University of Sheffield

Department of Chemical and Biological Engineering

Predicting Functional Alterations Caused By Non-synonymous Variants in CHO Using Models Based on Phylogenetic Tree and Evolutionary Preservation

Author: Qixun Fang
Supervisor: David James
Second Supervisor: Robert Falconer

Submitted for the degree of Doctor of Philosophy
February, 2018

Acknowledgements

I would like to give my most sincere thanks to Dr Paul Dobson who offered help throughout the entire project, even after he had left to another university. This project would not be possible without his effort and expertise. I am also very grateful to Professor David James who continued the supervision after Dr Dobson had left and provided many inspirations and fun during my research. I am also very grateful to my parents who sponsored my PhD and supported me in every way they can.

I would also like to thank all of my colleagues in both Paul's group and David's group and my friends here in the UK. My life here is much better with all these nice people in it.

Abstract

Chinese Hamster Ovary (CHO) cell is a major manufacturing platform for one of the most valuable biopharmaceutical products: monoclonal antibodies. Being an immortal cell line adapted to different environments, CHO has been accumulating massive mutations in its genome. Continuous effort has been invested into building a computational model to predict CHO cell productivity. However, not much attention has been focused on its proteins which are surely effected by the mutations accumulated to some extent.

In this project, we focused on the functional effect caused by non-synonymous variants found in CHO genome. A tool was built to firstly identify these variants and then predict their potential function effect by preservation, a concept derived from evolutionary conservation. Firstly, the PANTHER subfamilies, which defined on the base of potential function change within gene trees, were extended by adding proteins from species not covered by PANTHER. Sequences within the same subfamily were then aligned and had Hidden Markov Models (HMMs) built on these alignments. The HMMs were used to identify homologs in CHO proteins. After that preservation were calculated in every site of the alignments, which was then used to predict the function alterations caused by mutations on every site.

Our tool was then validated using data from origin PANTHER subfamilies, PANTHER-PSEP which also calculated site preservation and BLAST, a well-accepted homolog searching algorithm. CHO protein sequences were then imported and analysed by our tool. For comparison, protein sequences from Chinese hamster were also analysed alone with two published CHO cell lines: CHO-K1 and CHO-K1GS. The predictions of proteins from these three genomes were then compared by mapping onto Gene Ontology (GO). Some detailed case studies were also demonstrated. Our tool showed good performance in validations, however, they failed to produce useful hypotheses that would motivate further experiments on bench. The potential causes are discussed at the end.

Contents

List of Abbreviations

CHO	Chinese Hamster Ovary
mAb	monoclonal antibody
PTM	post translation modification
HMM	Hidden Markov Model
PANTHER	Protein Annotation Through Evolutionary Relationship
BLAST	Basic Local Alignment Search Tool
IgG	Immunoglobulin G
Fab	Fragment antigen-binding
Fc	Fragment crystallisable
DNA	Deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
ddNTP	Dideoxynucleotides triphosphates
PCR	Polymerase chain reaction
HGP	Human Genome Project
HTS	High-throughput sequencing
NCBI	National Center of Biotechnology Information
MC	Markov chain
MCMC	Markov Chain Monte Carlo
ILS	incomplete lineage sorting
UPGMA	unweighted pair-group method based upon arithmetic averages
NJ	Neighbour Joining
ML	Maximum Likelihood
PAM	point accepted mutation
BLOSUM	Blocks Substitution Matrix
CASP	Critical Assessment of methods of protein Structure Prediction
TBM	template-based model
FM	free model
PSSM	position-specific scoring matrix
nsSNV	non-synonymous Single Nucleotide Variant
nsSNP	non-synonymous Single Nucleotide Polymorphism
ENA	European Nucleotide Archive
EMBL	European Molecular Biology Laboratory
DDBJ	DNA Data Bank of Japan
INSCO	International Nucleotide Sequence DataBase Collaboration
RefSeq	Reference Sequence
ORF	open reading frame
PDB	Protein Data Bank

GO	Gene Ontology
SNV	Single Nucleotide Variation
OMIM	Online Mendelian Inheritance in Man
HGMD	Human Gene Mutation Database
UniProt	Universal Protein Resource
EBI	European Bioinformatics Institute
SIB	Swiss Institute of Bioinformatics
PIR	Protein Information Resource
UniProtKB	UniProt Knowledgebase
UniParc	UniProt Archive
UniRef	UniProt Reference Cluster
UniMES	UniProt Metagenomic and Environmental Sequence
NMR	nuclear magnetic resonance
PSSM	position-specific score matrices
PANTHER-PSEP	PANTHER position-specific evolutionary preservation
KEGG	Kyoto Encyclopaedia of Genes and Genomes
MFA	metabolic flux analysis
FBA	flux balance analysis
FWI	Fowlkes-Mallows index
TP	true positive
FP	false positive
FN	false negative
FFT	Fast Fourier Transform
RAxML	Randomised axelerated maximum likelihood
FPKM	Fragments Per Kilobase of transcript per Million mapped reads

List of Tables

List of Figures

1 Introduction

1.1 Research background and motivations

Chinese Hamster Ovary (CHO) cell is a mammalian cell line which has been cultured for more than three decades. For different research or manufacturing purposes, it had been adapted to different culture environments and therefore derided many descendent cell lines with various characters. Some of them were adapted to inorganic environments of cheap, easy-to-control, scalable, massive culture for industrial production. One of its most valuable products is the monoclonal antibody (mAb), an important compound secreted by the human immune system. In spite of recent recession in the pharmaceutical market, the global sale of mAbs and their functional fragment maintains stable in its growth. However, the complexity of mAbs also laid down substantial restrictions on manufacturing platforms. One of the most significant requirement is human compatible post translation modification (PTM). CHO cell as a mammalian cell line inherently shared a big part of protein synthesising pathways with humans and therefore, became one of the most promising candidates of expression platform. However, expressing exogenetic product brings large stress to the cell and the productivity was extremely low at the beginning. After decade's development on both culture process and cell lines, the productivity has been largely improved by tens of times. However, most of the improvements were made by process optimisation and the knowledge of CHO cell biology remains fairly limited. The first genome of CHO was not made available until 2011 (Xu et al., 2011). It came from a CHO cell line parent to many production cell lines: CHO-K1. The development of sequencing techniques allow cheap and fast acquirement of genomic data, enabling further research on CHO biology. However, processing and interpreting the massive data generated by sequencing machine became the bottleneck for current research.

Bioinformatics is a technique developed for resolving this problem. The CHO industry community is longing for a computational model that can predict and improve CHO productivity before the culture starts. It requires integration of large amounts of various types of data ranging from nutrition in the media to the gene expression in the cell. The first consensus CHO metabolic model was merely published in late 2016 (Hefzi et al., 2016). While metabolic flux are extensively modelled, the proteins carrying related functions were not getting enough attention. As a cell line being engineered for different purposes, mutations are intensively accumulated in CHO genome. It is very

often that these mutations change the function of the CHO proteins, which is also the reason that CHO could quickly adapt to major changes in the culturing environment. Given the knowledge of how these mutations affect the function of protein related to production process, accurate protein efficacy can be predicted for the metabolic model. Also, as genome editing tools, such as CRISPR techniques (Sander & Joung, 2014), had been made available, it is possible to optimise cellular production process in the molecular level given that knowledge. However, it has rarely been explored on CHO. Current use of genome editing techniques, such as CRISPR/CAS9, only focus on transgene integration(J. S. Lee et al., 2016) and engineering on glycan profile (Sun et al., 2015).

The ration between mutations and protein function is mainly studied in the human genome (Wong & Zhang, 2014; Zeng et al., 2014; Wu & Jiang, 2013). It is not rare that genetic information is used for understanding cancer pathology and predictions of cancer (Shihab et al., 2013; Reva et al., 2011). Computational tools had been developed to analyse sequence variant effect on human disease in residue level (Wong & Zhang, 2014; Hu et al., 2007). These tools are mostly supported by data from clinic researches and all focus on disease related genes in the human genome. Transferring these tools on production or metabolism related genes on CHO would need to counter a major challenge of lack of supporting data. Additionally, the first draft human genome was published in 2001 (Venter et al., 2001), 10 years earlier than CHO. Since then, annotating human genome has been conducted intensively by research groups around the world making the human genome one of the best understood animal genomes. In comparison, the CHO genome was only assembled in the scaffold level and the annotation works have been conducted by the CHO community for a few years.

In this project, we aimed to develop a tool predicting function impact of sequential variant in CHO proteins. As a starting point, we were minimising hypotheses application in our tool to reduce bias and allow various hypotheses application in further development. Data from other species, especially rodents, would be used to compensate for the lack of data for CHO. To use these data from different species, phylogenetic information which is also used by tools predicting human diseases, were extensively applied. We identified types of mutations and developed models accordingly which would all contribute to the predictions. Related data and tools would be used for validation and comparison to evaluate the performance of our tools and finally, hypotheses would be made on the generated predictions.

1.2 Objectives

The main target of this project is developing a computational tool that could predict functional change of proteins. To achieve this target, it was broken down to several objectives:

- Identify the most suitable tools and data resource from the related literature. Well-developed tools should be used as the foundation of our tool. Therefore, performance surveys of tools built for different purposes need to be reviewed. The underlying assumptions adopted by tools used in our pipeline must not conflict with each other. In addition, although large amount of data would be required to generate reliable prediction, the data needs to be selected to reduce unnecessary noise in the system.
- Build the predicting tool (models) using related with minimum assumptions. In this process, underlying concepts and assumptions would also be identified. Assumptions should be derived from existing assumptions adopted by tools with good performance. Inherent limitations should also be identified.
- Validate the tool with proper data and related tools. Data for validation should be selected before the tool was built and isolated from the building process. Inherent difference between the tool built in this project and tools used for validation should also be identified and standard for evaluating tools should be made clear. Performance and limitation of the tool should be examined with the concepts and assumption adopted. Matrices used by the tool should also be adjusted in this stage.
- Apply the tool to CHO proteins and generated hypotheses. Methods of interpreting the prediction results into hypotheses should be described. Certain hypotheses should be made for comparison with existing knowledge to estimate the quality of other hypotheses made for inspiring research.
- Identify the limitations and potentials of the tools and its resulted hypotheses. Depend on the quality of the hypotheses the prediction could provide, assumptions and potential improvements of the tool should be discussed.

It is worth to mention that not all the objectives above were successfully accomplished in this thesis but actions or resources required to fully accomplish these objectives will be discussed at the end. After all, this project merely set up a starting point for further sophisticated development of models which could make significant contribution to CHO cell engineering.

1.3 Thesis structure

Firstly, the research background and knowledge required to understand the rationales of related tools is provided by reviewing related literatures. The importance of CHO cell as a manufacturing platform and its most valuable product mAbs is highlighted in the literature review. The main sequencing platforms are then acknowledged as the source of sequences intensively used in this project. Important tools for comparing sequences, calculating phylogenetic distance and topology and predicting protein function and structure are then detailed. Lastly in that chapter, databases storing and managing related data and relevant research on CHO are briefly introduced.

After that, a methodology chapter follows with detailed descriptions of the pipeline of the tools, accompanied by justifications in every step. An overview of the pipeline is provided at the beginning of the chapter. In short, we used Hidden Markov Models (HMMs) to integrate sequential information of every homolog groups and identify the best available homolog in CHO. Then site conservation was calculated and provided evidence to make predictions on function alteration. The detailed description of the step is started from data selection, alignment to phylogenetic tree construction and the final HMM construction and conservation calculation. A detailed description of external tools used by the pipeline will be provided in the end of the chapter.

The validation of the built tool is described in chapter 4. As the homolog selection we used were mainly derived from PANTHER subfamilies, subfamily members not used for building HMMs were used to validate the performance of the HMMs we built. On the other hand, we compared the preservation result calculated by another tool with ours. At last, we validated the site preservation results with functional site annotations from UniProt.

In chapter 5, analysis of sequence data of three published CHO related genome using the tool we developed is described. We randomly selected prediction results to compare with BLAST results. And then we used published CHO expression results to verify a hypothesis we made given the site preservation was correctly calculated. Lastly, we compared prediction difference between three CHO related genomes by mapping the prediction onto GO.

Several case studies are described in chapter 6. We studied mutations on TP53 of three CHO related genomes. Then we chose glycolytic process, apoptosis, DNA repair and glycosylation to inspect in

detail. A discussion of the achievements and limitations of the tool is provided in the final chapter with the potential future works.

2 Literature Review

In this chapter, background information will be first provided in order to explain the motivations and objectives of this project. Then related bioinformatics techniques, tools and databases will be introduced, where basic concepts, strategies and algorithms are described briefly suggesting the selection of tools in this project. Lastly, CHO related research and the data they generated will be reviewed, explaining the challenges and novelty of the works in this thesis.

2.1 Chinese Hamster Ovary Cells

Despite the recent economic recession the global pharmaceutical market has remained strong (EFPIA, 2015), with biopharmaceuticals - drugs manufactured using biological cells - showing consistent growth. Including monoclonal antibodies (mAbs), hormones, enzymes, blood-related proteins, vaccines, interferons and gene therapy-based products, biopharmaceuticals are used to treating conditions such as cancer, haemophilia, diabetes and rheumatoid arthritis. According to Walsh (2014), since the first approval of recombinant protein (human insulin) in 1982 there were 246 biopharmaceuticals approved in the United States and European Union up to 2014. In 2014 these medications generated sales of $140 billion (Walsh, 2014) and this is projected to maintain growth in the near future.

mAbs are dominant in both sales and sales growth. As shown in Table 2.1, half of the top 10 best-selling biologics in 2014 are mAbs. Together mAbs accounted for over $40 billion sales in 2013. The number one product (Humira), which generated global sales of $10 billion in 2013, has reached a 20% increase resulting in $12 billion sales in 2014 (Philippidis, 2015). Most of the top-selling mAbs were developed by Roche, along with others by AbbVie, Johnson & Johnson, Merck and Gilead Sciences. mAbs also accounts for 23% of biologics approved since 1995 (Walsh, 2014).

mAbs are antibodies secreted by the plasma cell in the immune system that target a specific antigen. Licenced therapeutic mAbs are all Immunoglobulin G (IgG), consisting of two heavy chains and two light chains, mAb can be divided into three fragments: two Fragment antigen-binding (Fab), responsible for specifically binding to the antigen, and one Fragment crystallisable (Fc) that triggers downstream immune activities by binding to receptors on the cell surface. Due to the specificity of

mAbs they are often designed for cancer treatment (Adams & Weiner, 2005) and autoimmune diseases (Feldmann & Maini, 2003).

Table 2.1 Top 10 best-selling drugs in 2014
Data comes from www.genengnews.com

Name	Drug type	Sponsor	Sale (billion)
Humira	Mab	AbbVie	$12.543
Sovaldi	Nucleotide analogue	Gilead Sciences	$10.283
Remicade	Mab	J & J and Merck	$9.240
Rituxan	Mab	Roche (Genentech) and Biogen	$8.678
Enbrel	Fusion protein	Amgen and Pfizer	$8.538
Lantus	Insulin analogue	Sanofi	$7.279
Avastin	Mab	Roche	$6.957
Hercepin	Mab	Roche	$6.793
Advair	Formulation	GlaxoSmithKline	$6.431
Crestor	Small molecule	AstraZeneca and Shionogi	$5.869

Therapeutic mAbs require an expression platform with proper (specifically, human-like) protein folding and post-translation modification mechanisms. Therefore, mammalian cell lines are preferable as less genetic engineering is required to reconstruct these mechanisms (that is, they natively fold and modify proteins in a human-like manner). Currently, available mammalian manufacturing platforms include baby hamster kidney cells, human cell lines Hek 293 and PERC.C6 (Swiech et al., 2012), and Chinese Hamster Ovary (CHO) cells, which are the most favoured platform for mAb production (Wlaschin & Yap, 2007). Up to 2014 35.5% of the total approved biopharmaceuticals are manufactured by CHO (Walsh, 2014), including the majority of mAbs and some mAb fragments, which account for most of the revenues. Compared to other platforms, which only recently succeeded in expressing functional mAbs, CHO has been manufacturing correctly folded, glycosylated (in human form) proteins for the past few decades. Apart from the main reason of their capacity to correct PTMs, other reasons behind CHO's popularity include 1) it has been proven safe as a medicine manufacturing cell line for over two decades, which makes it easier to be approved by drug administration agencies (Kim et al., 2012); 2) it resists most human viruses; 3) it can easily adapt to serum-free suspension conditions which allow large scale manufacturing; and 4) as a result of 20 years' optimisation by the biopharma industry, bioprocesses in CHO have been well

established. Due to the dominant position in drug production, CHO has become the production cell line of greatest interest for industry-related research and manufacturing.

Most CHO productivity improvements were achieved by optimising culture media and the process design (Hacker et al., 2009). Hacker et al. (2009) suggested that further improvements in protein yield can be achieved by high throughput transfection and screening. Although large-scale screening has been proven effective for obtaining productive clones, it largely relies upon uncontrollable factors that could lead to failure when the clones for screening are insufficient. While highly productive cell lines were mainly obtained by screening, function manipulation in CHO engineering is restricted by a lack of knowledge of the CHO genome's sequence and organisation. Despite great interest in CHO, the first CHO genome was only published a few years ago (Xu et al., 2011). While this enabled better understanding both across the whole CHO genome and of its details, allowing derived research to thrive in recent years, our knowledge of CHO biology remains relatively limited.

In contrast, human cancer genomes have been more intensively studied and better understood. Although undertaken for totally different purposes, research on human cancer genomes also speaks to CHO as CHO shares many genotype and phenotype characteristics with cancers (Lewis et al., 2013; Kojima et al., 2009; Tannock & Guttman, 1981). A characteristic shared by all cancer cell lines and CHO is genetic instability, which is described as a 'hallmark of the cancer cell' by Loeb and Loeb (2000). Genetic instability is present at multiple scales, from point mutations through to chromosomal aberrations (Roychowdhury & Chinnaiyan, 2016; Weinstein, 2012). It was found long ago that most of these mutations are located in non-coding regions (Stoler et al., 1999), but the regulation mechanisms of these non-coding regions were only discovered recently and our knowledge of these remains limited (Ling et al., 2015). Compared to point mutations, which affect only one gene, chromosomal aberration can result in large region rearrangements affecting multiple genes. Genetic information, including family history and genome profile, has been used for cancer prognosis, diagnosis and treatment response assessment (Roychowdhury & Chinnaiyan, 2016; Ling et al., 2015; Weitzel et al., 2011). To acquire related information, systems containing multiple cancer cell lines have been set up on different aspects (Barretina et al., 2012; Weinstein, 2012). A series of genes have been associated with certain cancers (Weitzel et al., 2011) but surprisingly some research has shown that mutated cancer genes are rarely shared even by cancers observed in the same organ except *TP53* (Podlaha et al., 2012). Cortés and Calvo (2014), however, pointed out that the same driver mutations are observed in cancer in different organs and proposed that cancers

with the same driver mutations express similar treatment responses. Their findings suggest that although universally-shared cancer mutations are extremely rare, certain mutation patterns could be found on cancers sharing common drivers. Research effort has also been invested into addressing cancer drivers and, more broadly, genetic variances which cause human diseases (Frousios et al., 2013; Gray et al., 2012). Therefore, while CHO shares many features with cancer cell lines, by understanding the common drivers behind CHO we might better comprehend which cancer research results are of most relevance.

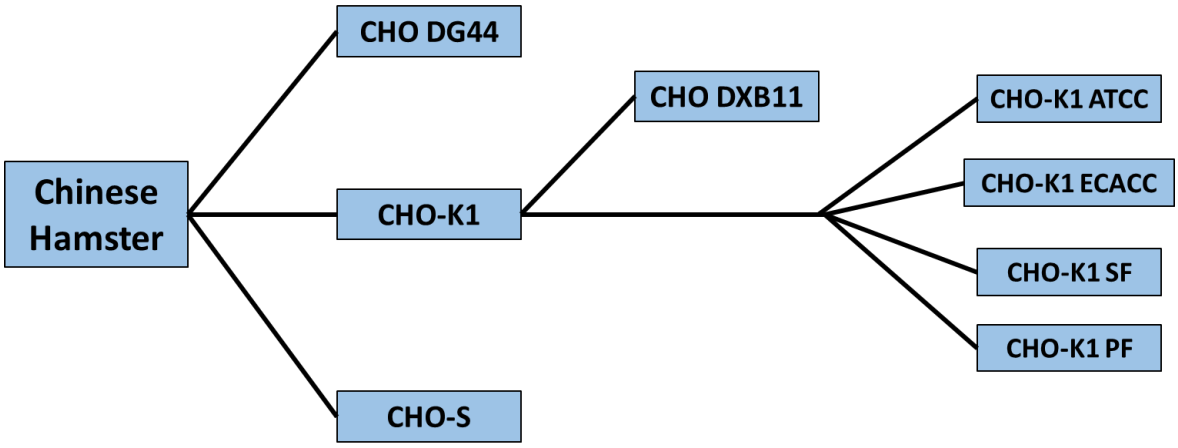


Figure 2.1 Phylogenetic tree of some most widely used CHO cell lines
The popular CHO cell lines presented are CHO DG44, CHO-K1 and CHO-S which are all derived from Chinese hamster separately. (Vishwanathan et al., 2017; Kaas et al., 2015; Lewis et al., 2013) The CHO-K1 cell line has been extensively cultured and deriving CHO DXB11 and another cell line which then separated into four different cell lines.

While there are many things we might learn about CHO from studying the Chinese hamster genome from which it is derived, it is important to note that after decades of engineering and being cultured in different environments, the CHO genome should be expected to be substantially different from the source organism. Therefore, referencing the host genome (note that the fully assembled Chinese hamster genome has not yet been available by the time this thesis is composed) may not be as beneficial as in human cancer research. However, compared to large numbers of human cancer cell lines, CHO cell lines in use nowadays are derived from only a few ancestor cell lines (Figure 2.1), which should have a simpler evolutionary history than cancer cell lines. Therefore, the evolution of CHO is easier to track once the cell lines in the lineage are sequenced. Transferring methods and concepts from cancer research, certain phenotypes of CHO might be diagnosed.

However, being restricted by the data available, information from non-coding regions may not support liable diagnosis, while information from coding regions, also known as the exome, could provide more supportive results.

2.2 High-throughput Sequencing Techniques

Genome sequencing was the basis of the genetic analyses mentioned in the previous section. Biologists have been trying to sequence DNA since it was found to possess a certain pattern. One of the most successful sequencing techniques, developed in 1977, is known as Sanger sequencing (Sanger, Nicklen, et al., 1977), which is based upon DNA replication and gel electrophoresis techniques. DNA replication techniques synthesise a complementary chain of the template DNA chain with DNA polymerase and four basic dNTPs. The concept of Sanger sequencing is to terminate the chain extension process of complementary DNA synthesis at certain points using ddNTPs - substances that can replace corresponding dNTPs but not accommodate another nucleotide at the next position. Controlling the concentration of dNTPs and ddNTPs in a PCR-like process enables the synthesis of DNA copies of different lengths (depending upon whether the terminator is incorporated or not), which can be separated using gel electrophoresis. Labelling four types of ddNTPs with different dyes (now usually fluorescent) then enables direct readout of the sequence after separation using imaging techniques. Using Sanger sequencing the first genome, bacteriophage phi X 174 genome, was sequenced (Sanger, Air, et al., 1977). Sanger sequencing was then improved to increase accuracy and read length, and reduce cost (Shendure & Ji, 2008) before being applied in the Human Genome Project (HGP) to sequence an entire human genome. The HGP eventually cost 15 years and about three billion dollars with the collaboration of six countries (Collins et al., 2003). Despite the cost, Sanger sequencing is still used as a main approach of sequencing short DNA fragments due to its high accuracy and read length.

Genome sequencing required higher throughput to be cost effective. Next generation sequencing techniques, now often referred to as High-throughput sequencing (HTS), became the answer. Similar to Sanger sequencing, most of the HTS techniques adopt the concept of sequencing by synthesis. They achieve high throughput by employing a shotgun strategy to parallelise the sequencing process. In a comprehensive review of HTS techniques by Loman et al. (2012) the sequencing process was summarised in four steps: fragmentation, tagging, amplification and

sequencing. In brief, the target genome is broken down into short fragments that are then immobilised separately on a surface where the fragments are amplified and sequenced. The sequencing process conducted by HTS is also different from Sanger sequencing. Instead of permanently terminating the extension process, reversible terminators (or even no terminator) are employed to pause the process after one base being added to the new molecule. Then the added base will be readout, via a series of reactions culminating in a fluorescent signal, before the extension continues. In practice, the fragmentation and tagging are often finished in usual labs with no sequencing machines, and then the samples (usually referred to as libraries) are sent to sequencing facilities where amplification and sequencing are performed by experts. The most popular HTS platforms include HiSeq/MiSeq from Illumina (Turcatti et al., 2008), 454 series from Roche (Margulies et al., 2005) and SOLiD 5500 series from Life Technologies (Shendure, 2005). Although similar strategies are applied, different chemistries are employed by different platforms, which lead to differences in performance (Table 2.2). HTS technologies have reduced the cost of sequencing the human genome to the order of a few thousand dollars (Loman et al., 2012) which allows routine whole genome sequencing. Most genomes now available online were sequenced by the HTS technologies introduced above (Ellegren, 2014).

Table 2.2 Performance comparison of DNA sequencing technologies.
Three most widely used high throughput platforms are represented showing difference in time length and amount of data generated for each run, error rate and length of reads.

Platform	Run time	Gb per run	Quality	Read-length(bp)
HiSeq 2000/2500	11 days (regular)	600	10^{-2}-10^{-3}	2*100
454 GS FLX+	23 hours	0.7	10^{-3}-10^{-4}	700~800
5500xl SOLiD	8 days	150	10^{-2}-10^{-3}	75+35

Not every sequencing technology has been described here since they have not made significant contribution to the genome database. For example, another technology known as nanopore sequencing, whose commercial product was announced in 2012 (Loman et al., 2012), uses transverse tunnelling current to detect nucleobases when the DNA strand is driven through the nanopore (Branton et al., 2008). There is no tagging or amplification required for this technology and it is found to be able to sequence DNA methylation in current research (Simpson et al., 2017).

While the approach holds much promise, it is mentioned only briefly here as it has yet to contribute significantly to genome databases.

2.3 Bioinformatics Techniques

With the capacity to acquire large sequence data sets, annotating and analysing these data becomes the next challenge. For related proteins, which is the main focus of this project, most annotation and analyses are centred upon their function and structure. Novel functions and structures of proteins can be obtained by laboratory experiments, while computational tools can annotate proteins without experimental evidence based upon sequential similarity or phylogenetic distance. The most popular tools for this will be introduced in this section.

2.3.1 Sequence Alignment

Building a sequence alignment usually is the first step of computational analysis applied right after the sequencing data is obtained. It lays down the foundation for downstream processes such as homologue analysis and phylogeny analysis. Therefore, the quality of the alignment would profoundly affect the performance of these afterward process. In protein sequence alignment, ideally sequences should be aligned on their function carrying structure so that functional annotation could propagate through alignments (Aniba et al., 2010; Thompson et al., 1999). Based on such goals, some benchmarks created wholly and partly by manual alignment according to protein structural information acquired from X-Ray crystallography, such as BAliBASE (Thompson et al., 1999), are used as gold standards to test the performance of automatic aligners. Although these benchmark databases contain high quality multiple sequence alignments, they are limited in alignment number and fail to cover all protein types (Le et al., 2017).

At the very beginning, alignment algorithms were focused upon pairwise alignment. Some classic algorithms, such as Needleman-Wunsch (Needleman & Wunsch, 1970) and Smith-Waterman (Smith & Waterman, 1981), are still widely used in small scale searching. The most widely used searching tool is the Basic Local Alignment Search Tool (BLAST) which has been made available on NCBI and Ensembl website. To achieve fast search on large databases, BLAST employs heuristic methods wherein short sub-sequences are used instead of full length sequences (Altschul et al., 1990, 1997). After decades of development, BLAST was specialised for a variety of purposes such as blastp and

blastn (Ye et al., 2012; Morgulis et al., 2008) and a slower but more accurate version, known as Position-Specific Iterated BLAST (PSI-BLAST), is also available.

Based on the results of sequence search, more intensive aligning between similar sequences has also attracted a great deal of research interest. Specified aligning tools were built to align multiple sequences to identify features such as genetic distance and sequence conservation. After decades of continuous effort, a great number of multiple sequence aligning tools have been developed for accurately aligning multiple sequences at relatively high speed. The most popular aligners include CLUSTAL W (Thompson et al., 1994), T-coffee (Notredame et al., 2000), MUSCLE (Edgar, 2004), MAFFT (Katoh & Standley, 2013; Katoh et al., 2002) and Probcon (Do et al., 2005). In general, the above methods employ the same aligning strategy laid down by Thompson et al. (1994) in the development of CLUSTAL W. They firstly calculate the distance between involved sequences, which is then used to construct a guide tree. After that, progressive alignment is performed according to the guide tree and the alignment is optimised on certain criteria in this process. The main differences between these methods are how alignments are evaluated and internal details of the aligning algorithms, which result in different alignment accuracy and time cost. By default, CLUSTAL W scores the alignments with PAM (Dayhoff & Schwartz, 1978) along with assigning weights to different sequences. Higher penalties are assigned when starting a gap while expending the gap costs less in scores. T-coffee inherits many features from CLUSTAL W but instead of aligning columns indistinguishably, T-coffee uses a library of high confidence short alignments as seeds and aligns the whole sequence by extending these seeds. MUSCLE also applies similar features to CLUSTAL but jointly improves the guide tree and alignment in a progressive optimisation phase. MAFFT vectorises residues on their chemical characters and uses fast Fourier transform to locate homologous regions. Lastly, Probcon adopts Hidden Markov Model (HMM) and Bayesian statistics to assess alignments. The results are then used to build a guide tree and refine the alignments.

With multiple high performance aligners available, the best of them remains arguable. In general, Probcon seems to create more accurate results in some aspects but it is also the most time consuming, while MAFFT creates equal or slightly worse results but at a much higher speed (Thompson et al., 2011; Blackshields et al., 2006). MAFFT also is one of a few methods that can create large alignment with about 10,000 sequence in an acceptable time period (Sievers et al., 2011).

2.3.2 Hidden Markov Models

Hidden Markov Models (HMM) are an advance version of Markov chain (MC) models. In simple words, MC is a model of a sequence of events where the order of events can be varied but the selection of the next event is solely dependent upon the current event regardless of the previous event sequence. In statistics, MC can be used to calculate the probability of a certain sequence of events or the probability of certain events occurring at a certain position of the sequence. One of its most important applications in Statistics is Markov Chain Monte Carlo (MCMC), which can be used to approximate Bayesian posterior probability when the actual posterior is hard or impossible to calculate. In the MCMC process, the concept of MC is only used to generate samples as every sample in the process is generated on the base of the previous sample and certain random variables. More details of posterior calculation with MCMC will be introduced in the tree construction section.

