proteins from the data, and correctly arranges formatting to collapse them into a single matrix where they are aligned with the expected values based on the experimental design.

RemoveZeros() removes any rows containing 1 or more zeros.

PeptidesQuants() generates a table of spectral counts for: all peptides in a sample, just the spike-in proteins, with and without 0-values removed. It also counts the number of proteins identified and the total number of fractions the samples was measured across (ie. relating to the degree of LC separation)

ScalingMatrix() calculates, based on the expected and observed quantifications from the data, the scalar that is needed to be applied to the spike-in data to make it equal to the expected values. This enables iterative investigation of the spiked in proteins, enabling systematic experimental operator error to be measured by comparing the same mix within the final experiment after it had been exposed to successive permutations (initial mixing,

dilution, etc).

ScaleData() applies the scaling matrix to the data in the manner described above.

In the main body of the code, initially the names of the spike-in proteins are given, then the summary table of counts is produced. After this, a series of plots are generated, showing the data after successive manipulations have been performed on it. Finally, the individual plots that make up the final plot from the previous section – where all transformations have been applied to the data – are produced, in both linear space and log space.

In the post-section of the code, a number of variations on the plots are produced. These highlight other features that were explored during the data processing in order to produce simpler graphics, but were ultimately not included in the thesis. DOI: `http://10.15131/shef.data.5327866`

Chapter 8

Appendices

8.1 Deliverable 7.1

CyanoFactory Confidential Report

***'Base case* Synechocystis – *omics analysis and Base network assembly of this data for the "Burrows" and the "Baebprasert" conditions'**

This deliverable was considered in two parts:

1. Generate a 'base case' -omics data set for the Uppsala strain of *Synechocystis*.

As iTRAQ 8-plex labels are used for quantification, 2 labels are reserved from each iTRAQ to collect additional data for the 'base case'. The initial results of this data collection are reported here, but will become more comprehensive as additional comparative proteomic analyses are run. Metabolic and transcriptomic data will be combined with this as they become available. A full list of confidently (1% FDR with at least 2 unique) identified proteins observed so far are included as an appendix ("raw-data").

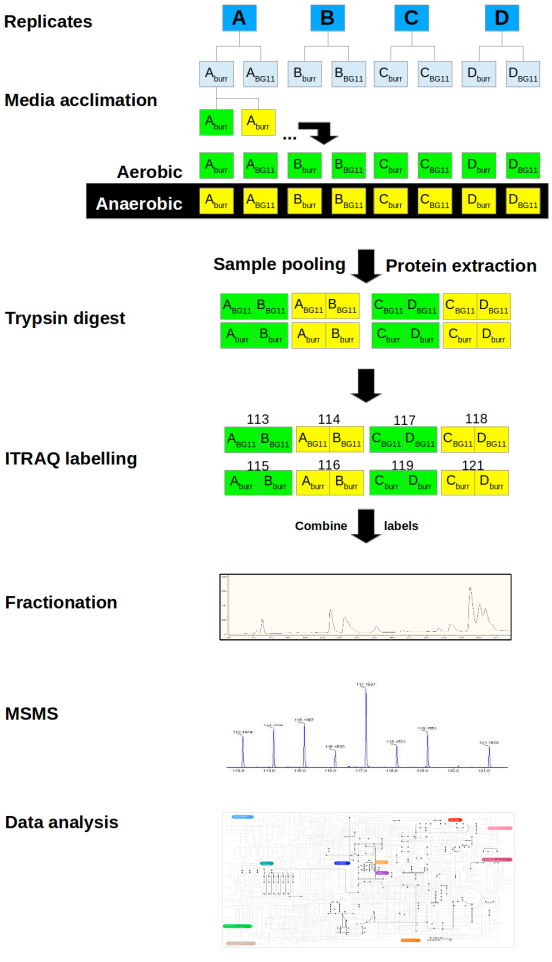
2. Assemble network trends observable in this data for the "Burrows" and "Braebprasert" conditions.

The assembly of networks is considered at a pathway-wide level, indicating increase or decrease of particular cellular functions from the protein data in conditions shown to increase hydrogen production. As the Braebprasert mutants are no longer available for analysis, the network assembly was performed on the "Burrows" media conditions in two different backgrounds to produce a deeper analysis:

- Standard growth conditions
- Conditions shown to induce hydrogen production

Methods

To increase the robustness of the iTRAQ labelling experiment, 4 biological replicates were performed for each condition. Two biological replicates were pooled for each iTRAQ label, to increase the detectability of consistent up and down regulation in individual proteins. Separately tagged replicates were included to increase statistical robustness for the investigation, as shown in the work-flow.



Four biological replicates were grown into mid-log phase (OD of 1.2) in separate 250 ml shaker-flasks. The samples were washed and used to inoculate flasks containing the experimental media conditions, BG11 and Burrows (OD of 0.075). They were grown to mid-log phase (OD of 1.2) to acclimate the cells to the target media. They were then spun down, split and transferred to serum bottles at equal OD, half in aerobic and half in anaerobic conditions. The experiment was stopped when hydrogen production was detected by gas chromatography in the anaerobic conditioned samples. The cell samples were harvested by centrifugation and the cell pellet processed for protein analysis.

The samples were lysed with bead-beating. The iTRAQ protocol implemented was according to manufacturer's instructions: protein reduction and alkylation, proteolytic

digestion with trypsin to generate peptides. Each sample was labelled with a different iTRAQ label (8 plex), samples were combined and fractionated by preparative HPLC using a porous graphitic carbon column (Hypercarb, Thermo). Fractions were analysed with nano flow LC MSMS, using a Q-Star XL QTOF tandem mass spectrometer.

Protein identification was done using EasyProt (Phenyx) against the Uniprot database entry *Synechocystis* PCC 6803 (uploaded December 2013) to find a list of significantly identified proteins. These were normalised by median division, then filtered to a 95% significance level to find a list of significantly increased or reduced protein quantifications, using the Mascat and Libra statistical methodologies from EasyProt. Metabolic network investigation was done using KEGG pathway database to find patterns in the metabolic network combined with a Uniprot search for gene ontology terms and keywords.

Results

From the proteomic analysis, 345 proteins were confidently identified at a 1% FDR with at least 2 unique peptide matches. A full list of these proteins is available as an appendix. Of these, 335 contained 2 or more unique iTRAQ labelled-peptides for quantification, with over 206 unique proteins being quantified as significantly different in the entire study. These values were obtained with a single injection from 30 fractions. Comparative KEGG maps, indicating the pathways covered by this study (black) and highlighting changes (green = up, red = down) for each comparison are available as supplementary materials.

Base Case

The 'base case' strain of *Synechocystis* investigates the strain that the consortium are using as a baseline for all of our investigations. Here we have reported a list of confidently identified proteins for this strain that can be used to compare against future experimental runs. This will ensure that all proteomic work performed during CyanoFactory is comparable.

Another part of this ongoing deliverable is to maximise data output per run. Here we report use of a porous graphitic carbon column for HPLC, a method that has not been performed in *Synechocystis* before, based on a literature review of proteomic studies in *Synechocystis*. This has shown significant improvement over other popular methods, such as strong cation exchange.

In anaerobic dark conditions, 76 proteins were found to be differentially expressed. There are reductions in the pentose phosphate and carbon fixation pathways, with an increase in some parts of the TCA cycle. There is also a strong reduction in the phycobiliproteins, but an increase in the enzymes relating to NADP and NADPH to compensate for electron transport associated with hydrogen production.

In future work, the 'base case' will be expanded to include other details such as native flux rates, metabolite concentrations and transcript data. Additional target areas of investigation for the project include further -omic analyses, such as identifying the phospho-proteome for our strain of *Synechocystis*.

“Burrows” condition

The “Burrows” condition media was compared to BG11, the standard media used by the consortium. Under standard conditions, 137 proteins were found in concentrations significantly different to BG11 when compared across all biological replicates.

Reduced protein quantification was observed across amino-acid biosynthesis pathways and nitrate-related pathways. This is concordant with the absence of nitrate in the media and general nitrogen starvation. We

observed reduced abundance in proteins involved in metal chelation or with metal ligand properties, as were other sulphur-rich proteins. This was an expected observation, as sulphur is a limiting factor in the "Burrows" condition.

An increase was observed in phycobiliproteins and enzymes relating to electron transport. These are perhaps responsible for the increase hydrogen production observed in the Burrows media conditions. Large sequences of pathways in the central carbon metabolism, including carbon fixation and the pentose phosphate pathway were found in lower quantities, however individual proteins between these points in carbon metabolism were also significantly increased. Ribosomal proteins (non-network) were made up 39% (18/46) of the identified proteins with significantly increased levels with a fold-change greater than 1.5, indicating a high level of protein turnover.

In anaerobic dark conditions shown to produce hydrogen, we observed 141 differentially quantified proteins between BG11 and "Burrows" conditions. Large portions of differentially quantified proteome remain similar to the aerobic investigation. Of particular interest is the further increase in "Burrows" conditions of phycobiliproteins during anaerobic-dark conditions, as the opposite effect is observed in BG11. There is also further reduction in proteins in the central carbon pathway.

Directly comparing "Burrows" in aerobic and anaerobic-dark conditions produced 53 differentially quantified proteins. The differences between these two conditions were less pronounced, although a further increase in phycobiliproteins and enzymes relating to NADP and NADPH, further validating the observed increase in antenna proteins for electron transport. There was also a further reduction in carbon fixation and the pentose phosphate pathways, similar to BG11 in the same conditions.

Conclusions

The major non-intuitive protein differences between "Burrows" and BG11 lie in the carbon metabolism pathways and electron-carrier pathways, and how they respond to hydrogen producing conditions.

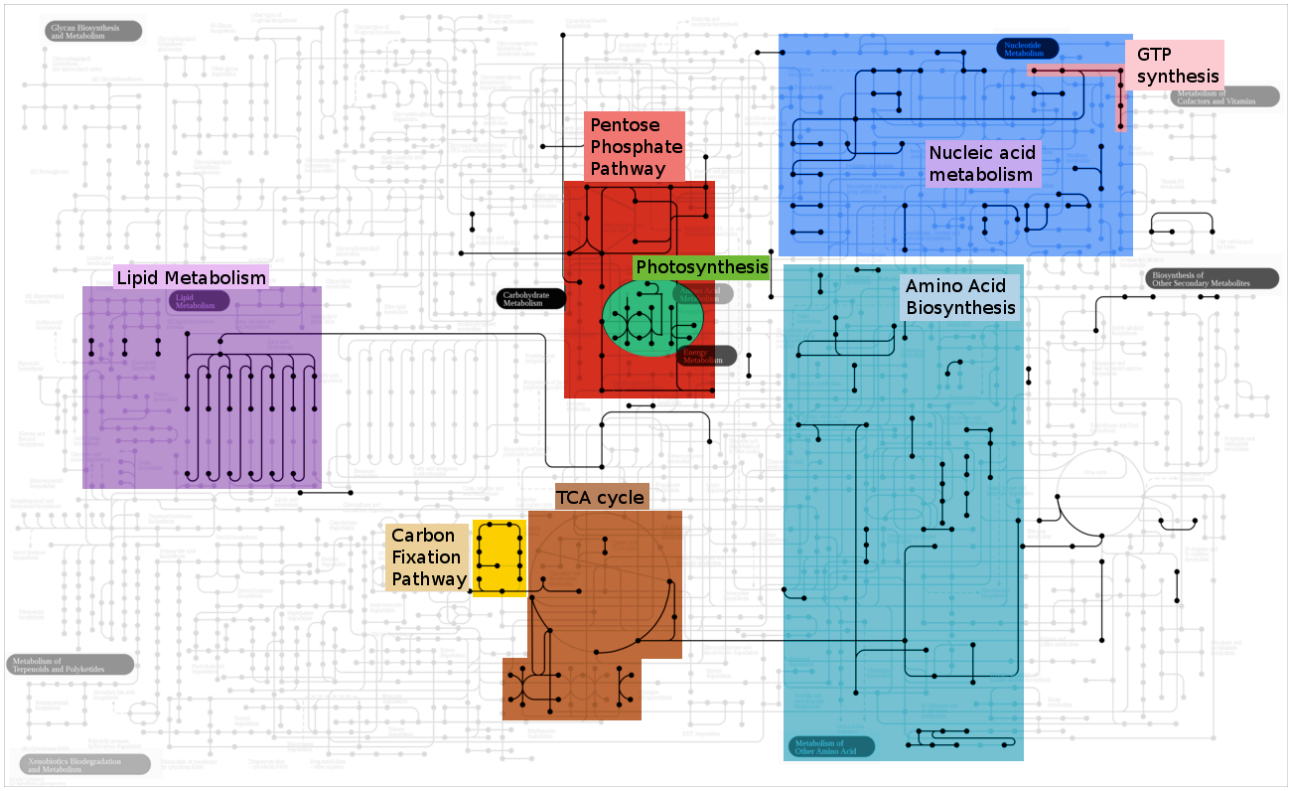
Higher resolution data is required to accurately identify individual bottlenecks in the carbon metabolism pathways, however data will be enriched with metabolite fingerprinting (Milestone, 18M) and carbon flux analysis (**D7.2**).

The differences in redox and electron carrier proteins between BG11 and "Burrows" has also been highlighted as a potential bottleneck in the hydrogen production process.

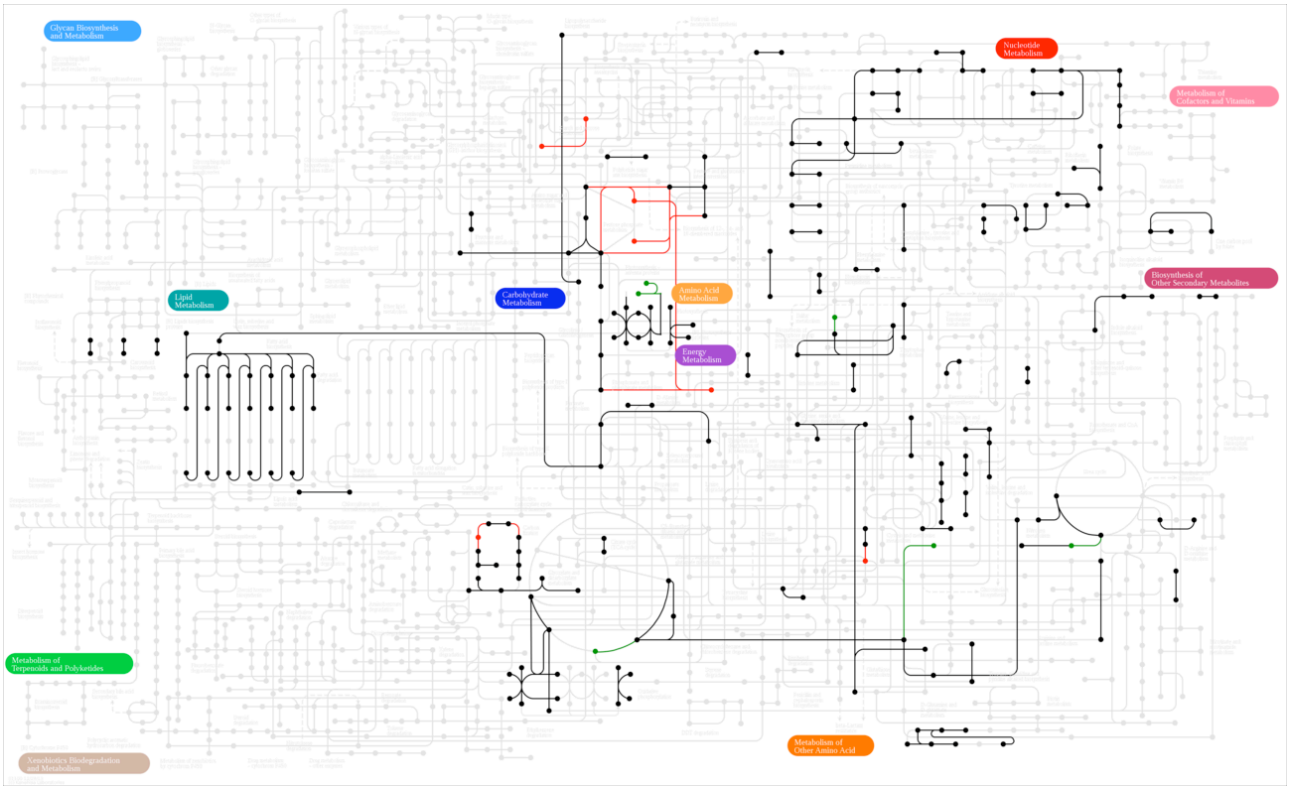
Appendixes

1. Identified proteins and metabolic pathways visualised using a KEGG map, summary followed by iindividual growth conditions.
2. Summary and characteristics of all identified proteins, "raw-data".

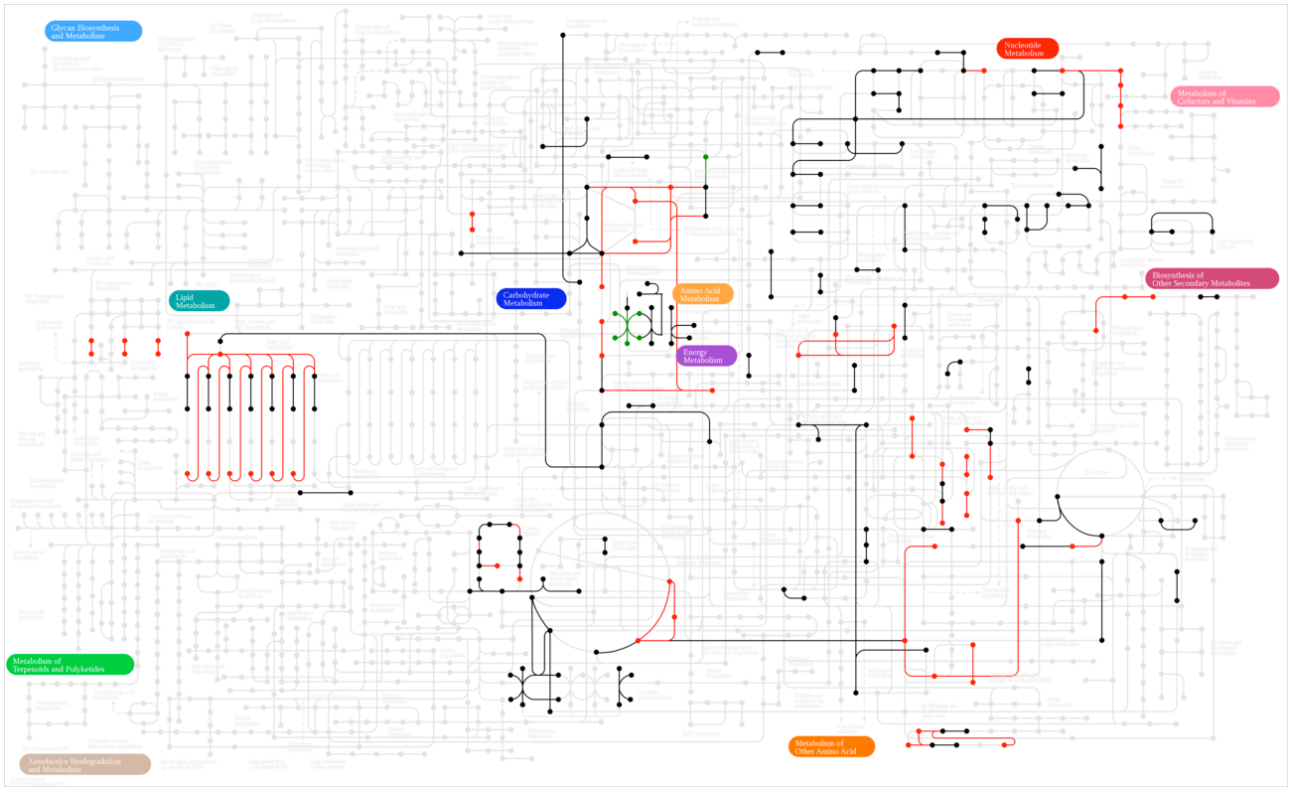
Graphical presentation of a summary identified proteins and metabolic pathways using a Kegg map



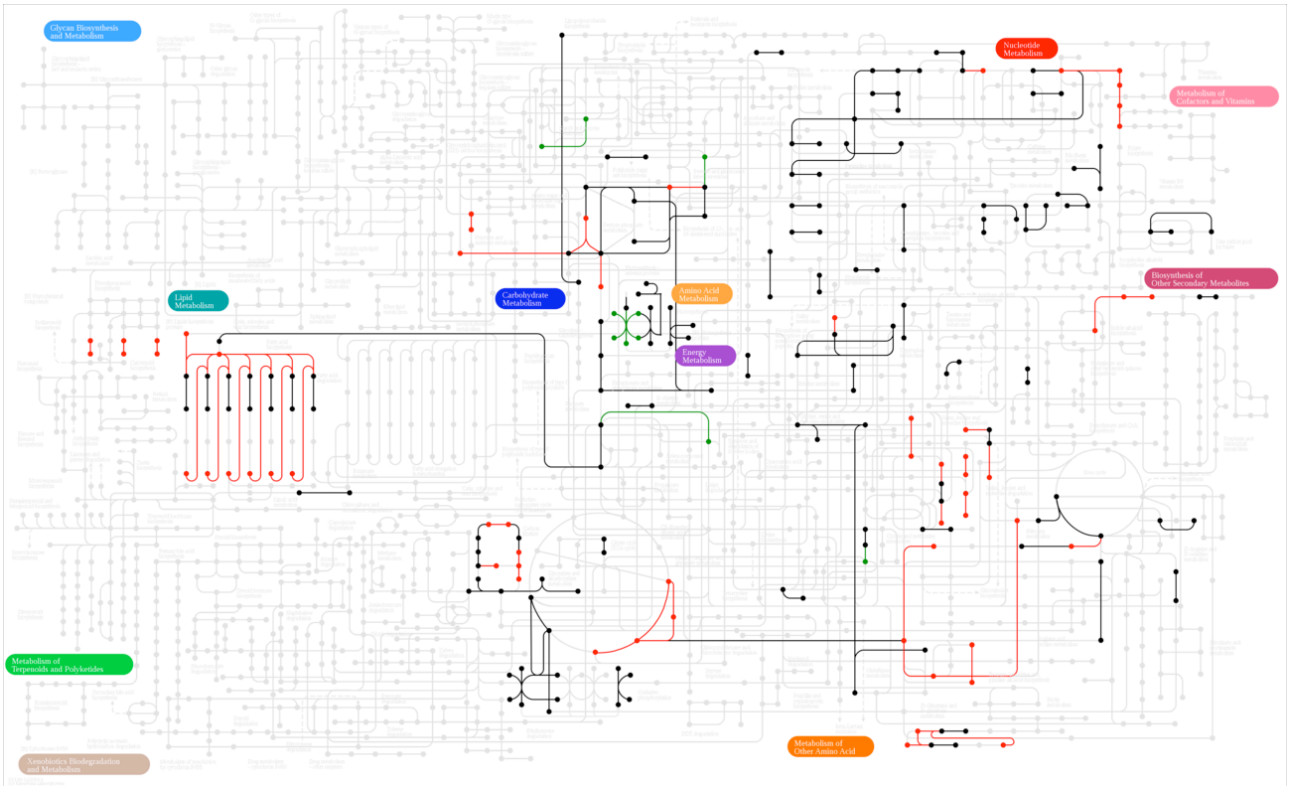
BG11 medium aerobic light condition versus Burrows medium anaerobic dark condition



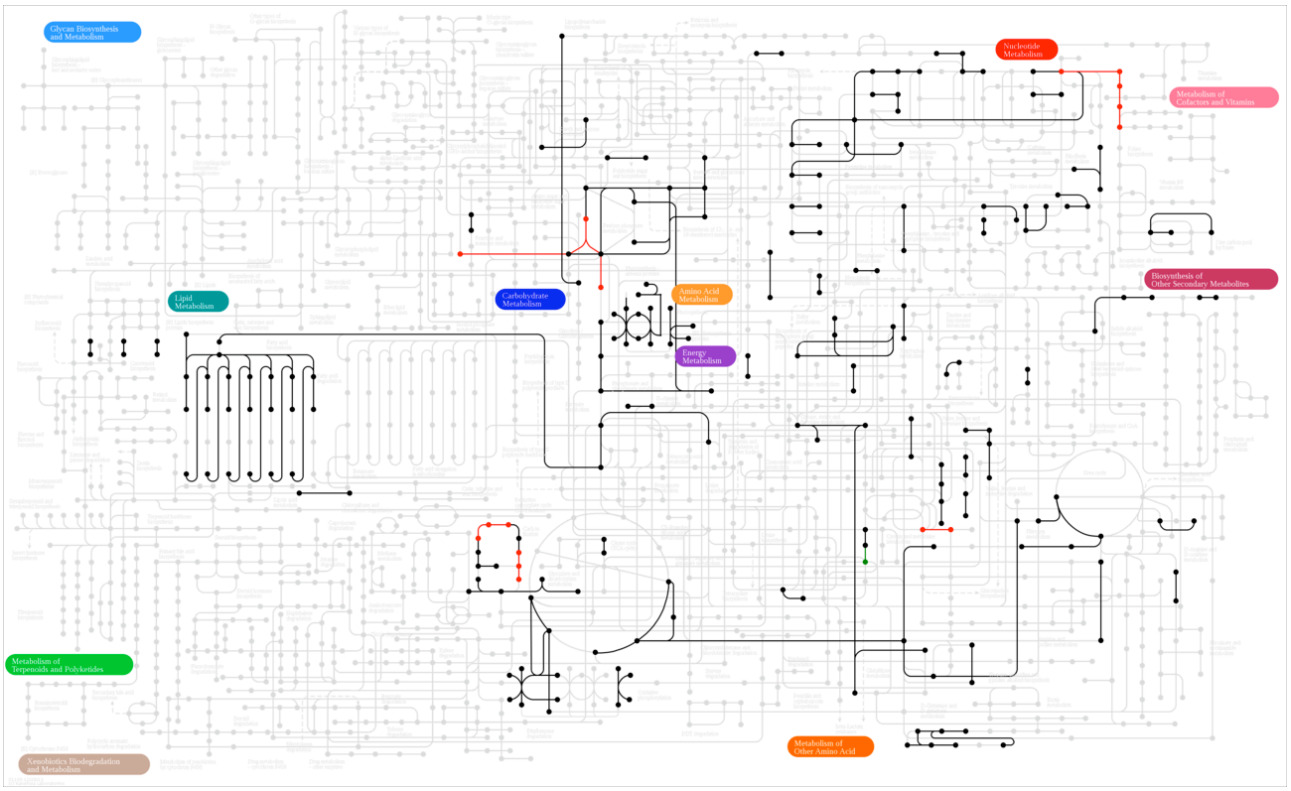
BG11 medium aerobic light condition versus Burrows medium aerobic light condition



BG11 medium anaerobic dark condition versus Burrows medium anaerobic dark condition



Burrows medium aerobic light condition versus Burrows medium anaerobic dark condition



Protein Summary

Rank	AC	ID	Description	Protein Score	% Coverage	Protein Seq	#PSMs	#Peptides	#Ambiguous Pr	Ambiguous Prc	#Sub-Prots	Sub-Prots	Protein PI	Protein Mass (I	Database	GO terms	Keywords	Job IDs
1	P22358	DNAK2_SYN	Chaperone pr	354,054	26,1	MGKVVGIDL(	40	19	0		0		4,74	67614,221	Synechocystis	ATP binding ((	ATP-binding; (	[1390935826213]
2	Q55366	Q55366_SYI	Ferredoxin--ni	248,233	26,49	MANKFETVK/	26	13	0		0		6,3	55927,131	Synechocystis	4 iron, 4 sulfur	4Fe-4S; Com	[1390935826213]
3	P74227	EFTU_SYNY3	Elongation fac	243,781	30,58	MARAKFERT	23	13	0		0		5,21	43733,045	Synechocystis	cytoplasm (GC	Complete prol	[1390935826213]
4	P29107	ILVC_SYNY3	Ketol-acid red	226,607	38,37	MARMYYDQI	27	12	0		0		4,97	35821,872	Synechocystis	coenzyme bin	Amino-acid bi	[1390935826213]
5	Q55118	PPI3_SYNY3	Putative thylal	226,464	30,87	MQIIKTPLGII	24	11	0		0		4,77	41385,521	Synechocystis	thylakoid lume	Complete prol	[1390935826213]
6	Q54714	PHCB_SYNY3	C-phycocyanii	213,284	46,51	MFDVFTRVV	175	9	0		0		5,17	18126,475	Synechocystis	phycobilisome	3D-structure;	[1390935826213]
7	Q55318	FENR_SYNY3	Ferredoxin--N	208,108	24,21	MYSPGYVAT	18	12	0		0		5,82	46359,554	Synechocystis	phycobilisome	Complete prol	[1390935826213]
8	P23353	AROC_SYNY3	Chorismate sy	203,274	30,66	MGNTFGSLF	21	11	0		0		5,89	39287,632	Synechocystis	chorismate sy	Amino-acid bi	[1390935826213]
9	Q55436	Q55436_SYI	Slr0848 protei	195,803	37,18	MTRRESANP	16	10	0		0		4,88	31790,463	Synechocystis		Complete prol	[1390935826213]
10	P73971	AMPA_SYNY3	Probable cyto	193,748	25	MQIRGTDYTA	23	11	0		0		5,05	52165,885	Synechocystis	cytoplasm (GC	Aminopeptida	[1390935826213]
11	P73317	RL2_SYNY3	50S ribosoma	192,431	32,97	MGIRNYRPM	16	9	0		0		11,29	30433,158	Synechocystis	large ribosom	Complete prol	[1390935826213]
12	P73853	P73853_SYN	IMP dehydroq	181,707	25,58	MNITIGRGKT	19	10	0		0		5,32	40234,734	Synechocystis	oxidoreductas	Complete prol	[1390935826213]
13	P73722	P73722_SYN	SOS function	181,246	41,38	MEPLTRAQKI	34	10	0		0		6,21	22744,208	Synechocystis	DNA binding (	Complete prol	[1390935826213]
14	P80505	G3P2_SYNY3	Glyceraldehyc	180,048	28,49	MTRVAINGFG	17	10	0		0		6,19	36512,469	Synechocystis	cytoplasm (GC	Complete prol	[1390935826213]
15	P73458	P73458_SYN	Carboxyl-term	179,382	25,06	MLKQKRSLIL	12	10	0		0		5,16	46832,931	Synechocystis	outer membra	Complete prol	[1390935826213]
16	Q05971	CH10_SYNY3	10 kDa chape	178,711	66,99	MAAISINVST	19	8	0		0		5,38	10859,425	Synechocystis	cytoplasm (GC	Chaperone; C	[1390935826213]
17	P22034	CH602_SYN	60 kDa chape	173,871	22,1	MSKLISFKDE	15	10	0		0		4,92	57774,336	Synechocystis	cytoplasm (GC	ATP-binding; (	[1390935826213]
18	P73304	RS5_SYNY3	30S ribosoma	171,095	53,18	MAKRRKTSR	14	8	0		0		10,73	18241,106	Synechocystis	small ribosom	Complete prol	[1390935826213]
19	P74689	ILVD_SYNY3	Dihydroxy-aci	168,078	18	MSNNPRSQV	20	10	0		0		5,15	58646,222	Synechocystis	4 iron, 4 sulfur	4Fe-4S; Amin	[1390935826213]
20	P73282	P73282_SYN	Transketolase	165,345	15,37	MVVATQSLDI	16	9	0		0		5,4	71725,256	Synechocystis	metal ion bind	Calcium; Com	[1390935826213]
21	P73479	P73479_SYN	Succinate deh	164,321	18,78	MLEQDVVIV(	12	10	0		0		5,51	63702,213	Synechocystis	flavin adenine	Complete prol	[1390935826213]
22	Q54715	PHCA_SYNY3	C-phycocyanii	161,086	33,33	MKTPLTEAVS	48	6	0		0		5,79	17586,599	Synechocystis	phycobilisome	3D-structure;	[1390935826213]
23	P77972	ENO_SYNY3	Enolase	155,833	21,76	MLSKVPATIE	14	9	0		0		5,07	46528,656	Synechocystis	cell surface (C	Complete prol	[1390935826213]
24	P23349	RL7_SYNY3	50S ribosoma	155,05	42,19	MSAATDQILE	66	8	0		0		4,82	13259,308	Synechocystis	ribosome (GO	Complete prol	[1390935826213]
25	Q59978	GRPE_SYNY3	Protein GrpE	153,865	37,35	MNEDQVSLE	14	8	0		0		4,65	27567,598	Synechocystis	cytoplasm (GC	Chaperone; C	[1390935826213]
26	P73308	RL5_SYNY3	50S ribosoma	150,692	38,89	MTQRLKTLY(	14	7	0		0		9,81	20230,434	Synechocystis	ribosome (GO	Complete prol	[1390935826213]
27	P77969	HEM2_SYNY	Delta-aminole	149,742	24,46	MFPTIRPRRL	13	9	0		0		4,94	36163,346	Synechocystis	metal ion bind	Complete prol	[1390935826213]
28	P72839	P72839_SYN	Slr1301 protei	148,307	17,2	MLYLAEIKKQ	12	8	0		0		4,83	61692,317	Synechocystis		Complete prol	[1390935826213]
29	Q55707	Y617_SYNY3	Uncharacteriz	145,883	32,21	MGLFDRLGR	13	7	0		0		5,02	28905,339	Synechocystis		Coiled coil; Cc	[1390935826213]
30	P72761	CCMK2_SYN	Carbon dioxid	145,777	53,4	MSIAVGMIET	22	6	0		0		5,66	11134,696	Synechocystis		3D-structure;	[1390935826213]
31	Q55665	GSA_SYNY3	Glutamate-1-s	145,131	21,71	MVNATPFITT	20	8	0		0		5,21	45891,634	Synechocystis	cytoplasm (GC	Chlorophyll bi	[1390935826213]
32	P73728	Y1621_SYN	Putative perox	144,654	31,75	MTPERVPSV	18	7	0		0		5,04	21167,235	Synechocystis	peroxidase ac	Complete prol	[1390935826213]
33	P52231	THIO_SYNY3	Thioredoxin	141,198	67,29	MSATPQVSD	18	6	0		0		5	11748,533	Synechocystis	protein disulfic	Complete prol	[1390935826213]
34	Q01952	PHAB_SYNY3	Allophycocyar	141,198	44,1	MQDAITAVIN	25	6	0		0		6,25	17215,645	Synechocystis	phycobilisome	Bile pigment;	[1390935826213]
35	P73057	Y1847_SYN	Nucleoid-assc	140,73	55,26	MAQGKGFGF	19	7	0		0		4,71	12134,819	Synechocystis	bacterial nucl	Complete prol	[1390935826213]
36	P73530	RS1A_SYNY3	30S ribosoma	140,238	23,17	MVSQTSTATI	9	8	0		0		4,56	36570,064	Synechocystis	ribosome (GO	Complete prol	[1390935826213]
37	P73418	DBH_SYNY3	DNA-binding p	136,95	40	MNKGELVHA	42	7	0		0		9,04	10890,647	Synechocystis	DNA binding (	Complete prol	[1390935826213]
38	P74281	P74281_SYN	Soluble hydro	136,127	24,22	MDNKQMLMI	19	8	0		0		7,13	40768,131	Synechocystis	catalytic activi	Complete prol	[1390935826213]
39	P74229	RS7_SYNY3	30S ribosoma	132,427	46,15	MSRRGNVKK	19	7	0		0		10,57	17384,149	Synechocystis	small ribosom	Complete prol	[1390935826213]
40	P72870	PHAC_SYNY3	Allophycocyar	132,221	39,75	MSVVSQVILC	16	6	0		0		5,86	17923,451	Synechocystis	phycobilisome	Bile pigment;	[1390935826213]
41	P74392	P74392_SYN	Sll0274 protei	128,793	31,63	MLMNFLVICC	12	6	0		0		6,3	21473,395	Synechocystis		Complete prol	[1390935826213]
42	P72827	FUTA1_SYN	Iron uptake pr	128,589	20,28	MVQKLSRRL	13	8	0		0		4,92	39370,263	Synechocystis	plasma membr	3D-structure;	[1390935826213]
43	P52415	GLGC_SYNY3	Glucose-1-ph	127,816	19,13	MCCWQSRG	11	8	0		0		6,23	49366,439	Synechocystis	ATP binding ((	ATP-binding; (	[1390935826213]
44	P72704	PPI1_SYNY3	Probable pept	126,968	28,46	MRILPNISRA	18	6	0		0		5,1	26580	Synechocystis	outer membra	Complete prol	[1390935826213]
45	P73152	Y982_SYNY3	Uncharacteriz	124,302	46,09	MTTNNSTLL	8	5	0		0		4,95	14282,752	Synechocystis	thylakoid mer	Coiled coil; Cc	[1390935826213]
46	P74456	RRF_SYNY3	Ribosome-rec	122,287	30,77	MKLAELKDHI	23	7	0		0		6,01	20185,962	Synechocystis	cytoplasm (GC	Complete prol	[1390935826213]
47	Q01951	PHAA_SYNY3	Allophycocyar	121,168	31,68	MSIVTKSIVN/	37	5	0		0		4,98	17411,829	Synechocystis	phycobilisome	Bile pigment;	[1390935826213]
48	P10549	PSBO_SYNY3	Photosystem I	120,632	21,9	MRFRPSIVAL	21	7	0		0		4,89	29911,705	Synechocystis	cell outer mem	Complete prol	[1390935826213]
49	P73319	RL4_SYNY3	50S ribosoma	119,934	24,76	MVDCIVKNW	12	8	0		0		10,3	23355,84	Synechocystis	ribosome (GO	Complete prol	[1390935826213]
50	P74494	NDK_SYNY3	Nucleoside di	119,869	39,6	MERTFIMIKP	18	5	0		0		5,6	16692,117	Synechocystis	cytoplasm (GC	ATP-binding; (	[1390935826213]
51	P74410	RS16_SYNY3	30S ribosoma	116,577	51,22	MIKLRLKRFG	14	6	0		0		10,73	9556,168	Synechocystis	ribosome (GO	Complete prol	[1390935826213]
52	P73488	P73488_SYN	Sll1130 protei	113,437	46,96	MNTIYEQFD\	28	5	0		0		9,3	12933,084	Synechocystis	plasma membr	Complete prol	[1390935826213]
53	P73826	FABG2_SYN	3-oxoacyl-[ac	113,389	24,17	MLSLGLEDK\	17	7	0		0		6,76	25333,131	Synechocystis	3-oxoacyl-[ac	Complete prol	[1390935826213]

54	P73603	P73603_SYN	Slr1852 protei	113,238	31,91	MSSRKNYYL	12	6	0	0	5,15	21834,664	Synechocystis Complete prot [1390935826213]
55	P73320	RL3_SYNY3	50S ribosoma	109,664	32,39	MSIGILGTKL(	16	6	0	0	10,22	22741,131	Synechocystis ribosome (GO Complete prot [1390935826213]
56	Q55499	SSB_SYNY3	Single-strand	108,596	34,71	MSVNSIHLVG	13	5	0	0	7,86	13656,307	Synechocystis single-strande Complete prot [1390935826213]
57	P72864	P72864_SYN	Carboxysome	107,723	21,65	MIVVMKVGTI	10	7	0	0	6,1	37633,72	Synechocystis aldehyde-lyas Complete prot [1390935826213]
58	Q05972	CH601_SYN	60 kDa chape	106,719	12,75	MAKSIIYNDE	12	7	0	0	5,04	57652,746	Synechocystis cytoplasm (GO ATP-binding; ([1390935826213]
59	P49433	G3P1_SYNY	Glyceraldehyc	106,271	18,58	MLKIGINGFG	17	6	0	0	6,12	36146,339	Synechocystis cytoplasm (GO Complete prot [1390935826213]
60	P72854	SIR_SYNY3	Sulfite reducta	104,342	10,87	MVTTPTAAPF	7	7	0	0	8,74	71441	Synechocystis 4 iron, 4 sulfur 4Fe-4S; Com [1390935826213]
61	P74232	PUR2_SYNY3	Phosphoribos	104,156	16,23	MKVAVIGSGC	8	7	0	0	5,06	44016,57	Synechocystis ATP binding ((ATP-binding; ([1390935826213]
62	P73643	P73643_SYN	Sll1762 protei	104,01	18,06	MICSATPDRF	11	7	0	0	4,85	41985,163	Synechocystis outer membra Complete prot [1390935826213]
63	Q55759	GCH1_SYNY3	GTP cyclohyd	103,865	21,37	MTIASSHSINI	11	6	0	0	6,31	26639,633	Synechocystis cytoplasm (GO Complete prot [1390935826213]
64	P73312	RL29_SYNY3	50S ribosoma	103,15	69,86	MALPNIADAF	17	5	0	0	8,31	8545,648	Synechocystis ribosome (GO Complete prot [1390935826213]
65	P73922	FBSB_SYNY3	D-fructose 1,6	103,104	16,23	MDSTLGLEIIE	13	7	0	0	5,17	37074,514	Synechocystis fructose 1,6-bi 3D-structure; ([1390935826213]
66	P73911	KATG_SYNY3	Catalase-perc	102,823	9,68	MGTQPARKL	11	7	0	0	5,5	84445,906	Synechocystis catalase activi Complete prot [1390935826213]
67	P26290	UCRIB_SYNY3	Cytochrome b	102,427	32,78	MTQISGSPDY	8	5	0	0	5,02	18996,356	Synechocystis integral to mer 2Fe-2S; Com [1390935826213]
68	P73348	P73348_SYN	Rehydrin	102,378	26,54	MALQLGDVV	15	6	0	0	5,2	23559,723	Synechocystis antioxidant ac Complete prot [1390935826213]
69	Q55552	Q55552_SYN	IMP dehydrog	100,994	28,39	MSRTVGEVM	14	5	0	0	5,37	17270,028	Synechocystis adenyl nucleo Complete prot [1390935826213]
70	P42352	RL9_SYNY3	50S ribosoma	100,448	25,66	MAKRVKVVLI	21	5	0	0	9,39	16641,321	Synechocystis ribosome (GO Complete prot [1390935826213]
71	P74122	ARGJ_SYNY3	Arginine biosy	99,283	14,04	MADWQVIEG	8	6	0	0	5,43	43320,09	Synechocystis cytoplasm (GO Acyltransferas [1390935826213]
72	P73565	P73565_SYN	Sll0872 protei	98,098	33,08	MAGGILKTGI	17	5	0	0	7,72	13929,416	Synechocystis Complete prot [1390935826213]
73	P72586	P72586_SYN	GDP-D-mann	97,467	16,3	MSKSKVVLLT	12	7	0	0	6,14	41333,809	Synechocystis intracellular (C Complete prot [1390935826213]
74	P80046	IDH_SYNY3	Isocitrate deh	95,383	13,26	MYEKLQPPS	10	7	0	0	5,48	52274,791	Synechocystis cytoplasm (GO Complete prot [1390935826213]
75	Q55497	OTC_SYNY3	Ornithine carb	92,168	17,21	MGIKALAGRI	6	5	0	0	5,5	33615,934	Synechocystis cytoplasm (GO Amino-acid bi [1390935826213]
76	P73037	P73037_SYN	Peptidyl-proly	92,134	20,4	MRGRTHGIIR	17	5	0	0	9,22	21554,986	Synechocystis outer membra Complete prot [1390935826213]
77	P36239	RL19_SYNY3	50S ribosoma	91,567	34,43	MTMNAQAIIN	6	5	0	0	10,82	13786,243	Synechocystis ribosome (GO Complete prot [1390935826213]
78	P72673	Y729_SYNY3	Thylakoid-ass	90,358	45,54	MTSTTPEIEA	8	5	0	0	4,75	10946,61	Synechocystis thylakoid mem Complete prot [1390935826213]
79	P74226	RS10_SYNY3	30S ribosoma	89,709	48,57	MATLQQQKIF	8	5	0	0	9,94	12037,049	Synechocystis ribosome (GO Complete prot [1390935826213]
80	P74008	SAHH_SYNY3	Adenosylhom	89,269	14,82	MVATPVKQK'	10	5	0	0	5,55	46213,906	Synechocystis cytoplasm (GO Complete prot [1390935826213]
81	P73309	RL24_SYNY3	50S ribosoma	89,084	34,78	MTKTKPAPHI	7	5	0	0	10,17	12823,004	Synechocystis ribosome (GO Complete prot [1390935826213]
82	P73303	RL15_SYNY3	50S ribosoma	88,092	35,37	MNLSELSPKI	8	5	0	0	10,52	15194,626	Synechocystis large ribosom Complete prot [1390935826213]
83	P74694	P74694_SYN	Slr0455 protei	87,298	26,95	MKTTLFSVRI	6	4	0	0	5,17	15790,765	Synechocystis Complete prot [1390935826213]
84	Q55013	CY550_SYNY	Cytochrome c	86,307	25,62	MKRFFLVAIA	14	3	0	0	4,89	17884,109	Synechocystis plasma memb 3D-structure; ([1390935826213]
85	P73253	P73253_SYN	Sll1911 protei	86,052	39,06	MAGLFGLFG	7	5	0	0	9,33	14125,969	Synechocystis Complete prot [1390935826213]
86	P72740	DLDH_SYNY3	Dihydrolipoyl	84,55	12,03	MSQDFDYDL	7	5	0	0	5,41	50832,255	Synechocystis plasma memb Cell inner mer [1390935826213]
87	Q55641	Q55641_SYN	Ribonuclease	83,372	22,12	MPSADTLSQ	10	5	0	0	5,98	24403,059	Synechocystis 3'-5' exonucle Complete prot [1390935826213]
88	P54205	RBL_SYNY3	Ribulose bispl	82,886	12,34	MVQAKAGFK	7	6	0	0	5,86	52490,698	Synechocystis magnesium io Calvin cycle; ([1390935826213]
89	P72707	P72707_SYN	Sll0224 protei	82,872	17,11	MKKFACLALS	7	4	0	0	4,28	32829,904	Synechocystis transporter ac Complete prot [1390935826213]
90	Q55511	TIG_SYNY3	Trigger factor	82,711	11,89	MKVTQEKLPI	6	5	0	0	4,31	52610,253	Synechocystis cytoplasm (GO Cell cycle; Ce [1390935826213]
91	P73628	Y1769_SYNY	Probable thyla	82,327	37,14	MQNQVLQAF	10	4	0	0	4,54	11557,781	Synechocystis thylakoid lume Coiled coil; Cc [1390935826213]
92	P74485	P74485_SYN	Sll1863 protei	82,239	33,64	MRITFMSNP1	13	4	0	0	5,47	12353,832	Synechocystis Complete prot [1390935826213]
93	P73299	RS13_SYNY3	30S ribosoma	80,092	33,07	MARIAGVDLF	8	4	0	0	10,75	14571,723	Synechocystis ribosome (GO Complete prot [1390935826213]
94	P24602	BFR_SYNY3	Bacterioferritir	79,782	28,21	MKGKPAVLA(	8	4	0	0	4,78	18330,861	Synechocystis ferric iron bind Complete prot [1390935826213]
95	P73311	RS17_SYNY3	30S ribosoma	79,659	48,15	MAIKERVGIV	4	4	0	0	10,07	9288,824	Synechocystis ribosome (GO Complete prot [1390935826213]
96	P73660	HEM3_SYNY3	Porphobilinog	79,361	13,12	MTVSTSAPTV	12	5	0	0	5,6	34893,948	Synechocystis hydroxymethy Chlorophyll bi [1390935826213]
97	P19569	PSAD_SYNY3	Photosystem I	78,667	30,5	MTELSGQPP	9	4	0	0	9,16	15643,774	Synechocystis photosystem I Complete prot [1390935826213]
98	P73305	RL18_SYNY3	50S ribosoma	78,001	30,83	MKSTRKSAT(	7	4	0	0	11,14	13204,003	Synechocystis ribosome (GO Complete prot [1390935826213]
99	P54691	ILVE_SYNY3	Probable bran	77,517	16,72	MHKFLPIAYF	7	5	0	0	6,12	33950,938	Synechocystis L-isoleucine tr Amino-acid bi [1390935826213]
100	P52208	6PGD_SYNY3	6-phosphoglu	77,49	9,75	MQFNVAIMTI	8	5	0	0	5,16	52873,54	Synechocystis NADP binding Complete prot [1390935826213]
101	Q55531	Q55531_SYN	Sll0301 protei	77,133	26,04	MSSFLVFSW	6	4	0	0	4,77	18611,017	Synechocystis Complete prot [1390935826213]
102	P74729	P74729_SYN	HrEpiB	76,946	17,63	MIMKILITGG(	6	5	0	0	8,07	34951,479	Synechocystis catalytic activi Complete prot [1390935826213]
103	P73306	RL6_SYNY3	50S ribosoma	74,739	18,99	MSRIGKRPIP	8	5	0	0	10,27	19666,721	Synechocystis ribosome (GO Complete prot [1390935826213]
104	P72760	CCMK1_SYN	Carbon dioxid	73,263	32,43	MSIAVGMIET	14	4	0	0	6,74	12101,966	Synechocystis 3D-structure; ([1390935826213]
105	P73463	P73463_SYN	Slr1220 protei	73,072	15,14	MTIDNIRLRN	6	5	0	0	4,21	36009,011	Synechocystis Complete prot [1390935826213]
106	P74390	P74390_SYN	Negative aliph	72,736	9,19	MTNPFGRRK	8	5	0	0	4,86	48359,627	Synechocystis outer membra Complete prot [1390935826213]
107	P36265	NUSG_SYNY3	Transcription a	72,095	22,44	MSFTDDQSP	8	4	0	0	6,08	23415,828	Synechocystis DNA-depende Complete prot [1390935826213]

108	P74185	P74185_SYN	Slr1273 protei	70,269	24,66	MFLTALRSFL	5	4	0	0		6,56	15714,07	Synechocystis Complete prot [1390935826213]
109	P73294	RL13_SYNY3	50S ribosoma	69,72	23,18	MNKTVLPTID	11	4	0	0		10,09	16990,692	Synechocystis ribosome (GO Complete prot [1390935826213]
110	P73600	P73600_SYN	Sll1785 protei	69,016	16,79	MLLKVKLWG	5	4	0	0		5,23	30011,352	Synechocystis outer membra 3D-structure; I [1390935826213]
111	P26527	ATPB_SYNY3	ATP synthase	68,92	11,8	MVAVKEATN\	5	4	0	0		4,93	51733,036	Synechocystis plasma memb ATP synthesis [1390935826213]
112	P73960	LEU3_SYNY3	3-isopropylma	68,414	10,22	MSQTYNVTLI	7	5	0	0		4,72	38667,3	Synechocystis cytoplasm (GO Amino-acid bi [1390935826213]
113	P74070	EFTS_SYNY3	Elongation fac	68,182	19,72	MAEITAQLVK	7	4	0	0		5,5	24230,824	Synechocystis cytoplasm (GO Complete prot [1390935826213]
114	P73213	P73213_SYN	Ssr2857 prote	67,721	37,5	MTIQLTVPTIA	6	3	0	0		4,53	6685,548	Synechocystis copper ion bin 3D-structure; I [1390935826213]
115	P73789	PPI2_SYNY3	Peptidyl-prolyl	67,685	23,39	MMSKVFFDIT	10	4	0	0		5,52	18534,946	Synechocystis peptidyl-prolyl Complete prot [1390935826213]
116	P73335	Y1786_SYN\	Uncharacteriz	67,474	16,86	MHLVDTHVH	6	4	0	0		5,62	29258,318	Synechocystis endodeoxyribo Complete prot [1390935826213]
117	P73527	RISB_SYNY3	6,7-dimethyl-8	67,381	24,39	MTVYEGSFTI	5	4	0	0		6,27	17617,379	Synechocystis riboflavin synt Complete prot [1390935826213]
118	P73875	P73875_SYN	Ssl0467 prote	67,154	40,85	MSTQDKARE	15	3	0	0		4,71	8254,047	Synechocystis Complete prot [1390935826213]
119	Q55513	DAPA_SYNY3	4-hydroxy-tetr	66,849	13,29	MADFVSTSPI	6	4	0	0		5,02	31821,46	Synechocystis cytoplasm (GO Amino-acid bi [1390935826213]
120	P48949	RS21_SYNY3	30S ribosoma	65,411	45	MTQVVVGQN	5	3	0	0		11,78	7341,444	Synechocystis ribosome (GO Complete prot [1390935826213]
121	Q55585	GABD_SYNY3	Probable succ	65,322	9,47	MAIATINPATC	8	5	0	0		5,08	48748,802	Synechocystis oxidoreductas Complete prot [1390935826213]
122	Q6ZEL1	Q6ZEL1_SYN	Slr5119 protei	65,069	12,11	MFDNLCKFL/	8	4	0	3	Q6ZEQ3,P73(	5,6	29502,151	Synechocystis Complete prot [1390935826213]
123	P74551	APCF_SYNY3	Allophycocyar	65,043	23,08	MRDAVTTLIK	4	4	0	0		5,09	18892,452	Synechocystis phycobilisome Bile pigment; I [1390935826213]
124	P73596	P73596_SYN	Sll1307 protei	63,855	18,29	MLKFTFTPLL	4	3	0	0		6,72	17848,738	Synechocystis outer membra Complete prot [1390935826213]
125	Q55648	Q55648_SYI	Sll0314 protei	63,732	13,38	MLVFLTRFTP	6	4	0	0		5,89	35444,109	Synechocystis outer membra Complete prot [1390935826213]
126	P73599	Y1304_SYN\	Uncharacteriz	63,426	13,59	MISSPKIKFG'	4	4	0	0		5,79	32848,47	Synechocystis Complete prot [1390935826213]
127	P72817	Y1654_SYN\	Universal stre	63,315	22,29	MSINCWRSP	7	3	0	0		5,58	16769,378	Synechocystis response to st Complete prot [1390935826213]
128	Q55730	Q55730_SYI	Slr0650 protei	62,991	20,9	MFEDFEQDA	6	4	0	0		4,98	22501,444	Synechocystis Complete prot [1390935826213]
129	P73302	KAD1_SYNY3	Adenylate kina	62,456	18,92	MAKGLIFLGA	5	3	0	0		5,59	20251,24	Synechocystis cytoplasm (GO ATP-binding; I [1390935826213]
130	P73061	URE1_SYNY3	Urease subun	61,918	8,08	MSYRMDRHS	7	5	0	0		5,86	61037,598	Synechocystis cytoplasm (GO Complete prot [1390935826213]
131	P74035	RIMM_SYNY3	Ribosome ma	61,215	16,22	MAEPMTEQC	5	4	0	0		4,89	20715,757	Synechocystis ribosome (GO Chaperone; C [1390935826213]
132	P73310	RL14_SYNY3	50S ribosoma	60,59	26,23	MIQQQTYLN\	5	3	0	0		10,28	13294,538	Synechocystis large ribosoma Complete prot [1390935826213]
133	P72659	PNP_SYNY3	Polyribonucle	59,844	7,52	MQEFDKSISF	5	5	0	0		5,17	77831,425	Synechocystis cytoplasm (GO Complete prot [1390935826213]
134	Q55664	ALF2_SYNY3	Fructose-bisp	59,418	9,75	MALVPMRLLI	4	4	0	0		5,55	38971,984	Synechocystis fructose-bisph Complete prot [1390935826213]
135	Q01903	SUBI_SYNY3	Sulfate-bindin	59,153	11,08	MARSAFGW(	7	4	0	0		4,86	38127,811	Synechocystis outer membra Complete prot [1390935826213]
136	Q55385	RRP3_SYNY3	Probable 30S	59,036	31,25	MTTAEAAST\	7	3	0	0		5,24	12638,271	Synechocystis ribosome (GO 3D-structure; I [1390935826213]
137	Q55765	Q55765_SYI	RNA-binding p	58,009	31,13	MSIRLYVGNL	5	3	0	0		8,6	16616,542	Synechocystis nucleic acid bi Complete prot [1390935826213]
138	P74470	P74470_SYN	Ssl0242 prote	57,158	23,08	MYNPSLRRE	6	3	0	0		3,76	8976,591	Synechocystis Complete prot [1390935826213]
139	Q55113	Q55113_SYI	Slr0431 protei	56,726	14,8	MRPKFFSRR	5	4	0	0		8,92	27011,788	Synechocystis Complete prot [1390935826213]
140	P73283	FABF_SYNY3	3-oxoacyl-[acy	56,561	10,82	MANLEKKRV'	5	4	0	0		5,58	44004,116	Synechocystis beta-ketoacyl- 3D-structure; / [1390935826213]
141	P74795	P74795_SYN	Ssl0352 prote	56,432	44,83	MIFPGATVRV	3	3	0	0		6,12	6577,517	Synechocystis 3D-structure; I [1390935826213]
142	P27320	FER_SYNY3	Ferredoxin-1	56,086	17,53	MASYTVKLIT	44	3	0	0		3,78	10363,281	Synechocystis 2 iron, 2 sulfur 2Fe-2S; 3D-st [1390935826213]
143	P26533	ATPE_SYNY3	ATP synthase	55,886	24,26	MTLTVRVITP	5	2	0	0		5,46	14580,536	Synechocystis plasma memb ATP synthesis [1390935826213]
144	P73609	P73609_SYN	Stage II sporu	55,644	29,63	MAFNIESEIIN	5	3	0	0		4,8	11989,777	Synechocystis regulation of t Complete prot [1390935826213]
145	P73318	RL23_SYNY3	50S ribosoma	53,939	31,68	MSKVIDQRRI	5	3	0	0		10,18	11525,663	Synechocystis ribosome (GO Complete prot [1390935826213]
146	P72606	P72606_SYN	Slr1485 protei	53,407	10,32	MFKFAQIIFL/	6	4	0	0		4,67	37586,756	Synechocystis outer membra Complete prot [1390935826213]
147	P73289	RL25_SYNY3	50S ribosoma	53,244	26,53	MALSIECQQF	8	3	0	0		9,39	11140,952	Synechocystis ribosome (GO Complete prot [1390935826213]
148	P73602	Y1783_SYN\	Uncharacteriz	53,233	17,01	MRADFFLSDI	5	3	0	0		4,95	16822,206	Synechocystis Complete prot [1390935826213]
149	P73066	P73066_SYN	Ycf23 protein	53,13	14,92	MSANLAQLH	3	3	0	0		5,27	25366,097	Synechocystis catalytic activi Complete prot [1390935826213]
150	P32422	PSAC_SYNY3	Photosystem I	53,128	27,16	MSHSVKIYDT	4	3	0	0		6,51	8828,238	Synechocystis photosystem I 4Fe-4S; Com [1390935826213]
151	P74769	P74769_SYN	Ssr1528 prote	52,66	23,4	MANTTKGAD	7	3	0	0		8,05	10118,58	Synechocystis Complete prot [1390935826213]
152	P74308	P74308_SYN	Aldehyde redu	52,166	8,87	MQSFNRINSI	4	4	0	0		5,07	36014,105	Synechocystis oxidoreductas Complete prot [1390935826213]
153	P23350	RL10_SYNY3	50S ribosoma	50,923	18,5	MGRTRENKA	3	3	0	0		8,99	18675,593	Synechocystis ribosome (GO Complete prot [1390935826213]
154	P73201	SYS_SYNY3	Serine--tRNA	50,864	10,7	MLDLKQIREN	7	4	0	0		5,53	48037,921	Synechocystis cytoplasm (GO Aminoacyl-tRI [1390935826213]
155	P74233	P74233_SYN	Slr1160 protei	50,86	19,12	MLQRLVHILA	3	3	0	0		5,68	22261,29	Synechocystis outer membra Complete prot [1390935826213]
156	P73481	YHIT_SYNY3	Uncharacteriz	50,746	22,81	MAEDTIFSKII	7	3	0	0		6,27	12456,532	Synechocystis catalytic activi Complete prot [1390935826213]
157	P73244	P73244_SYN	Slr2025 protei	50,168	17,65	MTMASPTPE	2	2	0	0		4,35	16887,893	Synechocystis Complete prot [1390935826213]
158	P48946	RS18_SYNY3	30S ribosoma	49,358	35,21	MNYYRKRLS	4	3	0	0		11,02	8380,926	Synechocystis ribosome (GO Complete prot [1390935826213]
159	P74341	P74341_SYN	Sll1537 protei	49,236	20,86	MPPQFPLAT\	5	3	0	0		4,75	15937,264	Synechocystis hydrolase acti Complete prot [1390935826213]
160	P73055	Y3122_SYN\	Uncharacteriz	48,828	37,65	MISLDQVKQ(	4	3	0	0		5,8	9045,316	Synechocystis Complete prot [1390935826213]
161	P74135	P74135_SYN	Sll1873 protei	48,622	27,27	MLKKLFGAKI	10	2	0	0		10,09	12905,989	Synechocystis Complete prot [1390935826213]

162	P72985	P72985_SYN	Slr1600 protei	48,485	17,45	MDDTNTTLLI	7	3	0		0	5,07	17129,696	Synechocystis Complete prot [1390935826213]
163	P74061	RPE_SYNY3	Ribulose-phos	48,416	12,61	MSKNIVVAPS	6	3	0		0	5,37	24970,857	Synechocystis metal ion bind 3D-structure; I [1390935826213]
164	P74002	Y1322_SYN	Uncharacteriz	48,067	7,16	MPPTLLLSQL	3	3	0		0	5,48	53129,717	Synechocystis Complete prot [1390935826213]
165	P73204	PYR2_SYNY3	Phycobilisome	47,655	11,36	MTSLVSAQRI	3	3	0		0	9,51	30797,389	Synechocystis phycobilisome 3D-structure; I [1390935826213]
166	P73929	P73929_SYN	Slr2101 protei	47,645	20,28	MKFISSFFAL/	3	3	0		0	4,99	15348,676	Synechocystis Complete prot [1390935826213]
167	P72866	RS15_SYNY3	30S ribosoma	47,542	23,6	MSLTQIRKQE	6	3	0		0	11,16	10373,062	Synechocystis ribosome (GO Complete prot [1390935826213]
168	P72753	UPP_SYNY3	Uracil phosph	47,005	14,35	MASQLRVYV	3	3	0		0	6,2	23637,675	Synechocystis GTP binding (Allosteric enz [1390935826213]
169	P72871	METK_SYNY3	S-adenosylme	46,875	7,75	MRGLKTLSKI	5	3	0		0	5,14	45865,812	Synechocystis cytoplasm (GC ATP-binding; I [1390935826213]
170	Q55332	PSBU_SYNY3	Photosystem I	46,648	16,79	MKFISRLLVA	16	3	0		0	4,53	14245,243	Synechocystis extrinsic to me Complete prot [1390935826213]
171	Q6YRS8	Q6YRS8_SYN	Sll6017 protei	46,52	12,5	MFDNLCKFL/	6	3	2	Q6ZEP8,Q6Z	4	Q6ZEQ3,P72	5,18	29211,736 Synechocystis Complete prot [1390935826213]
172	Q55450	GATC_SYNY3	Glutamyl-tRN/	45,961	17,31	MLDQSQVQK	6	3	0		0		4,25	11615,791 Synechocystis ATP binding ((ATP-binding; I [1390935826213]
173	P54386	DHE4_SYNY3	NADP-specific	45,594	7,71	MAGSLFADA:	4	3	0		0		5,05	47312,712 Synechocystis cytoplasm (GC Complete prot [1390935826213]
174	P73354	HTRA_SYNY3	Putative serin	45,532	9,07	MSAQAVFPIA	4	3	0		0		5,42	47656,158 Synechocystis cell outer men Cell outer mer [1390935826213]
175	P73604	P73604_SYN	Slr1853 protei	45,495	24,78	MSEFKNAVLI	8	3	0		0		4,87	12110,863 Synechocystis peroxiredoxin Complete prot [1390935826213]
176	Q6ZEP2	Q6ZEP2_SYN	Slr5088 protei	45,451	12,61	MKLMVIGASI	5	3	0		0		5,1	26170,277 Synechocystis oxidoreductas Complete prot [1390935826213]
177	P73293	RS9_SYNY3	30S ribosoma	44,86	18,98	MQANDSSNI	6	3	0		0		10,22	15086,339 Synechocystis ribosome (GO Complete prot [1390935826213]
178	P72851	RL28_SYNY3	50S ribosoma	44,107	28,21	MARRCQLTG	3	3	0		0		12,02	8993,49 Synechocystis ribosome (GO Complete prot [1390935826213]
179	P73202	PYS1_SYNY3	Phycobilisome	43,398	30,12	MLGQSSLVG	6	2	0		0		9,8	9322,352 Synechocystis phycobilisome Complete prot [1390935826213]
180	Q55770	Q55770_SYI	Sll0185 protei	43,276	7,64	MAKEDRPSL	4	4	0		0		5,04	46994,054 Synechocystis DNA-depende Complete prot [1390935826213]
181	P73946	P73946_SYN	Slr1506 protei	43,25	5,14	MEKKITLRW(	4	4	0		0		5,21	68847,364 Synechocystis 1-alkyl-2-acet Complete prot [1390935826213]
182	P73328	P73328_SYN	Slr1900 protei	42,843	10,93	MGSYSQLDL	4	3	0		0		4,82	27406,459 Synechocystis Complete prot [1390935826213]
183	Q55447	Q55447_SYI	CheY subfami	42,641	19,33	MGSALVIDDS	5	2	0		0		4,97	13181,153 Synechocystis phosphorelay Complete prot [1390935826213]
184	P73601	P73601_SYN	Sll1784 protei	41,845	10,11	MKTLRLSPLL	4	3	0		0		5,07	29757,788 Synechocystis outer membra Complete prot [1390935826213]
185	P74645	KAIB_SYNY3	Circadian cloc	41,646	23,81	MSPFKKTYVI	3	3	0		0		7,9	11935,026 Synechocystis circadian rhyth 3D-structure; I [1390935826213]
186	P73597	P73597_SYN	Sll1306 protei	41,466	8,06	MQRRDLFKY	4	3	0		0		8,49	38270,621 Synechocystis outer membra Complete prot [1390935826213]
187	P72642	DAPB_SYNY3	4-hydroxy-tetr	41,225	12	MANQDLIPV\	3	3	0		0		5,45	29090,618 Synechocystis cytoplasm (GC Amino-acid bi [1390935826213]
188	P73296	RL17_SYNY3	50S ribosoma	41,108	21,55	MRHRCRVPC	4	3	0		0		11,22	13228,444 Synechocystis ribosome (GO Complete prot [1390935826213]
189	P72776	PDXJ_SYNY3	Pyridoxine 5'-p	40,832	10,74	MLTLGVNIDH	3	3	0		0		5,51	26518,63 Synechocystis cytoplasm (GC Complete prot [1390935826213]
190	Q55953	Q55953_SYI	Sll0781 protei	40,758	15,48	MTTPLVPPFT	5	2	0		0		6,63	18633,936 Synechocystis Complete prot [1390935826213]
191	P73921	P73921_SYN	Sll1979 protei	40,58	18,31	MTESVISPEI	2	2	0		0		5,01	15946,176 Synechocystis Complete prot [1390935826213]
192	P74102	OCP_SYNY3	Orange carote	40,302	10,41	MPFTIDSARC	3	3	0		0		5,03	34658,816 Synechocystis phycobilisome 3D-structure; I [1390935826213]
193	Q55602	Q55602_SYI	NifS protein	40,193	9,33	MERPLYFDNI	3	3	0		0		6,16	41682,432 Synechocystis catalytic activi Complete prot [1390935826213]
194	P74112	P74112_SYN	Ssr3341 prote	40,174	27,14	MSRFDSGLP	4	2	0		0		8,25	7830,983 Synechocystis 3D-structure; I [1390935826213]
195	P72848	HEM6_SYNY3	Coproporphyr	39,766	9,41	MTVSPTTQP(	6	3	0		0		5,65	38937,441 Synechocystis cytoplasm (GC Complete prot [1390935826213]
196	P73298	RS11_SYNY3	30S ribosoma	39,694	21,54	MARPTRKTG	3	2	0		0		11,53	13761,844 Synechocystis ribosome (GO Complete prot [1390935826213]
197	P48959	RL35_SYNY3	50S ribosoma	39,557	26,87	MPKLKTRKA/	4	2	0		0		11,6	7891,402 Synechocystis ribosome (GO Complete prot [1390935826213]
198	Q55629	Y782_SYNY3	Uncharacteriz	38,976	5,52	MVIRSGKTNI	6	3	0		0		5,25	51404,269 Synechocystis oxidoreductas Complete prot [1390935826213]
199	P72781	P72781_SYN	NarL subfamil	38,687	10,99	MGLSLLRPRI	4	3	0		0		5,45	31394,893 Synechocystis DNA binding (Complete prot [1390935826213]
200	P73746	P73746_SYN	Sll0854 protei	38,671	10,39	MAFFKEDPAI	4	3	0		0		5,08	34527,165 Synechocystis Complete prot [1390935826213]
201	P72699	Y230_SYNY3	UPF0045 prot	38,629	17,36	MCQNLGKFE	9	2	0		0		7,72	13305,347 Synechocystis Complete prot [1390935826213]
202	P72798	P72798_SYN	Ssr2998 prote	38,535	36,92	MTIEIGQKVK	5	2	0		0		9,16	7220,373 Synechocystis Complete prot [1390935826213]
203	P36237	RL11_SYNY3	50S ribosoma	38,271	16,31	MAKKVVALIK	6	3	0		0		9,63	14977,503 Synechocystis ribosome (GO Complete prot [1390935826213]
204	P74486	P74486_SYN	Sll1862 protei	38,114	17,48	MGSLQGNYL	3	2	0		0		6,1	15194,154 Synechocystis Complete prot [1390935826213]
205	P73605	P73605_SYN	Slr1854 protei	37,813	13,64	MTTTKGKIGV	4	3	0		0		4,79	22293,203 Synechocystis Complete prot [1390935826213]
206	P77961	GLNA_SYNY3	Glutamine syr	37,72	6,34	MARTPQEVL	5	3	0		0		5,02	53025,941 Synechocystis cytoplasm (GC 3D-structure; / [1390935826213]
207	Q55841	Q55841_SYI	Arabinofuranc	37,408	15,25	MKLLPLLPLF	2	2	0		0		9,39	20485,176 Synechocystis alpha-N-arabi Complete prot [1390935826213]
208	P74421	PGK_SYNY3	Phosphoglyce	37,291	6,73	MLSKQSIANL	3	3	0		0		5,09	41783,949 Synechocystis cytoplasm (GC ATP-binding; I [1390935826213]
209	P73954	Y1513_SYN	Membrane-as	37,171	17,27	MAKPANKLVI	3	2	0		0		7,67	12006,814 Synechocystis plasma memb Cell membran [1390935826213]
210	P73742	P73742_SYN	Sll0858 protei	36,932	13,71	MVTKRSPTG	3	3	0		0		10,65	22600,669 Synechocystis periplasmic sp Complete prot [1390935826213]
211	P73452	NRTA_SYNY3	Nitrate transp	36,664	6,28	MSNFSRSTR	3	2	0		0		5,27	48966,606 Synechocystis plasma memb 3D-structure; I [1390935826213]
212	P74746	P74746_SYN	Slr0600 protei	36,543	11,04	MKLSSKNLD/	3	3	0		0		5,75	35993,138 Synechocystis oxidoreductas Complete prot [1390935826213]
213	P73579	P73579_SYN	Slr0890 protei	36,53	22,9	MGLVYANIEL	2	2	0		0		5,1	14187,587 Synechocystis Complete prot [1390935826213]
214	Q6YRU9	Q6YRU9_SYN	Sll6055 protei	36,311	14,47	MVVNAQQFF	3	3	0		0		5,52	17443,927 Synechocystis Complete prot [1390935826213]
215	P74447	P74447_SYN	Ferredoxin	36,308	14,52	MTMPPLWN(	4	3	0		0		5,02	20624,746 Synechocystis electron carrie Complete prot [1390935826213]