Based on the MC model, Hidden Markov Model (HMM), was developed. Similar to the MC, HMM describes an event sequence. However, in HMM the true events cannot be directly or completely observed and therefore are considered hidden. The hidden event sequence can be estimated by the observed sequence it derives. Similar to MC, algorithms, such as forward and backward algorithm, have been developed to calculate the probability of certain hidden sequence and certain events in the sequence. In bioinformatics, profile HMM (Eddy, 1998) was developed to describe the profile of given multiple sequence alignment. In profile HMM, the possible residue distribution of each position can be observed while the biological or statistical content of the position is considered hidden (Eddy, 2004). To enable easy application of profile HMM, HMMer (Eddy, 1998, 2011; Finn et al., 2011) was developed. HMMer is a package of basic HMM tools including building HMM given a sequence alignment, searching sequences using a profile HMM, searching profile HMMs with a given HMM and alignment sequences according to given HMM. Based on the profile HMM, more tools have been developed for other purposes, such as HHalign (Söding, 2005) adopted by Clustal Omega(Sievers et al., 2011) which aligns distant homologues by aligning their profile HMM, and ProbCons (Do et al., 2005) which employs HMM to create accurate multiple sequence alignments. HMM is able to capture features of a given sequence groups and provides high sensitivity sequence search. This allows databases based upon HMM to be built for protein families and conserved domains. Popular, HMM-based databases include PANTHER (Mi et al., 2016), which focuses upon

functionally-related subfamilies and has been adopted by Ensembl (Yates et al., 2015), and Pfam (Finn et al., 2016) which is a UniProt (Bateman et al., 2017) member database focusing upon protein domains. Recently, HMM was even used to detect DNA methylation with nanopore sequencing (Simpson et al., 2017).

2.3.3 Phylogenetic Tree Construction

Phylogenetic tree construction is another important part of bioinformatics that forms the foundation of much biological research. Understanding the evolutionary history of genes and species can provide evidence for genetic and taxonomic research. Based upon sequence alignment, the genetic distance between two genes, which is often defined as the minimum number of edits required to transform one gene into another, can be assessed. However, errors could be introduced by many factors such as: incorrect alignment, incomplete search of possible tree topology and incomplete lineage sorting (ILS) (Figure 2.2) (Noutahi et al., 2016a; Mirarab et al., 2016; Boussau et al., 2013), which make constructing accurate trees a challenge.

The unweighted pair-group method based upon arithmetic averages (UPGMA) (Sneath & Sokal, 1973) and Neighbour Joining (NJ) (Nei et al., 1987) are two of the most widely used tree building methods. They both build binary trees by combining genetically close homologues. Where UPGMA assigns equal distance between the ancestral node and its two offspring as a simplification, NJ assigns a more accurate branch length. However, the simple assumptions they rely upon - that the edit distance between sequences is proportional to their genetic distance - does not include the fact that the evolution rate varies over different times and species. Despite it being well acknowledged that their performance highly depends upon the order of the input sequences, meaning that multiple topologies may be created from the same data depending upon the order in which they are presented (Backeljau et al., 1996), they are still used by many software, such as CLUSTAL (Higgins & Sharp, 1988), to create draft trees. By adopting parsimony methods, they are able to construct trees on a small amount of data with high speed, but they lack the ability to examine and optimise the accuracy of the trees.

In order to build optimised and accurate trees on large data, more advanced methods have been developed, of which the most accurate ones are likelihood based methods adopting Maximum

Likelihood (ML) (Felsenstein, 1981) or Bayesian approaches (Yang & Rannala, 1997; Rannala & Yang, 1996). The most popular methods includes RaxML (Stamatakis, 2014), PhyML (Guindon et al., 2010; Guindon & Gascuel, 2003), MrBayes (Ronquist et al., 2012) and PhyloBayes (Lartillot et al., 2009).

The likelihood of a phylogenetic tree is the product of the likelihood of the branches, which in turn are calculated based upon substitution models. Similar to substitution matrixes such as PAM (Dayhoff & Schwartz, 1978) and BLOSUM (Henikoff & Henikoff, 1992), substitution models describe the substitution rate between nucleotides or amino residues but more parameters may be involved apart from log-odds scores. Such likelihood has been used as a standard to find the best tree, which has yielded promising results, but the high cost of the required processing power is a drawback (Guindon et al., 2010; Stamatakis et al., 2005; Holder & Lewis, 2003). Additionally, exhaustively searching all possible trees for the best tree is only possible for small amount of data and therefore heuristic searching must be used in most cases. The strategy of achieving the best tree with heuristic methods, which is often described as hill-climbing, involves altering the tree with the current best likelihood until no further improvement on likelihood can be made. The confidence of the tree is then assessed by bootstrapping how well the observed data supports the mooted tree by resampling the sequences (and replacing sequences generated from the observed alignment) to reconstruct the same tree. Such a process brings an extremely heavy computational burden since the iterating likelihood calculation for the tree search would be conducted after each bootstrapping which in turn repeat multiple times. Note that bootstrap scores only show the likelihood that the same result can be acquired when more data is imported, but not the accuracy of the result (Alfaro et al., 2003; Holder & Lewis, 2003). Software adopting ML such as RaxML and PhyML have significantly improved the speed of the integrated process but intensive computation is still required for large families. For example, analysing a 16S dataset of nearly 28,000 sequence with RaxML requires more than 1,200 hours (K. Liu et al., 2011).

Bayesian inference uses posterior probability that correlates with likelihood to evaluate trees. According to the hypothesis made, posterior probability can involve parameters not included by traditional likelihood calculations, which increases the flexibility of the Bayesian approach (Holder & Lewis, 2003). A form of posterior probability that only involves substitutions in sequence alignments, substitution models and the tree branch length was described by Huelsenbeck and Ronquist (2001). However, the authors also pointed out that analytically calculating the posterior

probability is too complicated to achieve. Therefore, Markov chain Monte Carlo (MCMC) method is used to achieve the approximate posterior. The general idea of using MCMC is to randomly explore the space of all trees step by step using certain criteria, such as change of likelihood or prior between current step and the next step (Huelsenbeck & Ronquist, 2001), using a random variable to decide the next move. A conceptual chain is then formed by these steps as they tend to stay in certain regions (of high likelihood if increase of likelihood is used as moving criterion) of the space. The ratio of steps in the region to total steps can be used as a valid approximation of the posterior probability. Different models and hypotheses have been proposed and the calculation has been optimised for higher speed on tools such as MrBayes and PhyloBayes. Note that MCMC would create many more trees than bootstrapping but each tree bootstrapping created requires tree search which could test many trees before the final and each tree in MCMC process is created under certain restrictions by random. Therefore, the MCMC process may be faster than bootstrapping in some situations. Greater comparison of bootstrapping and MCMC is discussed by Alfaro et al. (2003) and Holder & Lewis (2003).

The methods described above mainly account for substitutions between sequences. However, current research shows that the information contained by the sequences is insufficient for constructing accurate trees and reconciliation with certain additional information, such as taxonomic trees, which could significantly improve accuracy (Noutahi et al., 2016a; Boussau et al., 2013; Nguyen et al., 2013; Akerborg et al., 2009). Note that combining information from gene trees and species trees invokes a circular problem as the construction of accurate species trees depends upon multiple accurate gene trees, whose reconstruction requires reference from species tree (Szöllosi et al., 2013; Boussau & Daubin, 2010). Boussau and Daubin (2010) suggested joint inference as a solution to the problem. However, while species trees need to be first constructed as reference, arguments are raised between two conflicting species tree constructing approaches: concatenation and coalescence. While concatenation considers the gene tree and species tree to be identical, coalescence takes incongruence between the gene tree and species tree (Figure 2.2), which is well known (Maddison, 1997), into consideration. Maddison (1997) summarised the biological causes of the discord between gene phylogeny and species phylogeny. These biological events lead to ILS, where some alleles fail to present in certain lineages, which in turn leads to an important error in concatenation trees. Therefore, there is an increasing preference for using coalescence methods

(Simmons & Gatesy, 2015; Liu et al., 2009) although some research shows that both approaches result in statistically equivalent accuracy (Tonini et al., 2015; Warnow, 2015).

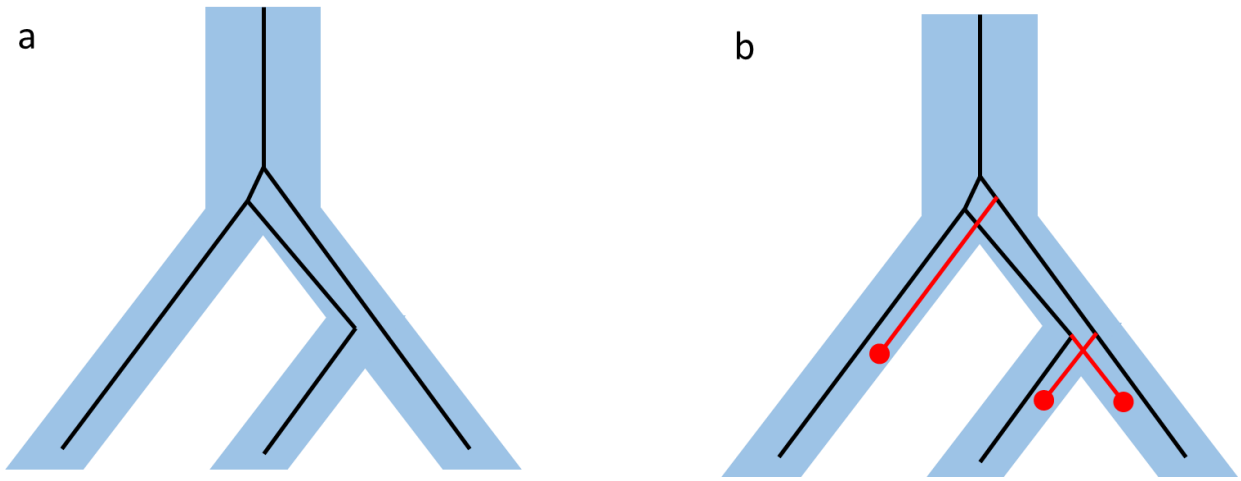


Figure 2.2 Incongruence between gene tree and species tree and ILS
An example of incongruence of gene tree and species tree is shown in a. The species trees are presented in blue bars and the gene trees are presented in black lines. The ILS that could cause such incongruence is illustrated in b. Red lines with dots at the end represent genes missed by sequencing or annotation. Losing these genes represented by red lines can also lead to the same incongruence.

Different methods have been developed for reconciliation between a species tree and gene tree to create a more accurate gene tree. In general, they focus on modelling evolutional events including duplication, loss, transfer (Boussau et al., 2013; Szöllosi et al., 2013; Rasmussen & Kellis, 2011; Akerborg et al., 2009) and occasionally speciation (Nguyen et al., 2013). The benefits to gene function inference of analysing these evolutionary events have been recognised by researchers (Blackstone, 2006; Eisen, 1998). While most methods focus upon gene trees, some of the methods, such as PHYLDOG (Boussau et al., 2013), also reconstruct a species tree when correcting gene trees. Bayesian approaches and MCMC are used to calculate likelihood or related posterior in these methods. Although these methods are able to increase the gene tree accuracy, the main drawback that most suffer are intensive computation and overfitting to the species tree (Noutahi et al., 2016a; Wu et al., 2013).

2.3.4 Protein Structure Prediction

It is common knowledge that certain protein structure is required for its function. Proteins with distorted structure, mostly due to incorrect folding, would be actively disassembled. The number of protein structures increased with the number of sequenced proteins, but still only a very small proportion of proteins have experimental structures (Moult et al., 2014). It is believed that the protein sequence alone contains enough information to determine the three-dimensional structure (Marks et al., 2012). Therefore, computational prediction of protein structure based upon sequence could potentially fill the gap. However, it has been a major challenge for more than four decades (Marks et al., 2012).

To facilitate better development of predictive models, the Critical Assessment of methods of protein Structure Prediction (CASP) tournament has been held to evaluate predictive models (Moult et al., 2014). According to CASP, there are two types of predictive models: template-based models (TBM) and free models (FM). TBM compares query sequence to (homologous) sequences with known structure and generates structure prediction by refining the known structure. In contract, FM usually predicts structure *ab initio*, although distant related templates may be used (Tai et al., 2014). Some predictors adopt both types of modelling, finding closely related templates or constructing novel structures when templates fail. The best predictive tools suggested by CASP include QUARK (Xu & Zhang, 2012, 2013), I-TASSER (Roy et al., 2010) and Phyre (Kelley & Sternberg, 2009).

In most cases, TBM achieves better accuracy than FM when available (Huang et al., 2014; Tai et al., 2014). Predictors with TBM are mostly based upon recursive improvement of profile alignments, where profiles can be in different forms such as position-specific scoring matrix (PSSM) and HMM. The modelling accuracy has largely improved in the last two decades, but only in relatively few cases can an accurate prediction be obtained (Moult et al., 2014). One consequence of the difficulty of predicting structure from sequence is that it remains very common to predict protein function directly from sequence without referencing the protein structure.

2.3.5 Non-synonymous Nucleotide Variant Impact

It is clear that changing protein structure can change function but the function could also be largely affected by the change of a few residues. Such changes are caused by non-synonymous Single Nucleotide Variant (nsSNV), which are also referred to as non-synonymous Single Nucleotide

Polymorphisms (nsSNPs) in some cases. These have attracted particular interest especially in the field of human disease. Experimentally these nsSNV can only be examined with very low throughput. Therefore, computational methods have been developed to predict the functional impact of these nsSNVs. The best predictors nowadays rely upon information such as phylogenetic conservation, 3D protein structure and chemical character of the residues. Protein structure could provide useful information for predicting nsSNV impact on protein function. For example, a nsSNV at the interaction site on a protein surface would significantly change the protein function (David et al., 2012). The most popular predictors include SIFT (Kumar et al., 2009), PolyPhen 2 (Adzhubei et al., 2010), PANTHER-PSEC (Thomas & Kejariwal, 2004; Thomas et al., 2003), PROVEAN (Choi et al., 2012), MutationAssessor (Reva et al., 2011), MutationTaster (Schwarz et al., 2010) and Condel (González-Pérez & López-Bigas, 2011). These are mostly focusing on predicting deleterious impact in human diseases. In general, these predictors can be divided into three types: 1) predictors based on certain model and hypotheses, mainly of genetic conservation and protein structure, such as SIFT, PANTHER-PSEC, PROVEAN and MutationAssessor; 2) predictors collecting various types of features and adopting machine learning classification methods such as naïve Bayes and random forest to make predictions, such as MutationTaster; 3) consensus predictors which summarise predictions of other predictors (usually predictors of first type) to generate their own predictions, such as Condel which integrates five other predictors including SIFT, PolyPhen2 and MutationAssessor. The last two types often include features involved by the multiple predictors in the first type. However, consensus predictors often generate integration scores using scores from other predictors with certain algorithms, while machine learning predictors assign different weight to the features depending on the training data, which means the same feature may be assigned different weight according to the training set.

The consensus predictors appear to outperform other predictors in performance surveys (Frousios et al., 2013; Gray et al., 2012) by taking advantage of including multiple features and hypotheses. In fact, features and hypotheses on multiple aspects enable machine learning based predictors to generate improved results (Wong & Zhang, 2014; Zeng et al., 2014). However, no single predictor generally outperforms others (Katsonis et al., 2014). Currently, Grimm et al. (2015) pointed out two circularities that could lead to erroneous conclusions on performance comparisons mentioned above. These circularities are caused by the fact that these predictors were trained and compared upon overlapping data sets and variants of the same genes are jointly labelled in variant databases. Based on that, Grimm et al. (2015) proposed using a subset of SwissVar to evaluate predictor

performance. In addition, Katsonis et al. (2014) suggested selection of predictors could be based not only on accuracy but also the hypotheses or features adopted.

Although structure information is useful for estimating the functional impact of a single residue change, using protein structure only is not practical since the structural information is not available for most proteins and the alternative way to access structure information, via structure predicting tools, is limited by accuracy. Also, a low number of nsSNVs are not likely to significantly change the protein structure and addressing interaction sites can be achieved by sequence alignments or HMMs with homologous annotation. Therefore, protein structure predictors are mostly unnecessary in estimating such functional impact.

2.4 Important Databases

To effectively make use of biological data and related annotations, sophisticated management of these information is required. Multiple databases were built and maintained by specialised institutes to facilitate the collection, storage and delivery of various types of biological data. Databases storing gene and protein related information were involved in this project and therefore reviewed in this section.

2.4.1 Gene databases

Genome sequences are commonly used as the base of molecular biology research. To keep these data up to date and available globally, three databases - GenBank built by the National Center of Biotechnology Information (NCBI), European Nucleotide Archive (ENA) operated by European Molecular Biology Laboratory (EMBL) and the DNA Data Bank of Japan (DDBJ) - collaborate with each other as partners of the International Nucleotide Sequence DataBase Collaboration (INSCO), exchanging data daily to maintain information consistency (Clark et al., 2015). Current sequencing techniques are not able to sequence genomes or chromosomes in the whole. The genome or chromosome sequence is obtained by sequencing fragmented nucleotide pieces then assembling these pieces with computational techniques. However, such a strategy is problematic in handling repeat content and duplicate regions in practice as assembling tools often fail to distinguish short reads from different copies. Pair-end sequencing techniques could help to improve the assembly

but still fail to bridge the entire duplication area (Alkan et al., 2011). Alkan et al. (2011) found substantial loss of these regions on genomes created solely by assembling short read generated by current high-throughput sequencing. Their research shows 93% of the exons were completely recovered while 43.7% of genes were underrepresented. Genome Reference Consortium improves the genome assembly using high-quality long sequences. However, they only work on a few genomes including two mammalian species: human and mouse.

Basing on specific genome assembly, annotations then can be made. Consisting of more than 340,000 species, GenBank also contains the corresponding annotations, which can be supplied by the authors who generate the sequence or the annotation pipeline provided by the manage institutes. Quality annotations are crucial for making use of the sequence data and can be generated in different ways with different quality and redundancy. NCBI initiated the Reference Sequence (RefSeq) project to collect and integrate such published annotations (Pruitt et al., 2014), while EMBL established Ensembl together with the Wellcome Trust Sanger Institute for the same purpose (Yates et al., 2015). Similarly, the annotation pipelines they provide rely on supporting evidence such as cDNA libraries, protein sequences and RNA-seq data. For annotating newly sequencing genomes which lack supporting data, sequences from related organisms will be used as reference. On this process, the annotation of existing sequence would transfer to newly discovered gene based on their sequential similarity. The annotations made by these pipelines include identifying open reading frames (ORFs), introns and exon and splice variances, homologous grouping and function inference based on homologue and assigning IDs to genes and proteins. One of the differences between the pipelines provided by NCBI and Ensembl is that the NCBI pipeline is more likely to create *ab initio* gene models when the supporting data is insufficient. Cross reference of external databases such as PDB and GO will be made when available. Sequence search and aligning tools are also available on their websites for easy access and comparison of interested data.

2.4.2 Single Nucleotide Variation (SNV) databases

SNV is frequently observed throughout the entire genome. It had been associated with human diseases and become a research hot spot of bioinformatics application in medicine (Ling et al., 2015; Cline & Karchin, 2011). Up until December 2017, data of 55,707 unique SNVs (referred as Single Nucleotide Polymorphisms, SNPs) had been collected by Genome-wide association studies (GWAS

http://www.ebi.ac.uk/gwas/home). Although SNVs in non-coding regions have been found to significantly contribute to disease in the last decade (Ling et al., 2015), most of the works have been focusing on SNVs on coding regions that change the protein sequence since it is believed that these non-synonymous SNVs (nsSNVs) are the major contributor (X. Liu et al., 2011). In some works, nsSNV is more often referred to as non-synonymous SNP (nsSNP). The main databases storing nsSNV data are: the Online Mendelian Inheritance in Man (OMIM) database (Hamosh, 2004), the Human Gene Mutation Database (HGMD) (Stenson et al., 2009) and the UniProt database (The UniProt Consortium, 2014). While OMIM and HGMD focus on deleterious nsSNVs, UniProt preserves both neutral and deleterious nsSNVs and therefore was used for training tools that are developed for predicting function impact of nsSNVs. For training purposes, neutral and deleterious SNV sets are often separated from databases mentioned above. For example, HumVar and HumDiv, were extracted from UniProt and specialised as neutral set and deleterious set respectively (Adzhubei et al., 2010). Among these specialised data sets, Grimm et al. (2015) showed that a unique subset of SwissVar (Mottaz et al., 2010), known as SwissVarSeleted is the most appropriate data set for comparing performance of the deleterious SNV predicting tools (Tang & Thomas, 2016). There are also database specified for results generated by popular predictors, instead of experiment proven results, for further research on potential candidates, such as dbNSFP (Liu et al., 2013).

2.4.3 Protein databases

Proteins are the gene product that facilitates most of the biological activities. In eukaryotic cells, one gene may encode more than one protein due to splicing, therefore the number of proteins in an eukaryotic cell could be much more than the number of genes. Compared to DNA, whose function mainly relies on the codes and subtle changes of certain nucleotides, proteins require certain three-dimensional structure and post-translational modifications (PTMs) to enact their functions. Accordingly, annotations stored by protein database involve more aspects than that of DNA. A summary database for proteins - Universal Protein Resource (UniProt) – contains the best available resources of these annotations. UniProt is maintained by the UniProt Consortium of three partners: European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). It consists of four databases: the UniProt Knowledgebase (UniProtKB), which will be the focus of this review; the UniProt Archive (UniParc); the UniProt

Reference Cluster (UniRef); and the UniProt Metagenomic and Environmental Sequence (UniMES) database (Apweiler et al., 2014). In short, UniProt is a high quality collection of all protein-related data that has been well organised for easy comprehension (The UniProt Consortium, 2014). Detailed annotations collected by UniProt include not only the basic sequence, family, function and homologue, but also PTM, sequence variance, mutation, expression, structure, cellular location, protein-protein interaction and related pathology. Manual curation is heavily involved in order to maintain the high quality and up-to-date aspect of the annotation. The manual curation is made on the experimental evidence together with structural information. Cross reference is also made with multiple databases including Protein Data Bank (PDB), Protein Annotation Through Evolutionary Relationship (PANTHER), Interpro and Pfam.

PDB is a protein structure database focusing on collecting, annotating and distributing three-dimensional coordinate data of protein structures obtained mainly from x-ray crystallography, nuclear magnetic resonance (NMR) and electron microscopy techniques. Along with the structures of proteins, information on ligands, binding sites and protein-protein interactions are also stored in PDB. Structure alignment tools are available on PDB for structure comparison and searching (Rose et al., 2013). PDB is intensively referenced by UniProt in terms of protein structure and manual curation. Pfam is a database of protein domain families. It consists of a collection of high quality models of protein domains (Sonnhammer et al., 1997). The foundation of Pfam is a huge number of profile HMM built upon multiple protein sequence alignments. Manual curation is applied to maintain the quality of the profile HMM conserved. Profile HMM built on alignments without curation were only generated after Pfam 29.0 and they do not represent potential domains (Finn et al., 2016). The manually curated Pfam domain annotation is one of the main sources referenced by Interpro. InterPro is a database of protein and domain families. Instead of using genetic distance like ordinary databases, it integrates information, which is termed signatures, including profile HMMs, position-specific score matrices (PSSM) and regular expression to classify proteins into families which allows prediction of domains and sites (Finn et al., 2017). To collect signature information, InterPro has extended its consortium from four databases to fourteen up to 2016 (Finn et al., 2017; Mitchell et al., 2014). Apart from protein search for the domains and site it contains, InterPro allows searching domain for proteins it composes. PANTHER is another database intensively referenced by InterPro. However, instead of focusing on domains and sites, PANTHER relies on traditional phylogeny-base protein classification to infer protein function. The protein phylogenetic trees preserved by PANTHER were constructed with reconciliation of species

phylogeny with GIGA (Thomas, 2010) to capture evolution events that may lead to function deviation such as duplication and transfer (Mi et al., 2016). Subfamilies were then defined according to these evolution events and manually reviewed. Ideally, members within the same subfamily should share the same function which is significantly different from other subfamilies within the family. Profile HMMs were built on every protein family and subfamily to provide predictions for query sequences. In addition, Gene Ontology (GO) enrichment analysis is also made available on PANTHER website. Recently, PANTHER-PSEP (position-specific evolutionary preservation) was introduced for predicting the potential function impact cause by non-synonymous single nucleotide variants based on its phylogenetic position (Tang & Thomas, 2016).

2.4.4 Function and pathway databases

As the number of sequenced genes increases, the amount of function annotation follows. Systematic organisation of annotated function is required for easy access and systematic analyses. Classification by function characters and by pathways are the most popular systems. As function characters can be set in arbitrary manners, classification on such base is highly flexible and able to easily achieve high coverage of known genes. A popular example is Gene Ontology (GO) which classifies functions into three main aspects. In contrast, classification on pathways is more centred on the nature of biological activities and able to present clearer pattern of the interaction network. However, it requires a high level of annotation on individual genes, which becomes the main restriction from including a large number of genes. Well known examples of such systems are the KEGG pathway and Reactome databases.

GO organises biological annotations with a tree-base system in three different aspects: molecular function, biological process and cellular component, which are the three ontologies in GO (Harris et al., 2004). Genes (or gene products) are attached to the tree through GO terms, which are certain biological concepts, and in turn every GO term is attached to their parent terms until the root of the tree. All GO terms are ultimately rooted on one of the three parent ontologies being the three aspects mentioned above. The backbone relation between these terms is “is a” relation, but other relations such as “part of”, “regulate” and “localized to” are also included. The quality of annotations and the accuracy of their relations are crucial. Therefore, similar to other high quality databases, manually creating high quality annotations from literatures is one of the main missions of GO (The Gene Ontology Consortium, 2013). Up to September 2014, 391,174 manually annotated genes and

gene products had been created in GO and the total number of annotations had reached 4,185,478 converged on 41,775 GO term covering 461,573 species (Blake et al., 2015). Although a large number of manual annotations had been made, the corresponding proportion remained low. The overall annotation quality is continually improved by cross-referencing and collaborating with many other databases such as PANTHER and UniProtKB. Tools such as AmiGo and GO slims are developed for easy use of GO data.

Kyoto Encyclopaedia of Genes and Genomes (KEGG) is a resource integrating information about genomes, biological systems, medicine and biochemistry on 15 main databases, of which 13 are entirely manually created by KEGG and the other two are created combining external databases with KEGG annotations (Kanehisa et al., 2012). The main focus of KEGG is the molecular network and pathway maps, which are manually generated diagrams containing rich information from experimental results about molecular interactions, reactions and locations. Mapping tools are also provided for mapping genes to the pathway and mapping transcripts abundance to analyse expression levels in the pathway context. The Colour Pathway tool in KEGG can present gene related values, such as expression levels, on top of the pathway diagram (Kanehisa et al., 2012). KEGG Pathway has been widely used in pathway level studies.

Similar to KEGG pathway, Reactome is another molecular reaction knowledge base. Compared to KEGG pathway, which is more metabolism orientated (Kanehisa et al., 2014), Reactome made efforts to cover both metabolic and signalling pathways. Most of the manual efforts were invested in human pathways, especially disease related pathways whose mutated forms are also covered by Reactome, while pathways from other species are inferred automatically from their human counterparts (Croft et al., 2014). In a recent update (Fabregat et al., 2015), the pathway browser has gone through significant changes to allow more interactive navigation. By referencing GO, the Reactome pathways are organised in a hierarchical structure derived from the biological process ontology. Compared to KEGG, Reactome offers a more interactive interface for pathway navigation and transcriptomic data of expression levels in different tissues are also presented if available.

2.5 Related Data Resource and Researches on CHO

CHO, as a cell line is widely utilised in mammalian cell biology research and biopharmaceutical manufacturing, it has drawn much interest from researchers in different fields. Many attempts at unveiling CHO cell biology profile have been performed before the CHO genome was sequenced.

However, not until the year 2011, when the first CHO genome was published by Xu et al. (2011), did most of the large-scale surveys become possible. To integrate CHO data resources in various perspectives, a Chinese hamster specific database, chogenome.org (Kremkow et al., 2015), was created. Up to the year this thesis is written (2017), four CHO related genomes were made available on NCBI: a CHO-K1 genome submitted by Xu et al. (2011), a CHO-K1GS genome submitted by Eagle Genomics Ltd and two Chinese hamster genomes submitted by Lewis et al. (2013) and Brinkrolf et al. (2013) respectively. Only the first CHO-K1 genome and the first Chinese Hamster genome, which both have accompanied transcriptomes reference, have been annotated by RefSeq and others were used to improve the annotation by NCBI. Annotation of the assembly of CHO-K1 was also performed by Ensembl, but the gene count on RefSeq annotation was different from that of the Ensembl version. Another annotation was done on the CHO-K1GS provided by Horizon Eagle and is available on Ensembl. Lewis et al. (2013) and Kaas et al. (2015) also conducted whole genome sequencing on CHO cell lines other than CHO-K1. However, these data are not available on public databases such as NCBI. Recently, the mitochondrial genome was intensively sequenced in multiple CHO cell lines to deepen understanding of the CHO energy metabolism (Kelly et al., 2017). The four CHO related genomes stored by NCBI are only assembled to scaffold level. Despite efforts of creating chromosomal map (Lewis et al., 2013; Brinkrolf et al., 2013; Xu et al., 2011), more evidence is required to order and orientate scaffolds to form chromosome map for Chinese Hamster or CHO. NCBI also continually improves the assemblies of these three genomes and integrates them into two reference assemblies, which are available on their ftp site, for Chinese Hamster and CHO-K1 respectively. Le et al. (2015) compared the RefSeq genomes released in 2012 and 2014 and the later shows significant improvement, indicating the importance of resequencing CHO related genomes.
Multiple sets of CHO related transcriptomics data are also available on NCBI. Compared to genomics data, expression data can be obtained not just by high-throughput sequencing but also by microarray. Up until 2017, 11 sets of expression profiling microarray data and five sets high throughput sequencing data had been made available in Gene Expression Omnibus (GEO) database on NCBI. Not all of this data was generated on CHO cell lines, for example, seven of the series were generated on hybrid cell lines which are also widely used in manufacturing.

Most of the research on CHO focuses on protein production processes: transgene expression, metabolism and protein secretion (Hefzi et al., 2016). To analyse and model these processes required input of transcriptome and metabolome. Some of the transcriptomic data was generated,

mostly with microarrays, even before the CHO genome was published. Full transcriptomic profile analyses were conducted on CHO cell lines at different growth rates (Doolan et al., 2013), in response to butyrate (De Leon Gatti et al., 2007), and at low temperature treatments (Kantardjieff et al., 2010) to understand cell metabolism in various environments. More broadly, transcriptomic data also supports specific interests focused upon miRNA (Hernández Bort et al., 2012) and product genes amplification (Vishwanathan et al., 2014). High throughput RNA sequencing was used more often after the first CHO genome. Multiple transcriptomic data have been generated on different CHO cell lines and different conditions to support research from different perspectives. Rupp *et al*. (2014) developed tools for integrating RNA-Seq data generated by different platforms to enable the reusing of this data for different purposes.

Another computational tool widely used to analyse CHO metabolism is flux models. Combining the genome and the metabolome could provide a systematic view of the metabolic network, upon which flux models can be built to predict the phenotype with metabolic flux analysis (MFA) and flux balance analysis (FBA) (Chen et al., 2012; Ahn & Antoniewicz, 2011; Goudar et al., 2010). These models are mostly focused on core metabolism and derive from models built for mice, so that they only partly cover the protein production process for CHO. Models that could accelerate cell engineering, clone selection and help optimise bioprocesses require genome scale multivariate input. Genome scale metabolic models have been proposed (Popp et al., 2016) and recently Hefzi et al. (2016) integrated models and data from different group and successfully constructed a consensus model for CHO accompanied by serval models specialised for widely used cell lines. They reported that the consensus model is able to predict metabolic features of CHO. The construction of these models is made possible by the availability of CHO related genomes and the transcriptome and proteome data derided from them.