216	Q55356	PSB28_SYNY	Photosystem	36,308	16,07	MAEIQFSKGV	7	2	0	0	5	12590,31	Synechocystis	photosystem I 3D-structure; [1390935826213]
217	P74386	URE2_SYNY3	Urease subun	36,29	20	MATMIPGEIT	3	2	0	0	4,76	11381,794	Synechocystis	cytoplasm (GC Complete prot [1390935826213]
218	P73424	P73424_SYN	Slr1540 protei	36,176	9,32	MRILIMGGTR	3	3	0	0	6,26	35092,64	Synechocystis	Complete prot [1390935826213]
219	P73637	P73637_SYN	5-oxo-1,2,5-tri	35,964	10,51	MVQRYVRIQA	2	2	0	0	4,4	30049,837	Synechocystis	catalytic activi Complete prot [1390935826213]
220	P74162	P74162_SYN	Sll1380 protei	35,797	12,15	MTTATHLLKL	2	2	0	0	9,34	23884,179	Synechocystis	outer membra Complete prot [1390935826213]
221	Q55541	Q55541_SYN	Slr0333 protei	35,706	16,04	MTLDKLGSAI	4	2	0	0	9,06	11406,143	Synechocystis	Complete prot [1390935826213]
222	P74396	P74396_SYN	Slr0280 protei	35,605	5,08	MALGDLLKM	3	3	0	0	9,85	65405,879	Synechocystis	Complete prot [1390935826213]
223	P73316	RS19_SYNY3	30S ribosoma	35,01	29,35	MGRSLKKGP	4	2	0	0	11,42	10290,017	Synechocystis	small ribosom Complete prot [1390935826213]
224	P73704	P73704_SYN	General secre	34,815	19,05	MASNFKFKLI	4	2	0	0	4,78	17573,691	Synechocystis	type II protein Complete prot [1390935826213]
225	Q55247	GLNB_SYNY3	Nitrogen regu	34,508	17,86	MKKVEAIIRP	4	2	0	0	7,95	12397,38	Synechocystis	enzyme regula 3D-structure; [1390935826213]
226	Q6ZEU7	Q6ZEU7_SYN	Sll5033 protei	34,42	12,29	MKSILSFIKS\	5	2	0	0	4,89	19892,094	Synechocystis	Complete prot [1390935826213]
227	Q55171	Q55171_SYN	Slr0476 protei	34,253	15,44	MTEEQQGPN	5	2	0	0	8,11	14394,113	Synechocystis	Complete prot [1390935826213]
228	P73654	P73654_SYN	Ssl3364 prote	33,918	27,03	MSNIQEKIEQ	6	2	0	0	4,14	8294,914	Synechocystis	Complete prot [1390935826213]
229	P74352	RPOZ_SYNY3	DNA-directed	33,89	28,95	MTKRSNLDS	3	2	0	0	5,77	8736,766	Synechocystis	DNA binding (Complete prot [1390935826213]
230	P72775	P72775_SYN	Alanine dehyc	33,52	6,39	MEIGVPKEIK	3	3	0	0	5,4	38267,216	Synechocystis	alanine dehyd Complete prot [1390935826213]
231	P73796	URE3_SYNY3	Urease subun	33,017	16	MQLSPQEKD	3	2	0	0	5,41	11055,78	Synechocystis	cytoplasm (GC Complete prot [1390935826213]
232	P74428	P74428_SYN	Slr0398 protei	32,921	13,66	MTTIMVNLAI	3	2	0	0	4,66	18079,558	Synechocystis	Complete prot [1390935826213]
233	P73495	P73495_SYN	Naphthoate s	32,791	9,82	MDWHIAKHY	4	3	0	0	6,6	30307,536	Synechocystis	1,4-dihydroxy- 3D-structure; [1390935826213]
234	P73406	CCMK3_SYN	Carbon dioxid	32,64	18,45	MPQAVGVIQ	2	2	0	0	6,56	11002,866	Synechocystis	Complete prot [1390935826213]
235	P72759	CCML_SYNY3	Carbon dioxid	32,622	18	MQLAKVLGT	2	2	0	0	9,52	10638,205	Synechocystis	3D-structure; [1390935826213]
236	P74591	AROE_SYNY3	Shikimate deh	32,611	11,03	MPSITGKTKL	2	2	0	0	7,1	31099,118	Synechocystis	NADP binding Amino-acid bi [1390935826213]
237	Q55561	Q55561_SYN	Sll0167 protei	32,344	14,02	MFTKIRDLAS	2	2	0	0	4	17865,116	Synechocystis	Complete prot [1390935826213]
238	P26287	CYF_SYNY3	Apocytochrom	32,002	6,71	MRNPDTLGL	4	3	0	0	5,17	35230,565	Synechocystis	cytochrome b(Complete prot [1390935826213]
239	P74426	P74426_SYN	Sll0359 protei	31,826	11,61	MLYFILCRLL	2	2	0	0	9,32	17218,594	Synechocystis	Complete prot [1390935826213]
240	P73824	P73824_SYN	Glutathione pe	31,592	14,94	MPLPTSLTTL	5	2	0	0	4,72	16645,96	Synechocystis	glutathione pe Complete prot [1390935826213]
241	P73439	P73439_SYN	Sll1461 protei	31,567	5,91	MVIAPHRTIIY	2	2	0	0	4,63	25524,816	Synechocystis	Complete prot [1390935826213]
242	P73126	Y997_SYNY3	Probable thyla	31,264	6,49	MAPYQSFHIC	2	2	0	0	4,32	37962,856	Synechocystis	thylakoid lume Complete prot [1390935826213]
243	P20804	ACP_SYNY3	Acyl carrier pr	31,185	19,48	MNQEIFEKVK	12	2	0	0	3,98	8589,581	Synechocystis	cytoplasm (GC Complete prot [1390935826213]
244	P74138	P74138_SYN	CheY subfami	30,968	13,01	MESMAKVLL\	2	2	0	0	5,32	13658,94	Synechocystis	phosphorelay Complete prot [1390935826213]
245	P36236	RL1_SYNY3	50S ribosoma	30,69	10,08	MTKKLSKRM	2	2	0	0	8,98	25851,796	Synechocystis	large ribosoma Complete prot [1390935826213]
246	Q55120	Q55120_SYN	Biotin carboxy	30,675	13,64	MAINFTELRE	2	2	0	0	4,34	16312,736	Synechocystis	acetyl-CoA ca Biotin; Compl [1390935826213]
247	P72890	P72890_SYN	Slr1612 protei	30,662	8,39	MSAITAQELE	2	2	0	0	4,68	33045,19	Synechocystis	Complete prot [1390935826213]
248	Q55744	Q55744_SYN	Sll0381 protei	30,601	9,52	MPPRKSMRS	2	2	0	0	4,76	30984,932	Synechocystis	Complete prot [1390935826213]
249	Q55149	Q55149_SYN	Slr0058 protei	30,482	12,68	MVFAMVDEK	2	2	0	0	5,07	16485,639	Synechocystis	Complete prot [1390935826213]
250	P74324	F16PA_SYNY	Fructose-1,6-l	30,143	7,49	MTVSEIHIPN	2	2	0	0	5,5	38262,298	Synechocystis	cytoplasm (GC Calvin cycle; [1390935826213]
251	P74060	Y821_SYNY3	Putative sulfu	30,141	16,51	MVSPALPRLS	5	2	0	0	4,71	11694,429	Synechocystis	Complete prot [1390935826213]
252	P73107	P73107_SYN	Sll1837 protei	29,674	15,17	MVNKGLWSA	5	2	0	0	4,8	15467,443	Synechocystis	outer membra Complete prot [1390935826213]
253	P72914	P72914_SYN	Ssr1766 prote	29,654	21,05	MSNLPSVQA	3	2	0	0	10,14	8670,288	Synechocystis	hydrolase acti Complete prot [1390935826213]
254	P74011	P74011_SYN	Sll1233 protei	29,634	8,5	MATYRIEHHT	2	2	0	0	5,22	33204,735	Synechocystis	Complete prot [1390935826213]
255	P72845	P72845_SYN	Sll1188 protei	29,471	12,2	MKKEINYWV	3	2	0	0	5	18816,275	Synechocystis	Complete prot [1390935826213]
256	P74367	PS11_SYNY3	Photosystem	29,38	14,93	MSFLKNQLSI	2	2	0	0	9,56	14785,839	Synechocystis	plasma membr 3D-structure; [1390935826213]
257	P73203	PYR1_SYNY3	Phycobilisome	29,377	7,56	MAITTAASRL	2	2	0	0	9,45	32520,679	Synechocystis	phycobilisome 3D-structure; [1390935826213]
258	Q55766	RPIA_SYNY3	Ribose-5-phos	29,086	9,36	MAELDAANLI	3	2	0	0	4,99	24752,741	Synechocystis	ribose-5-phos Complete prot [1390935826213]
259	P54206	RBS_SYNY3	Ribulose bispl	28,952	15,04	MKTLPKERR'	4	2	0	0	5,67	13239,043	Synechocystis	monooxygena Calvin cycle; [1390935826213]
260	P73410	CYSK_SYNY3	Cysteine syntl	28,865	7,37	MKIASNITELI	3	2	0	0	5,87	33173,448	Synechocystis	cysteine synth Amino-acid bi [1390935826213]
261	P55038	GLTS_SYNY3	Ferredoxin-de	28,736	2,31	MSFQYPLLAI	3	3	0	0	5,53	169499,084	Synechocystis	3 iron, 4 sulfur 3D-structure; [1390935826213]
262	Q55233	DRGA_SYNY3	Protein DrgA	28,676	9,52	MDTFDAIYQF	8	2	0	0	9,4	23703,443	Synechocystis	oxidoreductas Complete prot [1390935826213]
263	P74789	P74789_SYN	Sll0319 protei	28,464	7,07	MNWKFFPHF	2	2	0	0	4,94	32315,724	Synechocystis	outer membra Complete prot [1390935826213]
264	P74510	P74510_SYN	Dihydrolipoam	28,236	5,54	MIYDIFMPAL	2	2	0	0	6,04	44898,455	Synechocystis	transferase ac Acyltransferas [1390935826213]
265	P74220	Y1534_SYN	Putative carbo	28,231	8,97	MSSLPKNFL(	3	2	0	0	5,38	32906,437	Synechocystis	carboxypeptid Carboxypeptic [1390935826213]
266	P72807	P72807_SYN	Sll1663 protei	28,113	10,45	MSDSLTAIKA	3	2	0	0	4,54	24104,385	Synechocystis	Complete prot [1390935826213]
267	Q55734	Q55734_SYN	Sll0395 protei	28,045	8,49	MTLNLYFLRI	2	2	0	0	6,24	23751,95	Synechocystis	Complete prot [1390935826213]
268	P73045	P73045_SYN	Slr1767 protei	27,97	18,75	MNIWVDAQL	2	2	0	0	4,63	12464,317	Synechocystis	Complete prot [1390935826213]
269	Q55387	Q55387_SYN	Periplasmic bi	27,961	5,95	MNNLVGDFF	3	2	0	0	4,38	46672,172	Synechocystis	outer membra Complete prot [1390935826213]

270	P73056	YC64L_SYNY	Uncharacteriz	27,306	14,02	MNPETKARIE	4	2	0	0	4,53	11939,839	Synechocystis 2 iron, 2 sulfur 2Fe-2S; Com [1390935826213]
271	P73594	Y1409_SYN	Uncharacteriz	27,136	6,75	MRIFPVFLLTI	2	2	0	0	5,78	35759,527	Synechocystis outer membra Complete prot [1390935826213]
272	P48957	RL20_SYNY3	50S ribosoma	27,026	12,82	MTRVKRGNV	3	2	0	0	12,11	13553,122	Synechocystis ribosome (GO Complete prot [1390935826213]
273	P73093	P73093_SYN	Phycobilisome	26,997	8,09	MRVEGYEIGS	2	2	0	0	9,38	27392,019	Synechocystis phycobilisome Complete prot [1390935826213]
274	Q55547	Q55547_SYI	Sll0293 protei	26,959	9,41	MPNIRPLTAS	3	2	0	0	4,91	18583,19	Synechocystis Complete prot [1390935826213]
275	Q55484	DCDA_SYNY	Diaminopimel	26,891	5,97	MLSTEMPLP	2	2	0	0	5,35	50847,936	Synechocystis diaminopimela Amino-acid bi [1390935826213]
276	P74563	P74563_SYN	Sll0630 protei	26,736	13,1	MGYCRFLPT	3	2	0	0	6,58	16468,238	Synechocystis Complete prot [1390935826213]
277	Q55862	Q55862_SYI	Sll0588 protei	26,616	9,8	MSLFPLLTALI	5	2	0	0	6,21	16516,086	Synechocystis Complete prot [1390935826213]
278	P73130	P73130_SYN	Sll0995 protei	26,572	8,23	MTVALDREIV	2	2	0	0	4,7	26229,554	Synechocystis Complete prot [1390935826213]
279	P74267	RL27_SYNY3	50S ribosoma	26,512	10,34	MAHKKGTGS	3	2	0	0	11,47	9448,664	Synechocystis ribosome (GO Complete prot [1390935826213]
280	P74371	CPXT_SYNY3	Chromophore	26,438	8,67	MSHSTDLSAI	3	2	0	0	5,29	22578,511	Synechocystis lyase activity (Complete prot [1390935826213]
281	P74230	RS12_SYNY3	30S ribosoma	26,427	12,7	MPTIQQLIRS	4	2	0	0	11,37	14176,627	Synechocystis small ribosom Complete prot [1390935826213]
282	P73180	P73180_SYN	Slr1391 protei	26,28	8,33	MSDPRTIYFC	3	2	0	0	7,68	16806,079	Synechocystis Complete prot [1390935826213]
283	Q55887	Q55887_SYI	Slr0111 protei	26,194	10,4	MARKSLSDLI	2	2	0	0	4,86	19376,955	Synechocystis Complete prot [1390935826213]
284	P73807	HIS8_SYNY3	Histidinol-pho	25,997	5,44	MVSIRPSVRI	2	2	0	0	5,01	38702,127	Synechocystis histidinol-phos Amino-acid bi [1390935826213]
285	P48944	RS14_SYNY3	30S ribosoma	25,984	14	MAKKSMIERI	4	2	0	0	10,94	11853,651	Synechocystis ribosome (GO Complete prot [1390935826213]
286	Q6ZEQ0	Q6ZEQ0_SYN	Non-heme chl	25,929	9,85	MSTITTKDGT	2	2	0	0	5,43	30077,979	Synechocystis peroxidase ac Complete prot [1390935826213]
287	Q59994	TPIS_SYNY3	Triosephosph	25,798	8,68	MRKIIIAGNW	2	2	0	0	5,17	26158,648	Synechocystis cytoplasm (GO Complete prot [1390935826213]
288	P74741	PUR9_SYNY3	Bifunctional p	25,716	4,11	MARLALLSVS	3	2	0	0	5,29	54564,872	Synechocystis IMP cyclohydr Complete prot [1390935826213]
289	P73133	ARGD_SYNY	Acetylornithin	25,661	4,43	MTYSPVVES	3	2	0	0	5,17	46567,179	Synechocystis cytoplasm (GO Amino-acid bi [1390935826213]
290	Q55758	PYRR_SYNY3	Bifunctional p	25,581	10,11	MAAQIIEILSP	2	2	0	0	9,36	19946,349	Synechocystis uracil phosphc Complete prot [1390935826213]
291	Q6ZEQ5	Q6ZEQ5_SYN	Sll5075 protei	25,428	12,23	MKEDFGFTH	2	2	0	0	5,1	15621,719	Synechocystis Complete prot [1390935826213]
292	P73510	P73510_SYN	Sll1358 protei	25,419	5,33	MVNSVIGWLI	2	2	0	0	5,93	43150,995	Synechocystis outer membra 3D-structure; I [1390935826213]
293	P73732	P73732_SYN	Extracellular s	25,341	4,3	MLLNLPATVK	2	2	0	0	5,06	63606,86	Synechocystis transporter ac Complete prot [1390935826213]
294	P73426	ISPF_SYNY3	2-C-methyl-D-	25,306	12,42	MTALRIGNGY	2	2	0	0	6,22	17411,088	Synechocystis 2-C-methyl-D- Complete prot [1390935826213]
295	P73862	P73862_SYN	Rubisco operc	25,176	8,23	MQATLHQLK'	2	2	0	0	8,82	35696,25	Synechocystis DNA binding (Complete prot [1390935826213]
296	P73736	P73736_SYN	N-acetylmurar	25,166	3,7	MSRLPGFALT	2	2	0	0	9,9	70011,438	Synechocystis outer membra Complete prot [1390935826213]
297	P74338	NDHM_SYNY	NAD(P)H-quir	25,159	14,88	MLVKSTTRHV	2	2	0	0	4,64	14077,676	Synechocystis thylakoid mem Complete prot [1390935826213]
298	P74500	P74500_SYN	Slr1940 protei	25,059	4,77	MPQFPMGFC	2	2	0	0	4,8	51261,186	Synechocystis outer membra Complete prot [1390935826213]
299	Q55426	Q55426_SYI	Slr0841 protei	24,836	6,87	MRLNFSRTLI	3	2	0	0	4,92	31566,706	Synechocystis outer membra Complete prot [1390935826213]
300	P74142	RS1B_SYNY3	30S ribosoma	24,639	6,89	MPSSSNSAA	3	2	0	0	5,19	33794,99	Synechocystis ribosome (GO Complete prot [1390935826213]
301	Q55550	Q55550_SYI	Slr0169 protei	24,628	9,39	MGIELRSYVV	3	2	0	0	7,76	22745,132	Synechocystis Complete prot [1390935826213]
302	P74113	P74113_SYN	Slr1970 protei	24,529	9,25	MSIQEIFTKAI	2	2	0	0	4,5	19451,225	Synechocystis Complete prot [1390935826213]
303	P73636	RS6_SYNY3	30S ribosoma	24,295	11,5	MLVNSYELM'	2	2	0	0	8,92	13237,269	Synechocystis ribosome (GO Complete prot [1390935826213]
304	P74375	P74375_SYN	Slr0442 protei	24,245	3,6	MNTRFFLNFI	3	2	0	0	5,11	62907,721	Synechocystis Complete prot [1390935826213]
305	P73031	P73031_SYN	Ssl1918 prote	24,233	14,43	MLFTGTAMEI	3	2	0	0	5,1	11331,812	Synechocystis Complete prot [1390935826213]
306	P73053	THYX_SYNY3	Thymidylate s	24,097	9,77	MDVRFISLTK	2	2	0	0	8,19	25066,625	Synechocystis flavin adenine Complete prot [1390935826213]
307	Q55671	Q55671_SYI	Slr0013 protei	23,937	16	MKLIDSRGRI	2	2	0	0	9,3	18689,774	Synechocystis Complete prot [1390935826213]
308	P73145	SSBT_SYNY3	Thylakoid-ass	23,746	10,85	MNSFVLMAT'	3	2	0	0	4,76	14408,305	Synechocystis thylakoid mem Complete prot [1390935826213]
309	P72952	P72952_SYN	Sll0645 protei	23,609	7,84	MSNRPGKLL	2	2	0	0	5,11	27096,567	Synechocystis Complete prot [1390935826213]
310	P73730	P73730_SYN	Sll1620 protei	23,585	9,49	MGALLAVLLS	4	2	0	0	4,71	17177,557	Synechocystis Complete prot [1390935826213]
311	P74438	PYRC_SYNY3	Dihydroorotas	23,176	5,26	MEKLTITRPD	3	2	0	0	6,46	38126,957	Synechocystis dihydroorotase Complete prot [1390935826213]
312	P73595	Y1410_SYN	Uncharacteriz	22,989	6,59	MKHKFLVSFL	3	2	0	0	4,95	35989,824	Synechocystis outer membra Complete prot [1390935826213]
313	Q55167	Y461_SYNY3	Uncharacteriz	22,954	5,71	MTSDDAAAG	2	2	0	0	5,68	45772,305	Synechocystis cytoplasm (GO Complete prot [1390935826213]
314	Q55852	Q55852_SYI	Sll0596 protei	22,178	6,15	MQLHLSTWP	2	2	0	0	5,79	26325,429	Synechocystis Complete prot [1390935826213]
315	P73128	P73128_SYN	Sulfolipid bios	21,712	5,74	MRALVIGGDC	2	2	0	0	5,72	43202,211	Synechocystis catalytic activi Complete prot [1390935826213]
316	P73297	RPOA_SYNY3	DNA-directed	21,675	6,69	MAQFQIECVI	2	2	0	0	4,76	35003,781	Synechocystis DNA binding (Complete prot [1390935826213]
317	Q6ZEQ1	Q6ZEQ1_SYN	Sll5079 protei	21,67	6,99	MILVSLYPFR'	2	2	0	0	6,13	28689,691	Synechocystis oxidoreductas Complete prot [1390935826213]
318	P72802	P72802_SYN	Mitochondrial	21,628	5,64	MLILKNSMAS	2	2	0	0	5,21	29889,279	Synechocystis outer membra Complete prot [1390935826213]
319	P72805	P72805_SYN	Sll1665 protei	21,351	3,4	METLSLVLAV	2	2	0	0	3,44	63607,329	Synechocystis Complete prot [1390935826213]
320	P73222	P73222_SYN	Slr2005 protei	21,333	6,9	MKRRKFIRTA	2	2	0	0	9,84	27975,188	Synechocystis outer membra Complete prot [1390935826213]
321	P74656	P74656_SYN	Sll1549 protei	20,909	5,88	MKWANRLLL	5	2	0	0	4,98	26624,223	Synechocystis outer membra Complete prot [1390935826213]
322	P73679	P73679_SYN	Isopenicillin N	20,893	4,49	MADPVNLIPI	3	2	0	0	5,63	43999,011	Synechocystis catalytic activi Complete prot [1390935826213]
323	Q55564	Q55564_SYI	Sll0162 protei	20,76	7,73	MGLFDDVGR	3	2	0	0	5,9	22688,343	Synechocystis Complete prot [1390935826213]

324	P74507	GPMI_SYNY3	2,3-bisphosph	20,597	3,01	MAEAPIAPVV	3	2	0	0	5,6	57981,995	Synechocystis cytoplasm (GC Complete prot [1390935826213]
325	P80507	IPYR_SYNY3	Inorganic pyrc	20,59	6,51	MDLSRIPAQF	4	2	0	0	4,77	19087,978	Synechocystis cytoplasm (GC Complete prot [1390935826213]
326	P27179	ATPA_SYNY3	ATP synthase	20,291	4,37	MVSIRPDEIS	2	2	0	0	5,01	53965,584	Synechocystis plasma memb ATP synthesis [1390935826213]
327	Q55746	LPXA_SYNY3	Acyl-[acyl-car	20,082	5,8	MLTDNRLGEA	2	2	0	0	8,44	30002,25	Synechocystis cytoplasm (GC Acyltransferas [1390935826213]
328	Q6ZEI6	Q6ZEI6_SYN	Slr7012 protei	19,867	5,47	MLDSLKSQF(	2	2	0	0	6,1	36485,193	Synechocystis Complete prot [1390935826213]
329	Q55146	Q55146_SYI	Sll0064 protei	19,468	5,45	MLLKSAFTWI	2	2	0	0	4,4	30177,54	Synechocystis outer membra Complete prot [1390935826213]
330	P74598	Y1491_SYNY	Uncharacteriz	19,281	4,02	MNNYFPRLK	3	2	0	0	5,47	37347,023	Synechocystis outer membra Complete prot [1390935826213]
331	P73411	G6PD_SYNY3	Glucose-6-ph	19,273	3,73	MVTLLENPFF	2	2	0	0	7,07	57886,135	Synechocystis glucose-6-phc Carbohydrate [1390935826213]
332	P73098	DNAK3_SYN	Chaperone pr	19,055	1,95	MGKVVGIDL(	5	2	0	0	5,16	86030,118	Synechocystis ATP binding (A ATP-binding; A [1390935826213]
333	Q55544	APCE_SYNY3	Phycobiliprote	18,782	3,35	MSVKASGGS	2	2	0	0	9,29	100295,902	Synechocystis phycobilisome 3D-structure; I [1390935826213]
334	P77962	GLYA_SYNY3	Serine hydrox	17,48	3,28	MNQTNLDFL/	2	2	0	0	5,97	46259,743	Synechocystis cytoplasm (GC Amino-acid bi [1390935826213]
335	Q55469	MURE_SYNY3	UDP-N-acetyl	17,287	3,56	MVKLGQLLA	3	2	0	0	5,25	54601,124	Synechocystis cytoplasm (GC ATP-binding; A [1390935826213]

8.2 Deliverable 7.2

Partner 6 – The University of Sheffield

Deliverable D7.2

Intermediate stage identification of bottlenecks based on comprehensive –omics and flux analysis of: -omics and metabolite data together with assembled metabolic network for the best identified engineered system to date. Suggestions for improvements to system to potentially increase H2 production.

1. Introduction

In the USFD **18 month summary report**, a number of developments that had taken place between the completion of **D7.1** and the report were highlighted. To maintain and reflect continuity of progress through the project, this report begins by discussing how these achievements were built on, resulting in dynamic restructuring of our approach to achieving the longer-term objectives, tasks and deliverables. The main body of the report starts with detailed description of the in-house developments we have made to our proteomic pipeline (contributing to **O7.1, Task 7.2, Task 7.3** and **M7.1**), and then discusses each of our partner interactions that have gone towards completing this deliverable. Finally, the report concludes by highlighting the recommendations from our observations and improvements to the system for both improved bioengineering and hydrogen production.

In the **18 month summary report**, it was highlighted that the exogenous hydrogenase mutant in the hox background developed by UU would first be analysed in non-hydrogen producing conditions, to assess background effects that the transformation process might have on cell functional capacity. Investigation by UU on the mutant demonstrated that it was unable to produce hydrogen in a stable environment, eventually leading to their pioneering work on bicistronic design (BCDs) and the synthetic active site in *Synechocystis*. The generation of the non-native hydrogenase mutant was originally intended to provide the basis for the **D7.2** on suggestions for increased H_2 production, so while this was ongoing, our focus was redirected on insights gained into engineered system modifications to and analyses. This altered focus was designed to address concerns that the design may not be realised in time to carry out all the required omics investigations required for the USFD to complete the 'best engineered system' deliverables required from **WP7**. A number of other avenues were investigated, culminating in additional collaborations with both IBMC and RUB (**D7.2** - see below).

The **18 month summary report** suggested that a metabolic finger-print analysis under the Burrows media conditions would be carried out, however, feedback from the mid-term review meeting encouraged focus on CyanoFactory partner samples for the future deliverables, rather than samples generated at USFD. In addition, a detailed review of the approach indicated that, at this time, it would not directly contribute towards our deliverable requirement for measuring the metabolic flux. Through discussions with UPVLC, it became apparent that the original plan to run flux analysis on media variants under hydrogen-producing conditions was practically infeasible. The major limitation to this was that *Synechocystis* does not fix carbon when it is producing hydrogen, due to increased oxygen levels within the cells, and so monitoring an idealised state of carbon uptake would best be performed on either the best-engineered system available or the best bioreactor design. The high cost of labelled bicarbonate needed for each experiment made running the experiment in the large-scale photobioreactor impossible, and so a pragmatic decision was taken to instead limit analysis to the most optimal and readily available strain, the *olive* mutant from RUB, using the best practical bioreactor produced by KSD (more details of this experiment in **D7.3**). This provides a more scientifically useful outcome as a consequence.

In addition to the omic-level bottlenecks for features like carbon flow, an operational 'bottleneck' that was identified during the project was the need to understand how partners can transfer knowledge more effectively. Celso Gomes (a social sciences doctoral candidate from USFD) has

been interviewing CyanoFactory partners throughout the project and collecting data on how knowledge transfer takes place within EU consortia. UPVLC took steps to address this issue during the 24 month meeting by preparing a separate room for the early career members of the consortium to discuss the project collectively, whilst the PIs dealt with management matters. Celso's findings will be published in his doctoral thesis, which is expected to be completed in 2016.

2 Proteomics pipeline improvements

The findings in this section contribute collectively towards **O7.1**, the delivery of quantitative proteomics data together with recommendations for forward engineering. They achieve this by facilitating the work carried out in **Task 7.2** where new chassis-circuits have been designed by **UU** and **IBMC** and are analysed with proteomic methods; and **Task 7.3** where systems-level effects are analysed using quantitative proteomics. Clearly improved protein identification and quantification on both a localised and systems level analysis is key to generating the high quality data required for computational modelling work.

2.1 Upgrades to the standard lab protocol

As mentioned previously at consortium meetings, in periods where there was down-time between omic analyses of consortium partner biomass, USFD focused on improving our in-house proteomic pipeline. This has resulted in a significant improvement for *Synechocystis* related investigations, with the original capacity (**D7.1**) to confidently quantify 200 - 345 proteins with 2 or more unique peptides in *Synechocystis* (using the techniques described in the project BIOMODULARH2 – NEST - 043340) being improved by 6 to 8 fold to over 1850 confident quantifications from a single run. This number of quantifications is close to the practical maximum number of observable proteins (discussed in D 7.3). This was achieved by assessing the latest developments in the literature and highlighting the limitations that were often observed in production strain proteomics. Of particular note from this analysis were:

- A catalogue of techniques for cell disruption linked to the number of downstream identifications.
    - The processing method used for cell samples received by USFD has been updated to a combination of glass bead beating and sonication.
- An assessment of peptide liquid chromatography fractionation techniques
    - The USFD standard HILIC separation was found to be less effective compared to a porous graphitic carbon hypercarb column. Whilst requiring more cleaning cycles per use, it demonstrates a higher resolution of peptides, which translates to higher numbers of protein identified and quantified.
- Improved gel-free protein quantification methods
    - The highly chromophoric nature of many proteins associated with cyanobacteria causes disparities between measured protein concentrations from standard colourimetric protein assays, such as Bradford quantification. This is particularly notable when a change to the system produces a visible change in the cells, which occurs during nitrogen starvation conditions crucial to hydrogen production in *Synechocystis*. Using a lesser-known amino-acid independent UV method found in the literature:

      $$\mu g.ml^{-1} = 183[230_{nm}] - 75.8[260_{nm}]$$ **(Kalb et al, 1977)**

      This formula provides much more accurate assessment of protein that was less sensitive to these changes is now performed as standard practice. This improvement is important for quantification and identification purposes; as whilst global uneven loading on quantification can be corrected mathematically using techniques like

median correction, doing so makes assumptions about the data and can skew observations – particularly for very low and high abundance proteins.

- Standardised cell pellet requirements for proteomic samples
    - Following a case where insufficient cell pellet material was available for analysis from a CyanoFactory sample, USFD issued a standardised request for biological material for proteomic analysis. Based on improved protein quantification, it was found that the cells should have biomass at least equal to the equivalent of 50 ml volume with an OD_{730} value of 0.7.

Some of these findings were disseminated in an article in 'Trends in Biotechnology', featuring a focus on *Synechocystis* and its pioneering role in the field as a cyanobacterial cell factory candidate (**Landels et al, 2015**). The remainder will be published this year in the doctoral thesis of Andrew Landels (a chemical and biological engineering student from USFD), which is expected to be completed during 2016.

2.2 Upgrades to the standard data analytical techniques

Part of the improvements we made to the pipeline came from collaboration with partners UPVLC. As mentioned previously, we were in close communication with UPVLC about both our metabolic flux work (discussed in **D7.3**) and improving the ease of UPVLC to utilise our proteomic data for their modelling work. It became apparent that the data USFD were producing was limited in terms of application to computational modelling. Part of this was due to a limitation in the amount of information available from the USFD proteomic data processing pipeline; which whilst robust in generating proteomic quantifications and highlighting significant changes, was unable to provide comprehensive information on other details in a compact and efficient manner for ease of data integration. A number of different proteomic standard software packages were trialled against our USFD Phenyx server, including Mascot, Peaks, X!Tandem and MaxQuant.

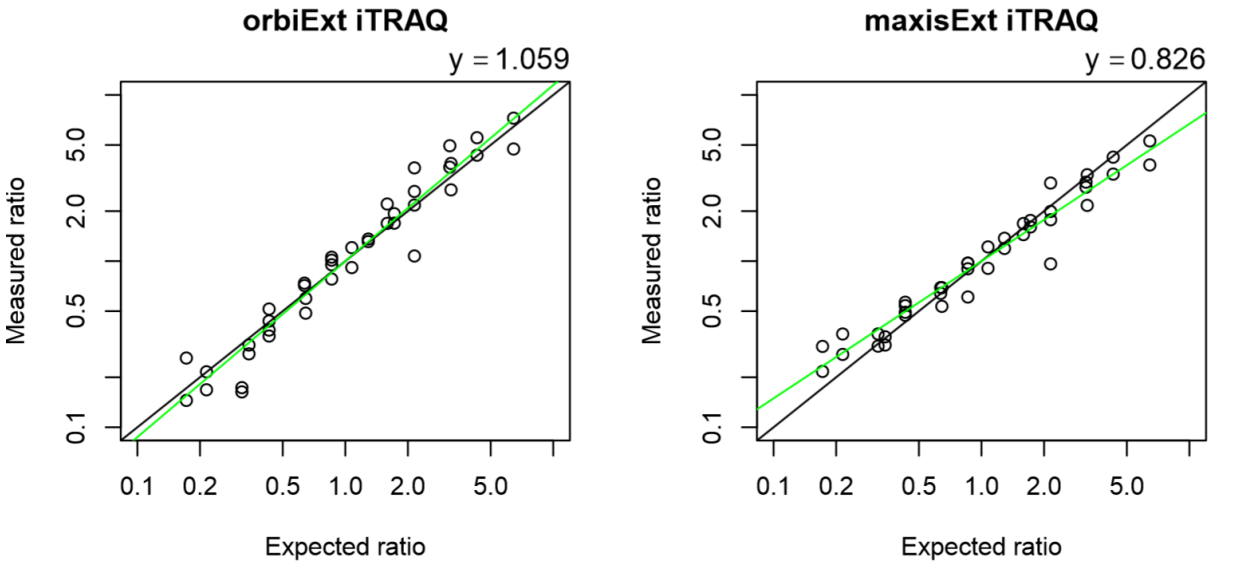
MaxQuant was selected as the best candidate, due to the combination of it being open-source software, the level of information it generated from the proteomic sample, the active development community it has and the efficient files it produces (with the data being stored in a number of compressible flat text files). Once this decision was made, key details concerning the level of the information that would now be available to other partners were disseminated to UPVLC and UM (appendix). Following this change, a number of parsing scripts were also written, to maintain compatibility between the previous data pipeline and our new data pipeline; as a result, all analyses performed by USFD on proteomic data throughout the CyanoFactory project have utilised consistent statistical techniques and generated SOPs.

2.3 Investigation into optimum isobaric tagging quantification protocols for proteome profiling

Finally on the proteomic pipeline development front, an analysis comparing the use of iTRAQ (isobaric tags for relative and absolute quantification) tagging reagents to TMT (tandem mass tags) tagging reagents for use in *Synechocystis* has been carried out. This analysis also presented the opportunity to benchmark the Q-Exactive (QE) mass spectrometer (a new addition to the USFD mass spectrometry facility) against the maXis (the best performing instrument prior to purchase of the QE) on the *Synechocystis* proteome analysis simultaneously.

Previous experimental work in UFSD on quantification analysis has used a mix of proteins at the same concentration, either protein standards or whole cell lysate, which is then diluted to a known concentration range and measured for fold change (Ow et al., 2009). One issue with this design is that it is not representative of the typical proteomic sample for purposes of CyanoFactory. It also means that typical methods used for correcting the data cannot be applied. We addressed these constraints by using pseudo-complex mixture of known standards, both in isolation and spiked into a *Synechocystis* proteome background at biologically relevant concentrations. The balanced design means that the data can be treated just like a normal proteomic sample, with a controlled set of proteins acting as a proteomic compression ruler within the sample. The spike concentration was chosen based on a modelled output of cumulative *Synechocystis* data collected by USFD looking at emPAI values (reflecting individual protein abundance) using an in-house script (Appendix).

Work on this has highlighted that more consistent results can be obtained with iTRAQ labels, rather than TMT labels, in *Synechocystis*. In addition, the QExactive spectrometer has shown more accurate quantification of the labels, whilst the maXis demonstrated systematic bias in label quantification contributing to the quantification compression effect described by Ow et al., 2009.



3 Partner interactions

3.1 CNR-ISE

As described in the **18 month summary report**, a lyophilised sample of *Synechocystis* grown in a large outdoor photo bioreactor was received from CNR-ISE. It was hoped that this sample would build towards further suggestions, however extensive analysis of the data produced very limited practical results. Of particular note was the way the high-level biological contamination of proteomic samples affected the ability to assess the proteome. During processing, the samples looked to be of high quality, with gels showing a large number of proteins under coomassie staining, however biological contamination skewed the results on 2 levels. Firstly, it reduced the overall concentration of *Synechocystis* protein, reducing the dynamic range and therefore the capacity to visualise low abundance proteins. The more significant issue was having proteins from an unrelated organism in the sample. Having identified the protein contaminant to be Crysophyceae, the possible benefit of meta-proteomic analysis of the mixed culture, however since it was a destructive relationship (the *Crysophyceae* preyed upon the *Synechocystis*) H_2 production was not positively impacted. This contamination issue could be effectively cleaned with a chemical approach (altering the pH), thus there was no benefit to a metaproteomic approach at this time. However, investigation into the possibility of meta-proteomic analysis resulted in inclusion of a section in our trends article (**Landels et al, 2015**). Following this, CNR-ISE sent replacement samples from their 1000 L outdoor photobioreactor, the results of this experiment are discussed in **D7.3**.

3.2 UU

Work with UU detailed in this section of the report contributes to **Task 7.2**, where new chassis circuits have been analysed and tested by USFD with the results being fed back to UU to encourage further development of the system. UU provided both full-proteome scale samples and individual gel-slice fractions to USFD, during progression of their engineering the advanced hydrogenase, which turned out to be pivotal for USFD troubleshooting recommendations for both UU hydrogenase transformations and our proteomic pipeline improvements. There were two main studies carried out between USFD and UU for this deliverable, namely the protein gel slice analysis and the transformed hydrogenase iTRAQ. Through mass spectrometry investigation of targeted gel bands from the proteome of Synechocystis mutants, USFD identified the presence of the exogenous hydrogenase-related proteins confirming that they had been successfully transformed into *Synechocystis*.

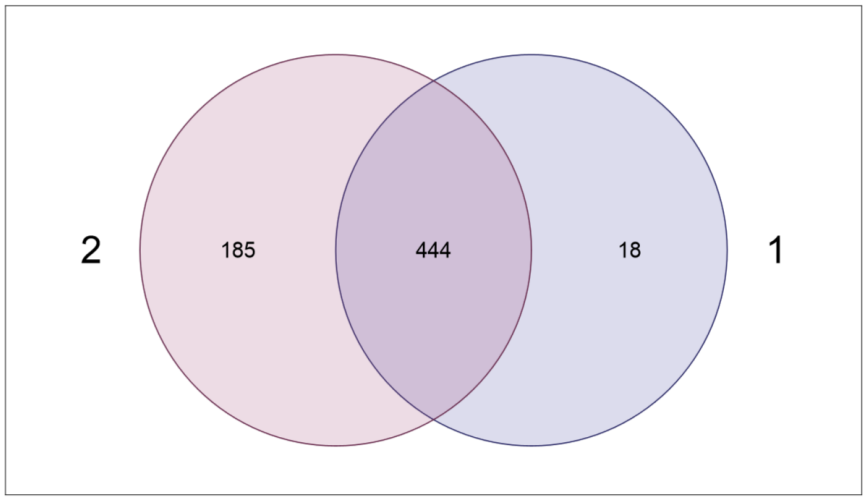
The iTRAQ analysis was carried out before month 18 of the project, which was prior to the pipeline improvements and thus resulted in limited dynamic range, with just 200 proteins identified and quantified. The statistically significant changes in the proteome reflected fluctuations in phycocyanins and ribosomal proteins – high abundance proteins with stable expression levels. We believe this reflects both protein over-estimation issues described in section 2.1, and the reduction in protein identification rate resulting from analysis of the membrane and soluble proteome represented in the sample. Literature searching confirmed that studies with the highest number of protein hits analysed the soluble and membrane fractions independently; however to gain the full benefit of this, the fractions would need to be kept separate throughout processing, which doubles the time and cost of the analysis. A practical “half-way” solution was to merge the extraction methods, by including a low concentration of detergent (such as SDS) in the lysis buffer and using a

more vigorous cell disruption method combining of both bead-beating and sonication for higher protein recovery and thus representation of the membrane and soluble fractions.

3.3 IBMC

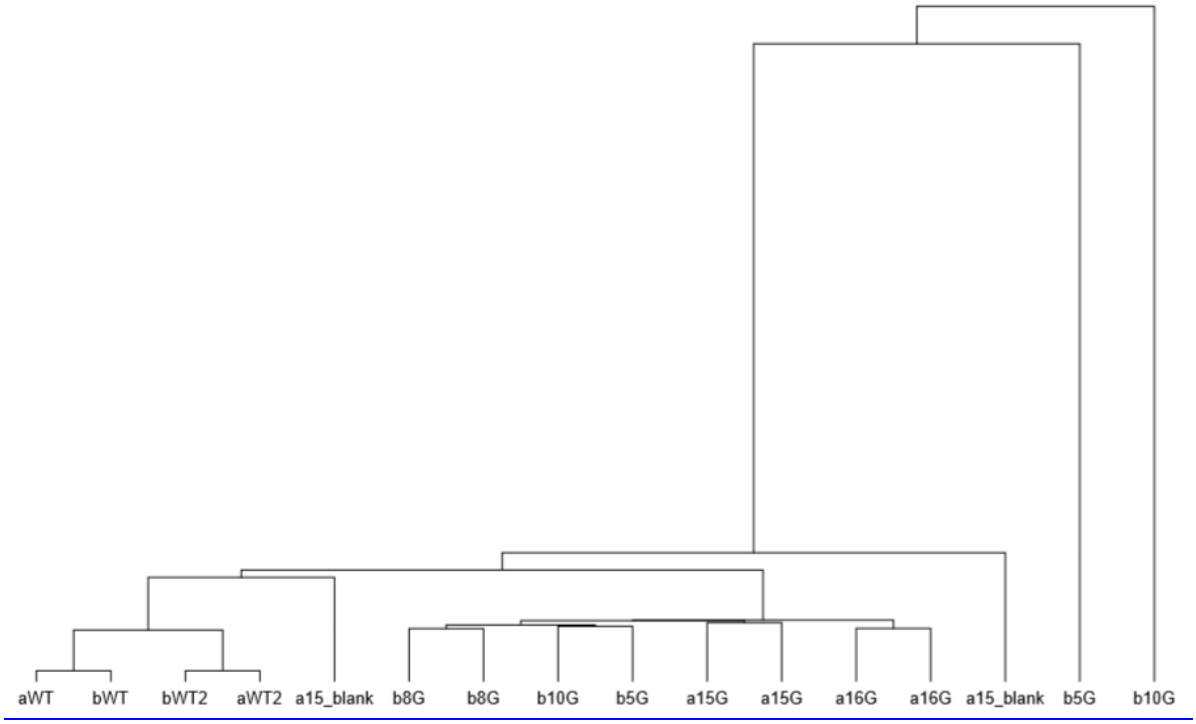
Work with IBMC detailed in this section of the report contributes to **Task 7.2**. Chassis developments by IBMC have been analysed by USFD, with the results being fed back to IBMC and UU to encourage further development of the chassis and provide a neutral genetic background for chassis-circuits to be developed in. USFD collaborated with the IBMC in their work on isolating genetically neutral sites for genetic modifications to the *Synechocystis* genome. These sites are crucial for the stable transformation of *Synechocystis*, as they limit unwanted background genetic side-effects. A combi-iTRAQ experiment was performed, merging the results of two separate iTRAQs with over-lapping samples into a single investigation of 5 potential neutral site candidates, including 2 controls, each with 2 experimental replicates. Our USFD methodology was to merge of data from individual iTRAQ data sets, extending the number of practically available labels from 8 up to 14. This was done by providing a 2 sample overlap between both experiments, cutting down the list of quantified proteins to only those isolated in both iTRAQs, and then normalising each of the protein ratios in the second iTRAQ to the mean of shared samples from the first iTRAQ. This represents a technical advance in iTRAQ data analysis. During the analysis, 639 proteins were quantified in total, with 444 proteins being common to both iTRAQs and carried forwards for further analysis.

(1) iTRAQ a 464 proteins (with 2 or more peptides)
(2) iTRAQ b 631 proteins (with 2 or more peptides)



Each of the sites was investigated for systematic proteomic changes, and the work was integrated with transcription analysis performed by IBMC and modelling work carried out by UPVLC. Three of the target sites showed tight clustering between paired replicates, however two sites had samples where the protein concentration varied significantly and resulted in a non-clustered effects. In addition, expression levels of GFP were quantified to analyse site production capabilities as well, these were found to be concordant with the expression levels identified by visual means. No systems-level proteomic changes were identified in any of the candidate neutral sites, suggesting that there would be no significant impediment to growth (and therefore hydrogen productivity) between each of the sites (figure). As a result, we recommend picking a site best tuned to the

expression level required for the genetic construct – which is dependent on future genetic engineering requirements. The work has been published (**Pinto et al, 2015**) with all experimental details available in the paper. USFD recommendations based on the iTRAQ work were compiled in a report and sent to IBMC as recommendations (full report attached as appendix).



4 Summary of recommendations

As detailed in this report, the following recommendations have been made to the CyanoFactory project:

1. A standardised form for sample delivery for proteomic analysis has been issued, improving data quality and uniformity of downstream analysis of samples.
2. A number of hydrogenase subunits were identified in UU samples, enabling trouble-shooting and further engineering of the hydrogen production machinery. This investigation identified that the inserted hydrogenase was being produced in the cells, although it was found through other investigation to be non-functioning.
3. An iTRAQ investigating background proteomic effects of the transforming the external hydrogenase into a *delta-hox* background, produced by UU, showed that there was no significant systematic changes to the proteome under standard growth conditions.
4. A technically advanced iTRAQ investigated 5 candidate neutral sites produced by IBMC. The sites didn't show any systematic changes to the proteomic network under normal growth conditions, but did have differing effects on the strength of expression of the inserted genetic construct. As a result, USFD recommends that the neutral sites chosen by partners in CyanoFactory for maximising hydrogen production should be tuned to the expression level required by the particular genetic construct – with details of the sites and their respective strengths available in **Pinto et al., 2015**.

8.3 Deliverable 7.3

Partner 6 – The University of Sheffield

Deliverable D7.3

Final design analysis combining optimised cells in optimised photobioreactor configuration: -omics data and assembled metabolic network for best identified engineered system in suitable photobioreactor to date. Suggestions for improvements to the system (bioreactor and cell) to potentially increase H2 production.

1. Introduction

For our final deliverable report, USFD present the most recent work carried out in conjunction with CyanoFactory partners. In addition, a list of the experimental work to take place over the next 2 months for omics level analysis is included, along with a roadmap for the publications that will be built from this work in the coming months. It is important to highlight that all samples presented in this report were made available to USFD in the final months of the project and so the research being presented here is bleeding edge. The recommendations presented here are valid and fulfil the requirements agreed to prior to beginning the project; and a more comprehensive analysis of this data will be carried out as this work gets converted into publications – in collaboration with UPVLC, CNR-ISE, RUB, UU, IMBC and UL. All initial data and findings presented in this report have been forwarded to the respective partners who sent samples, and will be passed to UM to be uploaded to the data warehouse once the full analysis is completed.

There are 4 major partnerships within CyanoFactory that have yielded the information presented in this report – investigation into the effects of the final design of the 1000 L outdoor photo-bioreactor (1000 L PBR) produced by CNR-ISE on *Synechocytis* over the course of a typical operational day, assessment of the reduced antenna *Olive* strain, isolated and characterised by RUB against the consortium WT strain using metabolic flux techniques, assessment of the hydrogenase active site complex investigation being carried out by UU, and investigation into the effects of key halo-tolerance genetic modifications produced by IBMC using RNAseq and proteomic techniques. Each case has a dedicated section of this report, which summarise the work carried out to date, highlight the conclusions drawn from assessment of the gathered data; and then identify the next stages of work to be carried out, indicating how publications from the work will be approached. The final section of this report ties together key findings into an idealised combination of all aspects, with suggestions for future experiments or practices that should maximise hydrogen production and growth efficiency in an industrial setting.

Publication Roadmap: Each of the sections mentioned in this report is building towards a CyanoFactory publication. The CNR-ISE proteomic time-course assessments on how the cells change either during an operational day or over the course of hydrogen production will either be bundled into a publication combining the comprehensive operational analyses carried out by CNR-ISE and published with an operational biotechnology focus, or else the proteomic dataset assessments will be re-focused for proteomics journals, and will include details of best practices for the advancement of proteomic analyses in biotech targeted investigations. The metabolic flux analysis of the *Olive* strain, which will feature collaboration between RUB, UPVLC and USFD, will be bundled with the proteomic assessment of the different *olive* strains (RUB, UU & USFD), forming an over-arching comparison between the consortium WT and the *olive* mutant. The proteomic assessment of the UU synthetic complex upon the external hydrogenase will be completed as a self-contained report to UU, who will decide what the next course of action to build towards their publication on the synthetic hydrogenase will be, either through further experiments or else analysing the data with a focus to complete a biological story. The RNAseq data and proteomic data collected for the halo-tolerance experiment conducted by IBMC will be assessed with the assistance of UPVLC, who will contribute the statistical expertise needed to generate concordant findings between the two datasets. The data should also provide additional information to improve the UPVLC *Synechocystis* model.

In addition to the work presented here, USFD is also in collaboration with UL to characterise proteomic changes in samples from an experiment of samples grown in controlled media versus standard environmental conditions (brook water). Through earlier discussions, USFD recommended UL make an analysis of the metals present in the water using inductively coupled plasma mass spectrometry, which they have completed. USFD have the samples in-hand from UL, which are currently queued to run in the mass spectrometer and should be completed by the end of this week. Findings of this investigation will be processed and circulated to the consortium via the data warehouse (UM) upon completion of analysis. The proteomic data for these samples will form part of a publication investigating the effects of *Synechocystis* released into the environment and the formation of biofilms.

2. 1000 L outdoor photo-bioreactor

The work in this section contributes towards **O7.1, Task 7.3 and M7.3**, by providing key proteomic datasets that feed into the best-case PBR design scenario. The findings in this section form the basis for this entire report, with output from the subsequent sections feeding back into it to keep the large-scale industrial applicability context of the CyanoFactory work in prime importance. The data here has not yet been circulated to the rest of the consortium, as further experimental processing on the samples will yield a more statistically significant and concordant dataset to upload to the data warehouse (UM).

CNR-ISE and USFD had a series of discussions and Skype calls following the consortium meeting at UPVLC with regard to the best way to assess, using -omic techniques, the interesting findings CNR-ISE presented during the meeting. It was decided that 2 experiments were of key interest to the consortium, the first was a time-course assessment of samples taken over a typical day of running the PBR and the second was an analysis of the changes that took place during large-scale hydrogen production. An analysis of the changes that took place in Synechocystis during a sharp spike in pH, such as the one used to remove external contamination from the bioreactor (please see **D7.2**), was also discussed – however it was felt that since there were multiple publications that already presented data on pH changes that this was a lower priority experiment to carry out.

The daily time-course experiment was conducted first, as having a basic understanding of the systematic changes that took place over the course of a typical day was deemed to be the most important set of information to feed back to the consortium. This was decided because it presented information that could not be gathered or accurately simulated in any other way in the lab, and was vitally important to understanding changes that took place when simulating other experimental conditions in the 1000L PBR. Three replicate samples were collected from a continuous cultivation on separate days, which were chosen based on having similar temperature and cloud cover to a typical summer day in the north of Italy. The samples were taken at 4 time points throughout the day: 0900, 1100, 1400 and 1700. The samples were collected from the reactor, spun down into a pellet of cells, and stored at -20 (-80 for long term storage). They were then mailed to USFD on dry ice – as per the standard delivery instructions devised during improvement of the proteomic pipeline. This is discussed in more detail in **D7.2**.

The 3 replicates were designed into a 2-iTRAQ experiment, similar to the one conducted with IMBC during the neutral site investigation (**D7.2**), in such a way that the iTRAQs could be assessed independently or combined together. The first 8-plex had one label per time-point sample for the first and second samples days, whilst the second included the second and third sample days. Each of these iTRAQs was considered to be robust and suitable for publication in their own right; but as discussed previously (**D7.2**; **Landels et al., 2015**), it is important to raise the quality of proteomics in industrial biotech publications to improve systems-level engineering approaches. For our work here, data from the first of these iTRAQs is presented. In this analysis, 1870 proteins were identified and quantified, with at least 2 confident unique peptide identifications. This represents ~85% of the entire observable proteome. The samples were initially assessed for significance in systematic progression (ie. 0900 vs 1100, 1100 vs 1400, etc.), however this did not produce significant findings. This was investigated by generating a heatmap and applying Ward clustering techniques to see how the different labels related to each other (figure 1).

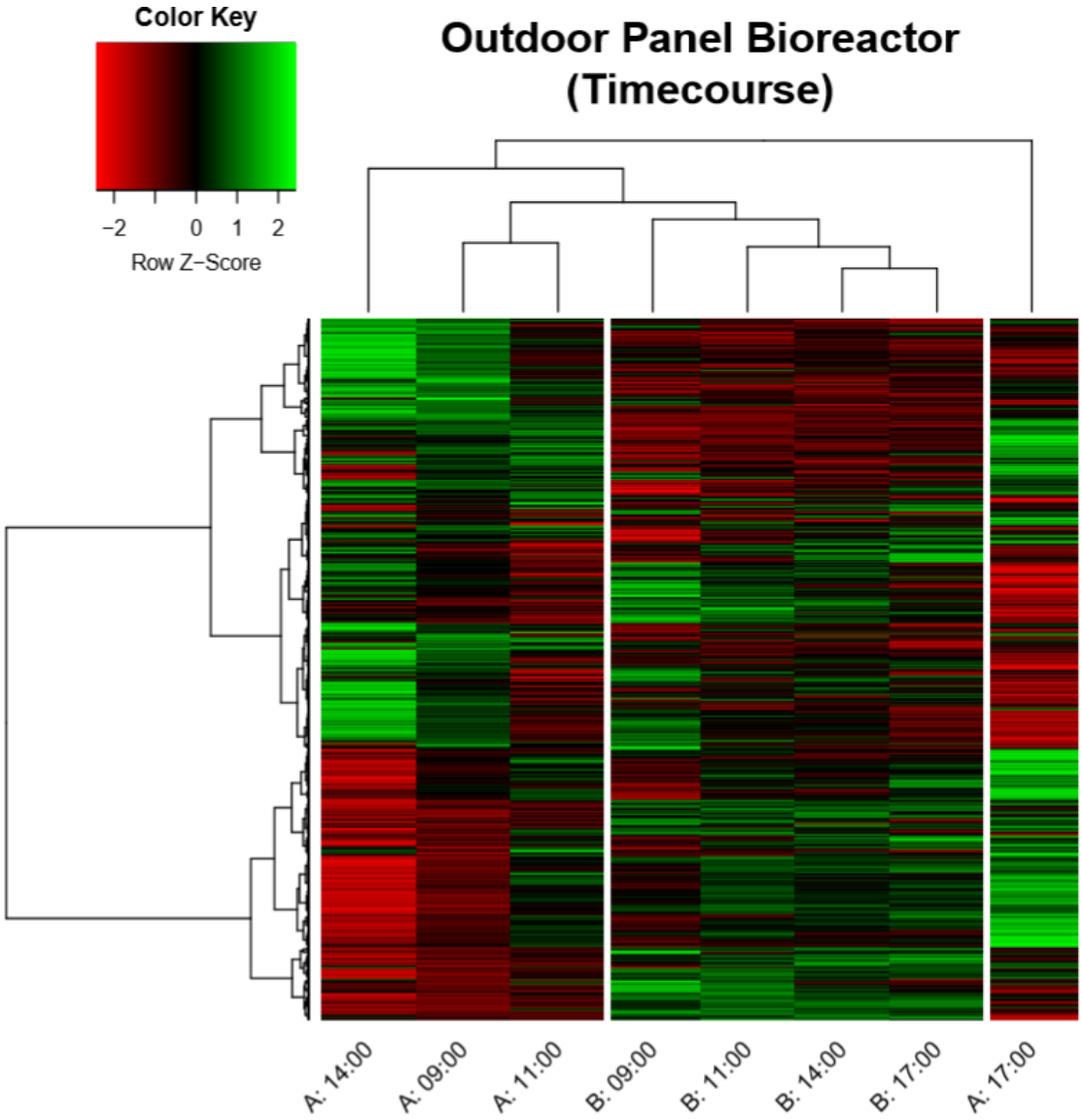


Figure 1: *A heatmap of the time-course data collected from the 1000 L PBR by CNR-ISE. The Ward-linkage clustering method shows high relatedness within sample B, leading increasingly similar expression profiles over the course of the day. Large variations are present in sample A, in time-points 1400 and 1700 in particular. To clarify groupings, white separators have been included in the diagram. The protein clustering profile has also been included in this diagram as a dendrogram on the left; this is also calculated using Ward linkage.*

Figure 1 clearly shows a disparity between the first and second replicate set. Whilst heatmaps are useful for looking at data trends, they are of limited use in quantifying that similarity – for this purpose USFD typically engages principal component analysis (PCA) techniques to identify relatedness. PCA is a useful technique as it collapses the data from a multi-dimensional problem, where finding similarities can be challenging, into a lower-dimensional problem, where most of the sample variation is collected in the first and second principal components (figure 2).

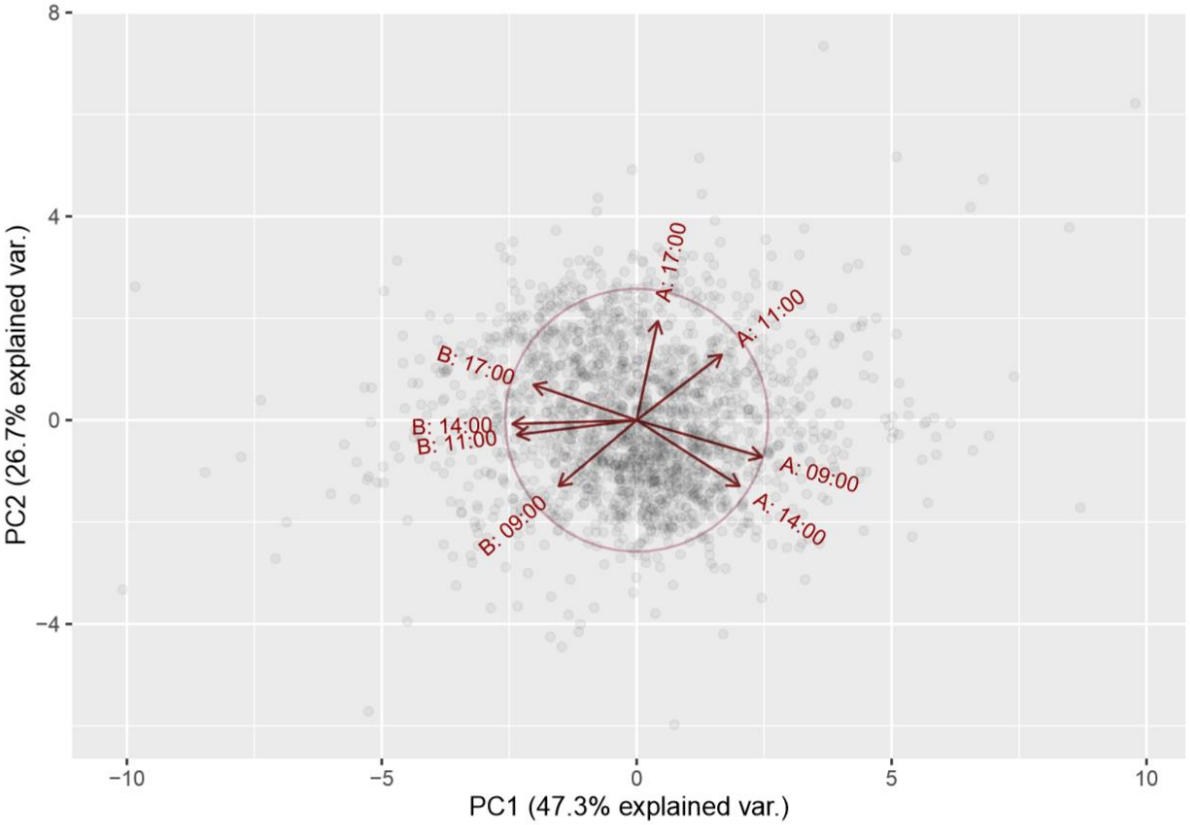


Figure 2: *A PCA plot showing all the samples plotted as reduced dimensional data-points. The first 2 components (presented as the x and y axes in this figure, respectively) are responsible for ~75% of all variance within the iTRAQ experiment. Each faded black circle represents where each protein falls in the plot and the vectors within the circle indicate the Eigenvectors for each of the conditions. In simple terms, the length of the arrow indicates the magnitude of change, and the angle of the arrow indicates the 'direction' of change, such that 2 arrows in a similar direction consist of a similar protein expression pattern, ie. B: 11:00 and B: 14:00 have almost equal orientation and magnitude, and so are ~75% – the variance explained by this figure – identical in protein expression patterns. Negative magnitude (B: 09:00 vs A: 11:00) indicates variance in the same set of proteins but in the opposite direction. It's important to mention at this point that the angle in PCA has no bearing on whether proteins in the set are in higher or lower concentrations, they either change in the same direction or opposite directions. For this reason, PCA alone cannot determine features about proteomic sets. Perpendicular arrows suggest that the vectors are changing independently of each other – in practical terms, the variance of protein expression in each of these samples is explained by different (independent) sets of proteins (ie. A: 09:00 and A: 17:00). A simplified variant of this figure is presented below, where the key findings are discussed further.*

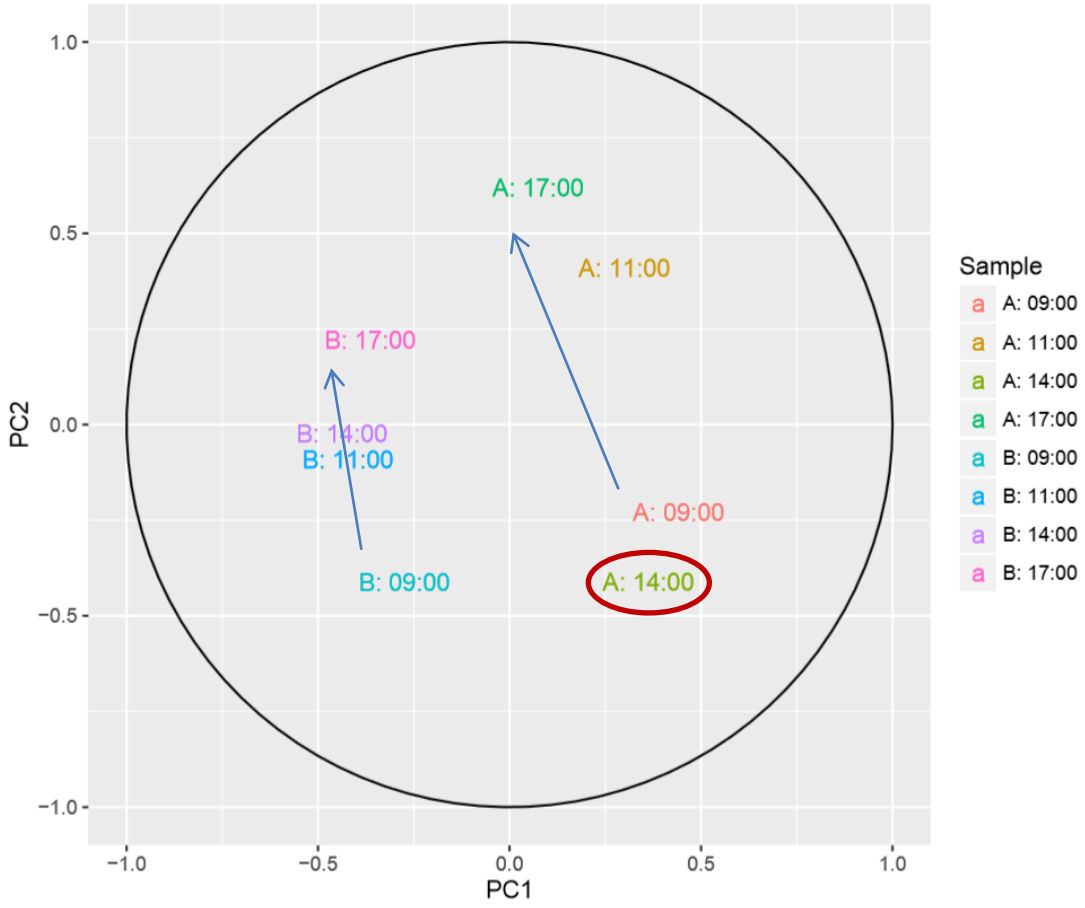


Figure 3: *This is the same PCA plot as presented above, although most of the features have been removed from it to make the explanation of effects within the sample simpler. As can be seen here, almost parallel vectors can be plotted running from 0900 for each sample through to 1700 (blue arrows). In replicate B, all the time-points fall near to this line, in the order of progression through sampling time. This is to be expected as systematic changes in proteome response to increased light and heat build-up in the system. In replicate A, sample 1400 (red circle) lies in the opposite direction to this vector. This suggests that there is a problem with this sample, enabling it to be excluded from the analysis at this time. This data also shows that there is a lot of independence between the two samples at the beginning of sampling, even though the direction of change across both samples is the same (in this case, upwards on PC2).*

Figure 3 highlights a potential 'change vector' running through the samples in the PCA plot. The 14:00 sample taken from replicate A falls in the opposite direction to this vector, discordant with all other samples in the experiment, and so is considered to be an outlier to be removed from further analysis. A comparison between the samples taken at 0900 and the samples taken at 1700 was made, to generate a list of proteins significantly affected over the time-course. The large disparity between the initial states of these two samples (separation on PC1 in PCA, large dendrogram separation between identical time-points in heatmap), supports the limited number of significant identifications made between the two time-points (Figure 4). Further analysis will increase the number of targets by applying more advanced statistical techniques to the dataset.