As the CHO specific model being available, bioinformatics research on CHO could focus on improving the existing model by providing detailed information. While the genome-scale metabolic CHO model covers multi-omics of CHO, properly analysing the related data with statistic method and integrating the results to improve the metabolic model could be one of the main aspects of CHO related bioinformatics research. Although significant improvements have been made, a big part of CHO metabolism, particularly the regulation mechanisms, remain unknown. Combining high-throughput experiments and bioinformatics techniques would provide useful tools to reveal these unresolved parts so that more accurate models can be built.

3 Methodology

In this chapter, the analysis pipeline will be introduced followed by detailed descriptions and rationales of every step. Important concepts and criteria will be described with the steps involved. Lastly, the algorithms and detail settings of external tools will be described.

The main task of this project is to develop computational tools that could identify proteins with altered functions. The related functions and proteins, which will be referred to as defective or altered according to likelihood of significant alteration in this work, are considered to either have completely lost function or suffered mutations that change their wildtype function and biological efficacy. We only investigate mutations within coding areas as proteins are the main operators of biological activities. Although non-coding regions have been found responsible for regulation, identifying functional impacts on mutations on these regions is limited by the knowledge of regulatory mechanisms. In addiction, as the CHO genome at this stage is mostly generated by short read sequencing techniques and not yet fully assembled, genes are likely to be underrepresented, while the coding region (exons to be more precise) are more likely to be completely recovered (Alkan et al., 2011). Protein mutations with functional impacts were divided into two types: 1) mutations such as truncation and chromosomal aberration affecting large portions of the protein sequence, and 2) mutations such as SNV affecting only a small part of the sequences. Correspondingly, two strategies were used for different types of mutations: HMMs for detecting large area mutations, and site analysis for spotting mutations at important sites. HMMs were only used to generate global scores for the entire protein, while site analysis was used only for scoring locally for single sites. In effect, HMMs capture function-related patterns from sequence alignments and site analysis captures evolutionary history of each column of the alignment. Both of these strategies rely on sequence alignments and phylogenetic information of the target protein families. Therefore, a series of publicly available tools were used to construct alignments and phylogenetic trees in optimised settings. The general process is shown in Figure 3.1.

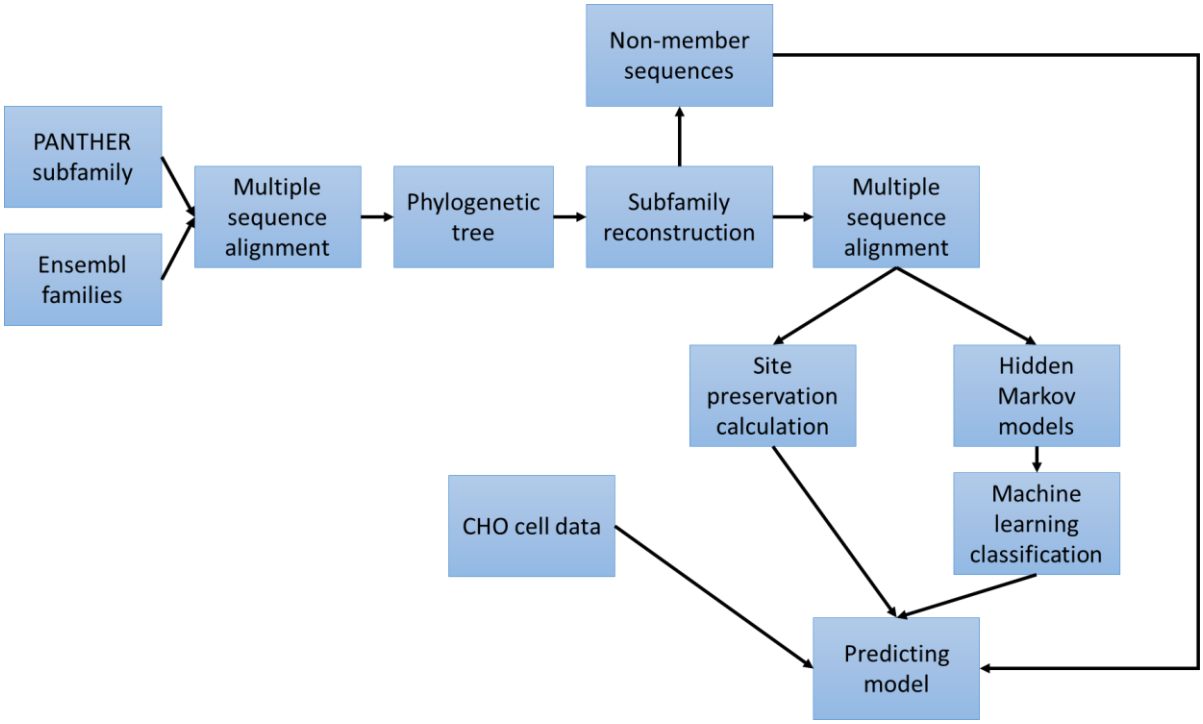


Figure 3.1 Flow chart of building the function predicting tool.
Firstly, homologous sequences from PANTHER and Ensembl were aligned. Phylogenetic gene trees were then built on these alignment referencing the species tree retrieved from Ensembl. According to the gene trees, new genes were added to the subfamilies forming an extended set of subfamily collection. The sequences of extended subfamilies were then aligned for building HMMs and preservation calculation. Sequences from the same tree but not in the subfamily were used for training machine learning algorithm which will contribute to prediction making with site preservation. Lastly, CHO cell data was analysed by the models.

3.1 Homolog Selection

Homologous sequences are frequently used for function inference for newly discovered proteins. Homologous proteins are often grouped together by sequential similarity forming families. Both NCBI and Ensembl have created a collection of protein families under slightly different criteria. These families are built on evolutionary relations without explicit reference to functionality. Although members within a short genetic distance are often expected to have similar functions, it is not rare that members within the same family encode completely different functions. To create groups based upon functional similarity, PANTHER proposed subfamilies on the basis of common protein families. They created subfamilies based on gene duplication events since these events

signal functional deviation between descendant linages (Mi et al., 2005) and manual curations were performed to improve the quality of the subfamilies. Such classifying methods largely inspired this research. However, this project is focused upon mammalian genomes and at the time this thesis is written (2017), only seven mammalian genome were included by PANTHER collection: human, chimpanzee, rhesus, mouse, rat, dog and cow.

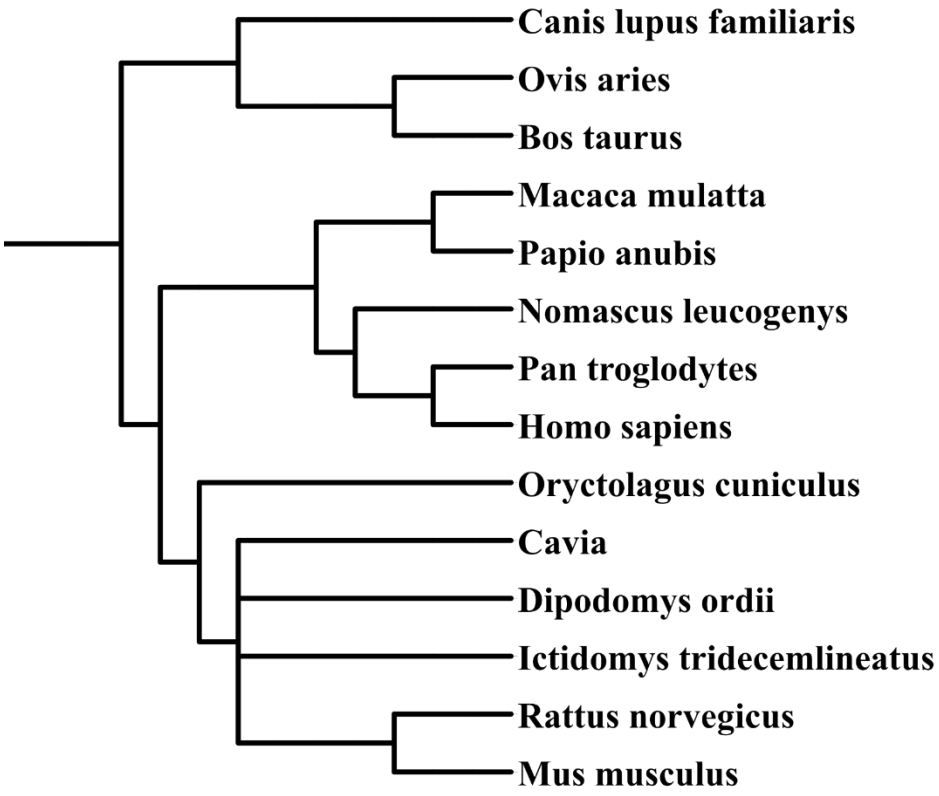


Figure 3.2 Mammalian species tree generate by iTOL 3
The figure was created using iTOL 3 (Letunic & Bork, 2016). The branch length is not proportional to the genetic distance. Only tree topology is presented. The tree topology was obtained by integrating data from Ensembl and NCBI.

To allow more accurate capturing of sequence patterns we integrated more mammalian genomes into PANTHER subfamilies to create an extended subfamily set. The organisms involved and their phylogenetic relations are shown in Figure 3.2. The special phylogenetic information was retrieved from NCBI and Ensembl. Data from fourteen genomes were used to construct HMMs, of which three of them: rhesus (*Macaca mulatta*), rat (*Rattus norvegicus*) and sheep (*Ovis aries*) were used as a test

set, while others were used to train models. The selection of these organisms was based on the species collection of PANTHER and their annotation level. It is clear that the extended set is centred on mouse and human since these are model organisms with the best annotated genomes. In addition, according to Ensembl, Chinese hamster and CHO-K1 genome are genetically very close to mouse. Therefore, a training set centred on mouse could yield better performance for CHO analysis.

Instead of scanning entire genomes for homologs, we use Ensembl protein families to guide homolog import into subfamilies. Since Ensembl protein families were built based upon sequence similarity, rebuilding these families using sequence searching tools such as BLAST should yield equivalent results. Therefore, using the existing family collection avoids the compute expensive recalculation. Compared to Homologene, the NCBI collection, Ensembl families cover all species selected for this project. By comparing Ensembl families and PANTHER subfamilies, we mapped subfamilies onto Ensembl families and extended the subfamilies by adding sequences in the same Ensembl families. For all the homologue collections we used, only members from selected organisms mentioned above were considered.

3.2 Aligning Sequences

All the multiple sequence aligning was performed by MAFFT 7 (version 7.313) (Katoh & Standley, 2013). As mentioned in the previous chapter, MAFFT is one of the most accurate and fastest aligners. It is also frequently used by PANTHER and data management institutes such as EBI as a component of their pipelines. In this project, MAFFT was used to align protein family members for phylogenetic tree construction, subfamily members for training HMMs and query sequences to corresponding alignments for identifying functional important sites. To preserve the integrity of conserved patterns, MAFFT was set to optimise local alignment even at the cost of global alignment score at all stages. When aligning new sequences to a given alignment, MAFFT provides options to avoid changes in the given alignment and generate additional map file for location look up. A detailed description of MAFFT is available in section 3.8.1. It is important to note that full-length protein sequences were targeted in the grouping and aligning process. Therefore, proteins created by splicing are considered fragmented. However, compared to assessing the intactness of wildtype function in individual

proteins, we were more focused on the function intactness of the proteome. Therefore, as long as full length sequences can be recovered with certain conservation in the proteome, we considered the function intact.

3.3 Phylogenetic Tree Construction

PANTHER uses conciliated gene trees generated by GIGA (Thomas, 2010), a tree building tool highlighting gene duplication events. Gene duplication is considered an important signal of function deviation in PANTHER. Compared to other gene tree building methods, GIGA is more reluctant to achieve high likelihood derived from sequential similarity while focusing on evolutionary events. Therefore, it is easily confused by sequence fragments and does not guarantee high sequence similarity within subfamily. In fact, the early version PANTHER (7.0) did include subfamilies with members of such low similarity that the HMMs built on the subfamily could not recognise these members (these subfamilies were removed on version 8.0) (Mi et al., 2013). In addition, most mammalian genomes involved in this project are mainly constructed by short read assembly which tends to underrepresent duplication regions. Therefore, duplication events may be missed in these genomes, which corrupts the performance of GIGA. In this project, we first built gene trees of Ensembl protein families using RAxML (version 8.2.11) (Stamatakis, 2014) and TreeFix (version 1.0.2) (Wu et al., 2013), then added new members to PANTHER subfamilies based on the corresponding gene tree. Alignments of members of extended subfamilies were then used to build HMMs for function prediction.

The combination of RAxML and TreeFIx was chosen since they are among the fastest and most accurate tree building tools (Noutahi et al., 2016b). Both tools are likelihood centred which could reduce the change of including distant homologs in the final subfamily collection and improve specificity of the HMM it derived. RAxML was first used to build preliminary unrooted binary trees from family member alignments generated by MAFFT. It is one of the most popular sequence-based gene tree constructors. Some surveys showed the RAxML outperformed other popular counterparts such as PHYML and MrBayes (Rasmussen & Kellis, 2011; Stamatakis et al., 2008). RAxML achieve accurate gene trees by maximising tree likelihood introduced by Felsenstein (1981). As mentioned in the previous chapter, maximising tree likelihood is highly computationally expensive. After years

of optimisation, the speed of RAxML has been significantly improved. RAxML provides optional substitutional models and is capable of automatically selecting the model that yields the highest likelihood. Here automatic model selection on protein families with less than 400 members was allowed. A GTRGamma model, which is one of the fastest models for RAxML, was assigned to protein families with more than 400 members. The most time expensive step for RAxML, the bootstrapping, was not performed at this stage. As tree modifications may be conducted at the next stage, examining tree confidence before the finalising step is unnecessary. The best trees generated by RAxML were taken for the next step.

Gene trees generated by RAxML were then rooted using the mid-point method. Mid-point rooting is the simplest rooting method that requires no additional parameter. As TreeFix requires a rooted tree but the root will not affect performance, mid-point method was sufficient. TreeFix can reconcile gene trees with a given species tree without significantly lowering the likelihood of the gene tree. It requires a gene tree, a species tree and a map of genes and species. All trees imported should be rooted and binary. The binary species tree was imported from Ensembl, and only species within the training set were included. TreeFix adopts similar likelihood calculation as RAxML but requires a specified substitution model. The GTRGamma model was used again and bootstrapping was performed in default settings.

3.4 Member Selection for Extended Subfamilies

Gene trees generated by TreeFix were used to create new subfamily collections. At this stage only the topology of the gene trees was used. Ideally all members of the same subfamily should share a common ancestor that in turn does not have descendants from subfamilies that include members of different lineages (Figure 3.3 a and b). Therefore, subfamily members were first mapped onto the trees to locate their closest common ancestors, then proteins under the ancestors were assigned to the same subfamily unless the proteins were included by another subfamily entirely derived from a descendent node of the common ancestor (for example, Figure 3.3 b). PANTHER subfamilies were defined using similar criteria. However, since tools with different tree constructing criteria were used, the new trees may conflict with PANTHER trees. Such conflicts would result in a situation similar to Figure 3.3 c where sequences of the different subfamilies are highly similar. Manual

curation after careful review of the alignments and related annotation could be a solution for this situation, but this was not feasible given the lack of human curator resources available. Therefore, two rules were used to resolve this problem: 1) when all the involved subfamilies contain less than four sequences, merge these subfamilies into a bigger subfamily; and 2) only add sequences one topology step away from the original member into the subfamily when the first role cannot apply. When two subfamilies both contain a small amount of highly similar sequences, HMMs built on their alignment would fail to distinguish members from these subfamilies. Therefore, merging these subfamilies could allow clearer separation between member and non-member sequences. In rare cases, a subfamily contains members from multiple Ensembl protein families so that such subfamilies would gather members from all related families. The new collection was intended to extend PANTHER subfamilies instead of changing them. As these subfamilies have been released, used and actively maintained for a long time, major errors of the collection had been corrected by manual curations and experimental evidences. Changing such collection using pure *in silico* results is not likely to lead to an improvement to the collection, hence the design of the second rule. The tree analysis was performed using the ete3 Python package at this stage.

After the member sequences were assigned, non-member sequences were selected as negative controls for final prediction. Since HMMs were allowed to report negative results, negative controls were used for identifying the score threshold between positive and negative result and to evaluate model performance. Non-member homologous sequences within the same protein family were first selected as negative controls. In less common cases, when non-member homologous sequences were not available, all non-member protein sequences from mice were used. As the mouse genome is well studied and relatively close to the CHO genome, being able to distinguish member sequences from other non-member mouse sequences indicates the model's capacity to perform well for CHO. As machine learning algorithms were used at later steps, the size of negative control sets was designed to be similar to the positive set unless not enough negative controls were available or the positive set was too large. When the number of reported negative controls is close to the number of positive results or reaches 200, the selection process ceases. As not every negative control could be reported by the models, the number of reported negative control could be far less than the positive results or even zero, in which case thresholds were not necessary since no evidence can be used to support the thresholding and we want to minimise artefact bias in this process.

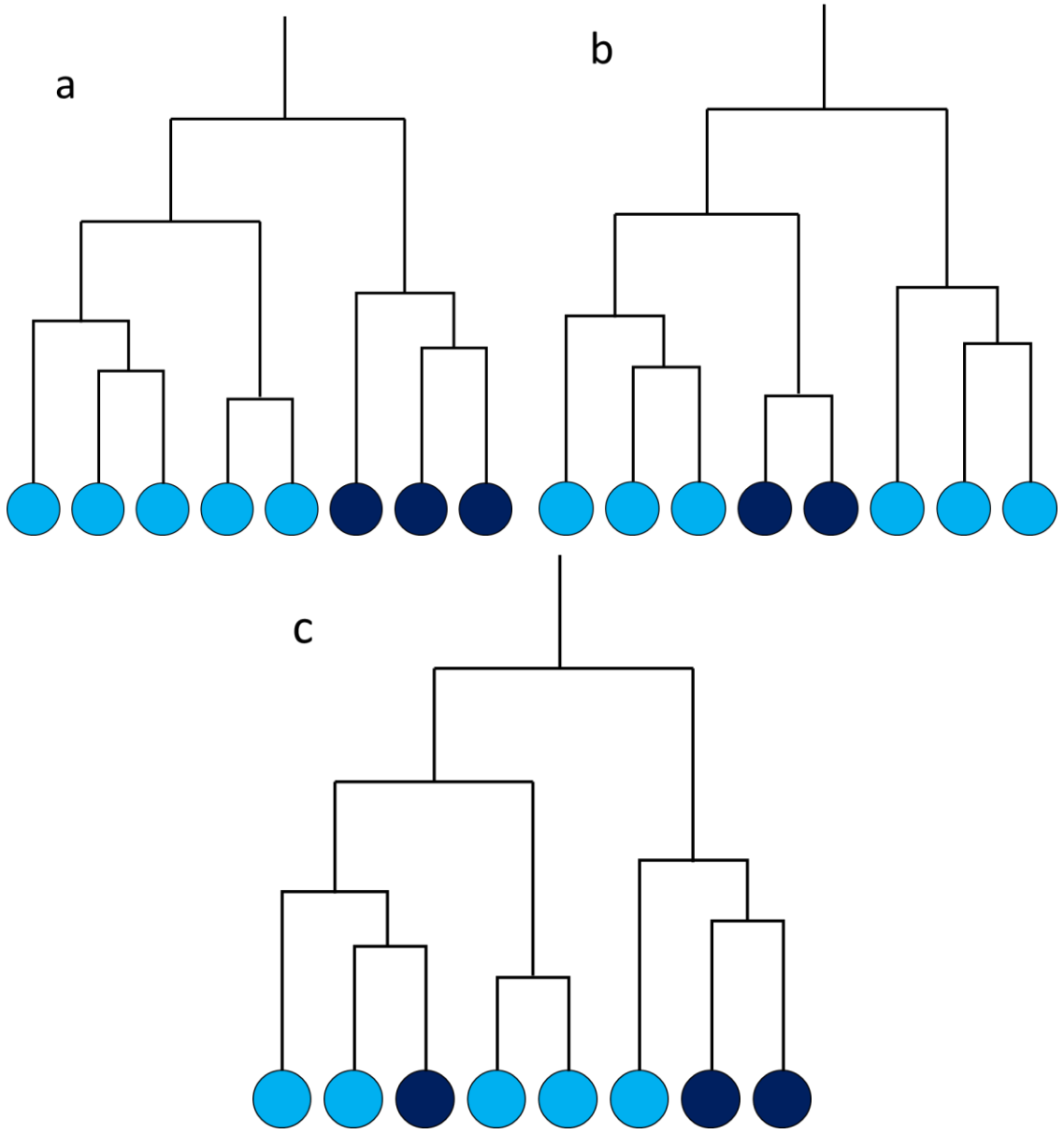


Figure 3.3 Samples showing homologs grouping with phylogenetic tree.
Three examples of phylogenetic tree with genes are presented. Balls of different colour representing genes of different PANTHER subfamilies. Grouping in a and b are acceptable while grouping in c is not.

3.5 Building HMMs

The sequences of new subfamilies were then aligned by MAFFT and HMMs were built on these alignments. The HMMs were built with HMMer 3 (version 3.1b1) on default settings. By default, HMMer evenly distributes weight to every variant observed regardless of the frequency. In addition, a built-in background residue pseudo-frequency would be used so that even if a residue is never observed in a certain position, the probability of observing that residue would not be 0. Such settings allow high sensitivity for detecting homologous sequences. At this stage we allowed our models to report non-member sequences as long as they could be distinguished from member sequences by scores.

3.6 Preservation Calculation

Non-synonymous SNV (nsSNV) impact on protein function has been widely studied for human cancer and diseases diagnosis. Therefore, most of the tools developed are biased towards humans and not suitable to be applied to CHO. The most popular tools that only rely upon alignment and phylogenetic information include SIFT (Ng & Henikoff, 2003) and PANTHER-PSEP (Tang & Thomas, 2016). As PANTHER-PSEP reported a significantly better performance, its concepts were largely adopted in this project. A core concept of PANTHER-PSEP is preservation, proposed by Marini et al. (2010), which means site conservation within related lineage. In PANTHER-PSEP, preservation was reported as the molecular age for which a site remains unchanged. Conservation usually refers to the frequency of residues observed in position, while preservation focuses on the genetic distance that certain residues have gone through.

In this project, we calculated site preservation as the number of species sharing the same variant. In PANTHER-PSEP preservation calculation, phylogenetic trees containing only orthologues were used. They constructed sequences for common ancestors and calculated preservation using those sequences. According to the estimation made by Ensembl, Chinese hamster and CHO shared a common ancestor with the mouse and the rat. Therefore, in the tree containing only species involved (Figure 3.4), only four ancestors (internal nodes) would be involved in the calculation. Although some gene trees may differ from the species tree, such incongruence should be less significant for orthologues from species separated by a long genetic distance. Thus, we focused on

three common ancestors (shown in Figure 3.4) with distinct distance, and divided species into three tiers accordingly so that we could calculate preservation according to the topology of the species tree. Molecular age was used by PANTHER-PSEP as a normalised measurement of mutation. As mutation rate can be affected by many factors such as population and generation time, molecular age is a rather approximate estimation. As we only compare preservation within subfamily, normalising between subfamilies was not required. Additionally, only three (out of four) distinct common ancestors were involved, which allows further simplifying of the preservation comparison into tier and species. Starting from mouse, the three tiers are referred to as tiers 1 to 3 with increasing genetic distance. We only calculate relative preservation. A variant shared by species only in tier 1 is less preserved than variants shared by species in both tier 1 and 2, and when variants are shared by species in the same tiers, variants shared by more species are more preserved. However, a variant shared by species in tier 1 and tier 3 is considered only preserved in tier 1. The maximum preservation that a variant can be is the number of tiers and species included by the subfamily. Based on these rules, the preservation of different sites of a protein can be easily compared. As the mouse genome is close to CHO and well annotated, preservation calculation was performed on mouse-related subfamilies, then transferred to CHO proteins using sequence alignment.

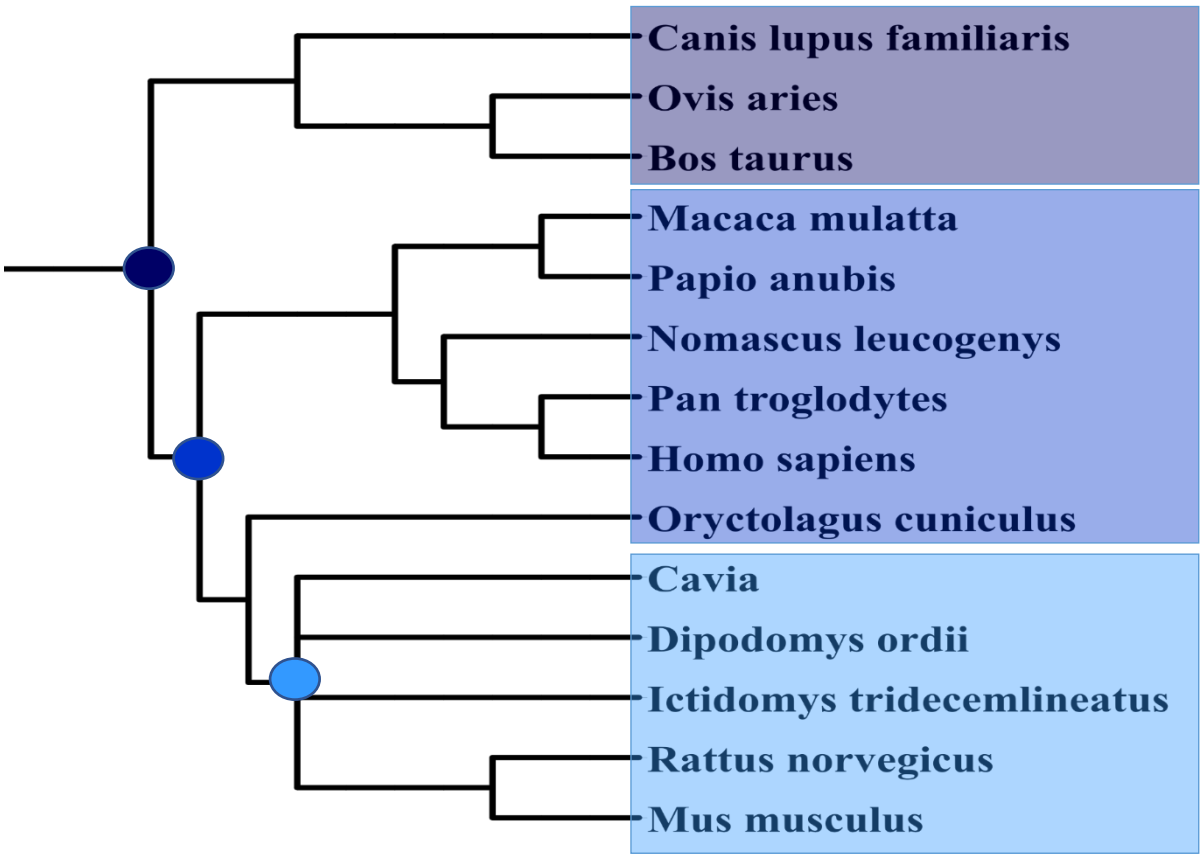


Figure 3.4 Species tree used for preservation calculation.
Species labeled in different colour box are assigned to different tiers representing the variants contained by ancetor node ladeled with the same colour. As CHO is genetically close to mouse, in calculating position preservation, mouse was used as a starting point to go through the marked nodes, from botton to top, and count species that share the same recidue in the same position in the process. The tiers mouse went though and the number of species sharing the same recidue in the same position would form the preservation level.

3.7 CHO Data Analysis and Prediction

After the models and the preservation were obtained, Chinese hamster and CHO sequences were imported to be analysed. HMMer contains BLAST like tools to compare models and sequences and report similarity scores: the E-value and bit score. These two scores are also used by BLAST and should be familiar to most biologists. In simple words, E-value is the expected probability of observing a sequence of equal or better similarity by chance, while bit score is a form of E-value

normalised by the length of query sequence. We used bit scores as it only relies on similarity. We searched for the best matching sequences in CHO using the models. As copy number variant was not considered in this project, finding the best available sequence for every sequence should obtain the same result as searching for the best fitting models for every sequence. After the CHO scores were obtained, protein sequences were used to build HMMs and the negative control sequences were also scored.

When negative control was available, a simple machine learning algorithm: nearest centroid classifier was used to identify the potentially defective or outgroup proteins. Given positive and negative controls, identifying such proteins can be considered as a classification problem, where being classified as negative means being out-of-group or potentially defective. Classification was done on one-dimensional data and thus a simple classification algorithm would be sufficient. Nearest centroid classifier is a straightforward tool for supervised classification available in Python package scikit-learn (Pedregosa et al., 2012). The main strategy of this classifier is gathering sample points that are close to each other which is consistent with our expectation of matching score distribution. It requires no parameters other than training data set and provides non-probabilistic predictions. However, it adopts an underlying assumption that the positive and negative training set share the same variance. The classifier was first trained by the positive and negative control sets, and then used to predict the classification (intact or potentially defected) of CHO proteins. The prediction was made according to the distance to the centroid (mean) of the training set.

The Fowlkes-Mallows index (FWI) was used to evaluate the classification performance. FWI is an evaluation index based on the number of true positive (TP), false positive (FP) and false negative (FN) without making assumptions. It can be calculated by:

$$FMI = \frac{TP}{\sqrt{(TP + FP)(TP + FN)}}$$

FMI is bounded on 0 to 1 where 1 means perfect classification and 0 means total disagreement with the expectations. Only classifications with FWI higher than certain threshold would be used for predictions.

CHO proteins deemed positive were then examined by preservation. To estimate mutation impact in function, Swiss-Prot annotations were also used with the preservation. Swiss-Prot is a subset of

the Uniprot Knowledgebase that has been reviewed by experts. Although Swiss-Prot covers most of the site annotations found, only a small proportion of sites were annotated. Preservation analysis could provide evidence for estimating the impact of unannotated sites. Although it is true that the number of mutation at highly preserved sites is not proportional to the functional impact they cause, high mutation rates at preserved site does suggest a higher chance of functional disruption. Therefore, the proportion of sites losing their preservation was used to support prediction of function alterations on CHO proteins. We defined function predictions as 1) defective, 2) altered or 3) intact (also referred to as normal in places), these being 1) completely lost and no full length proteins found, 2) likely being altered in function to some extent, and 3) functioning as wildtype, respectively. The detailed parameters were decided according to a preservation survey of annotated sites, described in the next chapter.

3.8 Related Software Description

In this project, we used multiple publically available tools. These tools are all well developed with powerful functions that may not be mentioned in this thesis. Comprehensive tutorials and documentations of these tools are available online. For readers who are not familiar with these tools, we provide brief introductions on their function related to this project.

3.8.1 MAFFT

In this project, MAFFT (version 7.313) (Katoh & Standley, 2013), one of the best performing aligners, was used as a primary aligning tool for quickly aligning with high accuracy. MAFFT applies a mathematical measure known as correlation, which can achieve character-by-character comparison but is significantly accelerated by Fast Fourier Transform (FFT). The correlation of two sequences works in a swiping card manner. The head of one sequence is firstly aligned with the tail of the other sequence where the overlapping area is compared. Then the overlapping area is then enlarged by moving one sequence toward the end of the other sequence step by step, and comparisons are made in every step during such process. In the case of alignment protein sequences, instead of the actual residues, vectors, which consist of the products of the volume values and of the polarity

values, are used in the comparison. To align multiple sequences, a guide tree is used to progressively increase the number of sequences involved. When high accuracy alignments are required, the guide tree is reconstructed according to the alignment and is used to reconstruct the alignment again. This process iterates to improve aligning accuracy until the maximum iteration number is reached or no changes are made on the new alignment.

MAFFT has multiple aligning algorithms built in for achieving different accuracy-speed requirements. In this project, the most accurate and time-consuming built-in algorithm was used for high quality alignments. Although applying the most time-consuming algorithm, MAFFT is still one of the fastest programs to achieve such quality of alignment. The actual command in this project was:

Mafft --localpair --maxiterate 2000 --anysymbol <raw_sequence_file> > <alignment_file>

where three options were used. Option "--anysymbol" instructs MAFFT to use not only the general one letter amino acid alphabet, which contains 20 usual residues but also the alphabet which contains unusual charters such as 'U' as selenocysteine. Option '--localpair' instructs MAFFT to focus upon achieving the best performance on conserved domains and allow long gaps added in variant regions when aligning sequences. MAFFT provides two general alignment strategies: 1) 'localpair' focuses on aligning the most similar part of the sequence, and 2) 'globalpair' focuses upon achieving best match for every residue in the sequence, even if that means adding more short gaps in the alignments. Using 'localpair' in this project is based on the hypothesis that critical sites spread across the peptide as short sequence islands and regions between these islands carry more mutations. Adding gaps to the critical sites to achieve better matching on non-conserved regions would mislead the models to reduce weight on highly conserved residues. Option '--maxiterate 2000' sets the maximum iteration number for refining alignments, avoiding excessive time cost and potential deadlocks.

The same options were used when adding new sequences to a given alignment with two additional options '--addfull' and '--mapout'. When '--addfull' option is applied, one sequence or more will be required. A guide tree containing all the sequences including the new sequences, then alignment calculation will be performed on nodes related to new sequences. '--mapout' instruct MAFFT to generate an additional file recording the residue position in original sequence and within the alignment. When '--mapout' is used, an additional option '--keeplength' is automatically activated

to avoid gaps added to the alignment and disable iterative alignment refinement so the given alignment will not be changed. Otherwise, gaps may be added to the given alignment even when alignment refinement is disabled. The command used in this project is:

Mafft --localpair –anysymbol –addfull <new_sequence_file> --mapout <reference_alignment_file> > <new_alignment_file>

When multiple cores are available, option '--thread' can be used to run aligning using multiple cores, which could significantly increase the speed. The number of core used can be specified after the option.

3.8.2 HMMer

HMMer is a software first developed by Sean Eddy (1998), who further improved its searching power in version 3 (Eddy, 2011). HMMer (the version used is 3.1b1) is based on the Hidden Markov Model (HMM) which derives from Markov Chain (MC). Both types of models describe a series of events or states occurring in order.

In HMMer 3, HMMs are built by a tool called *hmmbuild* based on the method introduced by Krogh *et al*. (1994). Three different kinds of states in alignments are modelled: match, insert and delete. The HMM architecture allows each site to be modelled by any of the three states. The emission probability of match states is subject to the observed distribution of residues on the site. Certain background residue distributions and pseudocounts, which essentially count residues that are not observed, are applied, so that the observed residues have a relative probability showing the significance of their observation, and other residues that are not observed have probabilities other than zero but subject to the background. Insertion and deletion events are modelled by insert and delete states respectively. The core HMM architecture used by HMMer is stringently linear, which complies with the linearity of biological sequences. The HMMs built to describe alignments are called profile HMMs (Figure 3.5 with core model in black and grey) in the literature (Eddy, 1998, 2011; Finn et al., 2011). By default, the background residue distribution is set to the residue frequency in Swiss-Prot 50.8, which is hardcoded into the software. Alternatively, the background can be set to a uniform distribution or subject to the profile (Eddy & Wheeler, 2013). However,

HMMer does not simply adopt the residue frequency from the alignment but instead, weights the sequences in the alignment first by default. It generally assigns more weight to sequences carrying more variants to retain maximum information.

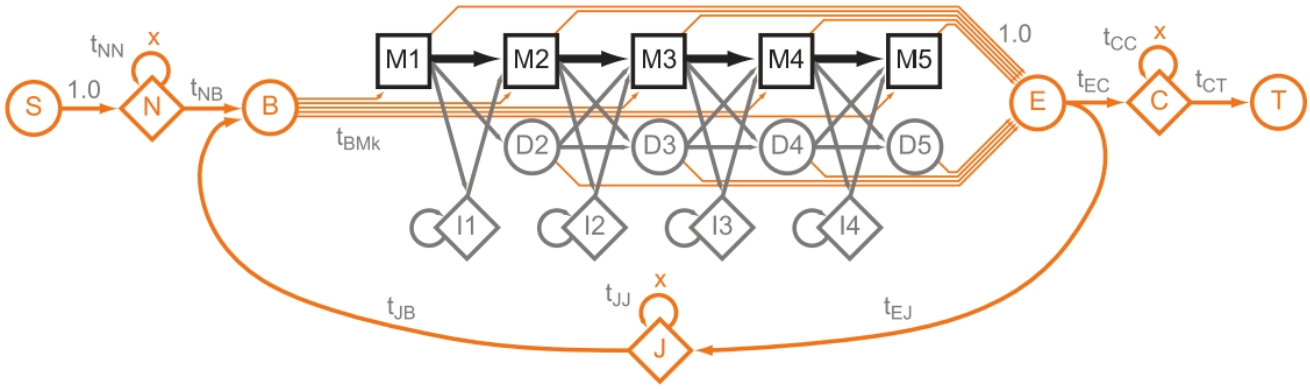


Figure 3.5 Profile HMM with searching structure
The figure was created by Eddy (2011). The core model is marked in black and grey. Ms stand for match state, Is for insertion and Ds for deletion. State S and T stand for the start and termination of the aligning process. State N, J, C, as N terminus, joining and C terminus, present the flank of core alignment. State B and E represent the beginning and ending of the core homologous region. Each match state is connected to an insertion state and a deletion state which could result in skipping in any number of match states.

One of the most crucial applications of HMMs is to search for matching sequences in a database or, conversely, matching a query sequence with one of the models in the HMM database. *Hmmsearch* and *hmmscan* are the tools designed for such purposes in the HMMer toolbox. To achieve database searching a searching engine is required. Essentially, matching model and sequence is to calculate their similarity. For HMMs, forward algorithm and Viterbi algorithm are two fundamental methods that can achieve such purpose. Forward algorithm is used to calculate the probability of achieving a certain hidden state given the observed sequence, while the Viterbi algorithm is used to calculate the most likely sequence of hidden state given the sequence of an observed state. Therefore, every possible hidden state in the same step would be considered when calculating the next step in the forward algorithm but only the optimum of these states is calculated in Viterbi. As a result, the Viterbi algorithm is less computationally expensive as the heuristic strategy is applied. In HMMer, searching with Viterbi algorithm is 3- to 9-fold faster than forward algorithm so that it is more preferable (Eddy, 2011)

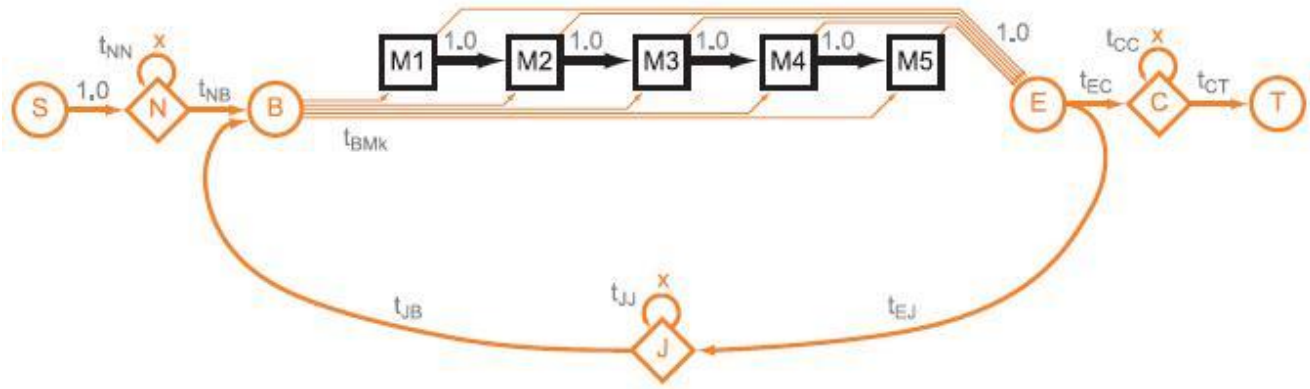


Figure 3.6 HMMer 3 HMM for comparing sequences
The figure was acquired from Eddy (2011). State S and T stand for the start and termination of the aligning process. State N, J and C are flank states standing for N terminus, joining and C terminus. State B and E stand for the beginning and end of homologous region. State M in the core process stands for matches in the alignment. Compared to the structure shown in Figure 3.5, deletion states and insertion states are removed forming ungapped alignment segments.

Whether implementing the forward or Viterbi algorithm in searching, the goal is to calculate the likelihood of the hidden sequence given the query sequence as observed. However, searching using ordinary methods are far more slower than BLAST, which makes it inappropriate to apply to database searching (Eddy, 2011). To extend the application of HMM, Eddy group developed a fast model-based searching method to allow searching with HMM to reach a comparable speed with BLAST. Another model is built on part of the HMM to guide the searching (Figure 3.6). In this model, the insertion and deletion states in the profile HMM are not considered, and only the match states are considered, which means it is the matched sequences without gaps that the program is searching for. The matching starts at state *B,* which can then transit to any match state. The model allows the matching to start at any point of the model and correspondingly any point of the query sequence. After the first match is found, the following sequence is then automatically compared to the following matching state until the similarity score of the match drops below a certain point, which is decided by the transition probabilities from state *E* to state *J* and *C*. Certain thresholds can be set so that short matched sequences that do not score highly enough are excluded. Generally, long matched sequences with high matching scores can continue with a few low score matches, which makes searching less sensitive towards discrete point mutations. As only a part of the HMM is involved in searching and it is effectively simplified, the computational power requirement is

highly reduced. Parallel versions of the method are also available for quick database searching with high performance computers.

Similar to BLAST, HMMer reports the matched sequences along with two statistical scores, the E-value and the bit score. The E-value, as mentioned previously, is the expected possibility of having a sequence with equal or better similarity by chance. It can be calculated as:

$$E = Kmne^{-\lambda S}$$

where *n*, *m* are the length of query sequence and of the sequences in the database, which also define the searching space, while *K* and λ are the scales of the size of the searching space and scoring system of the segments, and *S* is the similarity score set in the algorithm for discarding fragments with similarity below the score (Madden, 2013). The E-value relates to the length of the query sequence which is reflected by *n*, the size of the database which is reflected by *m* and the similarity score *S* which is affected by many factors of the scoring system. The lower the E-value, the less likely that the match occurs by chance, which equally means the more significant the match is statistically.

Compared to the E-value, bit scores only relate to the similarity score and the scoring system but not the length of the query and the size of the database. This is given by the formula:

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

the bit score *S'* can be considered a normalised version of *S,* which is the raw score generated directly by the scoring system as *K* and λ are derived from the scoring system. Such normalisation allows bit scores from different alignment to be compared.

3.8.3 RAxML

Randomised axelerated maximum likelihood (RAxML) (the version used in this project is 8.2.11) is one of the most widely used sequence-based phylogenetic tree building software. Like a lot of other tools, RAxML builds phylogenetic trees by maximising likelihood proposed by Felsenstein (1981). In

general, such likelihood is the product of prior probability derived from residue or base distribution, the probability of change of branches and the length of the branches. The computational complicity increases with the number of branches, therefore, calculating likelihood for trees involving 10,000 taxa would be time consuming. To maximise the likelihood of a tree, tree rearrangement would be conducted many times searching the potential solutions. As it is impossible to test all possible topology for trees of usual size, parsimony methods are used. A general strategy is that first generating a tree with simple algorithms which often do not encounter likelihood calculation, then rearranging a small part of the tree to reach local maximum in likelihood with hill climbing methods and then repeating the maximising process on other parts without changing the maximised branches. Random rearranging is used for generating proposals in hill climbing for maximum likelihood. In this way, since only a few branches are rearranged every step, likelihood calculation can be largely simplified as only the likelihood of changed branches need recalculation. To enable multi-thread computing on high performance computers (HPCs), specialised versions of RAxML are made available. The actual distribution used in this project is RAxML-PTHREADS-SSE3, a multi-thread vectorised computing version.

When calculating probability for change for likelihood, certain substitution models are required to obtain the substitution rate. In RAxML, multiple models are available, of which the GTRCAT is the fastest for large tree calculation. However, according to the manual provided by the developers, GTRCAT is not recommended when taxa is less than 50 and GTRGamma should be used in this case. GTRGamma is a general time reversible model based on Gamma distribution mutation rate. Automatic model selection is also available. In this case, the program selects the model which yields the highest likelihood. However, as tree topology is much more important than branch length in this project, a slight change in likelihood caused be model selection is acceptable.

After tree with maximum likelihood is acquired, bootstrapping can be used to evaluate how well the data supports the result. In this process, pseudo-sequences are created according to the input alignment and added to the sequence poll which will be resampled to generate a new sequence collection. Such collection will then be used to build the tree with maximum likelihood using the protocol described above. The resampling process will be repeated many times and the more the same tree is acquired, the better the tree is supported by the data. However, this bootstrapping is highly computationally expensive, although it has been significantly accelerated in RAxML.

3.8.4 TreeFix

TreeFix (used version 1.0.2) is one of the best reconciliation tools. Two series of TreeFix, TreeFix and TreeFix-DTL, are available for eukaryotic and prokaryotic genomes respectively. It requires a rooted species tree, a map of gene and species, a preliminary gene tree (preferably the ML tree) and the corresponding alignment as input to generate a reconciled gene tree. TreeFix reconcile species trees and gene trees by minimising reconciliation cost within the tree topology space where the likelihood is statistically equivalent. It assumes that minimum reconciliation cost can be reached in that tree space, but when the assumption is not applicable, ML trees with high reconciliation cost will be returned.

TreeFix calculates likelihood using RAxML package. It is recommended that the same substitution model should be used for both building ML trees as input and the reconciliation process. The model selected in this project is GTRGamma. To estimate the significance between different tree topologies, TreeFix conducts statistical tests on likelihood change between these trees, with the null hypothesis being all trees are equally supported by the alignments. TreeFix calculates reconciliation cost as the minimum evolutionary events, loss and duplications, required to solve the incongruence between the species tree and the gene tree.

Bootstrapping analysis is also available on TreeFix. However, TreeFix only resample gene trees with same topology as the input tree and corrected branch length as bootstrapping from ML gene tree construction would be too computationally expensive.

4 Model validation and performance

In this chapter, details of validations of HMMs and preservation calculations are described. After examining model performance, criteria and standards decided by the behaviours or performance of the models will be described and justified.

During the model construction process, 16,749 Ensembl protein families were involved, of which 15,620 had phylogenetic trees built, while the rest of the families contained fewer than three members from the selected genomes. These protein families derived 19,408 subfamilies where 342 original subfamilies were merged into others subfamilies in the final set. Equal number of HMMs were built on these subfamilies. Site preservation was calculated on these subfamilies, however, only 18,594 of them had at least one member annotated by Swiss-Prot. The performance of the HMMs was evaluated using gene sets from rat, rhesus and sheep, while the performance of preservation was investigated using PANTHER-PSEP results and Swiss-Prot site annotation.

4.1 HMM validation

The HMMs built in this project were used for two purposes: identifying the subfamilies to which query sequences belong, and identifying sequences with large area mutations including protein fragments. Amid these two purposes, the former can only be achieved by HMMs while the latter was aided by preservation calculated in other steps. Therefore, it is more important that the HMMs can correctly assign sequences to the subfamilies, while larger error on predicting protein defects is acceptable.

Subfamily collections of rat and rhesus were used to validate model performance on assigning sequences. These two genomes are included by the original PANTHER subfamilies but were not used for training the HMMs. Models built on subfamilies with no members from the corresponding genome were not involved in the validation as they will always assign proteins to subfamilies different from PANTHER. The rest of the models were used to find the best matched proteins on these genomes. Recovering members in this process would mean being successful in assigning the correct subfamilies. As a result, 16,258 of 17,634 models successfully recovered their member in

rat making up a correct rate of 92.2%. The numbers for rhesus were 15,385, 16,694 and 92.2%. We speculated that the main error came from subfamilies of the same families which are inherently highly similar. Therefore, we further investigated model accuracy on assigning proteins to correct families. The number of models correctly assigning families was 17,634 in rat and 16,385 for rhesus, making up 98.7% and 98.1% respectively. These numbers confirmed our speculation. Based upon these numbers and the numbers for subfamily assigning, we believe that our models are able to assign proteins to correct homologue groups.

In the prediction process, homologous sequences not included by the subfamily were used as negative controls. However, not all subfamilies have homologous sequences available from the selected genomes and 1,126 subfamilies fell into this situation. For these subfamilies, predictions were made solely based on site preservation. The predictions made by models would be largely affected by the scores of corresponding negative controls. To investigate the difference between scores of positive and negative controls, a t-test was performed on these two sets of scores for every subfamily with negative controls. In the 18,282 models with negative controls, 1,186 were not applicable for t-test since only one negative control yield a score; 1,836 have a two-tail P-value larger than 0.05 and the number increases to 2,531 for a P-value of 0.01. Therefore, we believe 15,260 (83.5%) models should provide good predictions at a significance level of 0.05. However, we brought all models to the next validating step. For the models without negative controls, positive predictions would be reported and passed to the next step.

In the validation of predicting protein intactness, a procedure similar to subfamily assigning validation was applied. Models of subfamilies not having members from the related genome were firstly removed and then the best matching sequences were used for validation. However, in this case, we used not only proteins from rat and rhesus, but also proteins from sheep which is not included by PANTHER subfamilies. Therefore, all models were used for sheep proteins. We expected most of the proteins to return positive results. Eventually, the models returned 1,682 (9.6%) negative predictions out of 17,606 reported predictions for rat proteins and the number for rhesus were 1,630 (9.8%) and 16,706. For sheep proteins, which have not been included by PANTHER, from the 19,373 reported models 2,673 (13.8%) of them reported negative results. To investigate the cause of significant higher negative rate on sheep proteins, we examined prediction results on rat and rhesus proteins using all models. It showed dramatic increase on both rat and rhesus, being 15.5% (2,989 out of 19,312) and 19.5% (3,759 out of 19,309) respectively. This result

showed that most of the models of subfamilies not involving the species yielded negative predictions. 1,307 negative predictions for rat came from 1,706 unrelated subfamilies and the numbers for rhesus were 2,129 from 2,603 subfamilies. As homologous sequences not in the subfamily were used as negative controls, when homologous sequences are identified by models of subfamilies of the same protein family, negative predictions will be made. Therefore, high negative rate caused by unrelated subfamilies can be considered as evidence showing that models work as expected. On the other hand, when unrelated subfamilies were removed, a significant proportion (about 10%) of proteins were predicted defected. It is highly likely that protein fragments are stored as a protein in these data set and they should be diagnosed as defected, but we were expecting a lower proportion. Since we used models to find the best matching sequences, only protein fragments not having whole protein homologs should be reported as defected. Therefore, we were only expecting a small percentage of negative prediction.

To address the cause of unexpectedly high rate of negative prediction, we further investigated the performance of classification by nearest centroid in the predicting process. Along with the predictions, the FMI was reported suggesting the quality of the predictions. A FMI close to 1 means perfect classification on training set while a FMI close to 0 stands for the opposite. One of the main causes of low FMI is indistinct separation of scores of positive and negative control. Therefore, the list of models with P-value higher than 0.05 in t-test was compared with the list of models with a FMI lower than 0.7. The FMI threshold of 0.7 was selected since it is the FMI obtained when all the samples are considered intact or defect and the number of positive and negative controls are equal. This comparison showed that 1,548 out of 1,836 models that failed t-test at a level of 0.05 have a FMI lower than 0.7. In total, there are 6,144 models with FMI lower than 0.7 reported on prediction for sheep proteins, accounting for 31.7% of the total models reported. After removing models with low FMI, only 515 models and 592 models, accounting for 2.9% and 3.5% of models of related subfamilies, reported negative for rat and rhesus proteins respectively. For sheep proteins which examined by all models, 909 (4.7%) models reported negative. It is important to note that when a model returns negative prediction, it does not necessarily mean the protein identified by this model is defected, although it could be the case. Instead, it actually means the protein carrying the function represented by the subfamily is defected in the corresponding species. The protein identified by the model could be identified by another model which reports a positive result and the subfamily whose model returns a negative result may not be needed by the species. Therefore,

for rat and rhesus, as only related subfamilies were accounted, negative predictions means the proteomes only contain fragmented proteins, probably caused by splicing effects but this could also be the result of incorrect gene or coding region boundary, or bad models caused by incorrect alignment or improper homolog collection. The negative results also means unrelated subfamilies for sheep proteins and also for CHO proteins in later sections. The prediction from models with low FMI were carried on to preservation analysis. Since the main function of HMMs is assigning subfamilies, the intactness prediction works as a filter at this stage. Therefore, only confident negative predictions would be cut off when CHO proteins were examined.

4.2 Preservation analysis

The preservation model is built for identifying functionally important sites. It can complement HMM predictions with local conservation information to generate more accurate results. Although summarising site preservation on whole sequence could achieve functional estimation like HMMs, in this project, site preservation was centred on the functional effect of single site. As the preservation analysis in this project is largely motivated by PANTHER-PSEP, we compare our preservation level with the molecular age calculated by PANTHER-PSEP for validation. Then we investigated preservation distribution on annotated sites whose information was obtained from Swiss-prot. The preservation effect on prediction was also discussed in this section.

In order to compare molecular age with our preservation level, the precomputed molecular age results were retrieved from PANTHER ftp site. PANTHER-PSEP automatically recognises the member most similar to the query sequence within its database and starts calculating molecular age from that member. However, in this stage, our approach requires a specified starting species from the selected species collection. Human was selected for this validation process since PANTHER-PSEP was built for recognising disease-causing site variant and human proteins are most well-annotated. However, when CHO proteins were applied to the model, mouse proteins were used as the starting point for preservation calculation. Our preservation level was presented with a tier number and a species number and the preservation increases with firstly tier number then with species number. The average molecular age of sites with same preservation level was used for the comparison. The result is shown in Figure 4.1. Note that since there was only one site preserved on two tiers with nine species, it was not shown in the Figure 4.1 due to lack of statistical

significance. The Figure 4.1 shows a trend of increase of molecular age with preservation level. However, for the levels of the beginning (especially first two levels) of every tiers, the molecular age increases unexpectedly. This could be the result of gene trees that are significantly different from the species tree or a more distant homolog not included by the selected species in this project. PANTHER includes much more species than that in this project, resulting in much longer linages. PANTHER-PSEP use reconstruction probability to build links between distant homologs, which means even when a variant is not observed in all internal notes, as long as it is observed at both ends of the linage it could be considered as preserved throughout the entire linage. Apart from these, the molecular age seems to correlate well with our preservation levels, though not linear. When we inspect the PANTHER-PSEP results, we found that although molecular age is used, the outcome ages are highly discrete and could actually be fitted in a level system. Overall, only 22 different ages are assigned to sites of human proteins (supplement 1). Therefore, although we used a different calculating protocol, the level system we used performs more similar to the molecular age from PANTHER-PSEP than we expected.

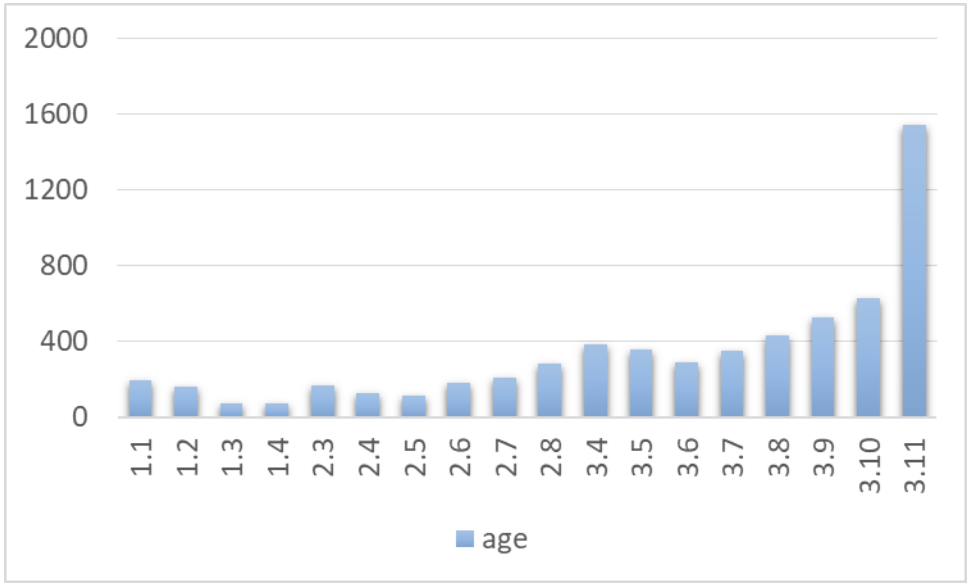


Figure 4.1 Average of molecular age on preservation levels.
The molecular age is shown in the Y axis while the preservation level is shown in the X axis. The average molecular age for sites of the preservation level is presented. A trend of increase of molecular age with preservation level is presented, showing correlation between two metrics. However, for the levels of the beginning (especially first two levels) of every tiers, the molecular age increases unexpectedly.

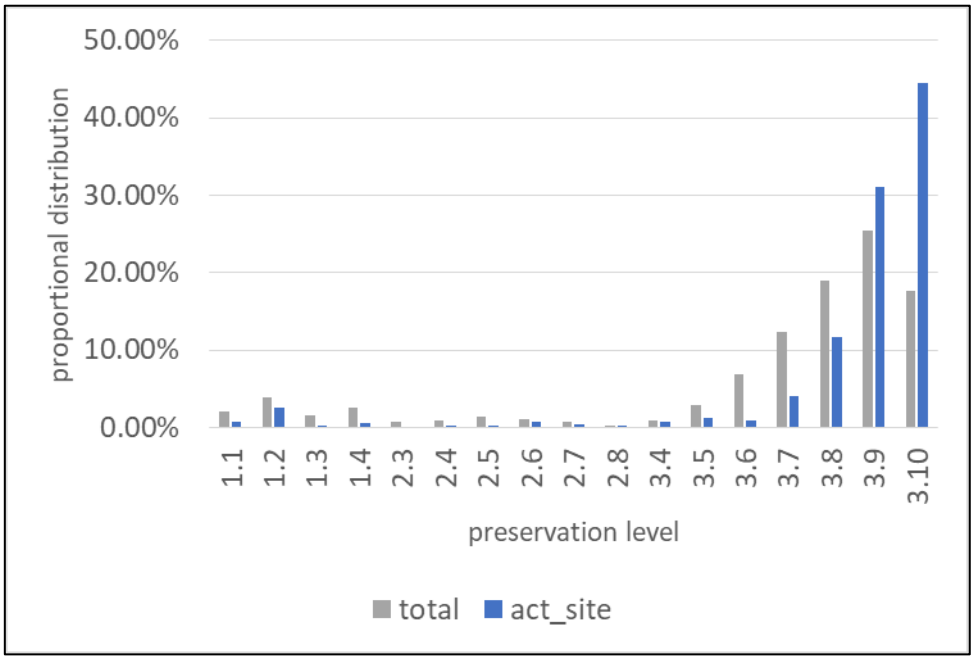
Table 4.1 The numbers of sites with different marks with different preservation level.
The sum of sites annotated by different market are shown at the bottom with the two-tail p-values calculated in t-test with the numbers in the second column as reference. The p-values are all smaller than 0.01, showing significant difference in distribution against the general distribution.

Preservation	total	act_site	np_bind	ca_bind	binding	metal	zn_fing	dna_bind	mutagen	variant
1.1	256520	11	328	210	19	31	10886	1330	187	2905
1.2	464289	42	558	74	93	96	36381	524	214	2531
1.3	196628	4	120	44	31	19	7097	299	95	1015
1.4	317858	10	167	56	17	14	12569	417	242	2041
2.3	84169	0	35	19	7	3	848	115	69	660
2.4	107730	2	34	68	3	2	892	200	95	398
2.5	165218	5	62	29	18	9	1330	503	142	797
2.6	126027	11	80	43	12	20	1538	103	115	760
2.7	81565	7	63	45	13	10	1201	85	69	416
2.8	34622	3	68	8	15	4	186	24	29	210
3.4	110580	13	209	53	21	39	917	211	127	565
3.5	345252	22	507	95	47	78	3149	1341	275	2751
3.6	830599	14	705	346	38	49	6215	2242	507	1933
3.7	1499380	69	1349	304	131	136	13580	3440	1343	4354
3.8	2312682	195	2641	691	405	352	20404	4285	2340	7635
3.9	3089517	521	5796	1415	913	1088	30690	10786	4412	16069
3.10	2138031	746	7192	1040	1536	1464	31297	10502	5555	22982
3.11	306	0	0	0	0	0	0	0	0	0
sum	12160973	1675	19914	4540	3319	3414	179180	36407	15816	68022
t-test pvalue		0.4182%	0.4232%	0.419%	0.4186%	0.4187%	0.4703%	0.4279%	0.4221%	0.4369%

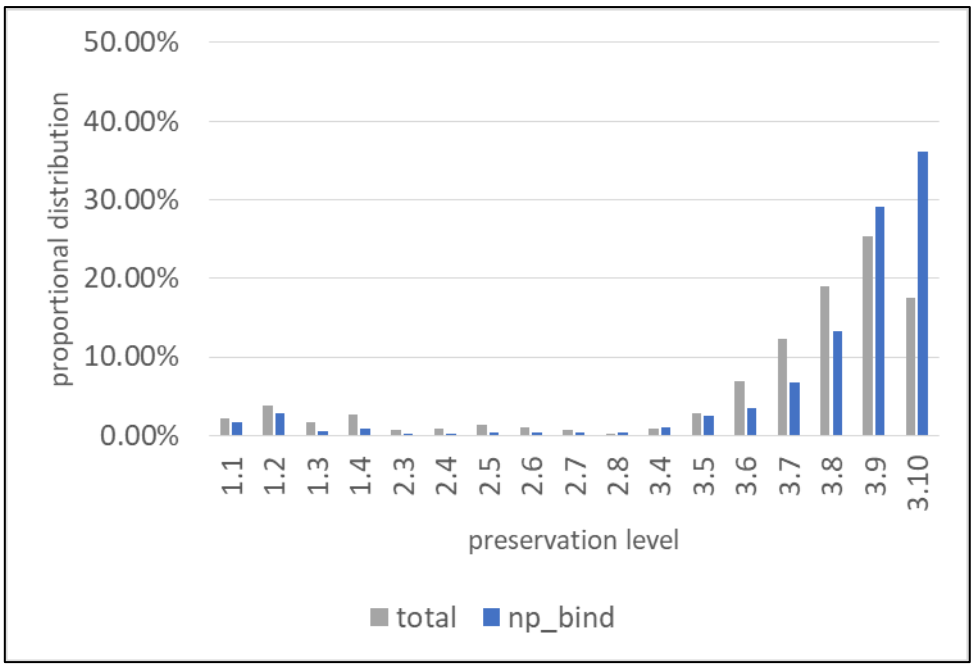
The preservation results were then aligned with Swiss-prot site annotations. As we wanted to investigate the preservation of functionally important sites, function related annotations were mainly involved. Sites annotated with marks of 'ACT_SITE', 'NP_BIND', 'CA_BIND', 'BINDING', 'METAL' AND 'ZN_FING' were selected as focus groups, of which 'ACT_SITE' and 'NP_BIND' standing for active sites and nucleotide, such as ATP and cAMP, binding sites respectively, were excepted to be most conserved. Apart from these function related sites, sites annotated with marks of 'MUTAGEN' and 'VARIANT' were also used as they were considered less conserved. Distribution of overall site preservation was used as base-line to highlight change in preservation. The numbers of

sites with different marks with different preservation level are shown in Table 4.1. The p-values of two-tails t-test against overall preservation level distribution was shown in the last raw, all showing significant difference from the base-line distribution.