The proteins identified as significantly upregulated are involved with reductive stress response, metal ion imbalances and nitrogen deficiency (list in appendix). These are well documented effects triggered by high-light intensity or reductive stress, where the light-harvesting structures are damaged by high intensity light and must be regenerated by the cells. Interestingly for CyanoFactory, the state the cells are in at the end of a day of high solar productivity is similar to the idealised state attained by creating artificial nutrient deficiencies or subjecting the cells to high intensity light. This in turn means that having a system where the cells are producing biomass during the day before being subjected to anaerobic hydrogen producing condition in the evening/overnight may be highly advantageous to efficient hydrogen production. Prior to publication, the remaining half of this iTRAQ experiment with the 2nd and 3rd replicates will be run on the mass spectrometer, to increase the statistical significance of these findings as well as providing a workaround for the outlying sample.

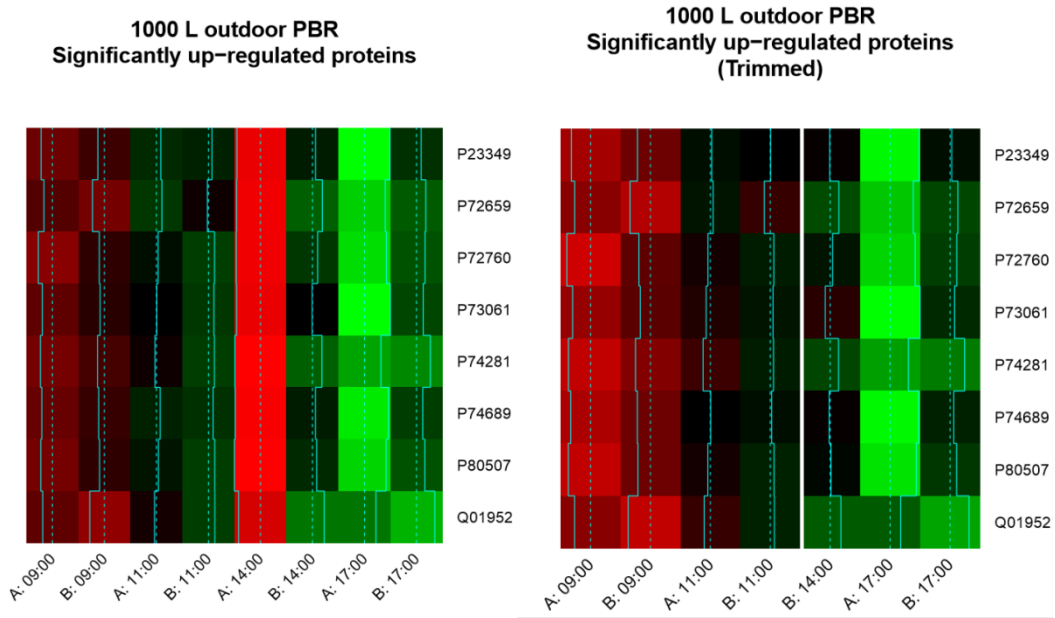


Figure 4: *Heatmap layout of the proteins found to be significantly up-regulated, outliers are excluded*

The next experimental samples to be analysed, which are now held in-hand by USFD, are the experiments from the large-scale hydrogen production experiment. In this experiment, CNR-ISE subjected the cells to nitrogen starvation conditions over 3 consecutive days, before running a 4 day anaerobic hydrogen production experiment. The observations from this experiment were very interesting; however the experimental design for analysing this data in iTRAQ 8-plex format is complex and involves 3 iTRAQ kits and a TMT 6-plex kit as a fail-safe. As in the 1st experiment, these data were collected in triplicate, with a morning (0900) and an afternoon (1700) measurement taken each day. The 2 days prior to hydrogen production tagged with 4 of the labels in each iTRAQ, and the first 2 days of hydrogen production tagged with the remaining 4 labels of each iTRAQ. The first time-point (2 days prior to anaerobic conditions) and the first time point after hydrogen production has begun (first anaerobic measurement) for each of the 3 replicates will be labelled with TMT tags and compared to provide a linking dataset between the 3 iTRAQ experiments. Although USFD have identified previously (**D7.2**) that there is a compression effect on TMT tags in Synechocystis, any bias introduced through the use of TMT tags can be accounted for with a conversion factor as a result of the iTRAQ vs TMT modelled analysis work described in **D7.2**.

3. Metabolic characterisation of the *Olive* strain vs WT

The work presented in this section covers the requirements for **M7.2, Task 7.2 and O7.3** by characterising a key chassis modification with metabolic flux analysis and proteomic investigation. The experimentation was carried out in the 5 L photo-bioreactor produced by KSD, as it was determined to be a suitable bioreactor system produced from the consortium. These samples were collected close to the end of the project, and so the integrated analysis of the fluxes to be carried out with in collaboration with UPVLC is still underway. The experimental data gathered and the time-course values already calculated will be presented in this section. All data presented in this section has already been circulated to both RUB and UPVLC, and will circulated to the rest of the consortium (via the data warehouse produced by UM) once the findings have been analysed more comprehensively.

Following on from the work in the previous section, it is apparent that one of the major contributing factors to any large-scale hydrogen production system in *Synechocystis* is how to utilise the effects of high-intensity solar radiation on the culture. The *olive* strain, isolated by RUB, has a reduced set of antenna proteins that allow increased transmission of light through the culture. These effects have been characterised by RUB and modelled by UPVLC, with details described elsewhere in the CyanoFactory project. A key future experiment would be to replicate the time-course experiment carried out in the 1000 L PBR with the *olive* strain, to see if the effects of solar radiation were equivalent, or if they had more rapid/slower onset. Due to the high daily variability between the initial states of the samples measured within the 1000 L PBR, it is important to further characterise metabolic features of the *olive* strain – particularly when making comparisons against the wild type – to understand how far the conditions of the reactor (ie. the solar radiation) and how far the metabolic response of the organism is responsible. Without understanding where the division of responsibility for these effects lies, it is difficult to make recommendations for further improvements to either the chassis or the bioreactor.

For this reason, the C^{13} labelling experiment was carried out on the *olive* strain and the consortium WT stain of *Synechocystis*, to determine if there were significant variations in the fluxes in the *olive* strain that might be important for future modelling work. This work is completely novel – no one else has attempted to characterise the metabolic fluxes of the *olive* strain to date. The experimentation was carried out in the 5 L KSD photo-bioreactor in RUB. This is the largest feasible bioreactor the experiment could be performed in without making the cost of the experiment impractically large. The cost for this work is almost completely absorbed by the large amounts of stably labelled carbonate needed for the experiment, with a single time-series replicate costing more than an 8-plex iTRAQ experiment.

As this is the first time an experimental analysis of stably labelled metabolite samples from photo-autotrophic conditions had been conducted in USFD, a number of preliminary experiments were required before the samples could be analysed. The first of these was a test with key metabolite standards, to identify how the metabolites would act individually, in a simple mixture, and in a complex background. A series of experiments with 5 standards chosen from the central carbon metabolism pathway were carried out, using the experimental technique described by Shastri and Young in their pioneering work on characterising the photo-autotrophic metabolic network in Synechocystis (Young et al., 2011). From this, elution times for the measured metabolites were

measured, along with the characteristic masses associated with the different metabolites to facilitate identification in a more complex mixture. These times were compared to a series of elution times using the same liquid chromatography method (identical column, gradient, buffers) utilised by Dr Shastri in her doctoral thesis work. This comparison enabled the inference of metabolite elution times, for compounds where standards were regrettably unavailable.

The next series of experiments carried out by USFD were with physical dry-runs of the sampling experiment using shake-flasks. By carrying out these runs on unlabelled samples, spiking in unlabelled carbonate to the mixture and running through the experiment, USFD were able to determine potential bottlenecks in the procedure. One of these was the grouping of samples together during extraction, to avoid losses caused by excessive resting time in methanol used for quenching. This quenching step is vital for the experiment, however if the entire experiment is completed without removing the quenching agent quickly, then the metabolites will be leeched from the sample and the experiment will fail. These samples were assessed with low resolution mass spectrometry in the same method used for the standards; however individual metabolites could not be confidently determined from the system due to high sample complexity. This necessitated the use of high resolution mass spectrometry, which was available but would require re-organisation of the system and was to be conducted at a time when the demand for the machine was lower. . Due to booking constraints, all high-resolution analysis was run at the same time, which took place in December 2015 after the final samples had been collected in RUB.

Once the experimental procedure had been characterised for completion in USFD, test runs using labelled carbonate were carried out on batch shake-flask cultures. These samples were assessed for the presence of ^{13}C labelling using low resolution mass spectrometry, to confirm increasing levels of labelling could be detected, although as stated above individual metabolites could not be confidently determined from the system, due to high sample complexity. Following confirmation that each part of the analysis could be completed at USFD, a PhD student from USFD travelled to RUB to translate the experiment to the KSD 5 L PBR in September 2015. During this visit, there was a key transfer of information both from RUB to USFD and vice versa. RUB shared extensive knowledge concerning both the running and maintenance of continuous culture bioreactor systems, as well as the growth features of the *olive* strain. USFD explained the experimental process and highlighted the key findings that the preliminary experimental work showed. Together, USFD and RUB devised a method for running the experiment, however due to an unforeseen culture crash with the cultures in the bioreactors the sampling could not be completed before the end of the visit. RUB further characterised the ^{13}C spike method and successfully produced the samples for analysis in November 2015, which were analysed using high resolution mass spectrometry by USFD – (data shown in Figure 5).

The next step for this analysis is to work with UPVLC to determine the fluxes of the identified metabolites, which will be done using the INCA framework devised by Jamey Young. Once these fluxes have been determined, a more kinetically accurate model of the *olive* strain can be computed. This will provide better understanding of the differences and similarities of these strains.

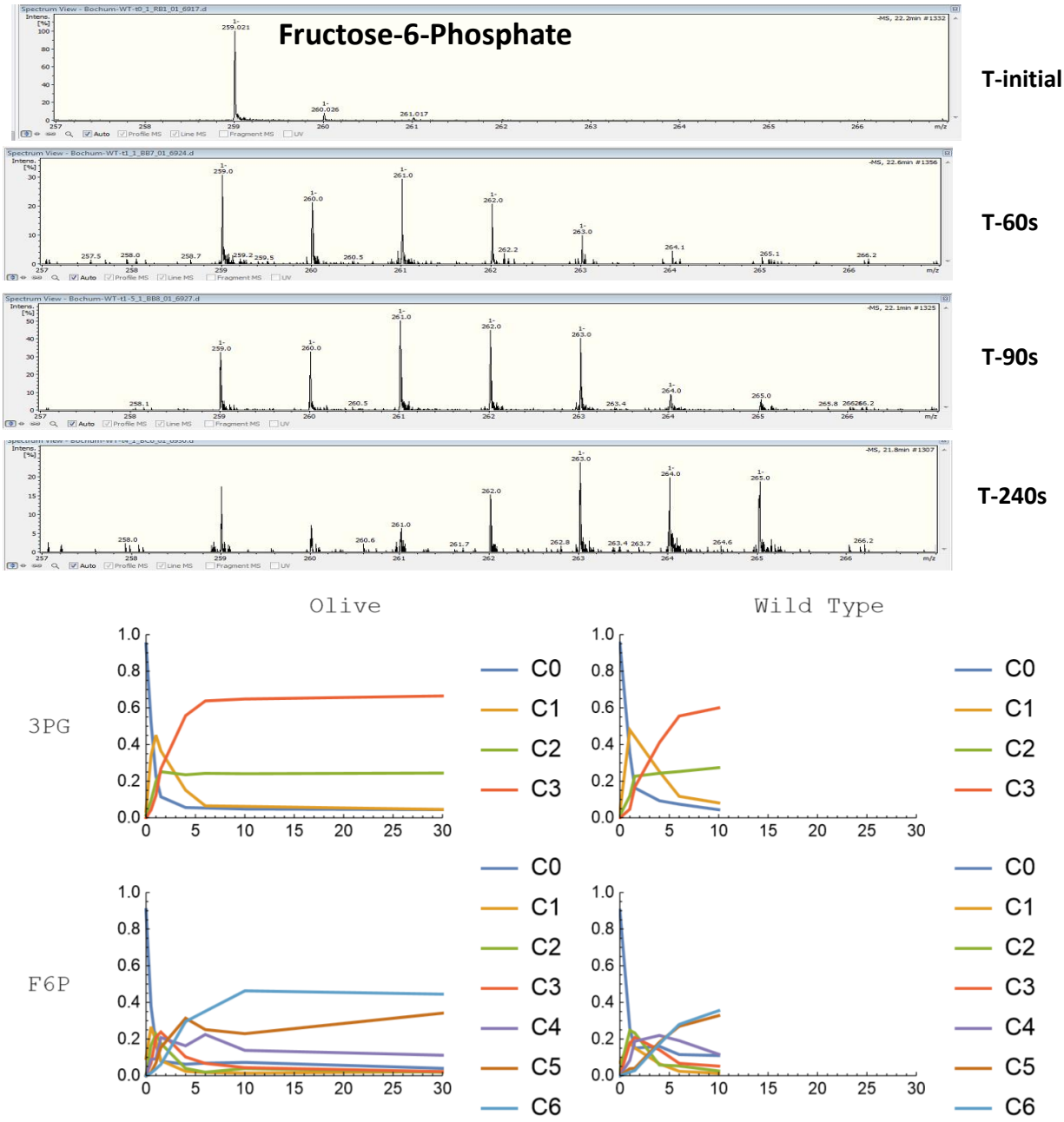


Figure 5: *Top: Isolated data for Fructose-6-Phosphate at 4 time-points in WT* Synechosytis*, shown as an example of raw data from the mass spectrometer. Bottom: the distribution of metabolites shown over the time series for 2 identified key metabolites in the central carbon metabolism, 3-phosphoglyceric acid (3PG) and fructose-6-phosphate (F6P). Due to a problem during sampling, the 30 minute time-point and 30 second time-point were unavailable for the Wild Type; however these are not needed to calculate the system fluxes.*

In addition to the work on metabolic flux analysis, protein samples have also been gathered by RUB and UU for both the UU WT strain and 2 genetically distinct *olive* strains. These samples have been processed in such a way that the membrane proteins and soluble proteins were collected separately. This will enable a final comparison between the cost-efficient merging of these fractions against the potential signal stability offered by processing the samples separately (**D7.2**). These protein samples are currently held in hand by USFD and are in the queue to be run on the mass spectrometer.

4. Characterising the external hydrogenase system

The work presented in this section contributes to **O7.1, Task 7.2, and M7.4**. This section highlights work that demonstrates a clear progression from the samples described in **D7.2**, where USFD gave feedback on the engineering work performed by UU. It is a key example of the iterative improvements to the system described in **M7.4**.

UU have done extensive work generating a mutant *Synechocystis* containing an exogenous hydrogenase enzyme. Having successfully transformed the HydA1 hydrogenase (apo enzyme) from *C. reinhardtii* into the *delta hox* background, it was found that the active site the hydrogenase required to function would not form. When a synthetic variant of this site was generated using synthetic chemistry methods combined with the enzyme it was shown to have the ability to produce hydrogen. USFD ran an 8-plex iTRAQ analysis to investigate the changes found to occur within the cells under hydrogen production conditions utilising this compound. The key comparison in the iTRAQ investigated the effect of the presence and absence of the synthetic complex (+ vs -) to the cells containing HydA1 in the delta hox background. A control comparison was also carried out to investigate systematic effects of the synthetic complex in a delta hox background where no HydA1 was expressed. 1980 proteins were quantified with 2 or more confidently identified unique peptides in the experiment, representing ~90% of the observable proteome. This is the largest number of unique protein quantifications of *Synechocystis* identified by USFD to date. The HydA1 protein was successfully identified to a high degree of confidence.

Initially, the data were plotted on a with a PCA unit circle to identify sample grouping and make sure that there were no unexpected effects that could cause problems for statistical analysis of the data (figure 6). For further details on this type of figure, please see section 2 of this report. The PCA plot confirmed that most of the systematic variance in the data occurred between the HydA1 mutants in the presence and absence of the synthetic compound. This comparison was run though our statistical pipeline, generating a list of 584 significantly up- and down-regulated proteins between the conditions. In addition to this, a protein 'trend effects' analysis was conducted (figure 7, figure 8), identifying 710 potential candidates for investigation into systematic effects. This highlights proteins that show systematic changes concordant with the statistically significant proteins, but lack the peptide-level evidence to generate statistically statistical leads. Both of these datasets have been included with this report as appendices.

Due to an issue with the KEGGmapper bioinformatics software, pathway maps are currently unavailable for this data. In lieu of that, the data here are presented with functions as general descriptors. These findings have been sent to UU to discuss potential proteomic leads, since pathway level information is not currently available and the list contains multiple protein systems. The data will also be further analysed in collaboration with UPVLC, to attempt to map them onto the best case metabolic pathway available to date. Discussions are underway to determine the next proteomic experiment necessary to further characterise the system and aid a publication of these findings.

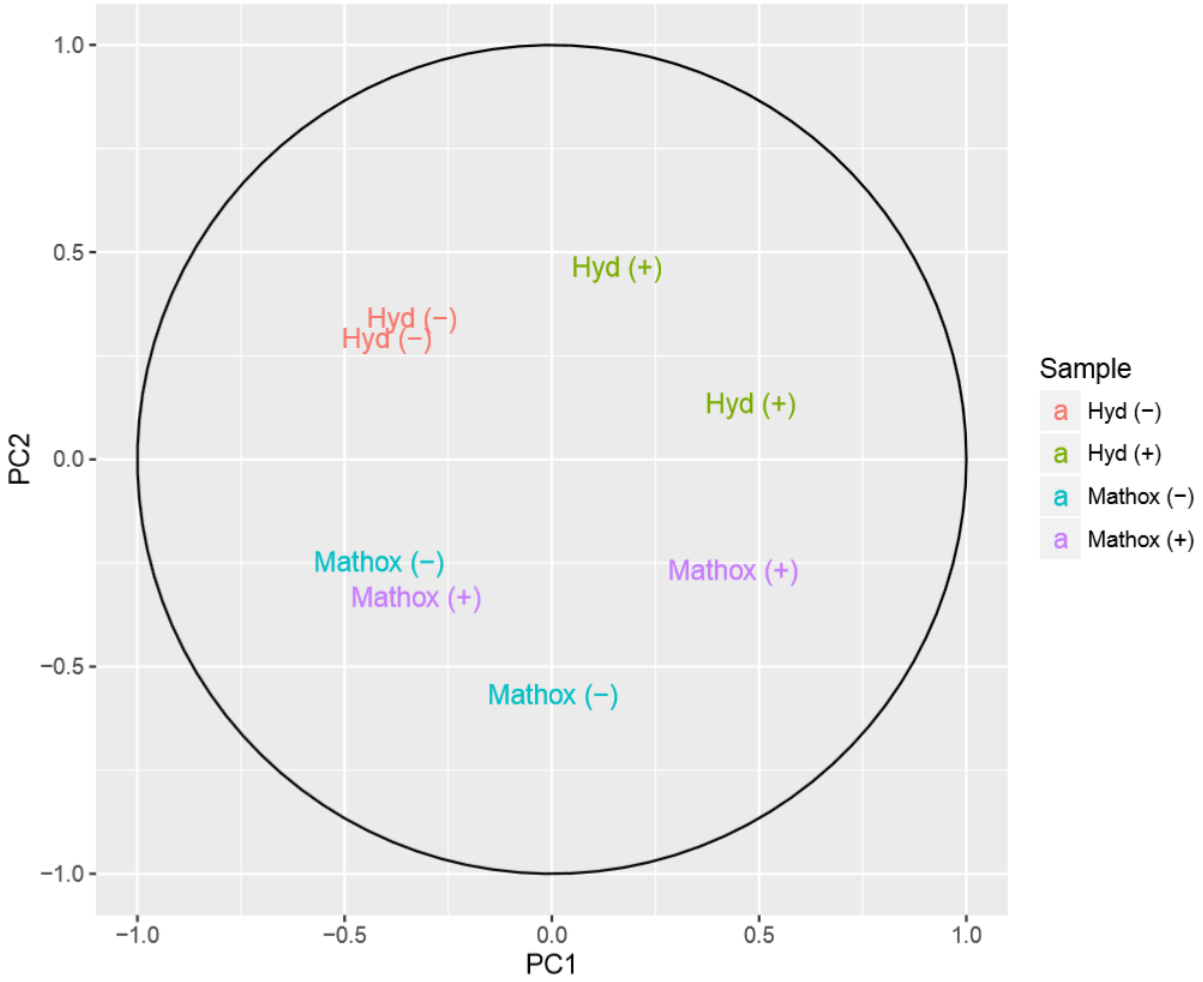


Figure 6: *A unit circle PCA diagram, showing the general grouping of the different conditions investigated in the 8-plex iTRAQ experiment. The principal components account for ~70% (PC1 - 45.8%, PC2 - 25.8%) of the cumulative variance in the sample. The HydA1 mutants are all separated from the delta hox background across the 2nd principal component, suggesting an underlying systematic change resulting from the presence or absence of the inserted hydrogenase. There is a clear separation across the 1st principal component for the presence or absence of the synthetic complex in the HydA1 mutants indicating a systematic difference in the cells that are able to produce hydrogen. This is not as clear in the delta hox background mutants, where the samples are more closely clustered by parallel experimental repeat. Due to the clustering effects, this plot suggests that the 2 comparisons of most interest from this data are HydA1(+) against Hyd1A (-), and all the Hyd1A mutants against all the delta hox mutants.*

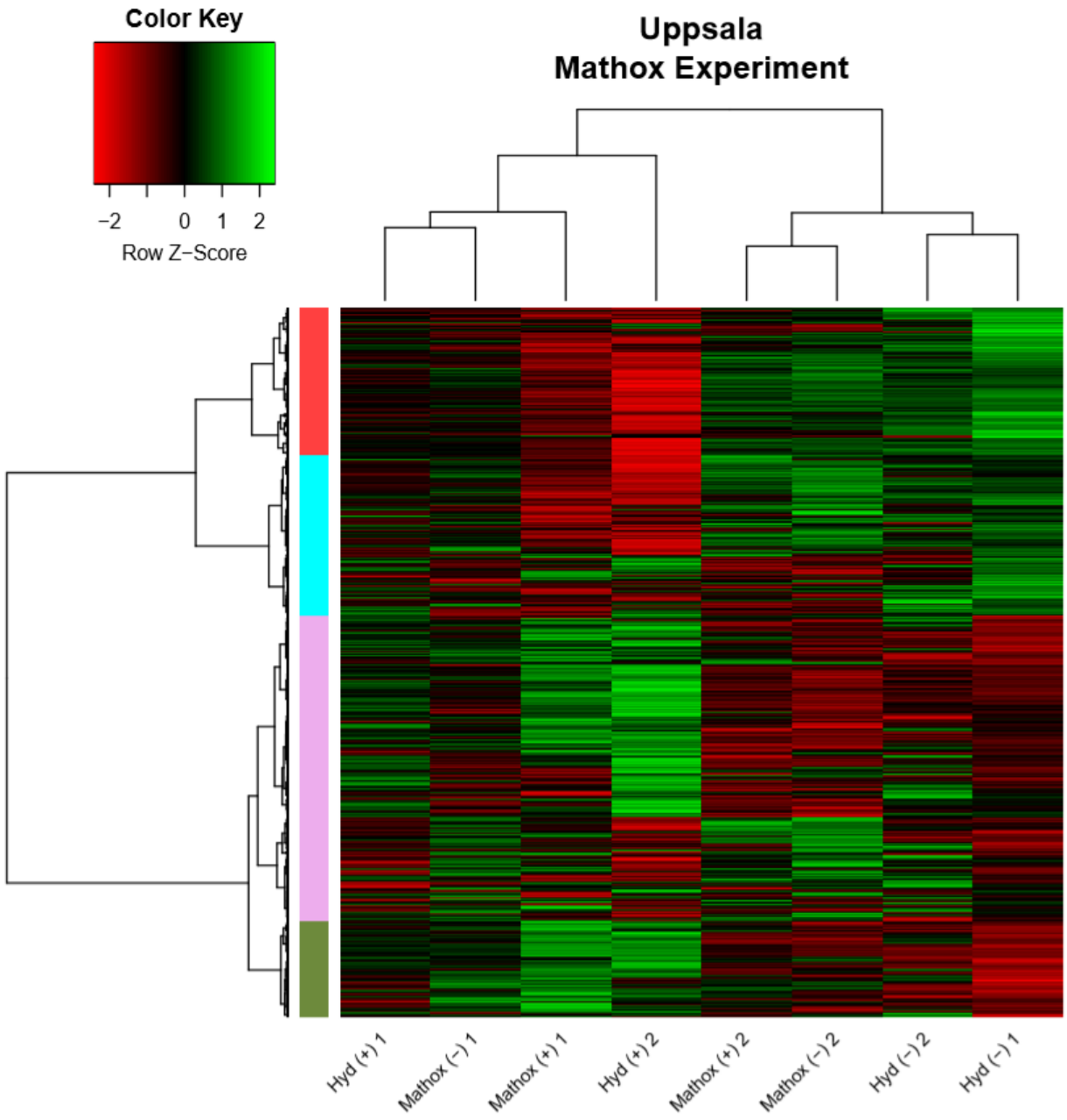


Figure 7: *A heatmap of the HydA1 proteomics 8-plex iTRAQ, the dendrograms on this heatmap use Ward linkage and the coloured bars on the left of the figure group proteins by cutting the dendrogram off at the point where there are 4 branches. This data is in agreement with the features seen in* ***Figure 6****, with close clustering of the hyd1A mutants in the absence of the synthetic complex. The coloured clusters on the left of the diagram were added as an attempt to categorise protein expression patterns by systematic expression trends, in the absence of a functional KEGG map tool.*

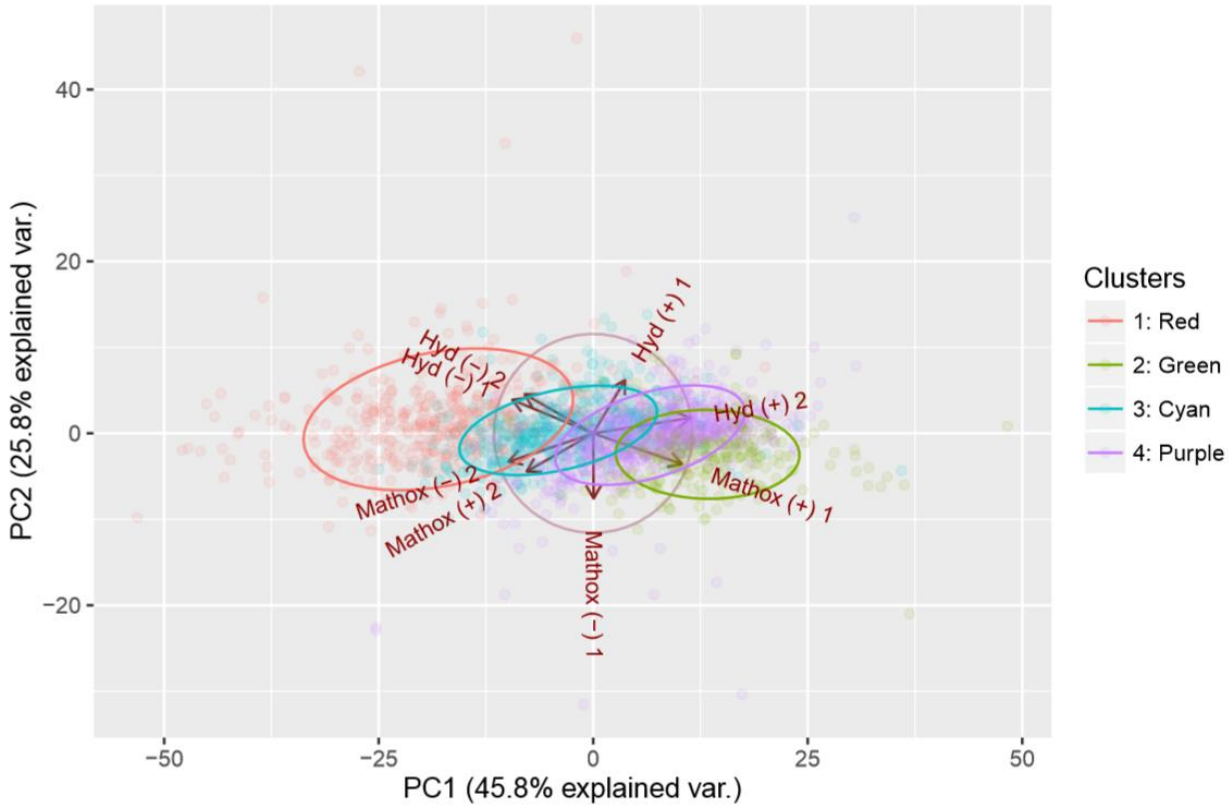


Figure 8: *A PCA plot showing the condition vectors describing the effects seen in the experiment within a unit circle (pink). The proteins are overlaid as faded circles, coloured according to which cluster they were localised in from* ***Figure 7****, with the area best representing each cluster highlighted with a coloured oval. This figure shows that the clusters separate the different clusters out across the first principal component. Proteins in the red and green clusters appear to be associated with the presence or absence of the synthetic complex. The PCA alone cannot determine direction changes, however when considering both* ***Figure 7 and 8*** *simultaneously it emerges that the red cluster contains proteins generally up-regulated in the presence of the complex, whilst the green cluster contains proteins generally down-regulated in the presence of the complex. This gives us a key insight into the proteins responsible for the variation observed on the 1st principal component.*

5. Investigation of halo-tolerance genetic modifications

The work presented in this section covers the requirements for **O7.2, Task 7.2 and M7.3** by characterising a genetic modification with RNAseq analysis and proteomic investigation. The transcriptomic analysis was initially planned to be carried out annually using microarrays, however an oversight in the budget assignment of 'consumables' caused a delay with this, as some of the funds needed to be categorised as 'services' to have the arrays read. This was correction was agreed at the mid-term meeting and confirmed at the 24 month point. During this delay, developments in RNAseq technology meant that for an equivalent cost, a more sensitive RNA analysis could be performed. This work was combined with a proteomic analysis, to generate the integrated –omic analysis mentioned in **O7.2**. All of the RNAseq samples were combined into a single, multi-condition comparative experiment covering both an investigation into genetic modifications to the chassis, as well as environmental changes to the bioreactor by adding NaCl, satisfying the requirements for **M7.3**.

Beyond modifications to the hydrogen production systems, it's important to consider the media conditions within the 1000 L PBR and whether it can be made more economical, by removing the requirements for 1000 L of fresh water for every run. Discussions with IBMC in Ljubljana during the 24 month consortium meeting lead to plans for an experiment investigating the effects of halo-tolerance gene knock-outs/over-expression in low and high salt conditions. The experiment investigates the effect the genes *glucosyl-glycerol-phosphate synthase* (*ggpS*) and *glucosyl-glycerol-phosphate phosphatase* (*ggpP*). Three genetic conditions are investigated: WT, Δ*ggpS,* and OE *ggpS* + *ggpP*. For each of these, 3 environmental conditions were investigated: 0% NaCl, 3% NaCl and 3% NaCl (9 day culture). Two experimental replicates were taken of each condition in the experiment, resulting in 18 samples. As there was only funding available for 16 samples, the 9 day culture NaCl was removed from Δ*ggpS*. A post-doc from USFD travelled to IBMC for a period of 1 week, to assist with the extraction of RNA and to provide continuity with the sample analysis. As with the ^{13}C metabolic flux analysis experiment, this is the first time USFD has performed an RNAseq analysis on *Synechocystis* and so a more detailed analysis of the data will follow as familiarity with the data structure improves. The initial investigation on these data was targeted to a list of genes provided by IBMC as potential targets of interest, one of these, the sucrose biosynthetic enzyme spp (slr0953), was found to have increased expression in the Δ*ggpS* conditions (Figure 9).