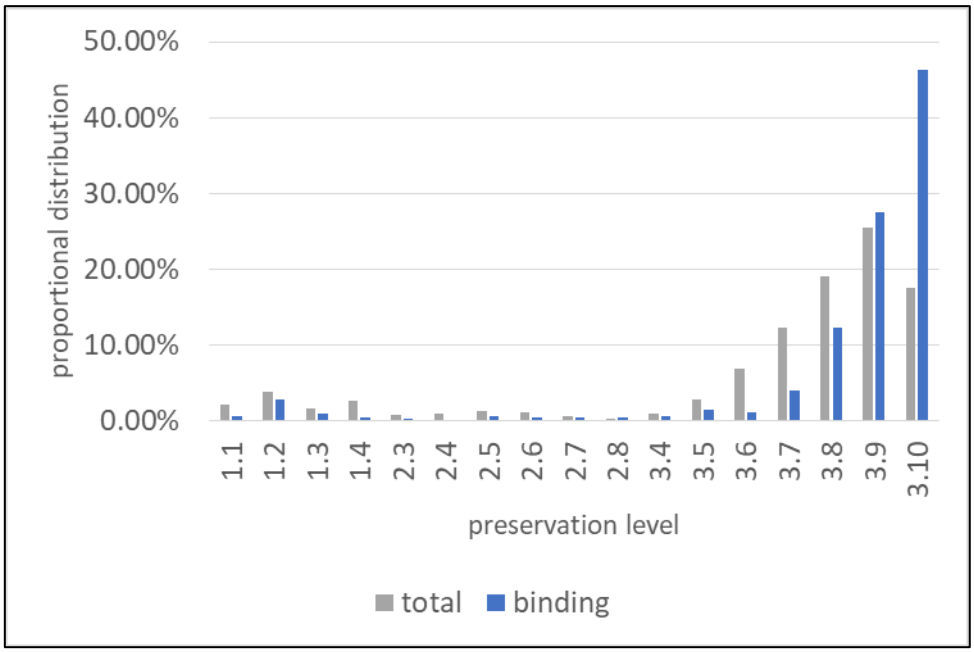
Further proportional distribution comparison with the base-line distribution is shown in Figure 4.2, active sites and nucleotide binding sites are mostly of high level of preservation which is consistent with our expectation. Surprisingly, general binding sites and metal binding sites also expressed high preservation throughout different species. Note that in Swiss-prot annotation, general binding sites stand for protein binding with ligands, substrates, products and cofactors. Although when binding with metal particles, related sites will be annotated as metal binding, some of these sites are also annotated as general binding sites. Therefore, as different species may react very differently to the same substances, especially when the immune system is involved, we expected general binding to be less preserved than some specific bindings such as calcium binding (marked 'CA_BIND') and nucleotide binding. However, the results show that the general binding sites show more preservation than the calcium and nucleotide binds. On the other hand, zin finger regions show the highest specificity where most of them are not shared by distant species. Mutagen and variant stand for site variants created on branch and detected in nature respectively. They are mostly disease related but we noticed some of these sites were further annotated as neutral. Therefore, we expected them to be close to the general distribution. However, they show significant bias of high preservation. Since these two marks contain various sites with or without functional effects, the cause of high preservation remains unknown.



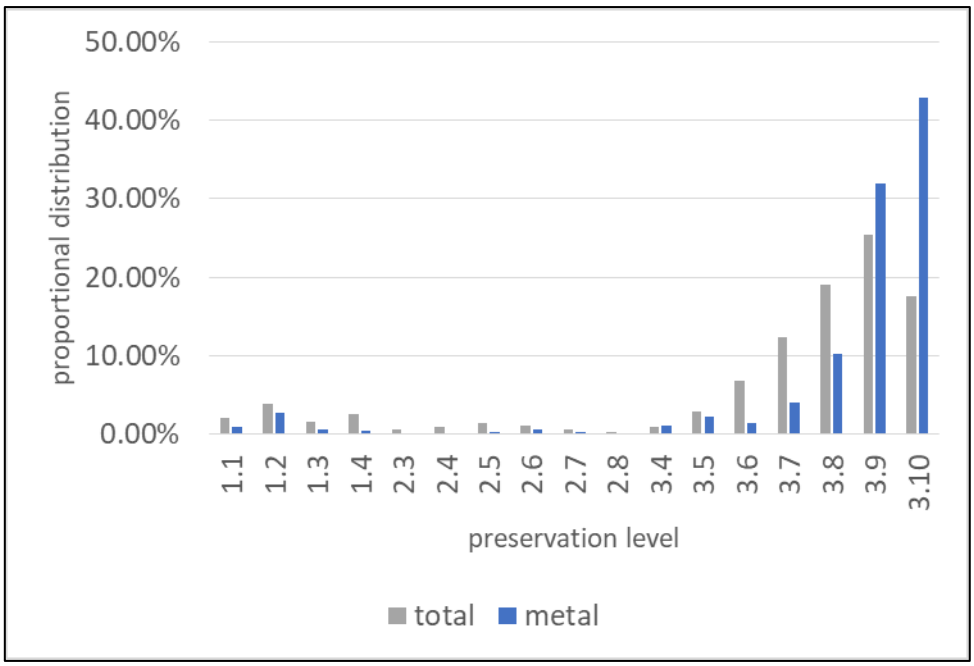
a


50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
proportional distribution
1.1
1.2
1.3
1.4
2.3
2.4
2.5
2.6
2.7
2.8
3.4
3.5
3.6
3.7
3.8
3.9
3.10
preservation level
total
np_bind

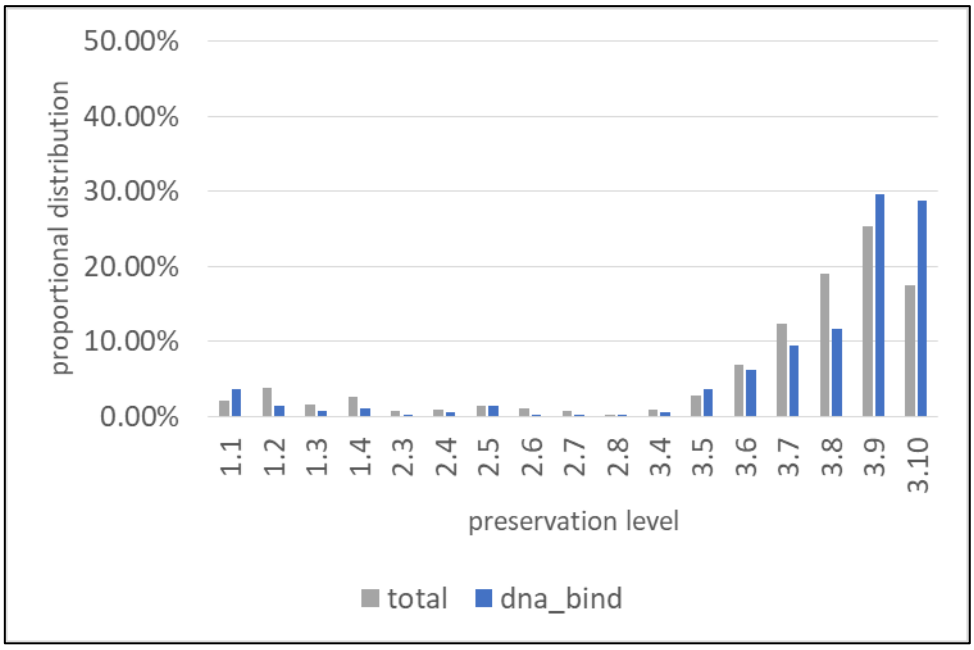
b


50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
proportional distribution
1.1
1.2
1.3
1.4
2.3
2.4
2.5
2.6
2.7
2.8
3.4
3.5
3.6
3.7
3.8
3.9
3.10
preservation level
total
binding

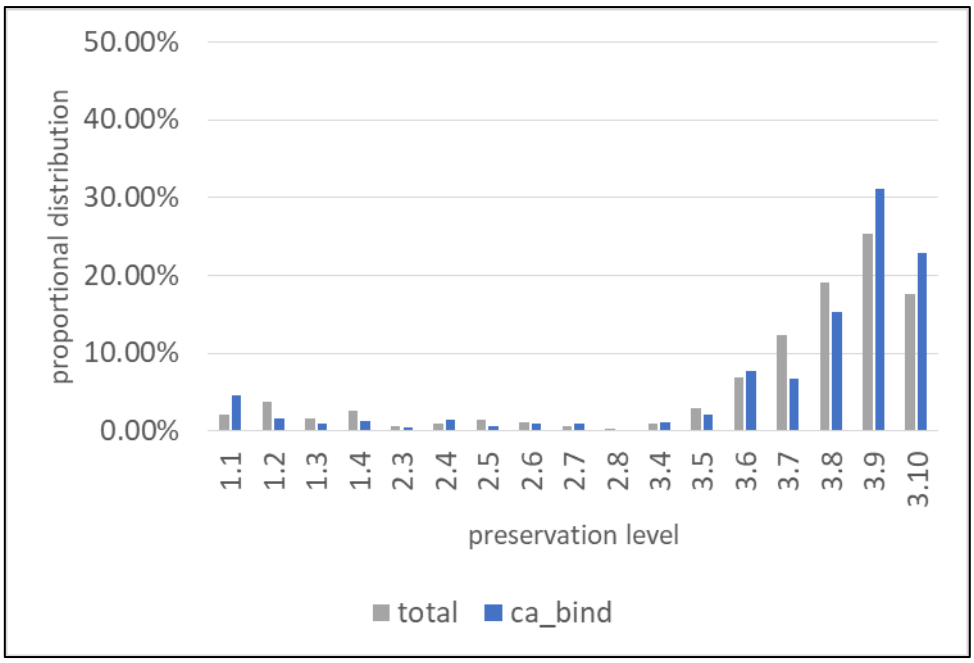
c


50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
proportional distribution
1.1
1.2
1.3
1.4
2.3
2.4
2.5
2.6
2.7
2.8
3.4
3.5
3.6
3.7
3.8
3.9
3.10
preservation level
total
metal

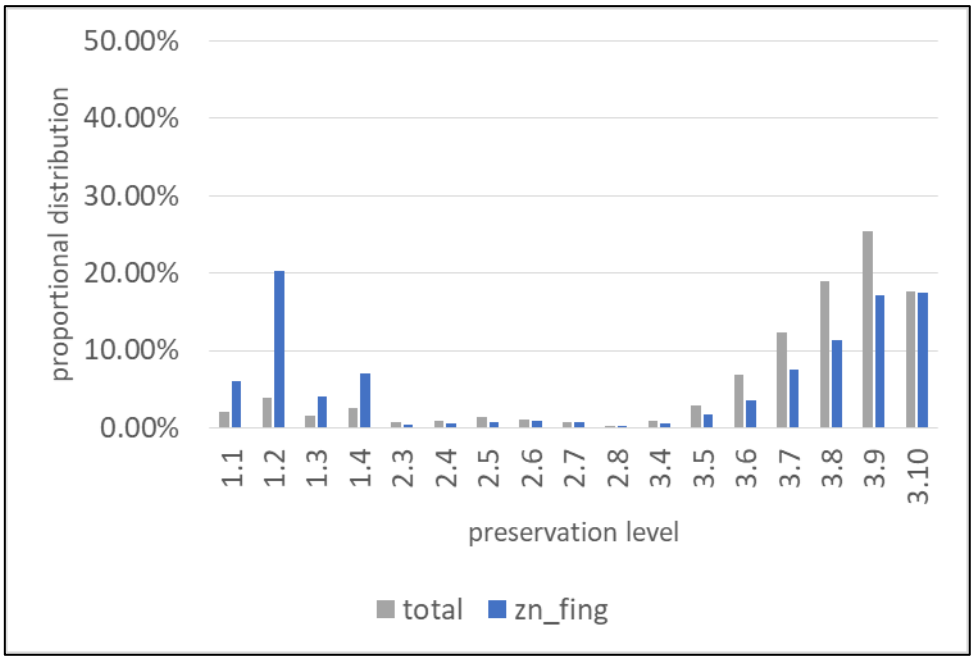
d


50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
proportional distribution
1.1
1.2
1.3
1.4
2.3
2.4
2.5
2.6
2.7
2.8
3.4
3.5
3.6
3.7
3.8
3.9
3.10
preservation level
total
dna_bind

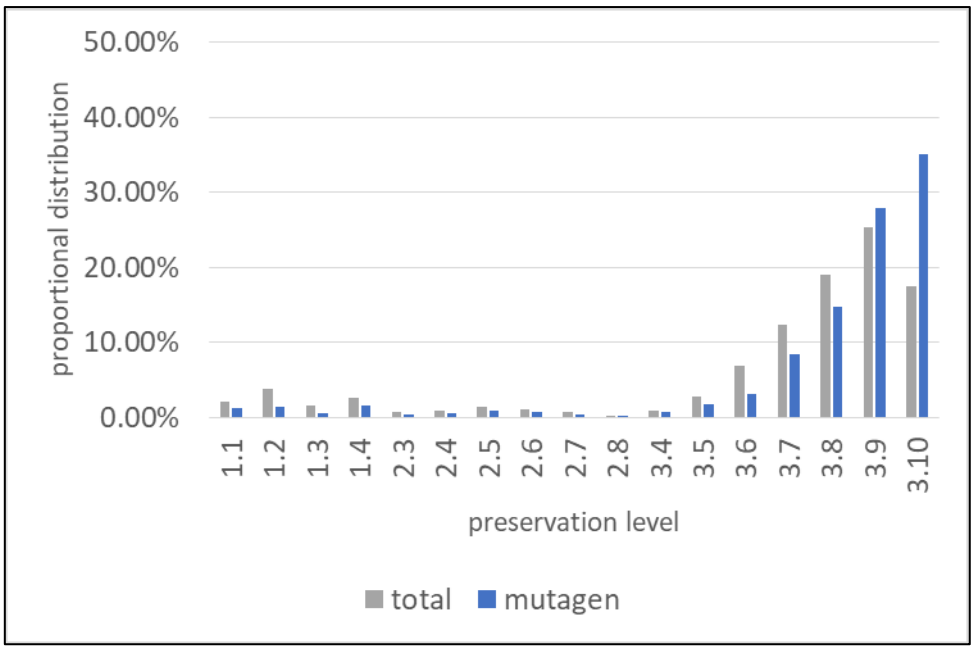
e


50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
proportional distribution
1.1
1.2
1.3
1.4
2.3
2.4
2.5
2.6
2.7
2.8
3.4
3.5
3.6
3.7
3.8
3.9
3.10
preservation level
total
ca_bind

f


50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
proportional distribution
1.1
1.2
1.3
1.4
2.3
2.4
2.5
2.6
2.7
2.8
3.4
3.5
3.6
3.7
3.8
3.9
3.10
preservation level
total
zn_fing

g



h

Figure 4.2 Preservation distributions on different annotated sites.

Preservation levels are marked at the X axis and percentage of total number of sites related to the annotation marker is shown on the Y axis. The annotated sites shown in all the graphs are a: active sites, b: nucleotide phosphate binding sites, c: general binding regions, d: metal binding sites, e: DNA binding regions, f: Calcium binding sites, g: zing fingers and h: sites had been artificially mutated in experiments. All types of annotated sites presented a preservation level above the general preservation profile except the region related to zing fingers.

The preservation distribution above also indicated that the increase of preservation often presents as increase of number of sites with the highest preservation level (ie 3.9 and 3.10). Therefore, we further used the proportion of these sites, referred as sequential preservation in later sections, for function intactness prediction in sequence level. The proportion of sites with these preservation levels were first calculated from alignments with human or mouse as the last note of the linage. There was a significant proportion of subfamilies not containing any site of such high preservation level. The number of these subfamily were 6,453 calculated from human and 6,512 from mouse. This was mainly caused by the definition of these subfamilies. One of the drawbacks of using only high preservation sites is that it can only be applied to subfamilies with long history. However, there were also subfamilies containing only sites with high preservation. Then the rat proteins and rhesus proteins were added to these alignments in order to count the number of high preservation sites that remain preserved in these two species (supplement 2 and 3). In this process, they used the preservation calculated from close related species, which was mouse for rat and human for rhesus. Excluding subfamilies without site of high preservation, 385 subfamilies detected mutations in over 50% of these sites in rat proteins and the number for rhesus is 649. After overlapping with the negative predictions generated by HMMs, the number became 204 and 197 respectively. Note that within these overlapping results, a big portion of subfamilies (144 for rat and 124 for rhesus) were overlapped by negative results with FMIs higher than 0.7. The details of preservation and alignment of these results can be found on supplement 2 and 3. Relatively loss alignment confirmed these predictions. Therefore, we considered the function to be defected when the related subfamily reported negative results agreed by both high preservation sites and high quality HMMs. Also when the subfamily HMM failed to find matches, the function it carries would be considered defected. As high preservation sites provide more confident results with lower coverage, we extended the coverage by using sites with the two highest preservation levels from among the subfamily instead of the highest of all training data. This allows automatic adjustment of local high preservation standard while recovering the same results for subfamilies reported by using only preservation level 3.9 and 3.10. However, to maintain prediction confidence, only sites with preservation in at least 4 species across all 3 tiers would be considered as sites with high preservation. Functions

related to subfamilies reporting negative predictions by HMM or mutations on more than 25% of sites with high preservation of local standard were considered altered. These criteria for predicting defected and altered functions with subfamilies were applied to rat and rhesus protein with and without unrelated subfamilies. The result is shown in the Table 4.2. A substantial portion of subfamilies returned altered results with or without unrelated subfamilies indicating the models are very sensitive toward species-specific features and unrelated subfamilies are highly likely to return altered or defected predictions.

Table 4.2 Functional prediction results for rat and rhesus.
Number of subfamily with non-conserved predictions are shown in the table. The following numbers are the results obtained after subfamilies not related to the species were removed. Significant dropping of number after the removal suggested irrelevant subfamily to be one of the main sources of negative predictions.

subfamily	Rat	Rhesus
Defected	203	327
Altered	3105	4249
Defected(filtered)	142	121
Altered(filtered)	1911	2357

Apart from making prediction for defected or altered functions, preservation can be used to compare conservation between homologs or function pathways. For example, as mentioned above, some subfamilies contain only high preservation sites but not all these subfamilies yield the same results when the preservation was calculated from different species. High preservation means the same variant is shared by most, if not all, the species selected as training set in this project. Therefore, most of the results were identical regardless of the starting species. Thus, different high preservation site number in a highly conserved subfamily mean the starting species expressed species-specific features in some sites of the conserved protein. For example, a subfamily (PTHR11639:SF90) (supplement 1) containing regulation protein S100A10. Although the regulation function of this subfamily remains unclear, it had been found to be highly conserved within mammals. However, the mouse homolog of this subfamily contain eight sites (14^{th}, 18^{th}, 24^{th}, 25^{th}, 37^{th}, 73^{th}, 76^{th} and 90^{th}) different from at least 9 (out of 10) other mammals involved in the training process. Most of these sites have not been annotated except the 37^{th} residue which would go through PTM process and form covalent linkage with another protein SUMO2. Another example would be another subfamily

(PTHR21141SF55) of 60S acidic ribosomal proteins. The mouse homolog in this subfamily contains three variant sites compared to other homologs but none of them have been annotated yet. Sites of such kinds could be the result of genetic draft without any functional effects but also could be the genetic signature of the species. Interesting species features may be found on highly conserved proteins and these proteins are mostly related to core biological processes that could affect many downstream processes. However, being highly conserved in evolution means most modifications introduced by random mutations had been wiped out by natural selection due to the significant negative effects they caused. The current cell engineering strategies tend to minimise unnecessary effects on cell biology. Since these proteins are more likely to affect multiple processes, they are rarely the engineering target. However, since radical changes had been induced to CHO in the process of immortalisation and adaptation of different environments, mutations rejected by previous evolution might now be beneficial. Comparing CHO homologs of these proteins with our models could suggest potential advantageous mutations which are the engineering target.

After sequential preservation was obtained, we mapped the subfamilies along with their preservation onto GO using enrichment analysis tool available on PANTHER website. The sequential preservation was calculated from human. The PANTHER GO-slim of biological process is used and only results with a p-value lower than 0.05 were reported. The result table (Table 4.3) is presented in hierarchical order derived from GO structure. It shows significant positive enrichments on core processes, especially metabolic processes, shared by many species. While also negative enrichments on processes with species specificity such as response to stimulus, immune response and regulation is presented, which confirm our preservation calculation in processes levels.

Table 4.3 Part of result of enrichment analysis from PANTHER website.
The GO terms are presented in hierarchical order and the most distinct terms (lowest P-value) are shown in this table.

PANTHER GO-Slim Biological Process	No. of subfamily	Enrichment (+/-)	P value
sensory perception of chemical stimulus (GO:0007606)	70	-	0.00E00
sensory perception (GO:0007600)	80	-	3.02E-14
neurological system process (GO:0050877)	111	-	1.04E-11
system process (GO:0003008)	115	-	2.22E-11
single-multicellular organism process (GO:0044707)	146	-	7.16E-09
multicellular organismal process (GO:0032501)	147	-	6.65E-09
cell surface receptor signaling pathway (GO:0007166)	125	-	4.98E-13
signal transduction (GO:0007165)	215	-	1.38E-06
cell communication (GO:0007154)	248	-	3.69E-05
response to stimulus (GO:0050896)	294	-	1.31E-10
metabolic process (GO:0008152)	649	+	1.78E-04
nucleobase-containing compound metabolic process (GO:0006139)	230	+	2.49E-03
primary metabolic process (GO:0044238)	530	+	6.92E-05
immune response (GO:0006955)	27	-	5.18E-07
cell adhesion (GO:0007155)	25	-	3.59E-02
biological adhesion (GO:0022610)	25	-	3.59E-02
cellular amino acid metabolic process (GO:0006520)	65	+	1.87E-04
regulation of biological process (GO:0050789)	347	-	2.36E-09
biological regulation (GO:0065007)	385	-	1.33E-07
lipid metabolic process (GO:0006629)	98	+	1.22E-02
phosphate-containing compound metabolic process (GO:0006796)	117	+	1.07E-04

4.3 Summary

The HMM and preservation models were verified using well sequenced and annotated genomes. The HMM models demonstrated high accuracy in assigning query proteins to subfamilies but comparatively lower confidence in predicting intactness. The preservation

models generate similar result with PANTHER-PSEP. By examining the preservation level of critical function sites, we addressed the preservation levels highly correlated to these sites and used them to calculate sequential preservation which was then used to assist predictions. However, both HMM and preservation suffer restriction of conservation where both their accuracy and coverage reduce alone with conservation. Although the sequential preservation could only be applied to the relatively conserved subfamilies, it could capture species-specific features ranging from intra-sequence level to cellular function level.

5 Model Predictions on CHO Related Data

In this project, we involved all three CHO related genomes available online: a CHO-K1 genome, CHO-K1GS genome and a Chinese hamster genomes. An overview of the prediction is first presented reporting the number of different predictions in different genomes. Then 100 subfamilies reported homologs with normal function in CHO-K1 and 100 subfamilies reported defected CHO-K1 homologs were verified with BLAST. In the verification process, we blast mouse homologs used to train the model against the CHO-K1 genome for the best reciprocal hits.

5.1 Prediction Overview

Although the three CHO related genomes were annotated by different pipelines (CHO-K1 and Chinese hamster annotated by NCBI RefSeq and CHO-K1GS by Horizon Eagle), they were referred by each other during the annotating process. We first conducted analysis on subfamilies using models to identify best matching protein sequences. By applying the protocols and criteria described in the previous chapters, we identified a significant portion of altered function in proteins from all three genomes (Table 5.1). Only numbers are discussed at this point and more details of these proteins regarding their functions and biological effects will be discussed later in this chapter (on session 5.4). The number of subfamilies failed to find significant matches are also reported in the table. Unexpectedly, the latest CHO-K1GS proteins yield least defected or altered results, although more subfamilies reported altered or defected prediction in CHO-K1 than Chinese hamster as expected. This could be caused by the quality control of annotating pipeline which could remove sequences less similar to known sequences keeping only the comparatively conserved sequences. Note that on the Ensembl website, another annotation generated by Ensembl on Chinese hamster genome (probably CHO-K1 genome as they used the CriGri_1.0 assembly with was first published with the CHO-K1 genome) other than CHO-K1GS is published, containing only 19,617 coding genes which is significantly different from the RefSeq annotation containing 27,752 genes. The newer Ensembl annotation was not used in this project as the RefSeq annotation was applied. Such inconsistency between the annotations is likely to be the result of insufficient re-sequencing data for CHO. Therefore, significant improvements could be made on annotation when more data is available.

Table 5.1 Number of subfamily returned negative result on CHO related genomes.

Subfamily	CHO-K1	CHO-K1GS	Chinese Hamster
Defect	142	66	160
Alter	3695	2985	3374
No match	59	68	96
Total	3896	3119	3630

According to the validation on rat and rhesus proteins, unrelated subfamilies could contribute to a significant part of the altered and defected predictions. Therefore, we attempted to identify the unrelated subfamilies by using protein sequences to find the best matching subfamilies. Subfamilies whose best matching proteins matched better to another subfamily would be considered unrelated. An overview result of identifying best matching subfamilies is shown in Table 5.2. Accordingly, proteins matched with subfamilies returned altered or defected in the result above were deemed altered or defected correspondingly. Clearly, all the altered and defected proteins were converged to a smaller number of subfamilies. It is important to note that the predicted defect or altered proteins are not necessarily defected or altered since, they may belong to subfamilies not covered by this project. The validation result showed our subfamily set should be covering about 17,000 related subfamily for rodent and primate species. Given a total gene number projection of 24,000 to 30,000 for these species, the subfamily coverage could be arguably low. However, given both the subfamily model and the protein recognising each other as best match and the subfamily model reporting negative prediction, there should be a good chance that the protein is significantly damaged by the mutations.

Table 5.2 Functional prediction of proteins on CHO related genomes.
Numbers of protein with different prediction are shown followed by the number of subfamilies which the proteins with negative prediction belong to.

	CHO-K1	CHO-K1GS	Chinese Hamster
Defect protein	121	49	65
Altered protein	4361	2926	3786
Intact protein	26788	24975	25566
Subfamilies with defect protein	81	41	35
Subfamilies with altered protein	2404	1682	2119

5.2 Verification with BLAST

To further examine the quality of predictions generated by our models, we verified some of our results with BLAST. Clearly, our models perform differently compared to BLAST in many aspects such as scoring method, aligning algorithm and sensitivity. However, the main difference should not affect their agreement on extreme samples, which are either highly conserved homologs or completely different sequence pairs. Therefore, we randomly selected (using random number generated by the computer) 100 subfamilies reported intact and 100 subfamilies reported defected for this verification. These predictions were made by using subfamily models to find the best matches on CHO proteins. Thus, we selected a member of the subfamily to BLAST against all CHO proteins to obtain the best hit on CHO. Since mouse is genetically close to CHO and well annotated as a model organism, mouse homologs in the selected subfamilies were used to BLAST against the proteins derided from the CHO-K1 genome. We considered the percentage of identical site and the percentage of matching region the significant metrics on comparing the BLAST result of two sets of sequences (supplement 4 and 5). Default settings of blastp was used for this verification. As expected, the sequences of intact prediction significantly exceeded those of other groups on both percentage of identical site and matching region. On average, the intact group achieved an identical percentage of 89% while covering 92% of the best matching CHO protein. On the other hand, the number of defect group were 59% and 55%.

The difference between scores generated by BLAST and our models were compared in detail in Figure 5.1 and Figure 5.2. In these graphs, samples were plotted by their score generated by both BLAST and the preservation models. Although our method includes HMM models and preservation models, sequential preservation was used to represent scores generated by our models since the thresholds for HMM scores were determined accordingly by machine learning algorithms, while threshold for preservation score was set at 0.75. Meanwhile, the similarity score generated by BLAST can also be presented by two parameters: identity and coverage. Both of these parameters are scaled between 0 and 1. Low score on either parameter would lead to negative predictions from our models. Therefore, the lower value of the two parameter was used to represent BLAST similarity in the comparison.

Figure 5.1 shows the comparison on predicted intact proteins. 0.75 was used as a threshold for sequential preservation so that all proteins in the intact group have preservation scores higher than 0.75. In most cases, BLAST generated high scores agreeing with our models. However, if a same threshold of 0.75 was applied to predict intact proteins, 7 out of the 100 randomly selected samples would be predicted altered or defected. On the other hand, Figure 5.2 shows the scores of samples of predicted defeated proteins. If the sample threshold was applied on BLAST scores, 25 of them would be predicted differently. These samples showed BLAST agrees 93% of the positive prediction and 75% of the negative predictions making up an overall agreement of 84%.

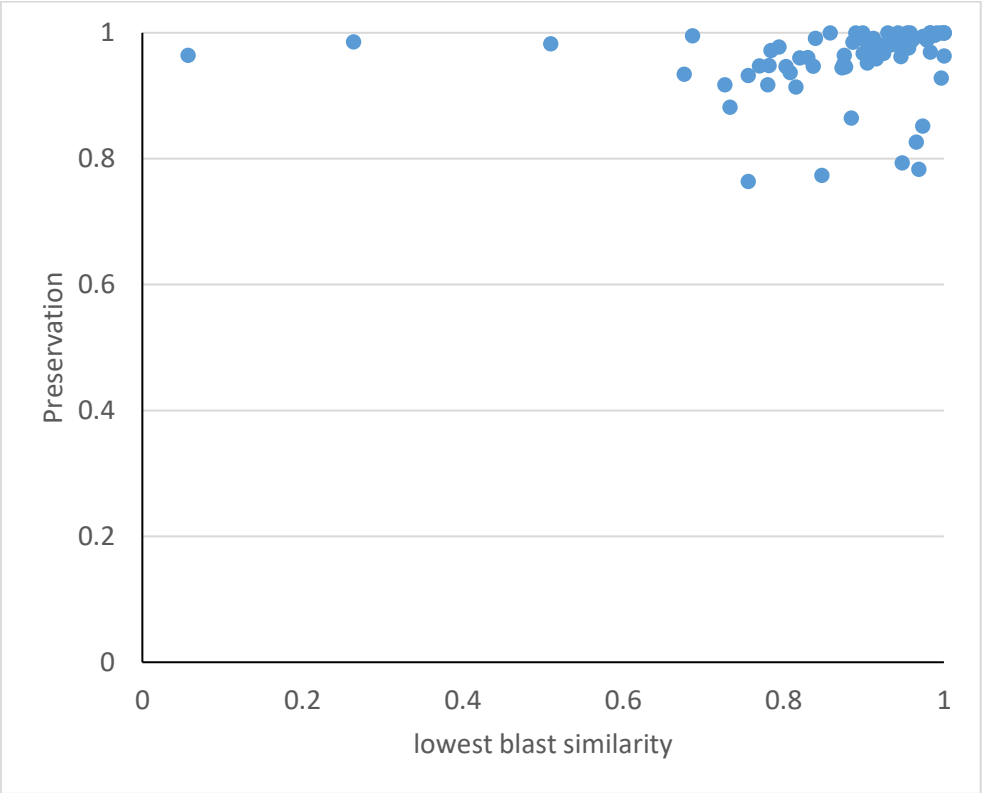


Figure 5.1 Comparison of BLAST similarity and preservation score of predicted intact proteins.
The lowest blast similarity is set to be the lower value of identity and coverage generated by BLAST. Most samples are concentrated on the top right corner showing that the two methods agree with each other in most cases.

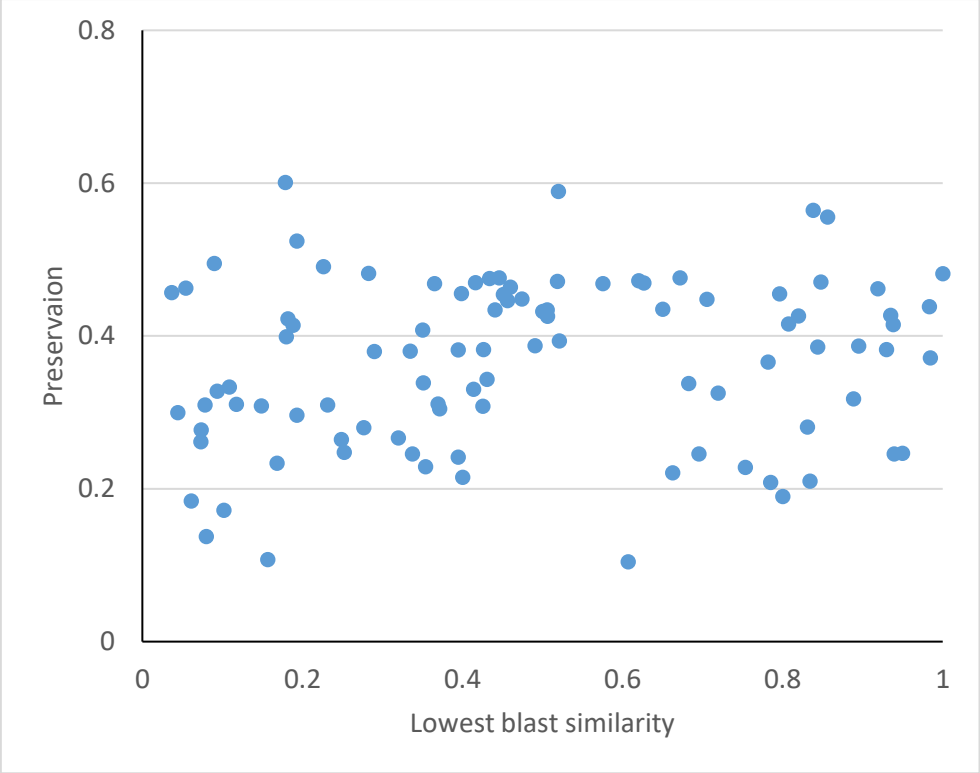


Figure 5.2 Comparison of BLAST similarity and preservation score of predicted defected proteins.
The lowest blast similarity is set to be the lower value of identity and coverage generated by BLAST. Samples are scattered horizontally showing that the correlation of scores generated by these two methods is not strong in this group. However, when threshold (0.75) is applied, 75% of the predictions are agreed by both methods.

Samples getting different results from BLAST and our models were investigated in detail. Although significant difference is presented between predicted defected CHO proteins and their mouse homologs, some mouse sequences were matched to CHO sequences in the defected group with identity and coverage both higher than 90%. Meanwhile the identity and coverage generated from BLAST could be very low for some of the sequences in the intact group. By checking the alignments and site preservation of subfamilies yielding negative prediction on high similarity alignment issued by BLAST, the cause of such disagreement was that the highly similar fragmented proteins were found on both mouse and CHO protein selection. Due to splicing, shorter versions of protein may be annotated as an individual protein and could be used as a representative of the subfamily for BLAST verification as we randomly select mouse homolog from the subfamily. All the HMMs built in this project were designed to model full length of the protein and that was also used in proportion calculation

for high preservation sites. Containing only the fragmented version could suggest loss of sequence parts or incorrect gene boundary in annotation. On the other hand, positive predictions could result in low identity and coverage in BLAST due to incorrect alignments created by BLAST. In most of these cases, the sequences of the subfamily contain two or more similar regions. BLAST mismatched these regions when attending local optimum, which lead to short alignments in the final result. In our models, alignments were created by MAFFT which is more accurate in making alignments than BLAST.

5.3 Expression and Prediction Correlation

Unlike human proteins, considering disease-related variants to be deleterious may not be applicable to CHO. However, we hypothesised that variants affecting protein function, even only changing the efficacy, in highly expressed genes would bring significant deleterious effect to the cell which would then lead to extinction of cells carrying these variants. As a result, highly expressed genes are more conserved, or in other words deleterious mutation proof. Based on such hypothesis, we investigated the relation between expression and sequential preservation. CHO expression data were obtained from NCBI GEO. Two sets of expression data based on RNA-Seq from van Wijk et al. (2017) and Lee et al. (2016) respectively were selected. In their results, the abundance of expression is presented as Fragments Per Kilobase of transcript per Million mapped reads, also known as FPKM value, and the gene ids are presented with gene names. To map subfamilies onto gene names, we used gene names from mouse homologs within the subfamily. When multiple mouse homolog were available within one subfamily, the gene name used by most was used to represent the subfamily. Gene names from human homologs were applied when mouse homologs were not available. Two replicas are reported for every sample in their data and we first average the FPKM values of the replicas as the abundance of the gene expression for the corresponding sample, then the maximum abundance was used to map with the sequential preservation. In our hypothesis, deleterious effects come from the cost of synthesising biomass required to fulfil sufficient protein efficacy. Such cost is high when synthesising large amounts of biomass regardless of

the growing phase the cells are experiencing. Therefore, we did not select data on specific phase but adopted maximum expression measured in these researches.

After the connection between FPKM value and sequential preservation via gene names, the genes were divided into five groups according to their maximum expression abundance. Genes with no expression detected in the involved researches were assigned to the same group which was excluded from the preservation distribution analysis. No expression detected in these researches indicates that these genes are likely expressed in a very low level which may largely increase in certain situations. Although a large proportion of genes fall into this category, as no evidence suggests the maximum expression of these genes, our hypothesis is not applicable to them leading to their removal from the related analysis. The rest of the genes with non-zero FPKM value were assigned to four groups whose range of FPKM value containing are shown in Table 5.3. Then the preservation distribution was plotted against the number of genes. The overall preservation distribution of all genes was used as a reference distribution which was then imported to the t-test with the distribution of different groups to evaluate the significance of being different. The significance level was set on two-tail p-value. The sequential preservation profiles of gene of different expression level are shown in Table 5.3. It can be observed that the number of genes significantly decrease with expression level and 65.5% of the genes involved express in the lowest level (<50 FPKM). On the other hand, apart from genes not covered by preservation models, gene numbers increase with the level of preservation and 64.8% of these genes have preservation score higher than 0.9. Genes of different expression level present similar trend but t-test show that preservation profile of every expression level is significantly different (in a level of 0.01) with the general profile.

The mean and variance of preservation in every expression level were investigated in Figure 2.1Figure 5.3. This shows that the mean preservation increases with the level of expression in the groups where FPKM value are lower than 5000, and a subtle decrease is observed in the last group. Meanwhile, the variance decreases with the increase of mean preservation. Such a result is consistent with our hypothesis showing a high preservation on genes with high expression. However, it needs to be clarified that such a result only provides statistic support for our hypothesis so it could be wrong for some individual genes. Therefore, it can only be considered as a general trend in the genome level.

Table 5.3 Sequential preservation profile of genes of different expression levels.
The expression levels are defined according to the FPKM value. The first level was defined as less than 50 and the following levels were defined with the increase of magnitude until the FPKM value reaches 5000. Total number of gene with sequential preservation level are presented to provide general preservation profile. T-test was performed for preservation profile of every expression level against the general profile and the results, including two-tail p-value, mean and variance, obtained are shown at the bottom.

Preservation	Overall gene	[0-50)	[50-500)	[500-5000)	>=5000
0.0	2541	1864	472	190	15
0.0-0.1	3	2	1	0	0
0.1-0.2	8	7	0	1	0
0.2-0.3	23	17	6	0	0
0.3-0.4	54	38	11	4	1
0.4-0.5	108	77	22	9	0
0.5-0.6	177	140	28	6	3
0.6-0.7	246	184	45	16	1
0.7-0.8	404	298	68	36	2
0.8-0.9	764	544	143	73	4
>0.9	7956	4875	1707	1284	90
total	12284	8046	2503	1619	116
P-value		1.324E-07	0.00521197	3.7348E-34	0.0040362
mean	0.73	0.70	0.76	0.84	0.83
variance	0.153	0.163	0.145	0.101	0.113

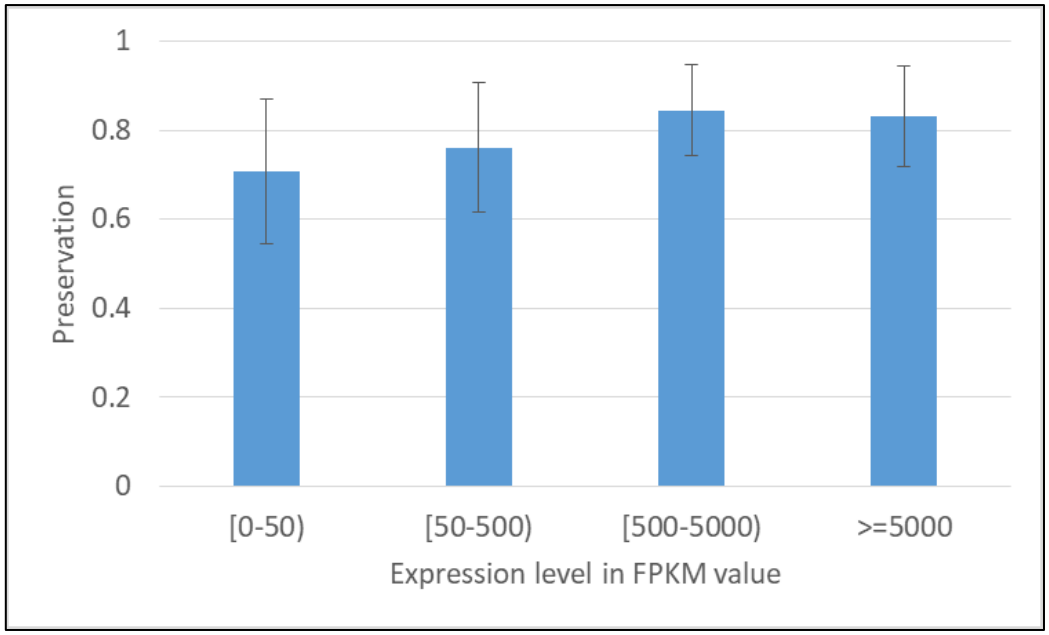


Figure 5.3 The mean and variance of sequential preservation in different expression levels divided by FPKM value.
Genes are classified by their expression levels according the FPKM values, and the corresponding means are presented with variance as error bars. Overall increase of mean preservation with expression level can be observed.

5.4 Comparison of Predictions of Different CHO Related Genomes

To identify the projection of subfamily prediction in the multi-gene function level, we compared prediction results of three CHO related genomes and mapped them onto GO using analysis tools available on the PANTHER website.

To identify the character of the three CHO related genomes, the number of subfamily issuing abnormal predictions were compared. As mentioned previously, the CHO-K1 protein set is closed to the Chinese hamster in terms of number of different type of predictions. In this comparison (Table 5.4), the same features were observed. The number of subfamily making abnormal prediction for each genome is shown in the diagonal line. The number of shared subfamily are reported on the upper right corner. It shows that most of the abnormal predictions shared by more than one genome and, similar to the previous result, CHO-K1 and Chinese hamster shared most of such predictions. Apart from the potential cause of annotation pipeline difference which is mentioned previously, it could also be evidence of CHO-K1GS having been significantly evolved during screening and the culture process so that

the genetic distance between CHO-K1GS and CHO-K1 is actually larger than that between CHO-K1 and the Chinese Hamster sequenced. We further found that most subfamilies with abnormal predictions shared by CHO-K1GS and either of the other two genomes are actually shared by all three CHO related genomes. The number of such subfamily was 2,285. With the doubt that these subfamilies may issue abnormal predictions due to model overfitting, we compared them with subfamilies reported abnormality for rat and rhesus protein in validation. As a result, 1,427 subfamilies were reporting abnormal results for all five genomes.

Table 5.4 Number of common subfamily reporting abnormal result

	Chinese hamster	CHO-K1	CHO-K1GS
Chinese hamster	3,534	2,834	2,362
CHO-K1		3,837	2,429
CHO-K1GS			3,051

The behaviour of these subfamilies could be due to improper selection of homologs or inherently low conservation within the family. To identify how these subfamilies would affect our later functional analysis with GO, they were mapped onto GO to identify terms these subfamilies cluster to. Since a different version (version 13.0) of PANTHER subfamily annotation was used by the online tool, direct mapping with subfamily ids would lead to substantial loss of mapped ids. The subfamilies were represented by gene names as they were mapped with the expression abundance. The list of representative gene names was uploaded to the website and overrepresentation test was performed with a list of gene names representing all subfamilies as reference using binomial model without Bonferroni correction. The PANTHER GO-slim for biological process was selected for the analysis. Such setting was carried on for all GO mapping of overrepresentation tests. The ten GO terms with lowest p-value are presented in Table 5.5 as examples of the result (more results available in supplement 7). With GO terms in the first column, the numbers of related protein in the reference are shown in the second column. Then the number of protein related in query sets and the expected number in the query set calculated by binomial model are presented. The Table 5.5 was sorted by p-values in the last column. In general, the representative genes converged into GO terms of multi-cellular functions or even higher level functions such as sensory perception. The only term shown mainly related to single cell function was G-protein coupled receptor signalling pathway. However, GO terms with highest statistical significance

often relate to a large amount of genes which makes it difficult to extract the details. However, these subfamilies being less related to terms of single cell biological process would be beneficial for analyses related to CHO cells.

Table 5.5 GO mapping for common altered functions

PANTHER GO-Slim Biological Process	Total proteins	Query proteins	Expected number	over/under-represent	fold Enrichment	P-value
sensory perception	823	108	48.62	+	2.22	2.87E-14
sensory perception of chemical stimulus	684	92	40.41	+	2.28	7.05E-13
sensory perception of smell	638	86	37.69	+	2.28	3.70E-12
neurological system process	1316	131	77.75	+	1.68	5.67E-09
system process	1408	136	83.18	+	1.63	1.69E-08
G-protein coupled receptor signaling pathway	793	86	46.85	+	1.84	9.13E-08
biological regulation	3424	263	202.28	+	1.3	2.90E-06
single-multicellular organism process	1979	165	116.92	+	1.41	4.76E-06
multicellular organismal process	1994	166	117.8	+	1.41	4.82E-06
cellular process	8073	549	476.94	+	1.15	7.54E-06

The functions predicted to be altered in all CHO related genomes excluding those in rat and rhesus were then mapped onto the same GO slim (Table 5.8). 862 subfamilies fell into this category and were mapped onto GO. Compared to the mapping with common altered functions in rat and rhesus included (Table 5.6), terms describing cellular processes, such as receptor signalling pathways, were more statistically significant. However, comparing terms associated with altered function specific in CHO related genomes (Table 5.8) and those shared by rat and rhesus (Table 5.7), 5 general terms are shared by their top list, which could be caused by the bias on large numbers in statistical tests. We also found that, by removing subfamilies predicting altered in rat and rhesus, specific GO terms with the number of gene manageable for manual inspection moved up on the list sorted by p-value. This suggests genes contained by these subfamilies are only related to general terms due to the lack of detailed knowledge. This highlighted the fact that the annotation level of individual genes could be the main restriction of making use of our model prediction, given the predictions were accurate.

Table 5.6 GO mapping of altered function shared by CHO related genomes

PANTHER GO-Slim Biological Process	Total proteins	Query proteins	Expected number	over/under-represent	fold Enrichment	P-value
sensory perception of chemical stimulus	684	158	58.26	+	2.71	1.32E-28
sensory perception	823	175	70.09	+	2.5	1.32E-27
sensory perception of smell	638	139	54.34	+	2.56	6.08E-23
G-protein coupled receptor signalling pathway	793	154	67.54	+	2.28	1.25E-20
neurological system process	1316	202	112.08	+	1.8	9.08E-16
biological regulation	3424	418	291.62	+	1.43	3.58E-15
system process	1408	207	119.92	+	1.73	2.58E-14
response to stimulus	3097	376	263.77	+	1.43	3.50E-13
regulation of biological process	2898	356	246.82	+	1.44	4.21E-13
single-multicellular organism process	1979	259	168.55	+	1.54	3.31E-12

Table 5.7 GO mapping of altered function shared by CHO related genomes, rat and rhesus

PANTHER GO-Slim Biological Process	Total proteins	Query proteins	Expected number	over/under-represent	fold Enrichment	P-value
sensory perception	823	108	48.62	+	2.22	2.87E-14
sensory perception of chemical stimulus	684	92	40.41	+	2.28	7.05E-13
sensory perception of smell	638	86	37.69	+	2.28	3.70E-12
neurological system process	1316	131	77.75	+	1.68	5.67E-09
system process	1408	136	83.18	+	1.63	1.69E-08
G-protein coupled receptor signalling pathway	793	86	46.85	+	1.84	9.13E-08
biological regulation	3424	263	202.28	+	1.3	2.90E-06
single-multicellular organism process	1979	165	116.92	+	1.41	4.76E-06
multicellular organismal process	1994	166	117.8	+	1.41	4.82E-06
cellular process	8073	549	476.94	+	1.15	7.54E-06

Table 5.8 GO mapping for altered proteins common in CHO related genomes but normal in rat and rhesus

PANTHER GO-Slim Biological Process	Total proteins	Query proteins	Expected number	over/under-represent	fold Enrichment	P-value
sensory perception of chemical stimulus	684	66	18.47	+	3.57	8.67E-19
G-protein coupled receptor signalling pathway	793	68	21.41	+	3.18	9.03E-17
sensory perception	823	67	22.22	+	3.02	1.87E-15
response to stimulus	3097	153	83.62	+	1.83	1.60E-14
sensory perception of smell	638	53	17.23	+	3.08	9.91E-13
biological regulation	3424	156	92.45	+	1.69	6.14E-12
regulation of biological process	2898	134	78.24	+	1.71	1.42E-10
multi-multicellular organism process	26	11	0.7	+	15.67	2.44E-10
mammary gland development	26	11	0.7	+	15.67	2.44E-10
cell surface receptor signalling pathway	1448	81	39.1	+	2.07	4.54E-10

To investigate the difference between CHO and Chinese hamster, functions predicted altered in CHO-K1 and CHO-K1GS but normal for Chinese hamster were analysed (Table 5.9). Only 124 subfamilies belong to this category and all the GO terms with a p-value lower than 0.05 were presented in Table 5.9. Unlike the previous mapping result, these subfamilies were more related to specific functions while showing negative enrichment on high level general functions such as biological regulation and sensory perception. Although the number of subfamilies were low and not many terms enriched significantly, interesting terms, such as protein folding, fatty acid metabolic process and lipid metabolic processes, are presented.

Table 5.9 GO mapping of altered functions common in CHO-K1 and CHO-K1GS but normal in Chinese hamster

PANTHER GO-Slim Biological Process	Total proteins	Query proteins	Expected number	over/under-represent	fold Enrichment	P-value
disaccharide metabolic process	2	1	0.01	+	88.23	1.13E-02
fatty acid beta-oxidation	21	2	0.12	+	16.81	6.49E-03
protein folding	83	3	0.47	+	6.38	1.20E-02

fatty acid metabolic process	168	5	0.95	+	5.25	2.80E-03
lipid metabolic process	407	6	2.31	+	2.6	2.87E-02
cell adhesion	350	5	1.98	+	2.52	4.95E-02
biological adhesion	350	5	1.98	+	2.52	4.95E-02
biological regulation	3424	12	19.4	-	0.62	3.59E-02
response to stimulus	3097	10	17.55	-	0.57	2.67E-02
nucleobase-containing compound metabolic process	2549	7	14.45	-	0.48	1.79E-02
sensory perception	823	1	4.66	-	0.21	4.99E-02
RNA metabolic process	1417	1	8.03	-	0.12	2.29E-03

The difference between CHO-K1 and CHO-K1GS was compared by inspecting altered function in either genome but not in the Chinese hamster genome (Table 5.10 and Table 5.11). When comparing these two CHO genomes, we assumed CHO-K1GS a descendent of CHO-K1 although in comparison of the two cell lines the inheritance direction only affects the hypotheses and interpretation of the result but not the result itself. Under such assumption, CHO-K1GS appeared to be more adapted to the suspension culture environment since more adhesion related functions are altered in CHO-K1GS while alteration in CHO-K1 still lingering on general terms (this could be the annotation problem however).

Table 5.10 GO mapping for proteins altered in CHO-K1 but normal in Chinese hamster

PANTHER GO-Slim Biological Process	Total proteins	Query proteins	Expected number	over/under-represent	fold Enrichment	P-value
sensory perception of smell	638	2	24.63	-	0.08	4.60E-09
sensory perception of chemical stimulus	684	3	26.4	-	0.11	7.95E-09
sensory perception	823	7	31.77	-	0.22	8.42E-08
G-protein coupled receptor signalling pathway	793	7	30.61	-	0.23	2.17E-07
mesoderm development	262	24	10.11	+	2.37	1.24E-04
response to stimulus	3097	87	119.55	-	0.73	4.37E-04
developmental process	1476	80	56.98	+	1.4	1.51E-03
neurological system process	1316	32	50.8	-	0.63	2.43E-03
cell surface receptor signalling pathway	1448	37	55.89	-	0.66	3.56E-03
blood circulation	19	4	0.73	+	5.45	6.72E-03

Table 5.11 GO mapping for proteins altered in CHO-K1GS but normal in Chinese hamster

PANTHER GO-Slim Biological Process	Total proteins	Query proteins	Expected number	over/under-represent	fold Enrichment	P-value
cell adhesion	350	18	5.91	+	3.04	3.78E-05
biological adhesion	350	18	5.91	+	3.04	3.78E-05
heart development	20	4	0.34	+	11.84	4.09E-04
metabolic process	5470	67	92.41	-	0.72	7.68E-04
response to stimulus	3097	32	52.32	-	0.61	7.71E-04
cell-cell signalling	571	21	9.65	+	2.18	8.43E-04
muscle organ development	27	4	0.46	+	8.77	1.24E-03
neurological system process	1316	10	22.23	-	0.45	2.38E-03
G-protein coupled receptor signalling pathway	793	4	13.4	-	0.3	2.42E-03
sensory perception of chemical stimulus	684	3	11.56	-	0.26	2.85E-03

5.5 Summary

In this chapter, protein sequences of three CHO related genomes were analysed by our pipeline. The predictions were verified against a widely used tool BLAST and our pipeline showed better accuracy in results disagreed by BLAST. Although only part of normal and defected predictions were verified, it suggests a high precision of these two types of predictions. We then related the gene preservation with expression which showed significant correlation as expected. However, when comparing predictions between three CHO related genomes, unexpected similarity between CHO-K1 and Chinese hamsters suggest the genetic distance between CHO-K1 and CHO-K1GS is further than that to Chinese hamster, but the difference of annotations toward the same assembly from different source was also concerning. The model predictions were intensively mapped on GO. Although most of the terms suggested by statistical significance were related to many genes, interesting terms with manually manageable number of genes were suggested which may motivate further research in greater detail in the future.

6 Case Studies and Hypotheses

Several cases were selected to be inspected in greater detail in this chapter. Firstly, the mutation on TP53, a marker gene for cancer cell lines. Secondly, some unexpected alterations on proteins related to glycolysis identified in the results described in chapter 5. After that, some pathways had been extensively studied on CHO or other cancer cell lines were studied in detail. As only general terms with large number of genes presented in the GO mapping reported in chapter 5, the mapped genes are mostly scattered and not well annotated. We could not choose cases from those results.

6.1 Mutations on TP53

TP53 is a well-known cancer suppressor whose mutation is observed in all types of cancer. It is possible that CHO cell lines share such feature. Therefore, we investigated prediction of TP53 in all three CHO related genomes. Unexpectedly, both our HMM and preservation model predicted TP53 being normal in CHO-K1 and CHO-K1GS, but altered in Chinese hamster. The matched sequence id and sequential preservation are shown in Table 6.1. We then BLAST the mouse version of TP53 against Chinese hamster related RefSeq data. The BLAST result (Table 6.2) confirmed our predictions. The TP53 homolog in CHO-K1 scores higher in not only bit score, but also in query cover and percentage of identity than the homolog in Chinese hamster. Our sequential preservation analysis showed that 10% of sites with preservation level higher than 3.8 mutated from their preserved residues in CHO-K1 and CHO-K1GS, while the proportion for Chinese hamster homolog is 48% (supplement 8 and 9). Such unexpected mutation on Chinese hamster homolog could be caused by incorrect assembly or annotation. Details of homologs from CHO-K1 and CHO-K1GS were further inspected. Our annotated preservation maps of all sites of the matched sequence (supplement 10 and 11) addressed multiple high preservation sites responsible for DNA binding mutated from the most preserved residues in both CHO-K1 and CHO-K1GS, suggesting alteration in protein function. As we used a fairly arbitrary threshold of 0.75 on prediction using sequential preservation, this result indicated that this threshold used could be too conserved to miss the true negatives.

Table 6.1 TP53 preservation in CHO related genomes

	Chinese hamster	CHO-K1	CHO-K1GS
Sequence id	XP_007606822	NP_001230905	ENSCGRP00001011268
Sequential preservation	0.52	0.90	0.90

Table 6.2 BLAST result of TP53 in CHO genome from NCBI

Description	Max score	Total score	Query cover	E value	Ident	Accession
cellular tumor antigen p53	603	603	100%	0.0	76%	NP_001230905.1
PREDICTED: tumor protein 63 isoform X3	270	270	72%	3e-86	49%	XP_003495647.1
PREDICTED: tumor protein p73 isoform X4	271	271	85%	2e-85	44%	XP_007606822.2
PREDICTED: tumor protein 63 isoform X2	268	268	72%	2e-85	49%	XP_007640645.1
PREDICTED: tumor protein p73 isoform X2	270	270	85%	2e-84	44%	XP_007606823.2
PREDICTED: tumor protein p73 isoform X1	269	269	66%	4e-84	50%	XP_016818894.1
PREDICTED: tumor protein p73 isoform X3	269	269	85%	5e-84	44%	XP_007606821.2
PREDICTED: tumor protein 63 isoform X1	268	268	72%	2e-83	49%	XP_003495644.1

6.2 Glycolytic Process

Glycolytic process is the core metabolic process providing energy for cell activities and precursors for amino acid synthesis. It is conserved across species in different kingdoms and therefore CHO-K1 was not expected to be an exception. However, the subfamilies reporting abnormal result in all validating and query genomes contained 6 (out of 26 were modelled) gene involved in the glycolysis. Although all of these results were believed to be false and the corresponding subfamilies were excluded by later analyses, we examined these 6 false

negative results for cause of false prediction. We found that all 6 negative results were issued by HMMs. The corresponding sequential preservation calculated on CHO-K1 proteins and FMI for HMMs were shown in the Table 6.3. As a threshold of 0.75 was applied to sequential preservation for prediction, all of these genes were predicted normal by the preservation model. However, only a spliced version of ENO1 was found on CHO-K1 proteins, resulting in a lower preservation ratio. Moreover, FMI of most corresponding HMM are lower than 0.7. Only HMM with FMI higher than 0.7 could issue defected predictions, however, any HMM could issue prediction of altered. According to the model performance on these 6 genes, sequential preservation appeared to be a more reliable metric for making predictions and HMMs predictions must be considered with its quality metrics to avoid false results.

Table 6.3 Details of prediction results for 6 glycolysis related genes

Gene name	Subfamily id	Sequential preservation	FMI
HK1	PTHR19443SF36	0.94	0.62
PGK2	PTHR11406SF21	0.97	0.79
PGK1	PTHR11406SF20	0.99	0.63
PFKFB2	PTHR10606SF59	0.94	0.65
ENO1	PTHR11902SF26	0.77	0.61
LDHA	PTHR43128SF4	0.99	0.54

Other results presented previously are worth mentioning as another energy related core process of fatty acid beta-oxidation appeared in mapping altered functions shared by two CHO-K1 genomes but not Chinese hamster. The reason this term appeared is the subfamily containing ACAA1 homologs issued abnormal prediction. This protein was not preserved enough for sequential preservation calculation and the HMM responsible has a FMI value of only 0.62. Therefore, this prediction is likely to be false and more importantly, the both forms in mouse ACAA1A and ACAA1B which were modelled by other subfamilies were predicted normal, indicating this process is not affected by mutations.

6.3 Apoptosis

Apoptosis is a process with many tumour related genes such as BCL2 and BAK. It has been extensively studied in CHO cell for prolonging cell life span and improving productivity. Lewis

et al. (2013) identified 101 anti- and pro-apoptosis proteins making up to 82 genes after merging the variant of the same gene. 39 of them were modelled and had their names mapped with the subfamilies in this project while others may be named differently or missed by the homolog collection. Consistent with Lewis et al. (2013), most of these genes were predicted normal by our models and only 4 had found to be altered in related CHO cell (Table 6.4). Of these 4 genes, only CIDEC were predicted to be altered in both CHO-K1 and CHO-K1GS while others, DFFA, NFKB1 and ATM were only predicted altered in CHO-K1. They also reported another 4 genes: CASP10, IL3RA, IL3 and IL1A to be missing in CHO related genomes. Our models agreed with most of these results except IL1A which was found in all three CHO related genomes with high confidence.