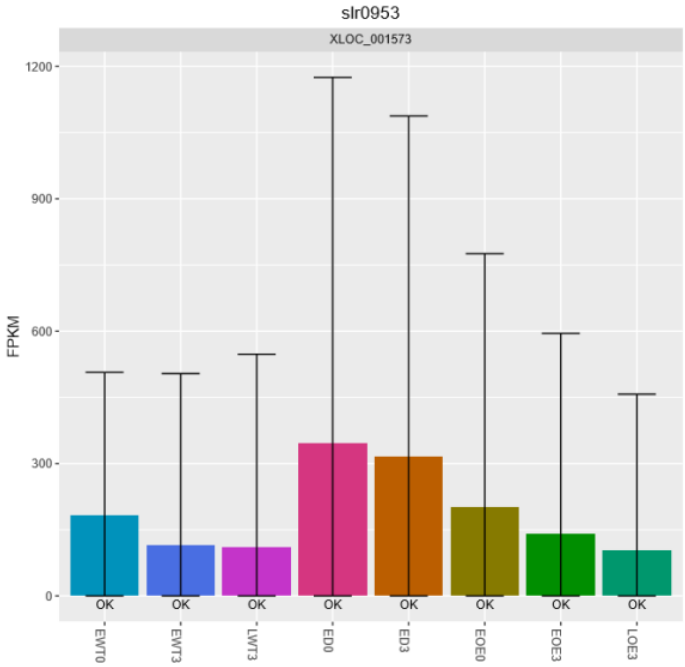


Figure 9: *Absolute transcript levels in fragments per kilobase pair for the gene slr0953. The levels were found to be higher in the* ΔggpS *conditions.*

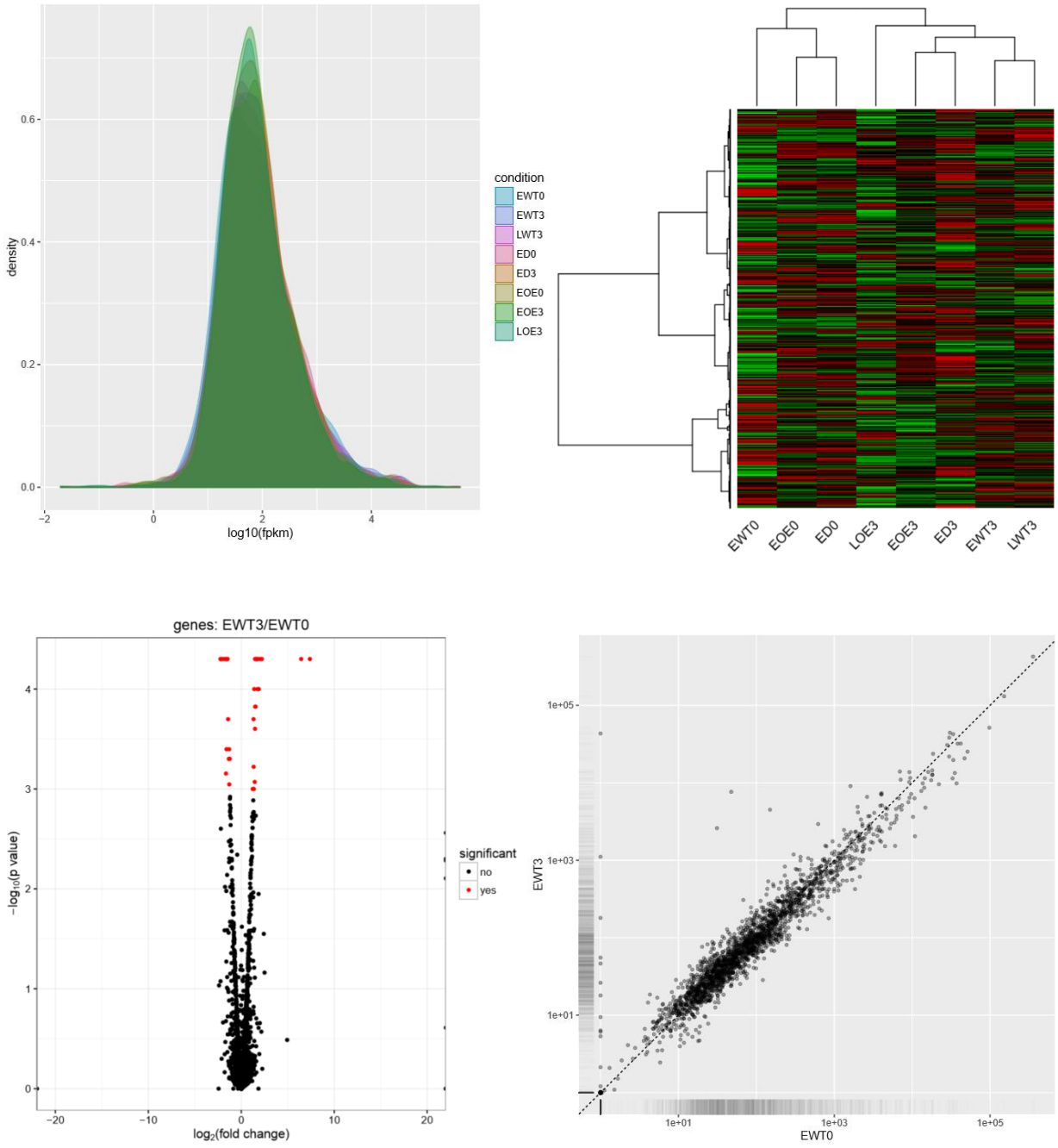


Figure 10: *Summary analyses performed on the RNAseq data. Top Row: A density plot and heatmap of all genes identified during the experiment. The density plot shows that there is an even distribution of transcript intensities between all samples after normalisation. The heatmap shows the general clustering between the samples exposed to 3% salt stress versus the ones not exposed to salt stress. Samples that have been genetically perturbed, either as a KO or OE, appear to cluster apart from the WT samples. Bottom Row: A volcano plot and scatter plot comparing WT* Synechocystis *in 0% NaCl against3% NaCl conditions. Genes which show statistically significant differential expression are highlighted in red in the volcano plot. This is 1 of 28 comparisons made between the different experimental conditions, the full list of which are available as appendices to this report. The graphics in this figure were generated with R, using the cummeRbund RNAseq data analysis package.*

The RNA analysis identified 2214 genes, with 4264 isoforms, 4188 transcription start sites, 3562 coding sequences and 61992 promoters within the genome over the 2 replicates of 8 conditions. There is a large amount of analysis still to take place on this extremely comprehensive dataset, which will take place over the coming months as our bioinformatics experts become more familiar with the data structure and utilise more of the features in the data. This RNA expression dataset details important information about the consortium strain, and is a useful complementary addition to the genome data of the consortium strain analysed by UM. This data will be passed to UM to be uploaded to the data warehouse.

In addition to the RNA analysis performed on the experimental set up, additional samples were collected to perform proteome analysis. As there are 8 samples with 2 replicates, the clearest sample combination approach was to run 2 separate iTRAQ experiments, with replicates being run on separate iTRAQs. The reason for this design is that previous experimental experience at USFD has shown that as long as there is data from similar conditions present in both iTRAQS to normalise the absolute values against; the scale of variation and direction of change of protein expression can be accounted for. The first iTRAQ experiment has been run, however as there is currently no replication data and therefore no variance, no statistical analysis has been performed on this data to date. In this first iTRAQ, 907 proteins were confidently identified and quantified, with 2 or more unique proteins (appendix). Integration of the proteomic and transcriptomic data is recognised in the literature as a challenging endeavour (*Haider and Pal, 2013*). When preparing the experimental design with IBMC, certain features such as concurrent sampling for RNA and protein from the same flasks, was agreed upon to make the data more suited to this task. The final output of this work will be presented as a publication in collaboration with IBMC.

6. Summary of final recommendations to the consortium:

A number of areas of interesting future research are highlighted in this report. Whilst there are still several datasets waiting to be run and analysed, due to a high volume of samples from the consortium arriving the last quarter of 2015, there are clear patterns for future investigation emerging from the data that we have analysed so far:

- Exposure to the sun over the course of the day builds up reductive stress and activates the proteomic profile for hydrogenase expression.
- Collected data suggests that timing the O_2 removal for after 5 pm following an aerobic growth phase during a sunny day should trigger significantly increased H2 production.
- Additionally, the cells have naturally lower levels of nitrogen as they are repairing the photosystems which coincides with a wealth of established literature on H_2 production.
- It would also be interesting to measure what levels of intracellular oxygen might be present with biological modelling of light vs dark phase growth.
- USFD recommend running the 1000 L PBR time course experiment with the *olive* strain by RUB to compare against WT and to provide important data on the scaled up growth of *olive* in outdoor conditions.
- The UU synthetic site doesn't appear to generate a systematic effect on *delta hox Synechocystis* without a hydrogenase apo-enzyme.
- The UU synthetic site produces a large number of systematic changes in *delta hox Synechocystis* with a hydrogenase apo-enzyme, suggesting the possibility of a phenotypic rescue. This suggests that the synthetic site may act as a suitable positive control for structurally accurate genetically engineered expression of the active site.

8.4 Neutral sites analysis report

Neutral site proteomics report

Mathematica data analysis by Andrew Landels

Experimental description

2 iTRAQ experiments were run as part of a single investigation into 5 neutral sites in the *Synechocystis* PCC6803 genome, denoted 5, 8, 10, 15 and 16. Controls in these experiments were the wild-type unmodified *Synechocystis* and a blank insert (at site 15). In all other cases GFP was inserted into the genome at the site of interest.

To link the two iTRAQ experiments together, the same WT *Synechocystis* samples were used in both cases. These experimental repeats are individually labelled in the experiment to differentiate them, this same differentiation is not done with any other samples.

The labels were assigned as follows:

label	iTRAQ a	iTRAQ b
113	aWT	bWT
114	aWT2	bWT2
115	a15-GFP	b5-GFP
116	a15-GFP	b5-GFP
117	a16-GFP	b8-GFP
118	a16-GFP	b8-GFP
119	a15-blank	b10-GFP
121	a15-blank	b10-GFP

Where a or b indicates the iTRAQ, the number indicates the site and the value following the hyphen indicates the treatment (GFP is shortened to G in the diagram labels).

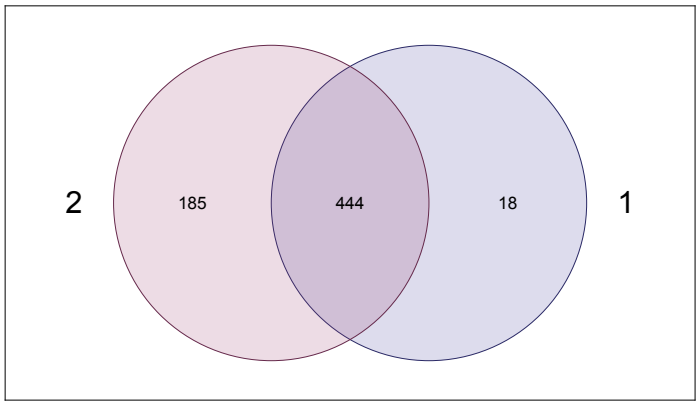
Standard proteome pipeline investigation

Initially, we analysed both iTRAQs individually with our proteome pipeline (Khoa et al. Proteomics. 2010 Sep;10(17):3130-41. doi: 10.1002/pmic.200900448) to generate a list of proteins that appeared to be significantly up or down regulated based on a P-value cut-off. Unfortunately, due to a loss of some experimental material from 1 of the iTRAQs (iTRAQ a) a direct comparison using the standard pipeline was difficult to acheive.

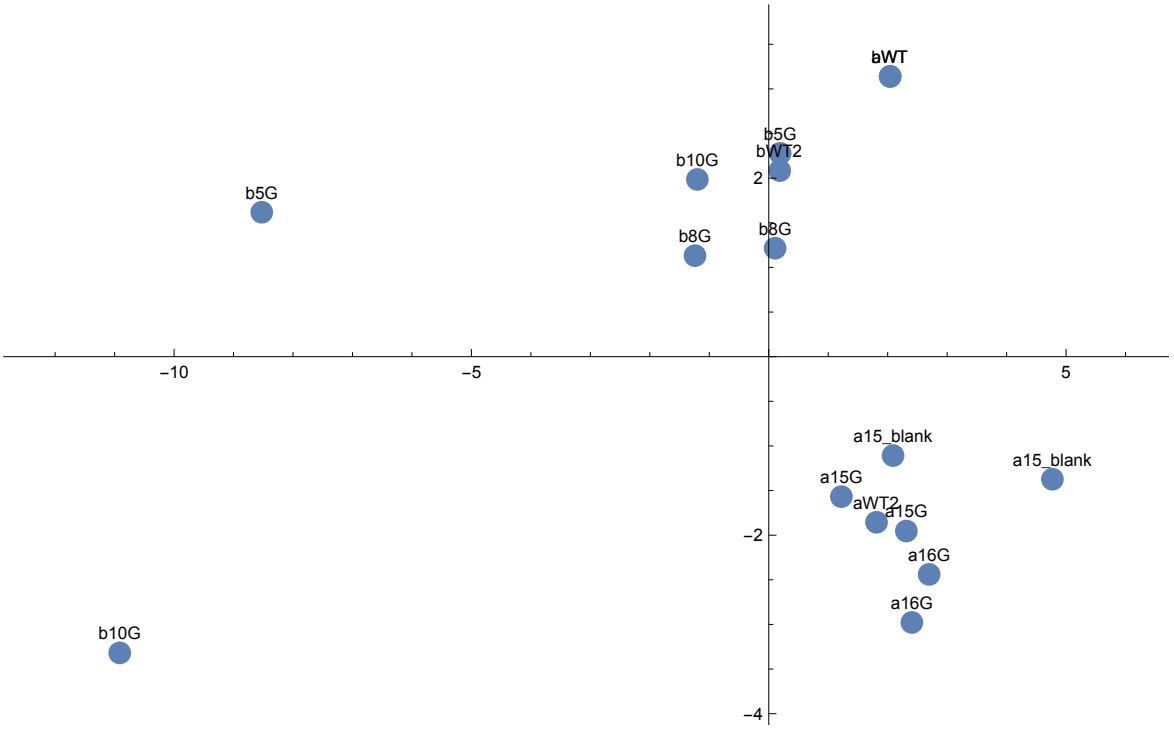
We produced a series of initial figures assessing the iTRAQs individually for relatedness and sent a list of proteins from signifiQuant which passed the statistical stringency test for being up or down regulated.

To generate a more informative analysis, we decided to investigate the effects of the combination of both iTRAQs in more detail, particularly from a clustering point of view. The first step was to work forward from the data that we had gathered from our proteomics pipeline.

(1) iTRAQ a 464 proteins (with 2 or more peptides)
(2) iTRAQ b 631 proteins (with 2 or more peptides)



We found an intersection of 444 proteins that had been successfully quantified by the standard method, so decided to solely use these proteins to assess clustering with PCA.



The values in this investigation were all generated relative to tag 113 (WT), which forms a joining point between the two datasets, however with the exception of the linking tag the two iTRAQs cluster separately. This is particularly notable for WT2 and is likely due to either the additional information present in iTRAQ b, or the manner in which the standard pipeline handles 0-intensity values.

As a result, it is difficult to draw clear conclusions about the features of the sites in general by meta-analysis of the terminal data alone.

Manually normalising the data

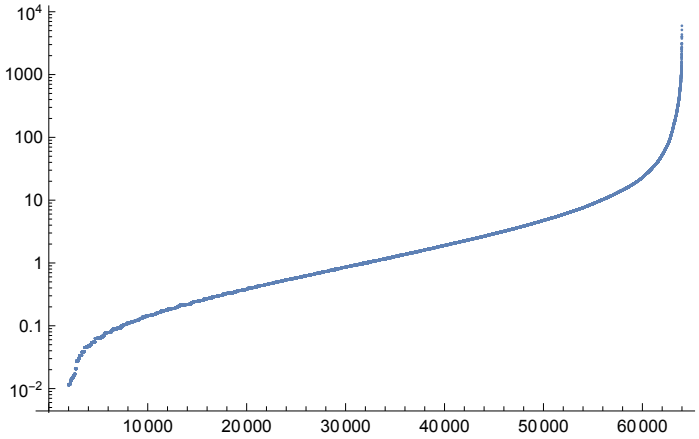
The first step was to look at the raw peptide information. The number of peptide spectral matches (PSMs) was compared.

iTRAQ	PSMs	Unique Sequences	Proteins
a	7999	1923	540
b	14814	3306	722

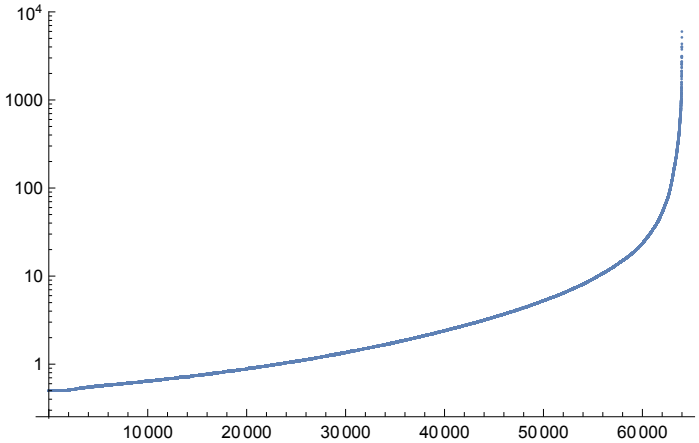
In total, **749** unique proteins had been identified with at least 1 PSM

We assessed both Channel sum - equalising the sum total intensity for each label - and median correction as normalisation methods for this data. Median correction was used as it produced a better ranking effect.

As certain proteins (notably GFP) were either below the level of detection or not present in all samples, a value α was added to all the label intensities. This has two benefits, firstly as we're interested in ratio data it removes 0s from the analysis which removes the infinite value (divide by 0) issue from the analysis. Secondly, as α is a fixed value it masks low-intensity labels which can contribute to abnormally large ratio readings. This is shown below for iTRAQ a, but the graph for iTRAQ b is almost identical (with the exception of the x-axis being almost double the range).



Prior to α addition, x-axis = ranked PSMs by reporter intensity, y-axis = intensity.



Following α addition. α is set as 0.5, which removes the inaccurate low-intensity readings whilst not affecting the high-intensity readings.

The next step was to select only high-quality proteins for comparative analysis. As a result, the datasets were filtered by peptides twice - only proteins that had both 3 or more peptides for quantification AND at least 2 unique sequence matches were retained. A total of **552** conidently identified and quantified unique proteins were found in the entire investigation, of which 365 overlapped between the two experiments.

iTRAQ	PSMs	Unique Sequences	Proteins	Intersecting Proteins
a	7656	1719	371	365
b	13327	3090	546	365

The intensity values were converted to natural log form and the geometric mean was calculated for each of the common proteins. The final step in normalisation was to generate ratios. To avoid having a single label generating all the ratios and therefore having all values on that label collapsed down to 1, all labels were divided by the mean of both the WT samples in each iTRAQ.

Cluster Analysis

The first step in assessing the normalised data is to quickly look at all the values to check consistency with a heatmap.

At a glance (not shown), the WT labels show very similar patterning which is a good sign. On close examination there is a checker - board type effect, indicating that the experimental samples are varying in similar directions rather than clustering by experimental replicate.

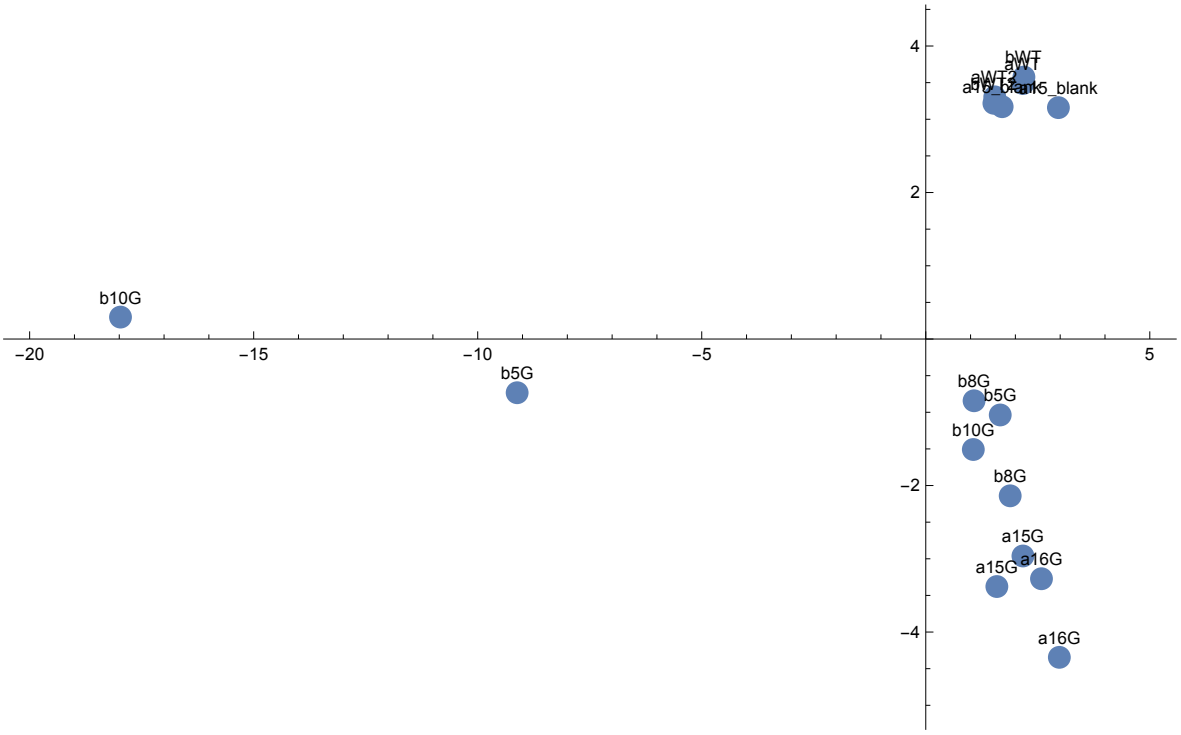
The test conditions also look like they have the correct patterning. The telltale sign of this is the GFP intensities (right - most column) which are visible for all samples that have GFP, but not for the WT (top 4 rows) or the blank (middle rows) controls. A second, larger version of this heatmap is provided as an appendix at the end of this report.

Over-interpretation of a heatmap can be misleading, however there do seem to be certain proteins in b5-GFP and b10-GFP that show similar patterning. This is clearer when using cluster tools such as a dendrogram or heatmap.

This dendrogram shows a clear relatedness between matching wild-type *Synechocystis* samples from the different iTRAQ experiments. It nicely demonstrates the amount of experimental variation (a vs b) as well as the experimental variation (WT vs WT2).

The GFP mutants all cluster clearly together away from the WT controls, although the blanks don't appear to be clearly separated in this graphic. The amount of experimental variation between paried samples for sites 8, 15 and 16 match the expected experimental variation (based on the WT controls). The variation beyond experimental variation for these samples appears to be minimal.

Notably, b5-GFP and b10-GFP have suppressed the diagram by being significantly different to all the other samples, reflecting the observation made earlier on the heatmap.

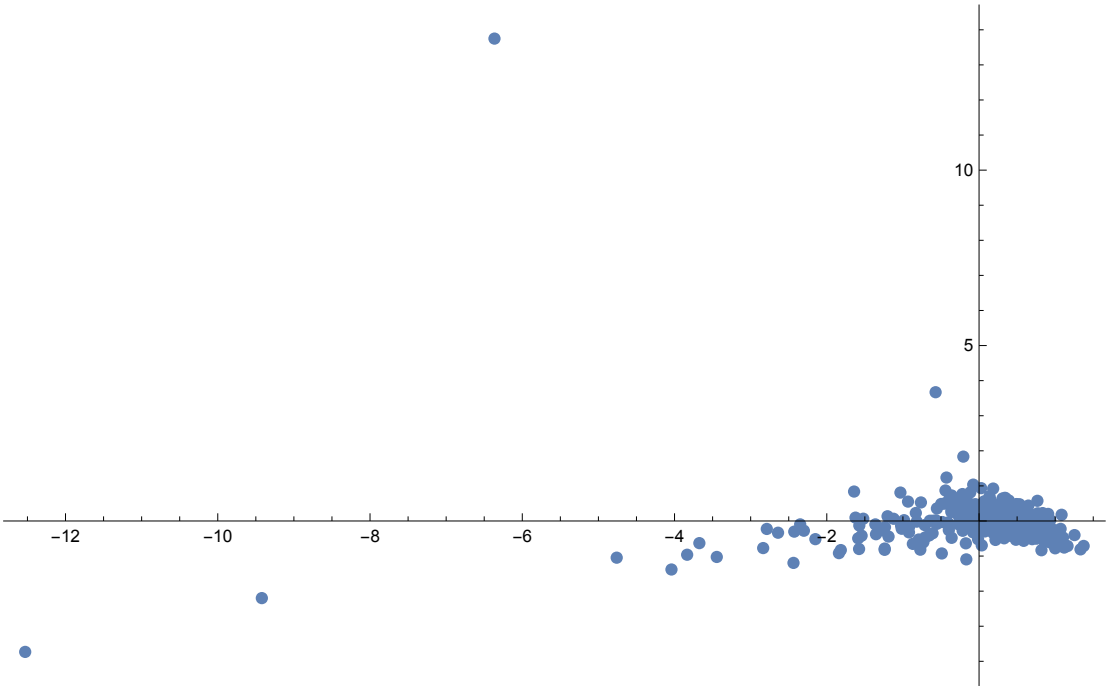


The pca plot clearly shows 3 clusters. For a simple interpretation of the axes, the y-axis relates to the amount of GFP detected, where lower values on the axis indicates an increased amount of GFP. The x-axis (responsible for the majority of the data variation) appears to be related to other background effects within the cell.

The top-right quadrant contains the control samples including the blank insertion control, the experimental-replicate clustering is clearer on the dendrogram as the samples are too crowded on the PCA, however it is interesting to note that the blank clusters closely to the WT, albeit with slightly more variation between the two experimental repeats than the unmodified cells.

The lower-right quadrant contains most of the GFP insertion mutants. The 15-blank cells show the same amount of x-axis spread-separation between the biological replicates as amongst all the clustered GFP-insertion mutants, suggesting that there are no major side-effects proteomically as a result of gene insertion into 3 of the neutral sites - 8, 15 and 16..

The left-half of the plot contains the final two samples, b10-GFP and b5-GFP. These show notable changes in the background variation axis, although as only an individual repeat has shown this effect in both cases it is difficult to assess statistically significant site-specific effects from the data we have available.



PCA plot of proteins. A version with annotation is in the appendix, however the labels are cluttered and difficult to read.

The protein PCA shows a relatively small number of proteins that stand out from the rest of the proteins identified in the experiment, however these don't shed much light on pathway-level changes withing the cell.

Here we've highlighted a couple of directions of change on the PCA, the trend towards the bottom left corner is purely centered around photosystems, energy and carbon fixation. The 3 proteins pushing away from the bulk of the proteins at a perpendicular angle towards the middle top are mostly uncharacterised and don't show any specific trend in function.

(top center)
P42212 = GFP

Direction 1 (Origin towards bot left, starting on the left)
P29254 = PS1
P29256 = PS1
m1m7g3 = phycobili prot
m1m7t6 = ribulose bisphosphate carboxylase
m1ml90 = fructose bisphosphate aldolase
m1lgt6 = atp synthase
m1mdr0 = uncharacterised aldehyde lyase
m1lzc3 = phycobili prot

Direction 2 (middle right, starting from the top)
m1mf82 = uncharacterised (hydrolase GO)
m1lhp6 = quinone oxireductase
F7URK1 = uncharacterised

Relative GFP levels

The table below shows the relative amounts of GFP that have been detected. It is important to note that these values have been calculated relative to a given α value and so values from non-GFP producing strains can be considered as the noise level. As a result, the variation between iTRAQ a and iTRAQ b make cross-iTRAQ investigation of these values impractical.

Each table contains two rows, the first is the Log of protein intensity, the second is the predicted ratio-linear values.

aWT	aWT2	a15G	a15G	a16G	a16G	a15_blank	a15
1.00393	0.996067	7.28006	6.80283	6.77149	8.11185	0.556368	0.8
2.72899	2.70761	1451.07	900.393	872.612	3333.73	1.74433	2.3

bWT	bWT2	b5G	b5G	b8G	b8G	b10G	b10G
0.98286	1.01714	5.01068	3.9141	5.99852	4.63505	5.25372	2.30489
2.67209	2.76528	150.007	50.104	402.832	103.033	191.276	10.0231

Sample a15-GFP seems to have the most consistent GFP production levels. Generally speaking, the protein production in neutral sites tends to vary by up to about a 4-fold amount between replicates. It is important to note that the outlying b10-GFP has much lower GFP intensity than its experimental counterpart, a 20-fold reduction with measured levels approaching the noise level.

b5-GFP production seems to be about half that of b8-GFP. This might be an indicator of genetic interference with protein production in these sites.

To generate more accurate data on these values, it might be worthwhile running a targeted proteomics experiment to get directly relatable spectral counts on all these proteins, coupled with some other form of measurement such as transcript analysis or fluorescence intensity.

Further work

These are things that I intended to do with this report, but had so much trouble re-working the data into a compatible form that I decided to send on what I had now just so you could see that we're making progress. Apologies again for the delay with this - I'm still learning a lot on the way! :-)

* Run signifiQuant with b10G and b5G vs WT to look for specific proteins that seem to be coming up in the heatmap.

* Re-work the α value to be relatively peptide specific to make the GFP values more stable. (I've done this on a previous analysis and I'm still undecided what the best way to go forward with it is)

* Generate a protein-labelled dendrogram-heatmap in the R, to look more closely at the protein-specific clusters in the data rather than just the label clustering effects. This might highlight protein families of interest that whilst not appearing on the signifiQuant analysis might make sense in GO terms or on a KEGG map.

Appendix - heatmap

The columns are (from left to right)
4x WT, 2x 15-GFP, 2x 16-GFP, 2x 15-blank, 2x 5-GFP, 2x 8-GFP, 2x 10-GFP

The outliers are the far-right column and 5th from the right.

List of Figures

List of Tables

Bibliography

Abdi, H. (2007). The bonferonni and šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, 3:103–107.

Abraham, P. E., Giannone, R. J., Xiong, W., and Hettich, R. L. (2014). Metaproteomics: extracting and mining proteome information to characterize metabolic activities in microbial communities. *Curr Protoc Bioinformatics*, 46:13 26 1–13 26 14.

Aggarwal, C. C., Wolf, J. L., Yu, P. S., Procopiuc, C., and Park, J. S. (1999). Fast algorithms for projected clustering. In *ACM SIGMoD Record*, volume 28, pages 61–72. ACM.

Ahrné, E., Glatter, T., Viganó, C., Schubert, C. v., Nigg, E. A., and Schmidt, A. (2016). Evaluation and improvement of quantification accuracy in isobaric mass tag-based protein quantification experiments. *Journal of Proteome Research*, 15(8):2537–2547.

Ahrne, E., Molzahn, L., Glatter, T., and Schmidt, A. (2013). Critical assessment of proteome-wide label-free absolute abundance estimation strategies. *Proteomics*, 13(17):2567–78.

Ajanovic, A. (2011). Biofuels versus food production: does biofuels production increase food prices? *Energy*, 36(4):2070–2076.

Allen, M. M. (1984). Cyanobacterial cell inclusions. *Annual Reviews in Microbiology*, 38(1):1–25.

Alonso-Gutierrez, J., Kim, E. M., Batth, T. S., Cho, N., Hu, Q., Chan, L. J., Petzold, C. J., Hillson, N. J., Adams, P. D., Keasling, J. D., Garcia Martin, H., and Lee, T. S. (2015). Principal component analysis of proteomics (pcap) as a tool to direct metabolic engineering. *Metab Eng*, 28:123–33.

Altelaar, A. F., Munoz, J., and Heck, A. J. (2013a). *Next-generation proteomics: towards an integrative view of proteome dynamics*, volume 14, pages 35–48. England.

Altelaar, A. M., Munoz, J., and Heck, A. J. (2013b). Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics*, 14(1):35–48.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402.

Anderson, N. L., Polanski, M., Pieper, R., Gatlin, T., Tirumalai, R. S., Conrads, T. P., Veenstra, T. D., Adkins, J. N., Pounds, J. G., Fagan, R., et al. (2004). The human plasma proteome a nonredundant list developed by combination of four separate sources. *Molecular & Cellular Proteomics*, 3(4):311–326.

Anemaet, I. G., Bekker, M., and Hellingwerf, K. J. (2010). Algal photosynthesis as the primary driver for a sustainable development in energy, feed, and food production. *Marine Biotechnology*, 12(6):619–629.

Anfelt, J., Hallström, B., Nielsen, J., Uhlén, M., and Hudson, E. P. (2013). Using transcriptomics to improve butanol tolerance of synechocystis sp. strain pcc 6803. *Applied and environmental microbiology*, 79(23):7419–7427.

Angermayr, S. A., van der Woude, A. D., Correddu, D., Vreugdenhil, A., Verrone, V., and Hellingwerf, K. J. (2014). *Exploring metabolic engineering design principles for the photosynthetic production of lactic acid by Synechocystis sp. PCC6803*, volume 7, page 99. England.

Appel, J., Phunpruch, S., Steinmuller, K., and Schulz, R. (2000). The bidirectional hydrogenase of *Synechocystis* sp. PCC 6803 works as an electron valve during photosynthesis. *Arch. Microbiol.*, 173(5-6):333–338.

Arike, L., Valgepea, K., Peil, L., Nahku, R., Adamberg, K., and Vilu, R. (2012). *Comparison and applications of label-free absolute proteome quantification methods on Escherichia coli*, volume 75, pages 5437–48. 2012 Elsevier B.V, Netherlands.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

Axmann, I. M., Kensche, P., Vogel, J., Kohl, S., Herzel, H., and Hess, W. R. (2005). Identification of cyanobacterial non-coding rnas by comparative genome analysis. *Genome biology*, 6(9):1.

Baebprasert, W., Jantaro, S., Khetkorn, W., Lindblad, P., and Incharoensakdi, A. (2011). Increased H2 production in the cyanobacterium *Synechocystis* sp. strain PCC 6803 by redirecting the electron supply via genetic engineering of the nitrate assimilation pathway. *Metab. Eng.*, 13(5):610–616.

Baebprasert, W., Lindblad, P., and Incharoensakdi, A. (2010). Response of h 2 production and hox-hydrogenase activity to external factors in the unicellular cyanobacterium synechocystis sp. strain pcc 6803. *international journal of hydrogen energy*, 35(13):6611–6616.

Bandhakavi, S., Stone, M. D., Onsongo, G., Van Riper, S. K., and Griffin, T. J. (2009). A dynamic range compression and three-dimensional peptide fractionation analysis platform expands proteome coverage and the diagnostic potential of whole saliva. *Journal of proteome research*, 8(12):5590–5600.

Bantscheff, M., Boesche, M., Eberhard, D., Matthieson, T., Sweetman, G., and Kuster, B. (2008). Robust and sensitive itraq quantification on an ltq orbitrap mass spectrometer. *Molecular & Cellular Proteomics*, 7(9):1702–1713.

Bantscheff, M., Lemeer, S., Savitski, M. M., and Kuster, B. (2012). Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and bioanalytical chemistry*, 404(4):939–965.

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. (2007). Quantitative mass spectrometry in proteomics: a critical review. *Analytical and bioanalytical chemistry*, 389(4):1017–1031.

Bardi, U. (2009). Peak oil: The four stages of a new idea. *Energy*, 34(3):323–326.

Batagelj, V. (1988). Generalized ward and related clustering problems. *Classification and related methods of data analysis*, pages 67–74.

Battchikova, N., Vainonen, J. P., Vorontsova, N., Keränen, M., Carmel, D., and Aro, E.-M. (2010). Dynamic changes in the proteome of synechocystis 6803 in response to co2 limitation revealed by quantitative proteomics. *Journal of proteome research*, 9(11):5896–5912.

Batth, T. S., Singh, P., Ramakrishnan, V. R., Sousa, M. M., Chan, L. J., Tran, H. M., Luning, E. G., Pan, E. H., Vuu, K. M., Keasling, J. D., Adams, P. D., and Petzold, C. J. (2014). A targeted proteomics toolkit for high-throughput absolute quantification of escherichia coli proteins. *Metab Eng*, 26C:48–56.

Beasley, V. R., Cook, W. O., Dahlem, A. M., Hooser, S. B., Lovell, R. A., and Valentine, W. M. (1989). Algae intoxication in livestock and waterfowl. *Veterinary Clinics of North America: Food Animal Practice*, 5(2):345–361.

Becker, J. and Wittmann, C. (2015). Advanced biotechnology: metabolically engineered cells for the bio-based production of chemicals and fuels, materials, and health-care products. *Angew Chem Int Ed Engl*, 54(11):3328–50.

Bellei, E., Bergamini, S., Monari, E., Fantoni, L. I., Cuoghi, A., Ozben, T., and Tomasi, A. (2011). High-abundance proteins depletion for serum proteomic analysis: concomitant removal of non-targeted proteins. *Amino acids*, 40(1):145–156.

Bentley, F. K., Zurbriggen, A., and Melis, A. (2014). Heterologous expression of the mevalonic acid pathway in cyanobacteria enhances endogenous carbon partitioning to isoprene. *Molecular plant*, 7(1):71–86.

Berg, J. M., Tymoczko, J., and Stryer, L. (2006). Biochemistry. *New York*.

Berg, P., Baltimore, D., Brenner, S., Roblin, R. O., and Singer, M. F. (1975). Summary statement of the asilomar conference on recombinant dna molecules. *Proceedings of the National Academy of Sciences*, 72(6):1981–1984.

Biemann, K. (1992). Mass spectrometry of peptides and proteins. *Annu. Rev. Biochem.*, 61:977–1010.