Table 6.4 Predictions on apoptosis related genes

Category	Normal	Altered or defected
Anti-apoptosis	TRAF2, OPTN, MYD88, CFLAR, BIRC7, BIRC3, AKT2, RIPK1, AKT1, XIAP, AKT3, IRAK1, IRAK2, IRAK3, PRKX, IRAK4, BCL2	DFFA, NFKB1, CIDEC
Pro-apoptosis	CASP6, CASP7, AIFM1, CASP3, BID, CHP1, TRADD, CASP9, CIB1, BAX, FADD, CASP8, CHP2, DFFB	ATM
Receptor	NTRK1, IL1R1	
Ligand	IL1B, NGF	

We further mapped the related predictions onto KEGG apoptosis pathway (Figure 6.1). All of the proteins in the pathway were predicted normal (marked with blue in the figure). However, detailed inspection on alignment of BCL2 from CHO-K1 and CHO-K1GS shows significant difference. BCL2 is strongly related to breast cancer so that it is used as one of the markers in diagnosis and treatment for related cancers. BCL2 was found to be overexpressed in most breast cancer cells. While BCL2 in CHO-K1GS is highly conserved to the BCL2 isoform alpha in mouse and human, the CHO-K1 version of BCL2 was found to be more close to the isoform beta which is about 35 residues shorter. However, the 3’ end sequence of BCL2 in CHO-K1 is not similar to that of isoform beta in mouse and human. Both homologs from CHO-K1 and CHO-K1GS are highly preserved on annotated function sites. For now, no evidence has suggested difference in function of the different isoforms. Another protein that attracted our attention was CASP9. Its homolog in CHO-K1 was found to be high conserved with that in

mouse. However, the CASP9 in CHO-K1GS was significantly less conserved and only 75% of its high preservation sites remain preserved. More importantly, one of its two active site annotated was found to be mutated suggesting significant alteration in protein function. However, such mutation could be the result of screening for high viability.

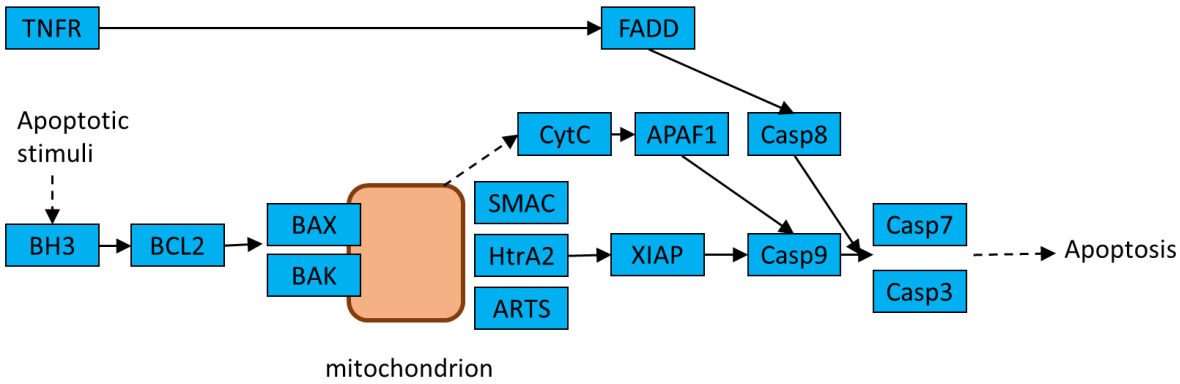


Figure 6.1 KEGG pathway of apoptosis.
All proteins included by the pathway were predicted normal in CHO related genomes.

6.4 DNA repairing

DNA repairing is also an essential process for maintaining genetic information within the cell. Defects or alterations in the process could lead to increase of the mutation rate and genetic instability. Therefore, cancer cell lines, which feature high mutation rate and chromosome instability, may adopt mutations on the related genes. However, as production cell lines, adopting these mutations could affect the stability of transfected sequences which reduce the productivity. As the related GO terms contain many repetition of annotations from various contributors, we adopted a gene selection commonly used in array analyses from QIAGEN website. It contains a total of 84 related genes categorised into five classes: Base Excision Repair (BER), Base Excision Repair (NER), Mismatch repair, Double-strand break repair and others. One of the genes involved, POLD3, was not modelled in this project. Our models predicted most of these genes to be normal in both CHO-K1 and CHO-K1GS, except Xpa, Xrcc5, Atm and Mgmt (**Error! Not a valid bookmark self-reference.**), of which Mgmt was actually predicted normal in CHO-K1GS and Chinese hamster but altered in CHO-K1. It shows

that mismatch repair and BER are perfectly conserved and only a few genes in NER and double-strand break repair presented evidence of significant alterations in at least one CHO cell line.

Table 6.5 Predictions on DNA repair related genes

Category	Normal	Altered or defected
Base Excision Repair	Apex1, Apex2, Ccno, Lig3, Mpg, Mutyh, Neil1, Neil2, Neil3, Nthl1, Ogg1, Parp1, Parp2, Parp3, Polb, Smug1, Ung, Tdg, Xrcc1	
Nucleotide Excision Repair	Atxn3, Brip1, Ccnh, Cdk7, Ddb1, Ddb2, Ercc1, Ercc2, Ercc3, Ercc4, Ercc5, Ercc6, Ercc8, Lig1, Mms19, Pnkp, Poll, Rad23a, Rad23b, Rpa1, Rpa3, Slk, Xab2, Xpc	Xpa
Mismatch Repair	Mlh1, Mlh3, Msh2, Msh3, Msh4, Msh5, Msh6, Pms1, Pms2, Trex1	
Double-Strand Break Repair	Brca1, Brca2, Dmc1, Fen1, Lig4, Mre11a, Prkdc, Rad21, Rad50, Rad51, Rad51c, Rad51b, Rad51d, Rad52, Rad54l, Xrcc2, Xrcc3, Xrcc4, Xrcc6	Xrcc5
Other Genes Related to DNA Repair	Atr, Exo1, Rad18, Rfc1, Top3a, Top3b, Xrcc6bp1	Atm, Mgmt

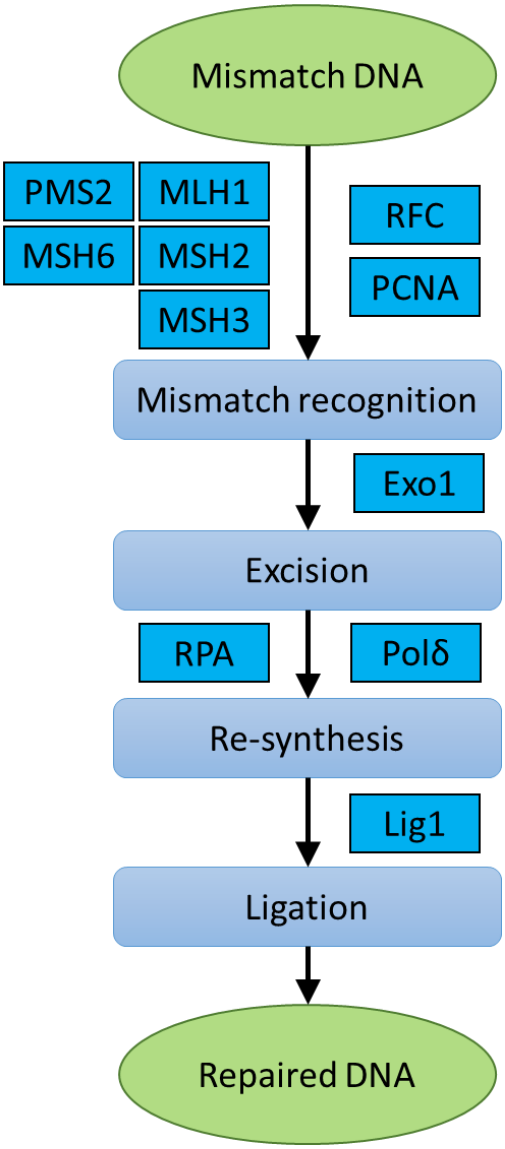


Figure 6.2 KEGG pathway of mismatch repair.
Related proteins were all predicted normal in this pathway in CHO related genomes.

Since pathways of the four different DNA repairing mechanisms were available on KEGG pathway, we further examined the details of these pathways. Most genes are shared by the array selection and pathway. Mapping of the mismatch repair pathway and the array both show perfectly normal in all genes despite the gene involved was quite different (Figure 6.2). In the NER pathway (Figure 6.3), TTDA was not covered by our models and, consistent with predictions on array selection, only XPA was considered altered. Such alteration was only observed in CHO-K1 and unexpectedly Chinese hamster while the CHO-K1GS homolog was highly conserved with the mouse homolog.

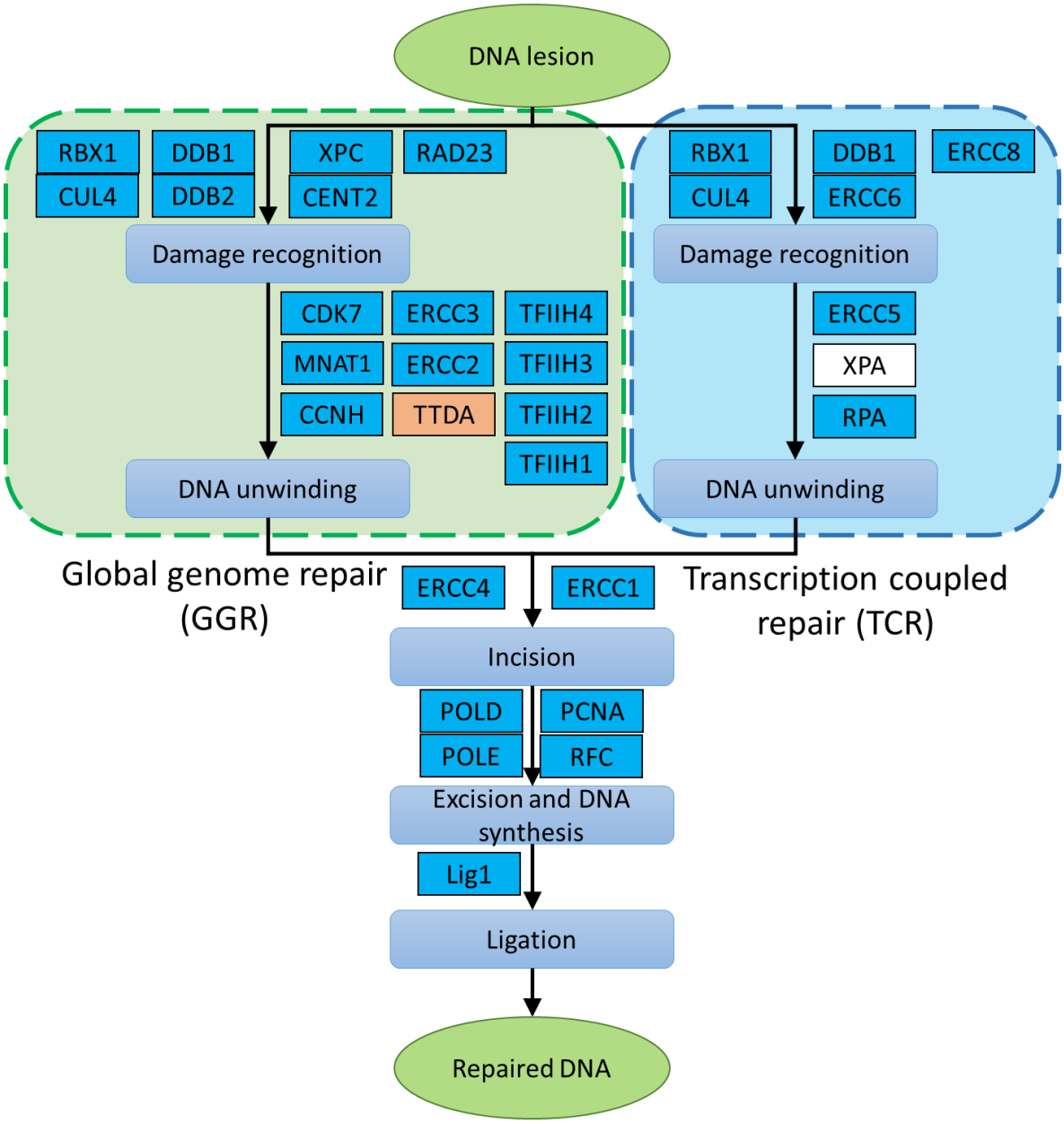


Figure 6.3 KEGG pathway of NER.
The protein in the red box, TTDA, was not covered by the models. The protein in the white box, XPA, was predicted altered or defected in at least one CHO related genomes.

No significant alteration was observed in the other pathways. In the long patch BER pathway (Figure 6.4), gene Mug was predicted to be altered in all three genomes by HMMs with high FMI. The corresponding alignment showed no full length protein found on these genomes, although the fragmented proteins were highly similar to the related part of sequences from other species, suggesting they could be active spliced proteins of the gene. The only two genes reported not normal in double-strand break repair were SYCP3 and BLM (Figure 6.5). However,

the subfamily related to SYCP3 contain only human sequences and thus was not reliable. On the other hand, CHO-K1 homolog of BLM was predicted to be altered as it is significantly shorter than homologs even in CHO-K1GS and Chinese Hamster which were predicted to be normal.

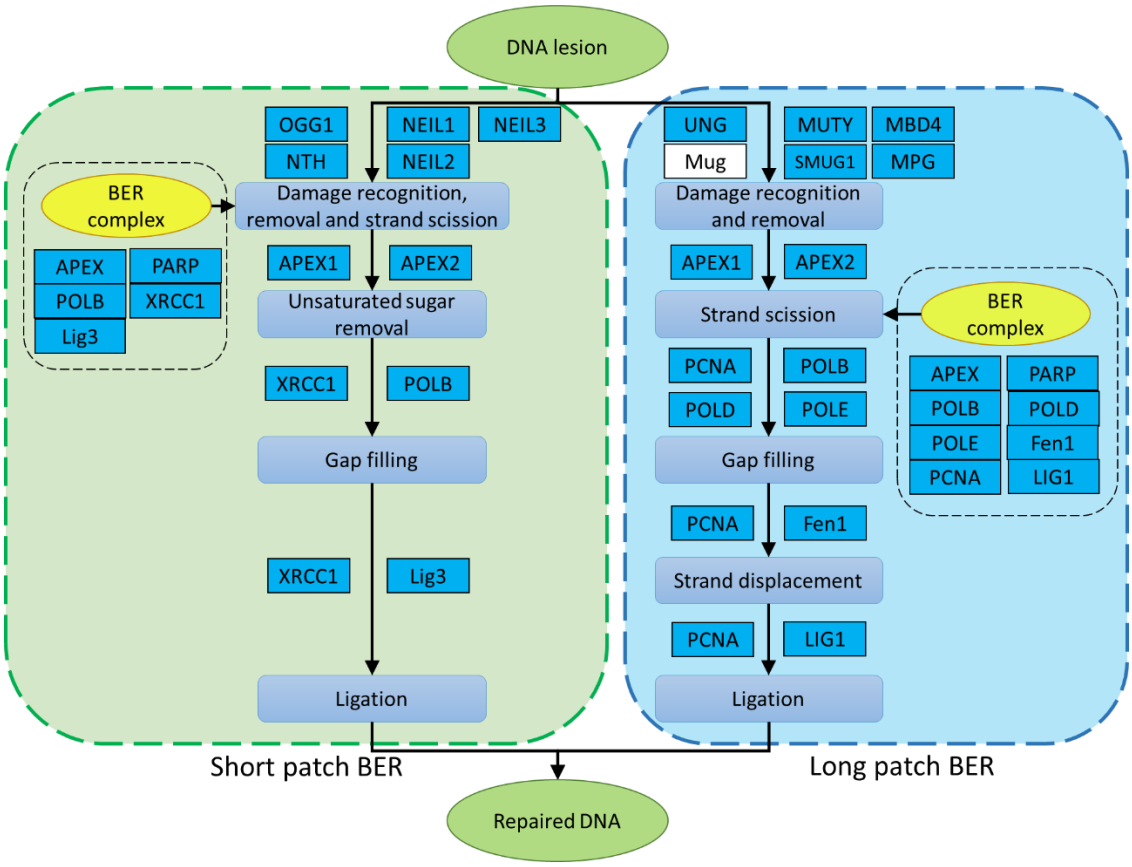


Figure 6.4 KEGG pathway of BER.
The pathway includes short patch and long patch BER and BER complex. Only one related protein, Mug in the white box, was predicted abnormal in at least one of CHO related genomes.

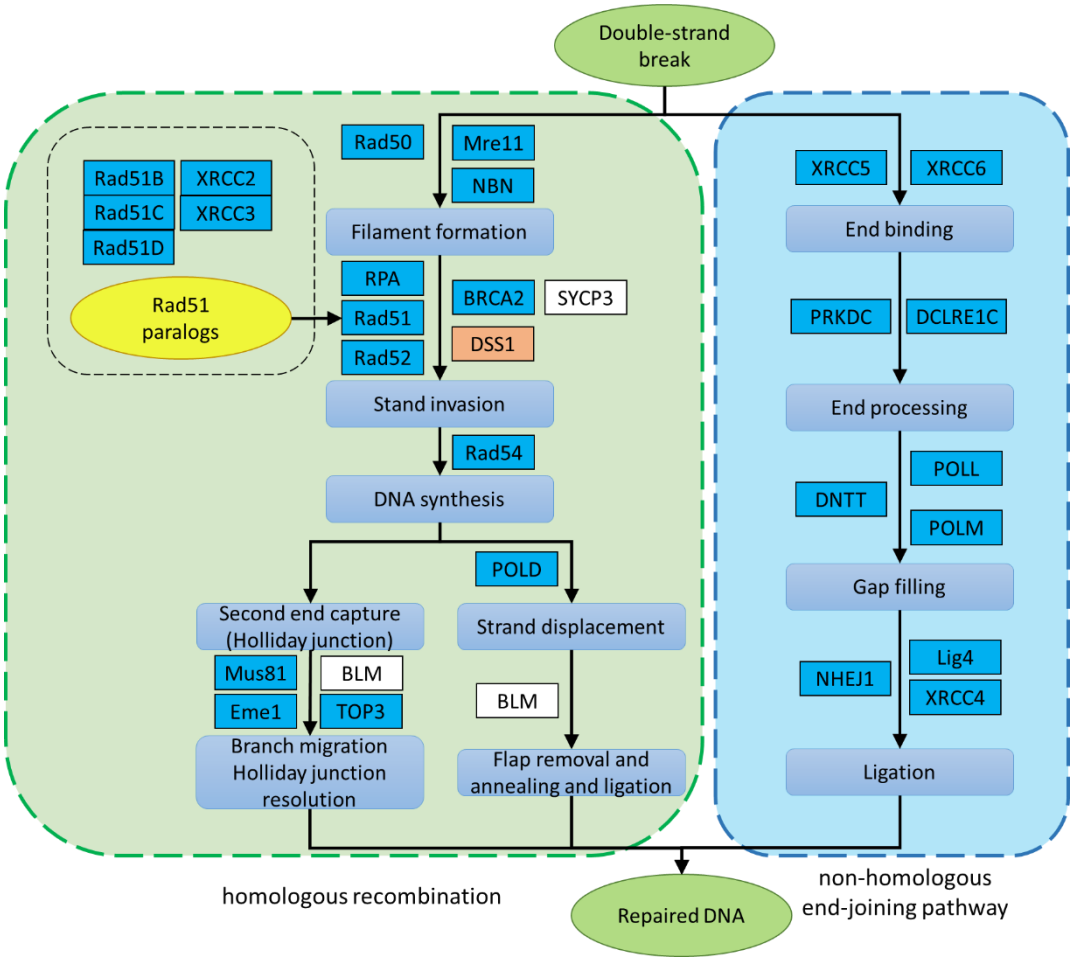


Figure 6.5 KEGG pathway of double-strand break repair.
Two mechanisms, homologous recombination and non-homologous end-joining, are included by the pathway. One protein, DSS1 in the red box, was not included by the models. Two proteins related to homologous recombination pathway, SYCP3 and BLM in white boxes, were predicted abnormal in at least one of the CHO related genomes.

Although only a few genes appeared to be affected by mutations, their homologs in CHO-K1GS are constantly more close to their wildtype compared to CHO-K1. This indicates CHO-K1GS could be a more stable cell line for production.

6.5 Protein Glycosylation

In simple words, glycosylation is a co-translational PTM that adds specific oligosaccharides to proteins. It can be so widely observed in nature that, according to Apweiler (1999), half of the proteins known are glycosylated. For mAbs, one of the main biopharmaceutical products

manufactured using CHO, glycosylation is crucial for their efficacy: on Fab, glycosylation can affect the affinity to target antigen while on Fc, glycosylation affects the efficiency of binding the cell receptor and triggering downstream immune activities, such as antibody-dependant cellular cytotoxicity (ADCC) and complement-dependant cytotoxicity (CDC) (Jefferis, 2005a). Moreover, improper glycosylation can trigger an immune reaction against mAbs themselves. However, the glycosylation profile is highly species-specific and acts as a marker that distinguishes self-bodies from foreign bodies. Therefore, manufacturing mAbs requires the host cell to be able to glycosylate proteins in a human-compatible manner. Although CHO cells are popular as a manufacturing platform, as rodent origin cell lines, further genetic modifications are still required to improve product efficacy.

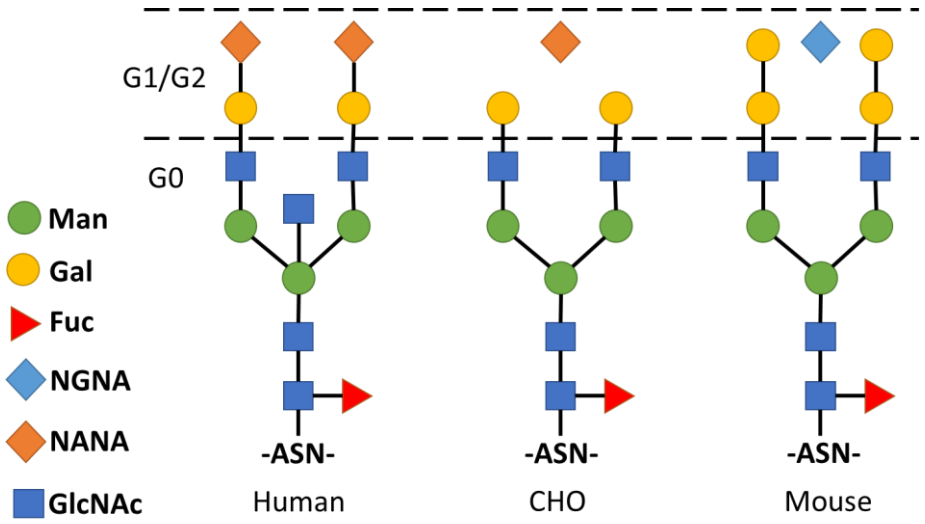


Figure 6.6 Species specific N-glycosylation on Asn.
Different glycosylation of Asn in human, CHO and mouse is compared. (Beck et al., 2008; Jefferis, 2005a)

In mAb IgG, glycosylation can occur at both ends of the residue but the N-glycosylation on the asparagine (Asn) residue is most studied. The Fc region is glycosylated at Asn-297 while multiple sites in the Fab region can be glycosylated, resulting in higher heterogeneity of the IgG glycoform (Jefferis, 2005b). The pathway responsible is shown in Figure 6.7 (Johnson et al., 2014). CHO is able to glycosylate Asn-297 in a slightly different way from humans and mice, as shown in Figure 6.6 (Jefferis 2005 and Beck *et al*. 2008). Although they all share the same core structure, a series of genetic engineering had been conducted on CHO. CHO originally

does not produce bisecting GlcNAc as human, and does not normally express ST6Gal activity to link sialic acid to galactose (Xu et al., 2011; Beck et al., 2008). This fact was confirmed by our models that the protein responsible for these processes, MGAT3 and ST6GAL1, were detected defeated in Chinese hamster genome. In order to improve the ADCC of product mAbs, GnTIII (MGAT3) was transfected into CHO to produce the bisecting GlcNAc branch (Umaña et al., 1999), while FUT8, which is responsible for adding fucose to the oligosaccharide, was knocked out for higher efficacy (Yamane-Ohnuki et al., 2004). Our models showed that although being predicted defect in Chinese hamster, MGAT3 and ST6GAL1 were predicted normal in CHO-K1 and CHO-K1GS, while another protein responsible for ST6GAL activity, ST6GAL1 was found normal in all three CHO related genomes. However, in the CHO-K1 genomes we examined, FUT8 was not knocked out and remained normal. Another critical gene in the pathway, MGAT1 was reported to be altered by our models, but further inspection showed that it is a prediction issued by HMM with low FMI and our preservation model considered it normal. Therefore, it was marked normal as other proteins involved in the pathway.

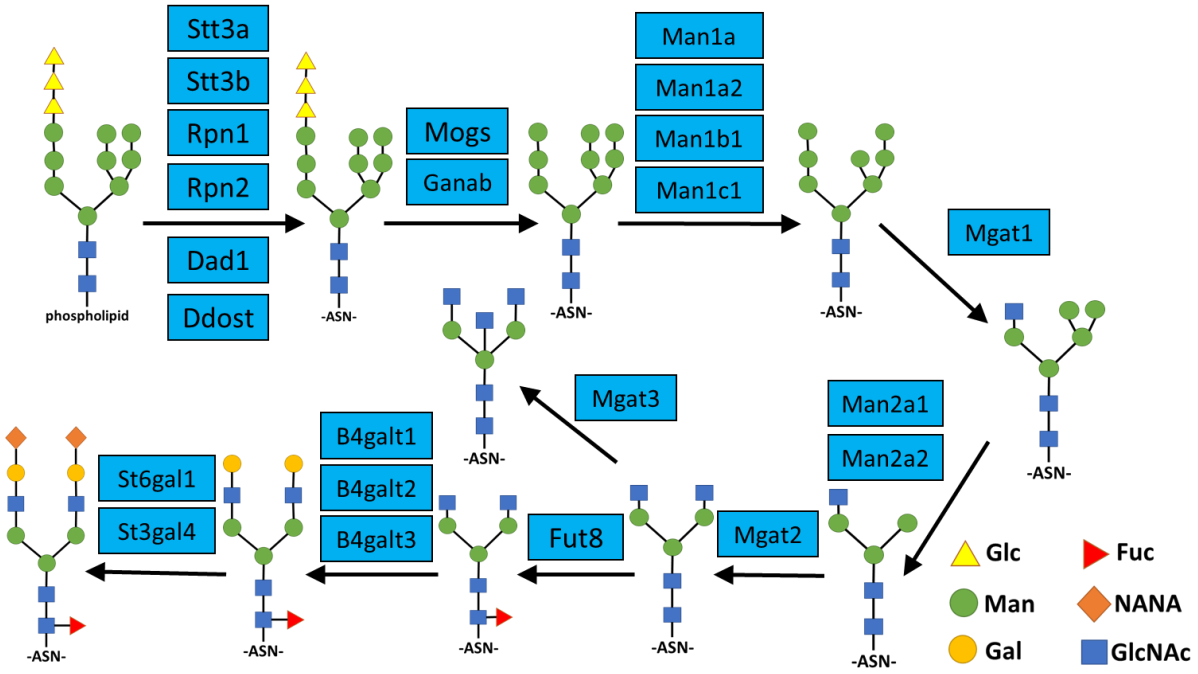


Figure 6.7 N-glycosylation pathway
The pathway is adapted from Johnson et al (2014). All the related proteins were predicted normal in CHO related genomes.

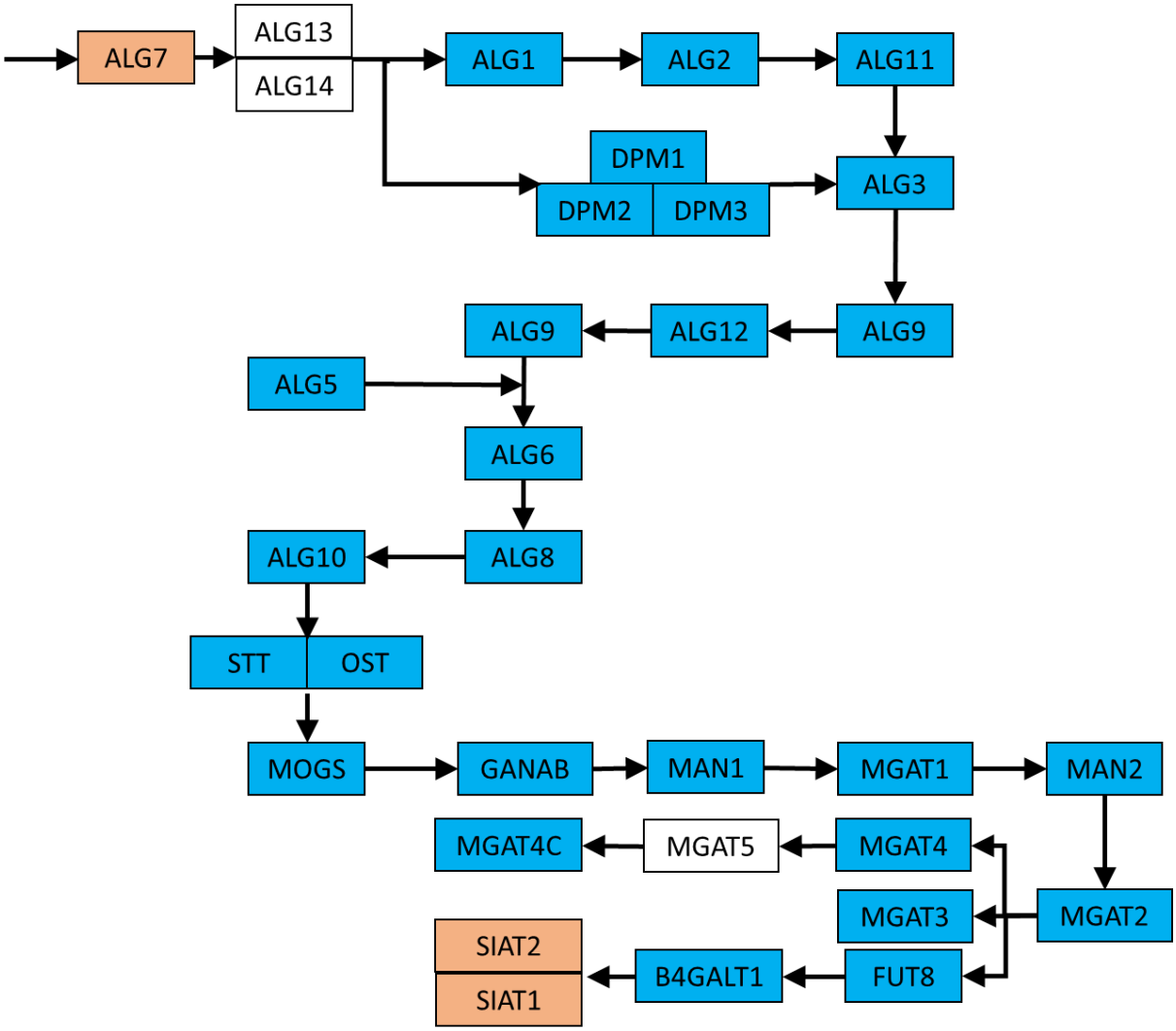


Figure 6.8 KEGG pathway of glycosylation.
Three proteins, ALG7, SIAT1 and SIAT2 marked by red boxes, were not covered by the models. Another three proteins, ALG13, ALG14 and MGAT5 in white boxes, were predicted altered or defected in at least one of CHO related genomes.