Blein-Nicolas, M. and Zivy, M. (2016). Thousand and one ways to quantify and compare protein abundances in label-free bottom-up proteomics. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1864(8):883–895.

Boaro, A. A., Kim, Y. M., Konopka, A. E., Callister, S. J., and Ahring, B. K. (2014). Integrated 'omics analysis for studying the microbial community response to a ph perturbation of a cellulose-degrading bioreactor culture. *FEMS Microbiol Ecol*, 90(3):802–15.

Bosma, R., de Vree, J., Slegers, P., Janssen, M., Wijffels, R., and Barbosa, M. (2014). Design and construction of the microalgal pilot facility algaeparc. *Algal Research*, 6:160–169.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical biochemistry*, 72(1-2):248–254.

Branco dos Santos, F., Du, W., and Hellingwerf, K. J. (2014). Synechocystis: not just a plug-bug for co2, but a green e. coli. *Frontiers in bioengineering and biotechnology*, 2:36.

Brenner, S., Jacob, F., and Meselson, M. (1961). An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, 190:576–581.

Brodbelt, J. S. (2015). Ion activation methods for peptides and proteins. *Analytical chemistry*, 88(1):30–51.

Burgard, A. P., Nikolaev, E. V., Schilling, C. H., and Maranas, C. D. (2004). Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome research*, 14(2):301–312.

Burrows, E. H., Chaplen, F. W., and Ely, R. L. (2008). Optimization of media nutrient composition for increased photofermentative hydrogen production by synechocystis sp. pcc 6803. *International Journal of Hydrogen Energy*, 33(21):6092–6099.

Burrows, E. H., Chaplen, F. W., and Ely, R. L. (2011). Effects of selected electron transport chain inhibitors on 24-h hydrogen production by *Synechocystis* sp. PCC 6803. *Bioresour. Technol.*, 102(3):3062–3070.

Burrows, E. H., Wong, W. K., Fern, X., Chaplen, F. W., and Ely, R. L. (2009). Optimization of pH and nitrogen for enhanced hydrogen production by *Synechocystis* sp. PCC 6803 via statistical and machine learning methods. *Biotechnol. Prog.*, 25(4):1009–1017.

Cai, T., Ge, X., Park, S. Y., and Li, Y. (2013). Comparison of synechocystis sp. pcc6803 and nannochloropsis salina for lipid production using artificial seawater and nutrients from anaerobic digestion effluent. *Bioresource technology*, 144:255–260.

Camsund, D., Lindblad, P., and Jaramillo, A. (2011). Genetically engineered light sensors for control of bacterial gene expression. *Biotechnology journal*, 6(7):826–836.

Casado-Vela, J., Martínez-Esteso, M. J., Rodriguez, E., Borrás, E., Elortza, F., and Bru-Martínez, R. (2010). itraq-based quantitative analysis of protein mixtures with large fold change and dynamic range. *Proteomics*, 10(2):343–347.

Chan, K. C., Lucas, D. A., Hise, D., Schaefer, C. F., Xiao, Z., Janini, G. M., Buetow, K. H., Issaq, H. J., Veenstra, T. D., and Conrads, T. P. (2004). Serum/plasma proteome. *Clinical Proteomics*, 1(1):101–225.

Chardonnet, S., Sakr, S., Cassier-Chauvat, C., Le Marechal, P.chal, P., Chauvat, F., Lemaire, S. D., and Decottignies, P. (2014). First proteomic study of s-glutathionylation in cyanobacteria. *Journal of proteome research*, 14(1):59–71.

Chen, G., Qu, S., Wang, Q., Bian, F., Peng, Z., Zhang, Y., Ge, H., Yu, J., Xuan, N., Bi, Y., et al. (2014a). Transgenic expression of delta-6 and delta-15 fatty acid desaturases enhances omega-3 polyunsaturated fatty acid accumulation in synechocystis sp. pcc6803. *Biotechnology for biofuels*, 7(1):1.

Chen, L., Wu, L., Wang, J., and Zhang, W. (2014b). Butanol tolerance regulated by a two-component response regulator slr1037 in photosynthetic synechocystis sp. pcc 6803. *Biotechnology for biofuels*, 7(1):1.

Chisti, Y. (2007). Biodiesel from microalgae. *Biotechnology advances*, 25(3):294–306.

Chisti, Y. (2010). Fuels from microalgae. *Biofuels*, 1(2):233–235.

Chisti, Y. (2013). Constraints to commercialization of algal fuels. *Journal of biotechnology*, 167(3):201–214.

Chisti, Y. and Yan, J. (2011). Energy from algae: current status and future trends: algal biofuels–a status report. *Applied Energy*, 88(10):3277–3279.

Chiverton, L. M., Evans, C., Pandhal, J., Landels, A. R., Rees, B. J., Levison, P. R., Wright, P. C., and Smales, C. M. (2016). Quantitative definition and monitoring of the host cell protein proteome using itraq–a study of an industrial mab producing cho-s cell line. *Biotechnology journal*.

Chong, P. K., Gan, C. S., Pham, T. K., and Wright, P. C. (2006). Isobaric tags for relative and absolute quantitation (itraq) reproducibility: Implication of multiple injections. *Journal of proteome research*, 5(5):1232–1240.

Christoforou, A. and Lilley, K. S. (2011). Taming the isobaric tagging elephant in the room in quantitative proteomics. *Nature methods*, 8(11):911–913.

Christoforou, A. L. and Lilley, K. S. (2012). Isobaric tagging approaches in quantitative proteomics: the ups and downs. *Anal Bioanal Chem*, 404(4):1029–37.

Ciferri, O. (1983). Spirulina, the edible microorganism. *Microbiological reviews*, 47(4):551.

Cohen, S. N., Chang, A. C., Boyer, H. W., and Helling, R. B. (1973). Construction of biologically functional bacterial plasmids in vitro. *Proceedings of the National Academy of Sciences*, 70(11):3240–3244.

Cojocari, D. (2016). Amino acids.

Collins, M. O., Yu, L., and Choudhary, J. S. (2007). Analysis of protein phosphorylation on a proteome-scale. *Proteomics*, 7(16):2751–2768.

Cologna, S. M., Crutchfield, C. A., Searle, B. C., Blank, P. S., Toth, C. L., Ely, A. M., Picache, J. A., Backlund, P. S., Wassif, C. A., Porter, F. D., et al. (2015). An efficient approach to evaluate reporter ion behavior from maldi-ms/ms data for quantification studies using isobaric tags. *Journal of proteome research*, 14(10):4169–4178.

Colyer, C. L., Kinkade, C. S., Viskari, P. J., and Landers, J. P. (2005). Analysis of cyanobacterial pigments and proteins by electrophoretic and chromatographic methods. *Anal Bioanal Chem*, 382(3):559–569.

Comelli, A. (2012). The green gold rush.

Compton, S. J. and Jones, C. G. (1985). Mechanism of dye response and interference in the bradford protein assay. *Analytical biochemistry*, 151(2):369–374.

Couto, N., Evans, C., Pandhal, J., Qiu, W., Pham, T., Noirel, J., and Wright, P. (2013). Making sense out of the proteome: the utility of iTRAQ and TMT. *Methods in Molecular Biology*.

Cox, J., Hein, M. Y., Luber, C. A., Paron, I., Nagaraj, N., and Mann, M. (2014). Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed maxlfq. *Molecular & Cellular Proteomics*, 13(9):2513–2526.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

Crick, F. H. (1968). The origin of the genetic code. *J. Mol. Biol.*, 38(3):367–379.

CyanoFactory (2012). Cyanofactory website.

Dahan, O., Gingold, H., and Pilpel, Y. (2011). Regulatory mechanisms and networks couple the different phases of gene expression. *Trends Genet.*, 27(8):316–322.

DeSouza, L. V., Taylor, A. M., Li, W., Minkoff, M. S., Romaschin, A. D., Colgan, T. J., and Siu, K. M. (2008). Multiple reaction monitoring of mtraq-labeled peptides enables absolute quantification of endogenous levels of a potential cancer marker in cancerous and normal endometrial tissues. *Journal of proteome research*, 7(8):3525–3534.

Dienst, D., Georg, J., Abts, T., Jakorew, L., Kuchmina, E., Börner, T., Wilde, A., Dühring, U., Enke, H., and Hess, W. R. (2014). Transcriptomic response to prolonged ethanol production in the cyanobacterium synechocystis sp. pcc6803. *Biotechnology for biofuels*, 7(1):1.

Doshi, A., Pascoe, S., Coglan, L., and Rainey, T. J. (2016). Economic and policy issues in the production of algae-based biofuels: A review. *Renewable and sustainable energy reviews*, 64:329–337.

Dost, B., Bandeira, N., Li, X., Shen, Z., Briggs, S. P., and Bafna, V. (2012). Accurate mass spectrometry based protein quantification via shared peptides. *Journal of Computational Biology*, 19(4):337–348.

Dowle, A. A., Wilson, J., and Thomas, J. R. (2016). Comparing the diagnostic classification accuracy of itraq, peak-area, spectral-counting, and empai methods for relative quantification in expression proteomics. *Journal of Proteome Research*, 15(10):3550–3562.

Drabik, D. (2011). The theory of biofuel policy and food grain prices. *Charles H. Dyson School of Applied Economics and Management Working Paper*, (2011-20).

Du, W., Liang, F., Duan, Y., Tan, X., and Lu, X. (2013). Exploring the photosynthetic production capacity of sucrose by cyanobacteria. *Metabolic engineering*, 19:17–25.

Ducat, D. C., Way, J. C., and Silver, P. A. (2011). Engineering cyanobacteria to generate high-value products. *Trends in biotechnology*, 29(2):95–103.

Duffy, J. E., Canuel, E. A., Adey, W., and Swaddle, J. P. (2009). Biofuels: algae. *Science*, 326(5958):1345–1345.

Duret, L. (2002). Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.*, 12(6):640–649.

Eberle, U., Müller, B., and von Helmolt, R. (2012). Fuel cell electric vehicles and hydrogen infrastructure: status 2012. *Energy & Environmental Science*, 5(10):8780–8798.

Eberle, U. and von Helmolt, R. (2016). Gm hydrogen4–a fuel cell electric vehicle based on the chevrolet equinox. *Fuel Cells: Data, Facts, and Figures*.

Eckert, C., Boehm, M., Carrieri, D., Yu, J., Dubini, A., Nixon, P. J., and Maness, P. C. (2012). Genetic analysis of the Hox hydrogenase in the cyanobacterium *Synechocystis* sp. PCC 6803 reveals subunit roles in association, assembly, maturation, and function. *J. Biol. Chem.*, 287(52):43502–43515.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461.

Energy Information Administration, U. (2013). Global energy usage data.

Englund, E., Pattanaik, B., Ubhayasekera, S. J., Stensjo, K., Bergquist, J., and Lindberg, P. (2014). *Production of squalene in Synechocystis sp. PCC 6803*, volume 9, page e90270. United States.

Escobar, J. C., Lora, E. S., Venturini, O. J., Yáñez, E. E., Castillo, E. F., and Almazan, O. (2009). Biofuels: environment, technology and food security. *Renewable and sustainable energy reviews*, 13(6):1275–1287.

Eurostat (2016). *Energy balance sheets, 2014 data: 2016 edition*. Luxembourg: Publications Office of the European Union.

Evans, C., Landels, A., Ow, S., Noirel, J., Couto, N., and Wright, P. (2013). Application of the iTRAQ workflows for peptide-based relative quantification of proteins. *submitted to Methods in Molecular Biology*.

Evans, C., Noirel, J., Ow, S. Y., Salim, M., Pereira-Medrano, A. G., Couto, N., Pandhal, J., Smith, D., Pham, T. K., Karunakaran, E., et al. (2012). An insight into itraq: where do we stand now? *Analytical and bioanalytical chemistry*, 404(4):1011–1027.

Evans, R. W. and Kates, M. (1984). Lipid composition of halophilic species of dunaliella from the dead sea. *Archives of microbiology*, 140(1):50–56.

Fischer, C. R. and Schaffer, S. (2014). Editorial overview: chemical biotechnology: the expansion of chemical biotechnology. *Curr Opin Biotechnol*, 30:v–vii.

Fischer, G., Hizsnyik, E., Prieler, S., Shah, M., and van Velthuizen, H. (2009). Biofuels and food security: Implications of an accelerated biofuels production.

Friedrich, B., Fritsch, J., and Lenz, O. (2011). Oxygen-tolerant hydrogenases in hydrogen-based technologies. *Current opinion in biotechnology*, 22(3):358–364.

Fröhlich, C. J. (2016). Climate migrants as protestors? dispelling misconceptions about global environmental change in pre-revolutionary syria. *Contemporary levant*, 1(1):38–50.

Fu, J. and Xu, X. (2006). The functional divergence of two glgp homologues in synechocystis sp. pcc 6803. *FEMS microbiology letters*, 260(2):201–209.

Fulda, S., Mikkat, S., Huang, F., Huckauf, J., Marin, K., Norling, B., and Hagemann, M. (2006). Proteome analysis of salt stress response in the cyanobacterium synechocystis sp. strain pcc 6803. *Proteomics*, 6(9):2733–2745.

Galinato, G. I. and Yoder, J. K. (2010). An integrated tax-subsidy policy for carbon emission reduction. *Resource and Energy Economics*, 32(3):310–326.

Gan, C. S., Chong, P. K., Pham, T. K., and Wright, P. C. (2007). Technical, experimental, and biological variations in isobaric tags for relative and absolute quantitation (itraq). *Journal of proteome research*, 6(2):821–827.

Gan, C. S., Reardon, K. F., and Wright, P. C. (2005). Comparison of protein and peptide prefractionation methods for the shotgun proteomic analysis of synechocystis sp. pcc 6803. *Proteomics*, 5(9):2468–2478.

Gao, L., Pei, G., Chen, L., and Zhang, W. (2015). A global network-based protocol for functional inference of hypothetical proteins in synechocystis sp. pcc 6803. *Journal of microbiological methods*, 116:44–52.

Gao, Q., Wang, W., Zhao, H., and Lu, X. (2012). Effects of fatty acid activation on photosynthetic production of fatty acid-based biofuels in synechocystis sp. pcc6803. *Biotechnology for biofuels*, 5(1):1.

Gao, Y., Xiong, W., Li, X.-b., Gao, C.-F., Zhang, Y.-l., Li, H., and Wu, Q.-y. (2009). Identification of the proteomic changes in synechocystis sp. pcc 6803 following prolonged uv-b irradiation. *Journal of experimental botany*, 60(4):1141–1154.

Gasparatos, A., Stromberg, P., Takeuchi, K., et al. (2013). Sustainability impacts of first-generation biofuels. *Animal Frontiers*, 3(2):12–26.

Gavrilescu, M. and Chisti, Y. (2005). Biotechnology—a sustainable alternative for chemical industry. *Biotechnology advances*, 23(7):471–499.

Gehrig, P. M., Hunziker, P. E., Zahariev, S., and Pongor, S. (2004). Fragmentation pathways of n g-methylated and unmodified arginine residues in peptides studied by esi-ms/ms and maldi-ms. *Journal of the American Society for Mass Spectrometry*, 15(2):142–149.

George, K. W., Alonso-Gutierrez, J., Keasling, J. D., and Lee, T. S. (2015). Isoprenoid drugs, biofuels, and chemicals-artemisinin, farnesene, and beyond. *Adv Biochem Eng Biotechnol*.

George, K. W., Chen, A., Jain, A., Batth, T. S., Baidoo, E. E., Wang, G., Adams, P. D., Petzold, C. J., Keasling, J. D., and Lee, T. S. (2014). Correlation analysis of targeted proteins and metabolites to assess and engineer microbial isopentenol production. *Biotechnology and bioengineering*, 111(8):1648–1658.

Germer, F., Zebger, I., Saggu, M., Lendzian, F., Schulz, R., and Appel, J. (2009). Overexpression, isolation, and spectroscopic characterization of the bidirectional [NiFe] hydrogenase from *Synechocystis* sp. PCC 6803. *J. Biol. Chem.*, 284(52):36462–36472.

Gingold, H., Dahan, O., and Pilpel, Y. (2012). Dynamic changes in translational efficiency are deduced from codon usage of the transcriptome. *Nucleic Acids Res.*, 40(20):10053–10063.

Glen, A., Gan, C. S., Hamdy, F. C., Eaton, C. L., Cross, S. S., Catto, J. W., Wright, P. C., and Rehman, I. (2008). itraq-facilitated proteomic analysis of human prostate cancer cells identifies proteins associated with progression. *Journal of proteome research*, 7(3):897–907.

Gluck, F., Hoogland, C., Antinori, P., Robin, X., Nikitin, F., Zufferey, A., Pasquarello, C., Fétaud, V., Dayon, L., Müller, M., et al. (2013). Easyprot—an easy-to-use graphical platform for proteomics data analysis. *Journal of proteomics*, 79:146–160.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333–351.

Griese, M., Lange, C., and Soppa, J. (2011). Ploidy in cyanobacteria. *FEMS microbiology letters*, 323(2):124–131.

Guo, J., Nguyen, A. Y., Dai, Z., Su, D., Gaffrey, M. J., Moore, R. J., Jacobs, J. M., Monroe, M. E., Smith, R. D., Koppenaal, D. W., et al. (2014). Proteome-wide light/dark modulation of thiol oxidation in cyanobacteria revealed by quantitative site-specific redox proteomics. *Molecular & Cellular Proteomics*, 13(12):3270–3285.

Gupta, R., Wang, Y., Agrawal, G. K., Rakwal, R., Jo, I. H., Bang, K. H., and Kim, S. T. (2015). Time to dig deep into the plant proteome: a hunt for low-abundance proteins. *Front Plant Sci*, 6:22.

Gutekunst, K., Phunpruch, S., Schwarz, C., Schuchardt, S., Schulz-Friedrich, R., and Appel, J. (2005). LexA regulates the bidirectional hydrogenase in the cyanobacterium *Synechocystis* sp. PCC 6803 as a transcription activator. *Mol. Microbiol.*, 58(3):810–823.

Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000). *Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology*, volume 97, pages 9390–5. United States.

Haldane, J. B. S. (1929). The origin of life. *Rationalist Annual*, 148:3–10.

Hall, D. O. and Rao, K. K. (1999). *Photosynthesis*. Cambridge University Press in association with the Institute of Biology, Cambridge.

He, Z., Huang, T., Liu, X., Zhu, P., Teng, B., and Deng, S. (2016). Protein inference: A protein quantification perspective. *Computational biology and chemistry*.

Heidorn, T., Camsund, D., Huang, H. H., Lindberg, P., Oliveira, P., Stensjo, K., and Lindblad, P. (2011). Synthetic biology in cyanobacteria engineering and analyzing novel functions. *Meth. Enzymol.*, 497:539–579.

Hernández-Prieto, M. A., Semeniuk, T. A., and Futschik, M. E. (2014). Toward a systems-level understanding of gene regulatory, protein interaction, and metabolic networks in cyanobacteria. *Frontiers in genetics*, 5.

Hernández-Prieto, M. A., Semeniuk, T. A., Giner-Lamia, J., and Futschik, M. E. (2016). The transcriptional landscape of the photosynthetic model cyanobacterium synechocystis sp. pcc6803. *Scientific reports*, 6:22168.

Hill, E. G., Schwacke, J. H., Comte-Walters, S., Slate, E. H., Oberg, A. L., Eckel-Passow, J. E., Therneau, T. M., and Schey, K. L. (2008). A statistical model for itraq data analysis. *Journal of proteome research*, 7(8):3091–3101.

Hill, J., Nelson, E., Tilman, D., Polasky, S., and Tiffany, D. (2006). Environmental, economic, and energetic costs and benefits of biodiesel and ethanol biofuels. *Proceedings of the National Academy of sciences*, 103(30):11206–11210.

Himmelblau, D. M. and Riggs, J. B. (2012). *Basic principles and calculations in chemical engineering*. FT Press.

Hirano, A., Ueda, R., Hirayama, S., and Ogushi, Y. (1997). Co 2 fixation and ethanol production with microalgal photosynthesis and intracellular anaerobic fermentation. *Energy*, 22(2):137–142.

Huang, H.-H., Camsund, D., Lindblad, P., and Heidorn, T. (2010). Design and characterization of molecular tools for a synthetic biology approach towards developing cyanobacterial biotechnology. *Nucleic acids research*, 38(8):2577–2593.

Huang, H.-H. and Lindblad, P. (2013). Wide-dynamic-range promoters engineered for cyanobacteria. *J. of Biol. Eng.*, 7(1):10.

Huang, L., Kim, D., Liu, X., Myers, C. R., and Locasale, J. W. (2014). *Estimating relative changes of metabolic fluxes*, volume 10, page e1003958. United States.

Huang, S., Chen, L., Te, R., Qiao, J., Wang, J., and Zhang, W. (2013). Complementary itraq proteomics and rna-seq transcriptomics reveal multiple levels of regulation in response to nitrogen starvation in synechocystis sp. pcc 6803. *Molecular BioSystems*, 9(10):2565–2574.

Huber, M. L., Sacco, R., Parapatics, K., Skucha, A., Khamina, K., Muller, A. C., Rudashevskaya, E. L., and Bennett, K. L. (2014). abfasp-ms: affinity-based filter-aided sample preparation mass spectrometry for quantitative analysis of chemically labeled protein complexes. *J Proteome Res*, 13(2):1147–55.

Hunt, D. F., Yates, J. R., Shabanowitz, J., Winston, S., and Hauer, C. R. (1986). Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences*, 83(17):6233–6237.

Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. (2005). Exponentially modified protein abundance index (empai) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Molecular & Cellular Proteomics*, 4(9):1265–1272.

Janković, V. and Schultz, D. M. (2016). Atmosfear: Communicating the effects of climate change on extreme weather. *Weather, Climate, and Society*, (2016).

Jers, C., Soufi, B., Grangeasse, C., Deutscher, J., and Mijakovic, I. (2008). Phosphoproteomics in bacteria: towards a systemic understanding of bacterial phosphorylation networks. *Expert review of proteomics*, 5(4):619–627.

Jin, S., Daly, D. S., Springer, D. L., and Miller, J. H. (2007). The effects of shared peptides on protein quantitation in label-free proteomics by lc/ms/ms. *Journal of proteome research*, 7(01):164–169.

Johnson, K. A. and Goody, R. S. (2011). The original michaelis constant: translation of the 1913 michaelis–menten paper. *Biochemistry*, 50(39):8264–8269.

Joseph, A., Aikawa, S., Sasaki, K., Matsuda, F., Hasunuma, T., and Kondo, A. (2014a). Increased biomass production and glycogen accumulation in apce gene deleted synechocystis sp. pcc 6803. *AMB Express*, 4(1):1.

Joseph, A., Aikawa, S., Sasaki, K., Teramura, H., Hasunuma, T., Matsuda, F., Osanai, T., Hirai, M. Y., and Kondo, A. (2014b). Rre37 stimulates accumulation of 2-oxoglutarate and glycogen under nitrogen starvation in synechocystis sp. pcc 6803. *FEBS letters*, 588(3):466–471.

Kalb, V. F. and Bernlohr, R. W. (1977). A new spectrophotometric assay for protein in cell extracts. *Analytical biochemistry*, 82(2):362–371.

Kämäräinen, J., Knoop, H., Stanford, N. J., Guerrero, F., Akhtar, M. K., Aro, E.-M., Steuer, R., and Jones, P. R. (2012). Physiological tolerance and stoichiometric potential of cyanobacteria for hydrocarbon fuel production. *Journal of biotechnology*, 162(1):67–74.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., and Tabata, S. (1996). Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions (supplement). *DNA Res.*, 3(3):185–209.

Karp, N. A., Huber, W., Sadowski, P. G., Charles, P. D., Hester, S. V., and Lilley, K. S. (2010). Addressing accuracy and precision issues in itraq quantitation. *Molecular & Cellular Proteomics*, 9(9):1885–1897.

Karthic, P. and Joseph, S. (2012). Comparison and limitations of biohydrogen production processes. *Research Journal of Biotechnology Vol*, 7:2.

Kasavi, C., Eraslan, S., Arga, K. Y., Oner, E. T., and Kirdar, B. (2014). *A system based network approach to ethanol tolerance in Saccharomyces cerevisiae*, volume 8, page 90. England.

Kazusa, Z. (2016). Cyanobase.

Keilhauer, E. C., Hein, M. Y., and Mann, M. (2015). *Accurate protein complex retrieval by affinity enrichment mass spectrometry (AE-MS) rather than affinity purification mass spectrometry (AP-MS)*, volume 14, pages 120–35. 2015 by The American Society for Biochemistry and Molecular Biology, Inc., United States.

Kelchtermans, P., Bittremieux, W., De Grave, K., Degroeve, S., Ramon, J., Laukens, K., Valkenborg, D., Barsnes, H., and Martens, L. (2014). Machine learning applications in proteomics research: how the past can boost the future. *Proteomics*, 14(4-5):353–66.

Keller, M. A., Turchyn, A. V., and Ralser, M. (2014). Non-enzymatic glycolysis and pentose phosphate pathway-like reactions in a plausible archean ocean. *Molecular systems biology*, 10(4):725.

Keshamouni, V. G., Michailidis, G., Grasso, C. S., Anthwal, S., Strahler, J. R., Walker, A., Arenberg, D. A., Reddy, R. C., Akulapalli, S., Thannickal, V. J., et al. (2006). Differential protein expression profiling by itraq-2dlc-ms/ms of lung cancer cells undergoing epithelial-mesenchymal transition reveals a migratory/invasive phenotype. *Journal of proteome research*, 5(5):1143–1154.

King, Z. A., Lloyd, C. J., Feist, A. M., and Palsson, B. O. (2015). Next-generation genome-scale models for metabolic engineering. *Current opinion in biotechnology*, 35:23–29.

Kirst, G. (1990). Salinity tolerance of eukaryotic marine algae. *Annual review of plant biology*, 41(1):21–53.

Kiss, E., Kos, P. B., and Vass, I. (2009). Transcriptional regulation of the bidirectional hydrogenase in the cyanobacterium *Synechocystis* 6803. *J. Biotechnol.*, 142(1):31–37.

Klaubauf, S., Narang, H. M., Post, H., Zhou, M., Brunner, K., Mach-Aigner, A. R., Mach, R. L., Heck, A. J., Altelaar, A. F., and de Vries, R. P. (2014). Similar is not the same: differences in the function of the (hemi-)cellulolytic regulator xlnr (xlr1/xyr1) in filamentous fungi. *Fungal Genet Biol*, 72:73–81.

Klein-Marcuschamer, D., Chisti, Y., Benemann, J. R., and Lewis, D. (2013). A matter of detail: Assessing the true potential of microalgal biofuels. *Biotechnol. Bioeng.*

Kopečná, J., Komenda, J., Bučinská, L., and Sobotka, R. (2012). Long-term acclimation of the cyanobacterium synechocystis sp. pcc 6803 to high light is accompanied by an enhanced production of chlorophyll that is preferentially channeled to trimeric photosystem i. *Plant physiology*, 160(4):2239–2250.

Kopf, M., Klähn, S., Scholz, I., Matthiessen, J. K., Hess, W. R., and Voß, B. (2014). Comparative analysis of the primary transcriptome of synechocystis sp. pcc 6803. *DNA Research*, page dsu018.

Korinko, P., Scogin, J., and Clarck, E. (2001). Development of aluminide coatings for hydrogen isotope permeation resistance. *Tsukaba, Japan: Tritium*.

Krey, J. F., Wilmarth, P. A., Shin, J. B., Klimek, J., Sherman, N. E., Jeffery, E. D., Choi, D., David, L. L., and Barr-Gillespie, P. G. (2014). Accurate label-free protein quantitation with high- and low-resolution mass spectrometers. *J Proteome Res*, 13(2):1034–44.

Kudoh, K., Kawano, Y., Hotta, S., Sekine, M., Watanabe, T., and Ihara, M. (2014). Prerequisite for highly efficient isoprenoid production by cyanobacteria discovered through the over-expression of 1-deoxy-d-xylulose 5-phosphate synthase and carbon allocation analysis. *Journal of bioscience and bioengineering*, 118(1):20–28.

Kurian, D., Phadwal, K., and Maenpaa, P. (2006). Proteomic characterization of acid stress response in synechocystis sp. pcc 6803. *Proteomics*, 6(12):3614–3624.

Kuzyk, M. A., Ohlund, L. B., Elliott, M. H., Smith, D., Qian, H., Delaney, A., Hunter, C. L., and Borchers, C. H. (2009). A comparison of ms/ms-based, stable-isotope-labeled, quantitation performance on esi-quadrupole tof and maldi-tof/tof mass spectrometers. *Proteomics*, 9(12):3328–3340.

Kwon, J.-H., Bernát, G., Wagner, H., Rögner, M., and Rexroth, S. (2013). Reduced light-harvesting antenna: consequences on cyanobacterial metabolism and photosynthetic productivity. *Algal Research*, 2(3):188–195.

Labarre, J., Chauvat, F., and Thuriaux, P. (1989). Insertional mutagenesis by random cloning of antibiotic resistance genes into the genome of the cyanobacterium *Synechocystis* strain PCC 6803. *J. Bacteriol.*, 171(6):3449–3457.

Landels, A., Evans, C., Noirel, J., and Wright, P. C. (2015). Advances in proteomics for production strain analysis. *Current opinion in biotechnology*, 35:111–117.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Latimer, L. N., Lee, M. E., Medina-Cleghorn, D., Kohnz, R. A., Nomura, D. K., and Dueber, J. E. (2014). Employing a combinatorial expression approach to characterize xylose utilization in saccharomyces cerevisiae. *Metab Eng*, 25:20–9.

Lee, D.-G., Kwon, J., Eom, C.-Y., Kang, Y.-M., Roh, S. W., Lee, K.-B., and Choi, J.-S. (2015). Directed analysis of cyanobacterial membrane phosphoproteome using stained phosphoproteins and titanium-enriched phosphopeptides. *Journal of Microbiology*, 53(4):279–287.

Leeper, F. (2000). Biosynthesis: aromatic polyketides and vitamins. *Berlin7 Springer*.

Lenz, O., Ludwig, M., Schubert, T., Bürstel, I., Ganskow, S., Goris, T., Schwarze, A., and Friedrich, B. (2010). H2 conversion in the presence of o2 as performed by the membrane-bound [nife]-hydrogenase of ralstonia eutropha. *ChemPhysChem*, 11(6):1107–1119.

Lesur, A., Ancheva, L., Kim, Y. J., Berchem, G., van Oostrum, J., and Domon, B. (2015). Screening protein isoforms predictive for cancer using immuno-affinity capture and fast lc-ms in prm mode. *Proteomics Clin Appl*.

Li, T., Yang, H.-M., Cui, S.-X., Suzuki, I., Zhang, L.-F., Li, L., Bo, T.-T., Wang, J., Murata, N., and Huang, F. (2011). Proteomic study of the impact of hik33 mutation in synechocystis sp. pcc 6803 under normal and salt stress conditions. *Journal of proteome research*, 11(1):502–514.

Li, Z., Adams, R. M., Chourey, K., Hurst, G. B., Hettich, R. L., and Pan, C. (2012). Systematic comparison of label-free, metabolic labeling, and isobaric chemical labeling for quantitative proteomics on ltq orbitrap velos. *Journal of proteome research*, 11(3):1582–1590.

Liese, A., Seelbach, K., and Wandrey, C. (2006). *Industrial biotransformations*. John Wiley & Sons.

Lindberg, P., Park, S., and Melis, A. (2010). Engineering a platform for photosynthetic isoprene production in cyanobacteria, using synechocystis as the model organism. *Metabolic engineering*, 12(1):70–79.

Lindblad, P., Lindberg, P., Oliveira, P., Stensjö, K., and Heidorn, T. (2012). Design, engineering, and construction of photosynthetic microbial cell factories for renewable solar fuel production. *Ambio*, 41(2):163–168.

Liu, J., Chen, L., Wang, J., Qiao, J., and Zhang, W. (2012). Proteomic analysis reveals resistance mechanism against biofuel hexane in synechocystis sp. pcc 6803. *Biotechnology for biofuels*, 5(1):1.

Liu, X., Brune, D., Vermaas, W., and Curtiss, R. (2010). Production and secretion of fatty acids in genetically engineered cyanobacteria. *Proceedings of the National Academy of Sciences*.

Lorenz, R. T. and Cysewski, G. R. (2000). Commercial potential for haematococcus microalgae as a natural source of astaxanthin. *Trends in biotechnology*, 18(4):160–167.

Lowry, O. H., Rosebrough, N. J., Farr, A. L., Randall, R. J., et al. (1951). Protein measurement with the folin phenol reagent. *J biol Chem*, 193(1):265–275.

Machado, I. M. and Atsumi, S. (2012). Cyanobacterial biofuel production. *J Biotechnol*, 162(1):50–6.

Maeda, T., Vardar, G., Self, W. T., and Wood, T. K. (2007). Inhibition of hydrogen uptake in escherichia coli by expressing the hydrogenase from the cyanobacterium synechocystis sp. pcc 6803. *BMC biotechnology*, 7(1):1.

Mahoney, D. W., Therneau, T. M., Heppelmann, C. J., Higgins, L., Benson, L. M., Zenka, R. M., Jagtap, P., Nelsestuen, G. L., Bergen III, H. R., and Oberg, A. L. (2011). Relative quantification: characterization of bias, variability and fold changes in mass spectrometry data from itraq-labeled peptides. *Journal of proteome research*, 10(9):4325–4333.

Maier, T., Schmidt, A., Guell, M., Kuhner, S., Gavin, A. C., Aebersold, R., and Serrano, L. (2011). Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Mol. Syst. Biol.*, 7:511.

Makarov, A. (2000). Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal. Chem.*, 72(6):1156–1162.

Mann, M., Kulak, N. A., Nagaraj, N., and Cox, J. (2013). The coming age of complete, accurate, and ubiquitous proteomes. *Mol Cell*, 49(4):583–90.

Martinez-Val, A., Garcia, F., Ximénez-Embún, P., Ibarz, N., Zarzuela, E., Ruppen, I., Mohammed, S., and Munoz, J. (2016). On the statistical significance of compressed ratios in isobaric labeling: A cross-platform comparison. *Journal of Proteome Research*, 15(9):3029–3038.

McAlister, G. C., Huttlin, E. L., Haas, W., Ting, L., Jedrychowski, M. P., Rogers, J. C., Kuhn, K., Pike, I., Grothe, R. A., Blethrow, J. D., and Gygi, S. P. (2012). Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Anal. Chem.*, 84(17):7469–7478.

McCrone, A., Moslener, U., d'Estais, F., Usher, E., and Grüning, C. (2016). Global trends in renewable energy investment 2016. *Bloomberg New Energy Finance*, (2016).

McDowell, G. S., Gaun, A., and Steen, H. (2013). ifasp: combining isobaric mass tagging with filter-aided sample preparation. *J Proteome Res*, 12(8):3809–12.

McIntosh, C. L., Germer, F., Schulz, R., Appel, J., and Jones, A. K. (2011). The [NiFe]-hydrogenase of the cyanobacterium *Synechocystis* sp. PCC 6803 works bidirectionally with a bias to H2 production. *J. Am. Chem. Soc.*, 133(29):11308–11319.

Merrill, A. E., Hebert, A. S., MacGilvray, M. E., Rose, C. M., Bailey, D. J., Bradley, J. C., Wood, W. W., El Masri, M., Westphall, M. S., Gasch, A. P., and Coon, J. J. (2014). *NeuCode labels for relative protein quantification*, volume 13, pages 2503–12. 2014 by The American Society for Biochemistry and Molecular Biology, Inc., United States.

Metz, B., Davidson, O., De Coninck, H., Loos, M., Meyer, L., et al. (2005). Carbon dioxide capture and storage.

Mikkat, S., Fulda, S., and Hagemann, M. (2014a). A 2d gel electrophoresis-based snapshot of the phosphoproteome in the cyanobacterium synechocystis sp. strain pcc 6803. *Microbiology*, 160(2):296–306.

Mikkat, S., Fulda, S., and Hagemann, M. (2014b). *A 2D gel electrophoresis-based snapshot of the phosphoproteome in the cyanobacterium Synechocystis sp. strain PCC 6803*, volume 160, pages 296–306. England.

Minogue, C. E., Hebert, A. S., Rensvold, J. W., Westphall, M. S., Pagliarini, D. J., and Coon, J. J. (2015). Multiplexed quantification for data-independent acquisition. *Anal Chem*, 87(5):2570–5.

Miranda, H., Cheregi, O., Netotea, S., Hvidsten, T. R., Moritz, T., and Funk, C. (2013). Co-expression analysis, proteomic and metabolomic study on the impact of a deg/htra protease triple mutant in synechocystis sp. pcc 6803 exposed to temperature and high light stress. *Journal of proteomics*, 78:294–311.

Mirsaleh-Kohan, N., Robertson, W. D., and Compton, R. N. (2008). Electron ionization time-of-flight mass spectrometry: historical review and current applications. *Mass Spectrom Rev*, 27(3):237–285.

Mischerikow, N., van Nierop, P., Li, K. W., Bernstein, H.-G., Smit, A. B., Heck, A. J., and Altelaar, A. M. (2010). Gaining efficiency by parallel quantification and identification of itraq-labeled peptides using hcd and decision tree guided cid/etd on an ltq orbitrap. *Analyst*, 135(10):2643–2652.

Mitchell, D. (2008). A note on rising food prices. *World Bank Policy Research Working Paper Series, Vol.*

Mo, R., Yang, M., Chen, Z., Cheng, Z., Yi, X., Li, C., He, C., Xiong, Q., Chen, H., Wang, Q., et al. (2015). Acetylome analysis reveals the involvement of lysine acetylation in photosynthesis and carbon metabolism in the model cyanobacterium synechocystis sp. pcc 6803. *Journal of proteome research*, 14(2):1275–1286.

Monod, J. (1949). The growth of bacterial cultures. *Annual Reviews in Microbiology*, 3(1):371–394.

Montagud, A., Gamermann, D., Fernández de Córdoba, P., and Urchueguía, J. (2013). Synechocystis sp. pcc6803 metabolic models for the enhanced production of biofuels. *Crit. Rev. Biotechnol*, 8551:1–15.

Montagud, A., Gamermann, D., Fernández de Córdoba, P., and Urchueguía, J. F. (2015). Synechocystis sp. pcc6803 metabolic models for the enhanced production of hydrogen. *Critical reviews in biotechnology*, 35(2):184–198.

Montagud, A., Navarro, E., de Córdoba, P. F., Urchueguía, J. F., and Patil, K. R. (2010). Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium. *BMC systems biology*, 4(1):156.

Morra, S., Arizzi, M., Allegra, P., La Licata, B., Sagnelli, F., Zitella, P., Gilardi, G., and Valetti, F. (2014). Expression of different types of [fefe]-hydrogenase genes in bacteria isolated from a population of a bio-hydrogen pilot-scale plant. *international journal of hydrogen energy*, 39(17):9018–9027.

Mota, R., Pereira, S. B., Meazzini, M., Fernandes, R., Santos, A., Evans, C. A., De Philippis, R., Wright, P. C., and Tamagnini, P. (2015). Effects of heavy metals on cyanothece sp. ccy 0110 growth, extracellular polymeric substances (eps) production, ultrastructure and protein profiles. *Journal of proteomics*, 120:75–94.

Mur, L., Gons, H., and Van Liere, L. (1977). Some experiments on the competition between green algae and blue-green bacteria in light-limited environments. *FEMS Microbiology Letters*, 1(6):335–338.

Neilson, K. A., Mariani, M., and Haynes, P. A. (2011). Quantitative proteomic analysis of cold-responsive proteins in rice. *Proteomics*, 11(9):1696–1706.

Nel, A. J., Garnett, S., Blackburn, J. M., and Soares, N. C. (2015). Comparative reevaluation of fasp and enhanced fasp methods by lc-ms/ms. *J Proteome Res*, 14(3):1637–42.

Niall, H. D. (1973). Automated edman degradation: The protein sequenator. In C. H. W. Hirs, S. N. T., editor, *Part D: Enzyme Structure*, volume 27 of *Methods in Enzymology*, pages 942 – 1010. Academic Press.

Nickelsen, J., Rengstl, B., Stengel, A., Schottkowski, M., Soll, J., and Ankele, E. (2011). Biogenesis of the cyanobacterial thylakoid membrane system–an update. *FEMS microbiology letters*, 315(1):1–5.

Noirel, J., Evans, C., Salim, M., Mukherjee, J., Yen Ow, S., Pandhal, J., Khoa Pham, T., A Biggs, C., and C Wright, P. (2011). Methods in quantitative proteomics: setting itraq on the right track. *Current Proteomics*, 8(1):17–30.

Noirel, J., Sanguinetti, G., and Wright, P. C. (2008). Identifying differentially expressed subnetworks with mmg. *Bioinformatics*, 24(23):2792–2793.

Oberg, A. L., Mahoney, D. W., Eckel-Passow, J. E., Malone, C. J., Wolfinger, R. D., Hill, E. G., Cooper, L. T., Onuma, O. K., Spiro, C., Therneau, T. M., et al. (2008). Statistical analysis of relative labeled mass spectrometry data from complex samples using anova. *Journal of proteome research*, 7(1):225–233.

Oliveira, P. and Lindblad, P. (2008). An AbrB-Like protein regulates the expression of the bidirectional hydrogenase in *Synechocystis* sp. strain PCC 6803. *J. Bacteriol.*, 190(3):1011–1019.

Olsen, J. V., Ong, S. E., and Mann, M. (2004). Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell Proteomics*, 3(6):608–614.

Olson, J. M. (2006). Photosynthesis in the Archean era. *Photosyn. Res.*, 88(2):109–117.

Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell Proteomics*, 1(5):376–386.

Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What is flux balance analysis? *Nature biotechnology*, 28(3):245–248.

Oswald, W. J. (1988). Large-scale algal culture systems (engineering aspects). *Micro-algal biotechnology. Cambridge University Press, Cambridge*, pages 357–394.

Ow, S. Y., Salim, M., Noirel, J., Evans, C., Rehman, I., and Wright, P. C. (2009). itraq underestimation in simple and complex mixtures:"the good, the bad and the ugly". *Journal of proteome research*, 8(11):5347–5355.

Ow, S. Y., Salim, M., Noirel, J., Evans, C., Wright, P., et al. (2011). Minimising itraq ratio compression through understanding lc-ms elution dependence and high-resolution hilic fractionation. *Proteomics*, 11(11):2341–2346.

Ow, S. Y. and Wright, P. C. (2009). Current trends in high throughput proteomics in cyanobacteria. *FEBS letters*, 583(11):1744–1752.

Pan, C. and Banfield, J. F. (2014). Quantitative metaproteomics: functional insights into microbial communities. *Methods Mol Biol*, 1096:231–40.

Pandhal, J. and Noirel, J. (2014). Synthetic microbial ecosystems for biotechnology. *Biotechnol Lett*, 36(6):1141–51.

Pascovici, D., Song, X., Solomon, P. S., Winterberg, B., Mirzaei, M., Goodchild, A., Stanley, W. C., Liu, J., and Molloy, M. P. (2015). Combining protein ratio p-values as a pragmatic approach to the analysis of multirun itraq experiments. *Journal of proteome research*, 14(2):738–746.

Pate, R., Klise, G., and Wu, B. (2011). Resource demand implications for us algae biofuels production scale-up. *Applied Energy*, 88(10):3377–3388.

Pei, G., Sun, T., Chen, S., Chen, L., and Zhang, W. (2017). Systematic and functional identification of small non-coding rnas associated with exogenous biofuel stress in cyanobacterium synechocystis sp. pcc 6803. *Biotechnology for Biofuels*, 10(1):57.

Pei, L., Gaisser, S., and Schmidt, M. (2011). Synthetic biology in the view of european public funding organisations. *Public Understanding of Science*, page 0963662510393624.

Pichler, P., Kőcher, T., Holzmann, J., Mazanek, M., Taus, T., Ammerer, G., and Mechtler, K. (2010). Peptide labeling with isobaric tags yields higher identification rates using itraq 4-plex compared to tmt 6-plex and itraq 8-plex on ltq orbitrap. *Analytical chemistry*, 82(15):6549–6558.

Picotti, P., Bodenmiller, B., Mueller, L. N., Domon, B., and Aebersold, R. (2009). Full dynamic range proteome analysis of s. cerevisiae by targeted proteomics. *Cell*, 138(4):795–806.

Pierce, A., Unwin, R. D., Evans, C. A., Griffiths, S., Carney, L., Zhang, L., Jaworska, E., Lee, C.-F., Blinco, D., Okoniewski, M. J., et al. (2008). Eight-channel itraq enables comparison of the activity of six leukemogenic tyrosine kinases. *Molecular & cellular proteomics*, 7(5):853–863.

Pinto, F., Pacheco, C. C., Oliveira, P., Montagud, A., Landels, A., Couto, N., Wright, P. C., Urchueguía, J. F., and Tamagnini, P. (2015). Improving a synechocystis-based photoautotrophic chassis through systematic genome mapping and validation of neutral sites. *DNA Research*, page dsv024.

Pinto, F., van Elburg, K. A., Pacheco, C. C., Lopo, M., Noirel, J., Montagud, A., Urchueguia, J. F., Wright, P. C., and Tamagnini, P. (2012a). Construction of a chassis for hydrogen production: physiological and molecular characterization of a *Synechocystis* sp. PCC 6803 mutant lacking a functional bidirectional hydrogenase. *Microbiology (Reading, Engl.)*, 158(Pt 2):448–464.

Pinto, F., van Elburg, K. A., Pacheco, C. C., Lopo, M., Noirel, J., Montagud, A., Urchueguia, J. F., Wright, P. C., and Tamagnini, P. (2012b). *Construction of a chassis for hydrogen production: physiological and molecular characterization of a Synechocystis sp. PCC 6803 mutant lacking a functional bidirectional hydrogenase*, volume 158, pages 448–64. England.

Pinto, F. L., Thapper, A., Sontheim, W., and Lindblad, P. (2009). Analysis of current and alternative phenol based rna extraction methodologies for cyanobacteria. *BMC molecular biology*, 10(1):1.

Pottiez, G., Wiederin, J., Fox, H. S., and Ciborowski, P. (2012). Comparison of 4-plex to 8-plex itraq quantitative measurements of proteins in human plasma samples. *Journal of proteome research*, 11(7):3774–3781.

Qiao, J., Wang, J., Chen, L., Tian, X., Huang, S., Ren, X., and Zhang, W. (2012a). Quantitative itraq lc–ms/ms proteomics reveals metabolic responses to biofuel ethanol in cyanobacterial synechocystis sp. pcc 6803. *Journal of proteome research*, 11(11):5286–5300.

Qiao, J., Wang, J., Chen, L., Tian, X., Huang, S., Ren, X., and Zhang, W. (2012b). Quantitative itraq lc-ms/ms proteomics reveals metabolic responses to biofuel ethanol in cyanobacterial synechocystis sp. pcc 6803. *J Proteome Res*, 11(11):5286–300.

Rauniyar, N. and Yates III, J. R. (2014). Isobaric labeling-based relative quantification in shotgun proteomics. *Journal of proteome research*, 13(12):5293–5309.

Raynie, D. E. (2006). Modern extraction techniques. *Analytical chemistry*, 78(12):3997–4004.

Redding-Johanson, A. M., Batth, T. S., Chan, R., Krupa, R., Szmidt, H. L., Adams, P. D., Keasling, J. D., Lee, T. S., Mukhopadhyay, A., and Petzold, C. J. (2011). Targeted proteomics for metabolic pathway optimization: application to terpene production. *Metabolic engineering*, 13(2):194–203.

Reinsvold, R. E., Jinkerson, R. E., Radakovits, R., Posewitz, M. C., and Basu, C. (2011). The production of the sesquiterpene β-caryophyllene in a transgenic strain of the cyanobacterium synechocystis. *Journal of plant physiology*, 168(8):848–852.

Ren, N., Li, J., Li, B., Wang, Y., and Liu, S. (2006). Biohydrogen production from molasses by anaerobic fermentation with a pilot-scale bioreactor system. *International Journal of Hydrogen Energy*, 31(15):2147–2157.

Ren, Q., Shi, M., Chen, L., Wang, J., and Zhang, W. (2014). Integrated proteomic and metabolomic characterization of a novel two-component response regulator slr1909 involved in acid tolerance in synechocystis sp. pcc 6803. *Journal of proteomics*, 109:76–89.

Rögner, M. (2013). Metabolic engineering of cyanobacteria for the production of hydrogen from water. *Biochemical Society Transactions*, 41(5):1254–1259.

Romano, A. and Conway, T. (1996). Evolution of carbohydrate metabolic pathways. *Research in microbiology*, 147(6):448–455.

Rosegrant, M. W. (2008). *Biofuels and grain prices: impacts and policy responses*. International Food Policy Research Institute Washington, DC.

Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., et al. (2004a). Multiplexed protein quantitation in saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & cellular proteomics*, 3(12):1154–1169.

Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., and Pappin, D. J. (2004b). Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Mol. Cell Proteomics*, 3(12):1154–1169.

Rowland, J. G., Pang, X., Suzuki, I., Murata, N., Simon, W. J., and Slabas, A. R. (2010). Identification of components associated with thermal acclimation of photosystem ii in synechocystis sp. pcc6803. *PLoS One*, 5(5):e10511.

Saffers, J.-B., Makarov, D., and Molkov, V. (2011). Modelling and numerical simulation of permeated hydrogen dispersion in a garage with adiabatic walls and still air. *international journal of hydrogen energy*, 36(3):2582–2588.

Saha, R., Verseput, A. T., Berla, B. M., Mueller, T. J., Pakrasi, H. B., and Maranas, C. D. (2012). Reconstruction and comparison of the metabolic potential of cyanobacteria cyanothece sp. atcc 51142 and synechocystis sp. pcc 6803. *PloS one*, 7(10):e48285.

Saifuddin, N. and Priatharsini, P. (2016). Developments in bio-hydrogen production from algae: A review.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467.

Savage, D. F., Way, J., and Silver, P. A. (2008). Defossiling fuel: how synthetic biology can transform biofuel production. *ACS Chem. Biol.*, 3(1):13–16.

Savakis, P. and Hellingwerf, K. J. (2015). Engineering cyanobacteria for direct biofuel production from co 2. *Current opinion in biotechnology*, 33:8–14.

Scheer, H. and Zhao, K.-H. (2008). Biliprotein maturation: the chromophore attachment. *Molecular microbiology*, 68(2):263–276.

Schellenberg, J. J., Verbeke, T. J., McQueen, P., Krokhin, O. V., Zhang, X., Alvare, G., Fristensky, B., Thallinger, G. G., Henrissat, B., Wilkins, J. A., Levin, D. B., and Sparling, R. (2014). *Enhanced whole genome sequence and annotation of Clostridium stercorarium DSM8532T using RNA-seq transcriptomics and high-throughput proteomics*, volume 15, page 567. England.

Schindler, D. W. (2006). Recent advances in the understanding and management of eutrophication. *Limnology and Oceanography*, 51(1):356–363.

Schmidt, M. (2012). The green gold rush.

Schneider, D., Fuhrmann, E., Scholz, I., Hess, W. R., and Graumann, P. L. (2007). Fluorescence staining of live cyanobacterial cells suggest non-stringent chromosome segregation and absence of a connection between cytoplasmic and thylakoid membranes. *BMC Cell Biol.*, 8:39.

Schreiber, S. (2000). Biosynthesis: aromatic polyketides, isoprenoids, alkaloids. *Berlin7 Springer*.

Schuergers, N., Lenn, T., Kampmann, R., Meissner, M. V., Esteves, T., Temerinac-Ott, M., Korvink, J. G., Lowe, A. R., Mullineaux, C. W., and Wilde, A. (2016). Cyanobacteria use micro-optics to sense light direction. *Elife*, 5:e12620.

Schultze, M., Forberich, B., Rexroth, S., Dyczmons, N. G., Roegner, M., and Appel, J. (2009). Localization of cytochrome b6f complexes implies an incomplete respiratory chain in cytoplasmic membranes of the cyanobacterium *Synechocystis* sp. PCC 6803. *Biochim. Biophys. Acta*, 1787(12):1479–1485.

Schwacke, J. H., Hill, E. G., Krug, E. L., Comte-Walters, S., and Schey, K. L. (2009). iquantitator: a tool for protein expression inference using itraq. *BMC bioinformatics*, 10(1):1.

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342.

Schwartz, J. C., Senko, M. W., and Syka, J. E. (2002). A two-dimensional quadrupole ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.*, 13(6):659–669.

Schwarz, R. and Forchhammer, K. (2005). Acclimation of unicellular cyanobacteria to macronutrient deficiency: emergence of a complex network of cellular responses. *Microbiology*, 151(8):2503–2514.

Sengupta, T., Bhushan, M., and Wangikar, P. P. (2013). Metabolic modeling for multi-objective optimization of ethanol production in a synechocystis mutant. *Photosynthesis research*, 118(1-2):155–165.

Shastri, A. A. and Morgan, J. A. (2007). A transient isotopic labeling methodology for ↑ metabolic flux analysis of photoautotrophic microorganisms. *Phytochemistry*, 68(16-18):2302–2312.

Shliaha, P. V., Jukes-Jones, R., Christoforou, A., Fox, J., Hughes, C., Langridge, J., Cain, K., and Lilley, K. S. (2014). Additional precursor purification in isobaric mass tagging experiments by traveling wave ion mobility separation (twims). *Journal of proteome research*, 13(7):3360–3369.

Singh, A. K. and Sherman, L. A. (2005). Pleiotropic effect of a histidine kinase on carbohydrate metabolism in synechocystis sp. strain pcc 6803 and its requirement for heterotrophic growth. *Journal of bacteriology*, 187(7):2368–2376.

Singh, P., Batth, T. S., Juminaga, D., Dahl, R. H., Keasling, J. D., Adams, P. D., and Petzold, C. J. (2012). Application of targeted proteomics to metabolically engineered escherichia coli. *Proteomics*, 12(8):1289–1299.

Slabas, A. R., Suzuki, I., Murata, N., Simon, W. J., and Hall, J. J. (2006). Proteomic analysis of the heat shock response in synechocystis pcc6803 and a thermally tolerant knockout strain lacking the histidine kinase 34 gene. *Proteomics*, 6(3):845–864.

Smith, P., Krohn, R. I., Hermanson, G., Mallia, A., Gartner, F., Provenzano, M., Fujimoto, E., Goeke, N., Olson, B., and Klenk, D. (1985). Measurement of protein using bicinchoninic acid. *Analytical biochemistry*, 150(1):76–85.

Song, Z., Chen, L., Wang, J., Lu, Y., Jiang, W., and Zhang, W. (2014). A transcriptional regulator sll0794 regulates tolerance to biofuel ethanol in photosynthetic synechocystis sp. pcc 6803. *Molecular & Cellular Proteomics*, 13(12):3519–3532.

Soufi, B., Jers, C., Hansen, M. E., Petranovic, D., and Mijakovic, I. (2008). Insights from site-specific phosphoproteomics in bacteria. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1784(1):186–192.

Soufi, B., Krug, K., Harst, A., and Macek, B. (2015a). Characterization of the e. coli proteome and its modifications during growth and ethanol stress. *Front Microbiol*, 6:103.

Soufi, B., Krug, K., Harst, A., and Macek, B. (2015b). Characterization of the e. coli proteome and its modifications during growth and ethanol stress. *Regulatory potential of post-translational modifications in bacteria*, page 65.

Spät, P., Macek, B., and Forchhammer, K. (2015). Phosphoproteome of the cyanobacterium synechocystis sp. pcc 6803 and its dynamics during nitrogen starvation. *Regulatory potential of post-translational modifications in bacteria*, page 20.

Stanier, R. Y., Kunisawa, R., Mandel, M., and Cohen-Bazire, G. (1971). Purification and properties of unicellular blue-green algae (order Chroococcales). *Bacteriol Rev*, 35(2):171–205.

Steen, H. and Mann, M. (2004). The ABCs (and XYZs) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.*, 5(9):699–711.

Stephens, E., Ross, I. L., King, Z., Mussgnug, J. H., Kruse, O., Posten, C., Borowitzka, M. A., and Hankamer, B. (2010). An economic and technical evaluation of microalgal biofuels. *Nature biotechnology*, 28(2):126–128.

Stodilka, D., Kherani, N., Shmayda, W., and Thorpe, S. (2000). A tritium tracer technique for the measurement of hydrogen permeation in polymeric materials. *International journal of hydrogen energy*, 25(11):1129–1136.

Stoscheck, C. M. (1990). Quantitation of protein. *Methods in enzymology*, 182:50–68.

Strohl, W. R. (1997). *Biotechnology of antibiotics*. M. Dekker.

Suzuki, I., Simon, W. J., and Slabas, A. R. (2006). The heat shock response of synechocystis sp. pcc 6803 analysed by transcriptomics and proteomics. *Journal of experimental botany*, 57(7):1573–1578.

Talamantes, T., Ughy, B., Domonkos, I., Kis, M., Gombos, Z., and Prokai, L. (2014). Label-free lc–ms/ms identification of phosphatidylglycerol-regulated proteins in synechocystis sp. pcc6803. *Proteomics*, 14(9):1053–1057.

Tamagnini, P., Axelsson, R., Lindberg, P., Oxelfelt, F., Wunschiers, R., and Lindblad, P. (2002). Hydrogenases and hydrogen metabolism of cyanobacteria. *Microbiol. Mol. Biol. Rev.*, 66(1):1–20.

Tamagnini, P., Leitão, E., Oliveira, P., Ferreira, D., Pinto, F., Harris, D. J., Heidorn, T., and Lindblad, P. (2007). Cyanobacterial hydrogenases: diversity, regulation and applications. *FEMS microbiology reviews*, 31(6):692–720.

Thompson, A., Schafer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Johnstone, R., Mohammed, A. K., and Hamon, C. (2003). Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Anal. Chem.*, 75(8):1895–1904.

Thompson, P. B. (2012). Synthetic biology needs a synthetic bioethics. *Ethics, Policy & Environment*, 15(1):1–20.

Tian, X., Chen, L., Wang, J., Qiao, J., and Zhang, W. (2013a). Quantitative proteomics reveals dynamic responses of synechocystis sp. pcc 6803 to next-generation biofuel butanol. *Journal of proteomics*, 78:326–345.

Tian, X., Chen, L., Wang, J., Qiao, J., and Zhang, W. (2013b). Quantitative proteomics reveals dynamic responses of synechocystis sp. pcc 6803 to next-generation biofuel butanol. *J Proteomics*, 78:326–45.

Ting, L., Rad, R., Gygi, S. P., and Haas, W. (2011). Ms3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nature methods*, 8(11):937–940.

Toepel, J., Illmer-Kephalides, M., Jaenicke, S., Straube, J., May, P., Goesmann, A., and Kruse, O. (2013). New insights into *Chlamydomonas reinhardtii* hydrogen production processes by combined microarray/RNA-seq transcriptomics. *Plant Biotechnol. J.*

Touloupakis, E., Cicchi, B., Benavides, A. M. S., and Torzillo, G. (2016). Effect of high ph on growth of synechocystis sp. pcc 6803 cultures and their contamination by golden algae (poterioochromonas sp.). *Applied microbiology and biotechnology*, 100(3):1333–1341.

Tran, H.-L., Hong, S.-J., and Lee, C.-G. (2009). Evaluation of extraction methods for recovery of fatty acids from botryococcus braunii lb 572 and synechocystis sp. pcc 6803. *Biotechnology and Bioprocess Engineering*, 14(2):187–192.

Tributsch, H. (2008). Photovoltaic hydrogen generation. *International journal of hydrogen energy*, 33(21):5911–5930.

Trotschel, C., Albaum, S. P., Wolff, D., Schroder, S., Goesmann, A., Nattkemper, T. W., and Poetsch, A. (2012). *Protein turnover quantification in a multilabeling approach: from data calculation to evaluation*, volume 11, pages 512–26. United States.

Vanwonterghem, I., Jensen, P. D., Ho, D. P., Batstone, D. J., and Tyson, G. W. (2014). Linking microbial community structure, interactions and function in anaerobic digesters using new molecular techniques. *Curr Opin Biotechnol*, 27:55–64.

Varman, A. M., Xiao, Y., Pakrasi, H. B., and Tang, Y. J. (2013). Metabolic engineering of synechocystis sp. strain pcc 6803 for isobutanol production. *Applied and environmental microbiology*, 79(3):908–914.

Vatsala, T., Raj, S. M., and Manimaran, A. (2008). A pilot-scale study of biohydrogen production from distillery effluent using defined bacterial co-culture. *International journal of hydrogen energy*, 33(20):5404–5415.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). The sequence of the human genome. *science*, 291(5507):1304–1351.

Vignais, P. M. and Billoud, B. (2007). Occurrence, classification, and biological function of hydrogenases: an overview. *Chemical reviews*, 107(10):4206–4272.

von Helmolt, R. and Eberle, U. (2014). Compressed and liquid hydrogen for fuel cell vehicles. *Encyclopedia of Applied Electrochemistry*, pages 245–253.

Walsby, A. (1972). Structure and function of gas vacuoles. *Bacteriological reviews*, 36(1):1.

Walsby, A. E. (1981). Cyanobacteria: planktonic gas-vacuolate forms. In *The prokaryotes*, pages 224–235. Springer.

Wang, H., Alvarez, S., and Hicks, L. M. (2011). Comprehensive comparison of itraq and label-free lc-based quantitative proteomics approaches using two chlamydomonas reinhardtii strains of interest for biofuels engineering. *Journal of proteome research*, 11(1):487–501.

Wang, J., Chen, L., Huang, S., Liu, J., Ren, X., Tian, X., Qiao, J., and Zhang, W. (2012a). RNA-seq based identification and mutant validation of gene targets related to ethanol resistance in cyanobacterial *Synechocystis* sp. PCC 6803. *Biotechnol Biofuels*, 5(1):89.

Wang, J., Chen, L., Huang, S., Liu, J., Ren, X., Tian, X., Qiao, J., and Zhang, W. (2012b). Rna-seq based identification and mutant validation of gene targets related to ethanol resistance in cyanobacterial synechocystis sp. pcc 6803. *Biotechnology for biofuels*, 5(1):1.

Wang, W., Liu, X., and Lu, X. (2013). Engineering cyanobacteria to improve photosynthetic production of alka (e) nes. *Biotechnology for biofuels*, 6(1):1.

Wang, Y., Chen, L., and Zhang, W. (2016). Proteomic and metabolomic analyses reveal metabolic responses to 3-hydroxypropionic acid synthesized internally in cyanobacterium synechocystis sp. pcc 6803. *Biotechnology for biofuels*, 9(1):209.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63.

Warburg, O. and Christian, W. (1941). Isolierung und kristallisation des gärungsferments enolase. *Naturwissenschaften*, 29(39):589–590.

Warner, J. R. (1999). The economics of ribosome biosynthesis in yeast. *Trends in biochemical sciences*, 24(11):437–440.

Washburn, M. P., Koller, A., Oshiro, G., Ulaszek, R. R., Plouffe, D., Deciu, C., Winzeler, E., and Yates, J. R. (2003). Protein pathway and complex clustering of correlated mRNA and protein expression analyses in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U.S.A.*, 100(6):3107–3112.

Wasinger, V. C., Cordwell, S. J., Cerpa-Poljak, A., Yan, J. X., Gooley, A. A., Wilkins, M. R., Duncan, M. W., Harris, R., Williams, K. L., and Humphery-Smith, I. (1995). Progress with gene-product mapping of the mollicutes: Mycoplasma genitalium. *Electrophoresis*, 16(7):1090–4.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

Weaver, L. J., Sousa, M. M., Wang, G., Baidoo, E., Petzold, C. J., and Keasling, J. D. (2015). A kinetic-based approach to understanding heterologous mevalonate pathway function in e. coli. *Biotechnology and bioengineering*, 112(1):111–119.

Wegener, K. M., Singh, A. K., Jacobs, J. M., Elvitigala, T., Welsh, E. A., Keren, N., Gritsenko, M. A., Ghosh, B. K., Camp, D. G., Smith, R. D., et al. (2010). Global proteomics reveal an atypical strategy for carbon/nitrogen assimilation by a cyanobacterium under diverse environmental perturbations. *Molecular & Cellular Proteomics*, 9(12):2678–2689.

Weinersmith, Z. (2016). Fossils.

Wenger, C. D., Lee, M. V., Hebert, A. S., McAlister, G. C., Phanstiel, D. H., Westphall, M. S., and Coon, J. J. (2011). Gas-phase purification enables accurate, multiplexed proteome quantification with isobaric tagging. *Nature methods*, 8(11):933–935.

Werner, T., Becher, I., Sweetman, G., Doce, C., Savitski, M. M., and Bantscheff, M. (2012). High-resolution enabled TMT 8-plexing. *Anal. Chem.*, 84(16):7188–7194.

Werner, T., Sweetman, G., Savitski, M. F., Mathieson, T., Bantscheff, M., and Savitski, M. M. (2014). Ion coalescence of neutron encoded tmt 10-plex reporter ions. *Anal Chem*, 86(7):3594–601.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wijffels, R. H. and Barbosa, M. J. (2010). An outlook on microalgal biofuels. *Science*, 329(5993):796–799.

Wilm, M., Shevchenko, A., Houthaeve, T., Breit, S., Schweigerer, L., Fotsis, T., and Mann, M. (1996). Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry. *Nature*, 379(6564):466–469.

Wisniewski, J. R. and Rakus, D. (2014). Multi-enzyme digestion fasp and the 'total protein approach'-based absolute quantification of the escherichia coli proteome. *J Proteomics*, 109:322–31.

Wisniewski, J. R., Zougman, A., Nagaraj, N., and Mann, M. (2009). *Universal sample preparation method for proteome analysis*, volume 6, pages 359–62. United States.

Wiśniewski, J. R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. *Nat. Methods*, 6(5):359–362.

Wolfgang, P. and Steinwedel, H. (1956). Verfahren zur trennung bzw. zum getrennten nachweis von ionen verschiedener spezifischer ladung. Translated as: Method for the separation and separate proof of ions of various specific charge. *Patent application*, (DE944900).

Wright, P. C., Jaffe, S., Noirel, J., and Zou, X. (2013). Opportunities for protein interaction network-guided cellular engineering. *IUBMB Life*, 65(1):17–27.

Wünschiers, R. (2016). Making-of synthetic biology: The european cyanofactory research consortium. In *Ambivalences of Creating Life*, pages 55–72. Springer.

Xiang, F., Ye, H., Chen, R., Fu, Q., and Li, L. (2010). N, n-dimethyl leucines as novel isobaric tandem mass tags for quantitative proteomics and peptidomics. *Analytical chemistry*, 82(7):2817–2825.

Yang, A. and Cui, Y. (2012). Global coal risk assessment: data analysis and market research. *World Resources Institute*.

Yates, J. R., r. (2013). The revolution and evolution of shotgun proteomics for large-scale proteome analysis. *J Am Chem Soc*, 135(5):1629–40.

Yost, R. A. and Enke, C. G. (1978). Selected ion fragmentation with a tandem quadrupole mass spectrometer. *J. of the Am. Chem. Soc.*, 100(7):2274–2275.

Young, J. D., Shastri, A. A., Stephanopoulos, G., and Morgan, J. A. (2011). Mapping photoautotrophic metabolism with isotopically nonstationary (13)C flux analysis. *Metab. Eng.*, 13(6):656–665.

Zerulla, K., Ludt, K., and Soppa, J. (2016). The ploidy level of synechocystis sp. pcc 6803 is highly variable and is influenced by growth phase and by chemical and physical external parameters. *Microbiology*, 162(5):730–739.

Zhang, L., Selão, T. T., Pisareva, T., Qian, J., Sze, S. K., Carlberg, I., and Norling, B. (2013a). Deletion of synechocystis sp. pcc 6803 leader peptidase lepb1 affects photosynthetic complexes and respiration. *Molecular & Cellular Proteomics*, 12(5):1192–1203.

Zhang, L.-F., Yang, H.-M., Cui, S.-X., Hu, J., Wang, J., Kuang, T.-Y., Norling, B., and Huang, F. (2009). Proteomic analysis of plasma membranes of cyanobacterium synechocystis sp. strain pcc 6803 in response to high ph stress. *Journal of proteome research*, 8(6):2892–2902.

Zhang, Y., Fonslow, B. R., Shan, B., Baek, M.-C., and Yates III, J. R. (2013b). Protein analysis by shotgun/bottom-up proteomics. *Chemical reviews*, 113(4):2343–2394.

Zhang, Y., Wen, Z., Washburn, M. P., and Florens, L. (2010). Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. *Analytical chemistry*, 82(6):2272–2281.

Zhi, X., Yang, H., Berthold, S., Doetsch, C., and Shen, J. (2010). Potential improvement to a citric wastewater treatment plant using bio-hydrogen and a hybrid energy system. *Journal of Power Sources*, 195(19):6945–6953.

Zhu, H., Ren, X., Wang, J., Song, Z., Shi, M., Qiao, J., Tian, X., Liu, J., Chen, L., and Zhang, W. (2013). Integrated omics guided engineering of biofuel butanol-tolerance in photosynthetic synechocystis sp. pcc 6803. *Biotechnology for biofuels*, 6(1):1.

Zou, X., Pham, T. K., Wright, P. C., and Noirel, J. (2012). Bioinformatic study of the relationship between protein regulation and sequence properties. *Genomics*, 100(4):240–244.