A pathway containing former synthesis reaction was found in KEGG pathway showing more glycosylation related proteins (Figure 6.8). However, not all the proteins involved were covered by our models possibly due to lack of annotation. Proteins marked with red (ALG7, SIAT1 and SIAT2) were not covered by our models and thus no prediction was made on them. Proteins marked with blue were predicted normal in CHO-K1 and CHO-K1GS while those marked with white were predicted not normal in at least one of the CHO genomes. Xu et al. (2011) reported lack of ALG13 homolog in the CHO-K1 genome. However, our models predicted ALG13 to be normal in CHO-K1 but altered or defeated in CHO-K1GS and Chinese

hamster. Meanwhile, a similar protein ALG14 was found normal in both CHO-K1GS and Chinese hamster but predicted altered in CHO-K1. Other proteins involved were mostly found normal in this project except MGAT5 which was predicted altered in all three genomes. Our results were consistent with most of the public results and engineering outcome, suggesting that both CHO cell lines are able to produce human-compatible proteins.

7 Conclusion

7.1 Result summary

In this project, we constructed a tool for predicting protein function alteration based on phylogenetic preservation. The tool works on a two steps system: 1) identify best available protein sequences in the genome using HMMs which would issue a primary prediction of the function alteration on the protein; 2) then calculate the site preservation on every site of the matched sequence which would become the base of sequential preservation for final prediction. Validation by step showed that HMMs were capable of finding the correct best available sequence and assigning sequences to the correct subfamilies in high precision. However, the HMMs' prediction of function impact based on sequence variants was not reliable on its own. In comparison, the preservation model was better in making such predictions. The validation shows our preservation model could produce results strongly correlating to the molecular age calculated by PANTHER-PSEP and the annotated function sites from uniprot. Moreover, in the validation with BLAST on selected CHO-K1 sequences, the normal and defected predictions made by our tool were mostly agreed by BLAST in terms of percentage of identity and alignment coverage. In comparison with gene expression data, our sequential preservation showed significant correlation with the maximum expression level. However, further comparison on predictions for different genomes showed that a small proportion of subfamily models are constantly making unreliable negative predictions. After removing these unreliable results, we still could not retrieve hypotheses specific enough to motivate further research on cell line development. The main reason was the biological processes that the hypotheses were built on were too general and involved too many genes so that could not provide clear directions for further research. In the case studies, we identified major source of false predictions of altered function being HMMs with low FMI, which stands for low accuracy in classifying proteins functions to be normal or not. Further investigation on pathways related to protein production showed that CHO-K1GS appear to be a better production cell line than CHO-K1, although the number of protein predicted to be different in function was not high between the two genomes.

7.2 Achievements and limitations

We identified HMM for search homologous sequence and preservation model for potential function altering site from the related research. They both demonstrated good performance for the purpose assigned. However, an arbitrary threshold used for making prediction was not well supported and therefore resulted in questionable predictions in later analyses. Supportive evidence and assumptions are required for the better threshold to be made. A metrics FMI made for evaluating classification capacity of HMMs was only used for predictions of defected function but not altered function which lead to many false predictions. Involving the FMI in all HMM prediction should significantly improve the accuracy. However, it would lead to significant shrink in coverage as subfamilies close to each other would result in low FMI on each other in our pipelines. As a matter of fact, coverage has been a significant limitation for our tool. Firstly, both HMM and preservation model require good quality of homologous sequence collection. To achieve good performance, HMMs required clear grouping on the homologous sequence collection while preservation required involvement of many related species. Secondly, the main source of sequence collection, the PANTHER subfamilies, merely covered less than 10,000 families making up more than 19,000 subfamilies in this project. Compared to the gene counts of more than 24,000 in annotations of mammalian genomes, such a number was insufficient. By the time this thesis was written, the number of family and subfamily included by PANTHER had been significantly improved. However, before proper investigation of the quality of new PANTHER subfamilies, it was not certain that they could improve the effective coverage of our models. The preservation model appeared to be promising in making predictions and identifying critical site mutations in the sequence given that most of the sites are conserved with the wildtype sequence. However, by definition of the PANTHER subfamily, it is normal that the preservation of some subfamilies is low. These subfamilies will never be covered by the preservation model unless we change the method of calculation. An option is to involve sequences within the same linage outside the subfamily. However, it would then conflict with the main objective of HMMs, which was identifying sequence between subfamilies.

Although validation results showed our models were capable of making accurate predictions, we failed to extract practical hypotheses from the prediction. One of the causes for that is GO mapping and enrichment calculation. In this project, we used the most significantly enriched

GO terms for making hypotheses. However, the statistical significance always favours large number so that specific terms with lower number of related gene will rarely be ranked on top. In addition, as the enrichment was calculated against all the genes involved, the calculation for each GO term was independent. However, when comparing GO terms with such enrichment, as GO terms were designed to be overlapping with each other, the assumption of independent does not apply. Therefore, we suggest that using statistical significance as a filter to highlight enriched GO terms is effective, but further sorting GO terms with the significance may not reveal the true order of enrichment significance. Once a significant enriched GO term is identified, its daughter term should be investigated under the condition that their parents are enriched to some extent. However, if one uses a GO-slim with terms independent to each other, such sorting should be preferable. Another cause of failure in generating useful hypotheses was insufficient prediction accuracy which could be improved by the approaches mentioned in previous paragraphs. It is worth mentioning that inaccurate genome annotations could also be the cause. Some of the unexpected results raised our doubts toward the annotations. However, we do not have sufficient evidence showing the annotations are problematic, although improvement on the annotation is absolutely required for better outcome.

7.3 Future works

To accomplish our final goal to accurately predict function impact of variants in CHO proteins, further works could be focused on several aspects:

- Further integration of HMM predictions and preservation using machine learning techniques. Using proper machine learning technique could assign optimised weights to the HMM result and preservation result for better integrative results. More features or metrics, such as expression level, can be added to the integration for better predictions.
- Improving subfamily collection and annotation. In this project, we extended PANTHER subfamily using in silico approaches. However, for higher quality subfamily collection, human review and curation could be required.

- Customising production related GO slim on CHO. Creating GO slim on specified on CHO cell engineering could highlight terms related to protein production and allow useful hypotheses to be made given accurate function predictions.
- Relating CHO phenotype information with sequence variant data. Achieving accurate function prediction relies on strong connection between phenotype and genotype. Therefore, coupling these two sets of data is highly necessary for further improvement of this project. Currently, data of CHO phenotype were rarely extracted with its gene profile but more with the parameters of culture environments.
- Sequencing different CHO cell line for a better consensus CHO genome assembly and annotation. The quality of gene annotation for protein sequences is always a determinative factor for accurate prediction. Improvements of the CHO genome annotation has been constantly made by related institutes such as EBI and NCBI. However, we suggest greater improvement is acquired.

Reference

Adams, G. P. & Weiner, L. M. (2005). Monoclonal Antibody Therapy of Cancer. *Nature biotechnology*, 23(9), 1147–1157.

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A Method and Server for Predicting Damaging Missense Mutations. *Nature Methods*, 7(4), 248–249. Retrieved from http://dx.doi.org/10.1038/nmeth0410-248

Ahn, W. S. & Antoniewicz, M. R. (2011). Metabolic Flux Analysis of CHO Cells at Growth and Non-Growth Phases Using Isotopic Tracers and Mass Spectrometry. *Metabolic Engineering*, 13(5), 598–609. Retrieved from http://dx.doi.org/10.1016/j.ymben.2011.07.002

Akerborg, O., Sennblad, B., Arvestad, L. & Lagergren, J. (2009). Simultaneous Bayesian Gene Tree Reconstruction and Reconciliation Analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 106(14), 5714–5719.

Alfaro, M. E., Zoller, S. & Lutzoni, F. (2003). Bayes or Bootstrap? A Simulation Study Comparing the Performance of Bayesian Markov Chain Monte Carlo Sampling and Bootstrapping in Assessing Phylogenetic Confidence. *Molecular Biology and Evolution*, 20(2), 255–266.

Alkan, C., Sajjadian, S. & Eichler, E. E. (2011). Limitations of Next-Generation Genome Sequence Assembly. *Nature Methods*, 8(1), 61–65.

Altschul, S. F., Madden, T. L., Schäffer, a a, Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic acids research*, 25(17), 3389–402. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=146917&tool=pmcentrez&rendertype=abstract

Altschul, S., Gish, W. & Miller, W. (1990). Basic Local Alignment Search Tool. *Journal of molecular …*, 403–410. Retrieved August 22, 2013, from http://www.sciencedirect.com/science/article/pii/S0022283605803602

Aniba, M. R., Poch, O. & Thompson, J. D. (2010). Issues in Bioinformatics Benchmarking: The Case Study of Multiple Sequence Alignment. *Nucleic acids research*, 38(21), 7353–63. Retrieved May 24, 2013, from

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2995051&tool=pmcentrez&rendertype=abstract

Apweiler, R. (1999). On the Frequency of Protein Glycosylation, as Deduced from Analysis of the SWISS-PROT Database. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1473(1), 4–8. Retrieved from http://www.sciencedirect.com/science/article/pii/S0304416599001658

Apweiler, R., Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alam-Faruque, Y., et al. (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, 42(D1), D191--8. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965022&tool=pmcentrez&rendertype=abstract

Backeljau, T., De Bruyn, L., De Wolf, H., Jordaens, K., Van Dongen, S. & Winnepennincks, B. (1996). Multiple UPGMA and Neighbor-Joining Trees and the Performance of Some Computer Packages. *Molecular Biology and Evolution*, 13(2), 309–313. Retrieved from https://academic.oup.com/mbe/article-lookup/doi/10.1093/oxfordjournals.molbev.a025590

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity. *Nature*, 483(7391), 603–607.

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., et al. (2017). UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Research*, 45(D1), D158–D169.

Beck, A., Wagner-Rousset, E., Bussat, M.-C., Lokteff, M., Klinguer-Hamour, C., Haeuw, J.-F., et al. (2008). Trends in Glycosylation, Glycoanalysis and Glycoengineering of Therapeutic Antibodies and Fc-Fusion Proteins. *Current pharmaceutical biotechnology*, 9(6), 482–501.

Blackshields, G., Wallace, I. M., Larkin, M. & Higgins, D. G. (2006). Analysis and Comparison of Benchmarks for Multiple Sequence Alignment. *In silico biology*, 6(4), 321–39. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16922695

Blackstone, N. W. (2006). Charles Manning Child (1869-1954): The Past, Present, and Future of Metabolic Signaling. *Journal of experimental zoology. Part B, Molecular and developmental evolution*, 306(1), 1–7.

Blake, J. A., Christie, K. R., Dolan, M. E., Drabkin, H. J., Hill, D. P., Ni, L., et al. (2015). Gene Ontology Consortium: Going Forward. *Nucleic Acids Research*, 43(D1), D1049–D1056.

Boussau, B. & Daubin, V. (2010). Genomes as Documents of Evolutionary History. *Trends in Ecology and Evolution*, 25(4), 224–232.

Boussau, B., Szollosi, G. J., Duret, L., Gouy, M., Tannier, E. & Daubin, V. (2013). Genome-Scale Coestimation of Species and Gene Trees. *Genome Research*, 23(2), 323–330. Retrieved from http://genome.cshlp.org/cgi/doi/10.1101/gr.141978.112

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., et al. (2008). The Potential and Challenges of Nanopore Sequencing. *Nature Biotechnology*, 26(10), 1146–1153.

Brinkrolf, K., Rupp, O., Laux, H., Kollin, F., Ernst, W., Linke, B., et al. (2013). Chinese Hamster Genome Sequenced from Sorted Chromosomes. *Nature biotechnology*, 31(8), 694–5. Retrieved March 26, 2014, from http://www.nature.com/nbt/journal/v31/n8/full/nbt.2645.html

Chen, N., Koumpouras, G. C., Polizzi, K. M. & Kontoravdi, C. (2012). Genome-Based Kinetic Modeling of Cytosolic Glucose Metabolism in Industrially Relevant Cell Lines: Saccharomyces Cerevisiae and Chinese Hamster Ovary Cells. *Bioprocess and biosystems engineering*, 35(6), 1023–33. Retrieved November 13, 2012, from http://www.ncbi.nlm.nih.gov/pubmed/22286123

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. (2012). Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS ONE*, 7(10).

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. (2015). GenBank. *Nucleic Acids Research*, gkv1276. Retrieved from http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv1276

Cline, M. S. & Karchin, R. (2011). Using Bioinformatics to Predict the Functional Impact of SNVs. *Bioinformatics*, 27(4), 441–448.

Collins, F. S., Morgan, M. & Patrinos, A. (2003). The Human Genome Project: Lessons from Large-Scale Biology. *Science (New York, N.Y.)*, 300(5617), 286–290.

Cortés, J. & Calvo, E. (2014). New Approach to Cancer Therapy Based on a Molecularly Defined Cancer Classification. *CA: a cancer journal …*, 64(1), 70–74. Retrieved from http://onlinelibrary.wiley.com/doi/10.3322/caac.21211/full

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., et al. (2014). The Reactome

Pathway Knowledgebase. *Nucleic Acids Research*, 42(D1), D472–D477. Retrieved from http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkt1102

David, A., Razali, R., Wass, M. N. & Sternberg, M. J. E. (2012). Protein-Protein Interaction Sites Are Hot Spots for Disease-Associated Nonsynonymous SNPs. *Human Mutation*, 33(2), 359–363.

Dayhoff, M. & Schwartz, R. (1978). A Model of Evolutionary Change in Proteins. *In Atlas of protein sequence and structure*, 345–352. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.145.4315

Do, C. B., Mahabhashyam, M. S. P., Brudno, M. & Batzoglou, S. (2005). ProbCons: Probabilistic Consistency-Based Multiple Sequence Alignment. *Genome research*, 15(2), 330–40. Retrieved March 1, 2013, from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=546535&tool=pmcentrez&rendertype=abstract

Doolan, P., Clarke, C., Kinsella, P., Breen, L., Meleady, P., Leonard, M., et al. (2013). Transcriptomic Analysis of Clonal Growth Rate Variation during CHO Cell Line Development. *Journal of biotechnology*, 166(3), 105–113. Retrieved May 30, 2013, from http://www.ncbi.nlm.nih.gov/pubmed/23651948

Eddy, S. (1998). Profile Hidden Markov Models. *Bioinformatics*, 14(9), 755–763. Retrieved November 7, 2013, from http://bioinformatics.oxfordjournals.org/content/14/9/755.short

Eddy, S. (2004). What Is a Hidden Markov Model? *Nature biotechnology*, 22(10), 1315–1316. Retrieved November 7, 2013, from ftp://ftp.cecalc.ula.ve/bioinfo/temp/Eddy-WhatisaHiddenMarkovModel.pdf

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS computational biology*, 7(10), e1002195. Retrieved October 25, 2012, from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3197634&tool=pmcentrez&rendertype=abstract

Eddy, S. R. & Wheeler, T. J. (2013). HMMER User ' s Guide. *HMMER Development Team*, (May).

Edgar, R. C. (2004). MUSCLE: A Multiple Sequence Alignment Method with Reduced Time and Space Complexity. *BMC bioinformatics*, 5, 113.

EFPIA. (2015). *The Pharmaceutical Industry in Figures*.

Eisen, J. A. (1998). Phylogenomics : Improving Functional Predictions for Uncharacterized Genes by Evolutionary ? Analysis Phylogenomics : Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. *Genome Research*, (1997), 163–167.

Ellegren, H. (2014). Genome Sequencing and Population Genomics in Non-Model Organisms. *Trends in Ecology and Evolution*, 29(1), 51–63. Retrieved from http://dx.doi.org/10.1016/j.tree.2013.09.008

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., et al. (2015). The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, 44(December 2015), gkv1351. Retrieved from http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv1351

Feldmann, M. & Maini, R. N. (2003). Lasker Clinical Medical Research Award. TNF Defined as a Therapeutic Target for Rheumatoid Arthritis and Other Autoimmune Diseases. *Nature medicine*, 9(10), 1245–1250.

Felsenstein, J. (1981). Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. *Journal of Molecular Evolution*, 17(6), 368–376.

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., et al. (2017). InterPro in 2017-beyond Protein Family and Domain Annotations. *Nucleic Acids Research*, 45(D1), D190–D199.

Finn, R. D., Clements, J. & Eddy, S. R. (2011). HMMER Web Server: Interactive Sequence Similarity Searching. *Nucleic acids research*, 39(Web Server issue), W29-37. Retrieved May 23, 2013, from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3125773&tool=pmcentrez&rendertype=abstract

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The Pfam Protein Families Database: Towards a More Sustainable Future. *Nucleic Acids Research*, 44(D1), D279–D285.

Frousios, K., Iliopoulos, C. S., Schlitt, T. & Simpson, M. A. (2013). Predicting the Functional Consequences of Non-Synonymous DNA Sequence Variants - Evaluation of Bioinformatics Tools and Development of a Consensus Strategy. *Genomics*, 102(4), 223–228. Retrieved from http://dx.doi.org/10.1016/j.ygeno.2013.06.005

González-Pérez, A. & López-Bigas, N. (2011). Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. *American*

Journal of Human Genetics, 88(4), 440–449.

Goudar, C., Biener, R., Boisart, C., Heidemann, R., Piret, J., de Graaf, A., et al. (2010). Metabolic Flux Analysis of CHO Cells in Perfusion Culture by Metabolite Balancing and 2D [13C, 1H] COSY NMR Spectroscopy. *Metabolic Engineering*, 12(2), 138–149.

Gray, V. E., Kukurba, K. R. & Kumar, S. (2012). Performance of Computational Tools in Evaluating the Functional Impact of Laboratory-Induced Amino Acid Mutations. *Bioinformatics*, 28(16), 2093–2096.

Grimm, D. G., Azencott, C. A., Aicheler, F., Gieraths, U., Macarthur, D. G., Samocha, K. E., et al. (2015). The Evaluation of Tools Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. *Human Mutation*, 36(5), 513–523.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3), 307–321.

Guindon, S. & Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, 52(5), 696–704.

Hacker, D. L., De Jesus, M. & Wurm, F. M. (2009). 25 Years of Recombinant Proteins from Reactor-Grown Cells — Where Do We Go from Here? *Biotechnology Advances*, 27(6), 1023–1027. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0734975009000974

Hamosh, A. (2004). Online Mendelian Inheritance in Man (OMIM), a Knowledgebase of Human Genes and Genetic Disorders. *Nucleic Acids Research*, 33(Database issue), D514–D517. Retrieved from https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gki033

Harris, M. a, Clark, J., Ireland, a, Lomax, J., Ashburner, M., Foulger, R., et al. (2004). The Gene Ontology (GO) Database and Informatics Resource. *Nucleic acids research*, 32(Database issue), D258-61. Retrieved August 6, 2013, from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=308770&tool=pmcentrez&rendertype=abstract

Hefzi, H., Ang, K. S., Hanscho, M., Bordbar, A., Ruckerbauer, D., Lakshmanan, M., et al. (2016). A Consensus Genome-Scale Reconstruction of Chinese Hamster Ovary Cell Metabolism. *Cell Systems*, 3(5), 434–443.e8.

Henikoff, S. & Henikoff, J. G. (1992). Amino Acid Substitution Matrices from Protein Blocks.

Proceedings of the National Academy of Sciences of the United States of America, 89(22), 10915–10919.

Hernández Bort, J. a, Hackl, M., Höflmayer, H., Jadhav, V., Harreither, E., Kumar, N., et al. (2012). Dynamic MRNA and MiRNA Profiling of CHO-K1 Suspension Cell Cultures. *Biotechnology journal*, 7(4), 500–15. Retrieved March 1, 2013, from http://www.ncbi.nlm.nih.gov/pubmed/21751394

Higgins, D. G. & Sharp, P. M. (1988). CLUSTAL: A Package for Performing Multiple Sequence Alignment on a Microcomputer. *Gene*, 73(1), 237–44. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3243435

Holder, M. & Lewis, P. O. (2003). Phylogeny Estimation: Traditional and Bayesian Approaches. *Nature Reviews Genetics*, 4(4), 275–284. Retrieved from http://www.nature.com/doifinder/10.1038/nrg1044

Hossler, P., Khattak, S. F. & Li, Z. J. (2009). Optimal and Consistent Protein Glycosylation in Mammalian Cell Culture. *Glycobiology*, 19(9), 936–49. Retrieved November 1, 2013, from http://www.ncbi.nlm.nih.gov/pubmed/19494347

Hu, P., Bader, G., Wigle, D. a & Emili, A. (2007). Computational Prediction of Cancer-Gene Function. *Nature reviews. Cancer*, 7(1), 23–34. Retrieved May 13, 2014, from http://www.ncbi.nlm.nih.gov/pubmed/17167517

Huang, Y. J., Mao, B., Aramini, J. M. & Montelione, G. T. (2014). Assessment of Template-Based Protein Structure Predictions in CASP10. *Proteins: Structure, Function and Bioinformatics*, 82(SUPPL.2), 43–56.

Huelsenbeck, J. P. & Ronquist, F. (2001). MRBAYES: Bayesian Inference of Phylogenetic Trees. *Bioinformatics*, 17(8), 754–755. Retrieved from https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/17.8.754

Jefferis, R. (2005a). Glycosylation of Recombinant Antibody Therapeutics. *Biotechnology progress*, 11–16. Retrieved August 1, 2014, from http://onlinelibrary.wiley.com/doi/10.1021/bp040016j/full

Jefferis, R. (2005b). Glycosylation of Recombinant Antibody Therapeutics. *Biotechnology Progress*, 21(1), 11–16.

Johnson, K. C., Yongky, A., Vishwanathan, N., Jacob, N. M., Jayapal, K. P., Goudar, C. T., et al. (2014). Exploring the Transcriptome Space of a Recombinant BHK Cell Line through

next Generation Sequencing. *Biotechnology and Bioengineering*, 111(4), 770–781.

Kaas, C. S., Kristensen, C., Betenbaugh, M. J. & Andersen, M. R. (2015). Sequencing the CHO DXB11 Genome Reveals Regional Variations in Genomic Stability and Haploidy. *BMC genomics*, 16, 160. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4359788&tool=pmcentrez&rendertype=abstract

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. (2012). KEGG for Integration and Interpretation of Large-Scale Molecular Data Sets. *Nucleic acids research*, 40(Database issue), D109-14. Retrieved August 6, 2013, from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245020&tool=pmcentrez&rendertype=abstract

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. (2014). Data, Information, Knowledge and Principle: Back to Metabolism in KEGG. *Nucleic Acids Research*, 42(D1), D199–D205. Retrieved from http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkt1076

Kantardjieff, A., Jacob, N. M., Yee, J. C., Epstein, E., Kok, Y.-J., Philp, R., et al. (2010). Transcriptome and Proteome Analysis of Chinese Hamster Ovary Cells under Low Temperature and Butyrate Treatment. *Journal of biotechnology*, 145(2), 143–59. Retrieved October 19, 2013, from http://www.ncbi.nlm.nih.gov/pubmed/19770009

Katoh, K., Misawa, K., Kuma, K. & Miyata, T. (2002). MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform. *Nucleic acids research*, 30(14), 3059–66. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=135756&tool=pmcentrez&rendertype=abstract

Katoh, K. & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular biology and evolution*, 1–9. Retrieved March 1, 2013, from http://www.ncbi.nlm.nih.gov/pubmed/23329690

Katsonis, P., Koire, A., Wilson, S. J., Hsu, T. K., Lua, R. C., Wilkins, A. D., et al. (2014). Single Nucleotide Variations: Biological Impact and Theoretical Interpretation. *Protein Science*, 23(12), 1650–1666.

Kelley, L. A. & Sternberg, M. J. E. (2009). Protein Structure Prediction on the Web: A Case Study Using the Phyre Server. *Nature protocols*, 4(3), 363–371. Retrieved from

http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19247286&retmode=ref&cmd=prlinks%5Cnpapers3://publication/doi/10.1038/nprot.2009.2

Kelly, P. S., Clarke, C., Costello, A., Monger, C., Meiller, J., Dhiman, H., et al. (2017). Ultra-Deep next Generation Mitochondrial Genome Sequencing Reveals Widespread Heteroplasmy in Chinese Hamster Ovary Cells. *Metabolic Engineering*, 41(February), 11–22. Retrieved from http://dx.doi.org/10.1016/j.ymben.2017.02.001

Kim, J. Y., Kim, Y.-G. & Lee, G. M. (2012). CHO Cells in Biotechnology for Production of Recombinant Proteins: Current State and Further Potential. *Applied Microbiology and Biotechnology*, 93(3), 917–930. Retrieved from http://link.springer.com/10.1007/s00253-011-3758-5

Kojima, H., Takeuchi, S., Uramaru, N., Sugihara, K., Yoshida, T. & Kitamura, S. (2009). Nuclear Hormone Receptor Activity of Polybrominated Diphenyl Ethers and Their Hydroxylated and Methoxylated Metabolites in Transactivation Assays Using Chinese Hamster Ovary Cells. *Environmental Health Perspectives*, 117(8), 1210–1218.

Kremkow, B. G., Baik, J. Y., MacDonald, M. L. & Lee, K. H. (2015). CHOgenome.Org 2.0: Genome Resources and Website Updates. *Biotechnology Journal*, 10(7), 931–938.

Krogh, a, Brown, M., Mian, I. S., Sjölander, K. & Haussler, D. (1994). Hidden Markov Models in Computational Biology. Applications to Protein Modeling. *Journal of molecular biology*, 235(5), 1501–1531.

Kumar, P., Henikoff, S. & Ng, P. C. (2009). Predicting the Effects of Coding Non-Synonymous Variants on Protein Function Using the SIFT Algorithm. *Nature protocols*, 4(7), 1073–1081.

Lartillot, N., Lepage, T. & Blanquart, S. (2009). PhyloBayes 3: A Bayesian Software Package for Phylogenetic Reconstruction and Molecular Dating. *Bioinformatics*, 25(17), 2286–2288. Retrieved from https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp368

Le, H., Chen, C. & Goudar, C. T. (2015). An Evaluation of Public Genomic References for Mapping RNA-Seq Data from Chinese Hamster Ovary Cells. *Biotechnology and Bioengineering*, 112(11), 2412–2416.

Le, Q., Sievers, F. & Higgins, D. G. (2017). Protein Multiple Sequence Alignment Benchmarking through Secondary Structure Prediction. *Bioinformatics*, 33(9), btw840. Retrieved from https://academic.oup.com/bioinformatics/article-

lookup/doi/10.1093/bioinformatics/btw840

Lee, J. S., Grav, L. M., Pedersen, L. E., Lee, G. M. & Kildegaard, H. F. (2016). Accelerated Homology-Directed Targeted Integration of Transgenes in Chinese Hamster Ovary Cells via CRISPR/Cas9 and Fluorescent Enrichment. *Biotechnology and Bioengineering*, 113(11), 2518–2523.

Lee, N., Shin, J., Park, J. H., Lee, G. M., Cho, S. & Cho, B. K. (2016). Targeted Gene Deletion Using DNA-Free RNA-Guided Cas9 Nuclease Accelerates Adaptation of CHO Cells to Suspension Culture. *ACS Synthetic Biology*, 5(11), 1211–1219.

De Leon Gatti, M., Wlaschin, K. F., Nissom, P. M., Yap, M. & Hu, W.-S. (2007). Comparative Transcriptional Analysis of Mouse Hybridoma and Recombinant Chinese Hamster Ovary Cells Undergoing Butyrate Treatment. *Journal of bioscience and bioengineering*, 103(1), 82–91. Retrieved October 19, 2013, from http://www.ncbi.nlm.nih.gov/pubmed/17298905

Letunic, I. & Bork, P. (2016). Interactive Tree of Life (ITOL) v3: An Online Tool for the Display and Annotation of Phylogenetic and Other Trees. *Nucleic acids research*, 44(W1), W242–W245.

Lewis, N. E., Liu, X., Li, Y., Nagarajan, H., Yerganian, G., O'Brien, E., et al. (2013). Genomic Landscapes of Chinese Hamster Ovary Cell Lines as Revealed by the Cricetulus Griseus Draft Genome. *Nature Biotechnology*, 31(8), 759–765. Retrieved October 6, 2014, from http://dx.doi.org/10.1038/nbt.2624

Ling, H., Vincent, K., Pichler, M., Fodde, R., Berindan-Neagoe, I., Slack, F. J., et al. (2015). Junk DNA and the Long Non-Coding RNA Twist in Cancer Genetics. *Oncogene*, 34(39), 5003–5011. Retrieved from http://dx.doi.org/10.1038/onc.2014.456

Liu, K., Linder, C. R. & Warnow, T. (2011). Multiple Sequence Alignment: A Major Challenge to Large-Scale Phylogenetics. *PLoS Currents*, 2(NOV), RRN1198. Retrieved from http://currents.plos.org/treeoflife/article/multiple-sequence-alignment-a-major-challenge-to-large-scale-phylogenetics

Liu, L., Yu, L., Kubatko, L., Pearl, D. K. & Edwards, S. V. (2009). Coalescent Methods for Estimating Phylogenetic Trees. *Molecular Phylogenetics and Evolution*, 53(1), 320–328. Retrieved from http://dx.doi.org/10.1016/j.ympev.2009.05.033

Liu, X., Jian, X. & Boerwinkle, E. (2011). DbNSFP: A Lightweight Database of Human Nonsynonymous SNPs and Their Functional Predictions. *Human Mutation*, 32(8), 894–

899.

Liu, X., Jian, X. & Boerwinkle, E. (2013). DbNSFP v2.0: A Database of Human Non-Synonymous SNVs and Their Functional Predictions and Annotations. *Human Mutation*, 34(9), 2393–2402.

Loeb, K. R. & Loeb, L. a. (2000). Significance of Multiple Mutations in Cancer. *Carcinogenesis*, 21(3), 379–385.

Loman, N. J., Constantinidou, C., Chan, J. Z. M., Halachev, M., Sergeant, M., Penn, C. W., et al. (2012). High-Throughput Bacterial Genome Sequencing: An Embarrassment of Choice, a World of Opportunity. *Nature Reviews Microbiology*, 10(9), 599–606. Retrieved from http://dx.doi.org/10.1038/nrmicro2850

Madden, T. L. (2013). The BLAST Sequence Analysis Tool. *The NCBI Handbook*. Retrieved March 21, 2015, from http://www.ncbi.nlm.nih.gov/books/NBK153387/

Maddison, W. P. (1997). Gene Trees in Species Trees. *Systematic Biology*, 46(3), 523–536. Retrieved from http://sysbio.oxfordjournals.org/content/46/3/523.abstract

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome Sequencing in Microfabricated High-Density Picolitre Reactors. *Nature*, 437, 376–380.

Marini, N. J., Thomas, P. D. & Rine, J. (2010). The Use of Orthologous Sequences to Predict the Impact of Amino Acid Substitutions on Protein Function. *PLoS Genetics*, 6(5), 3.

Marks, D. S., Hopf, T. A. & Sander, C. (2012). Protein Structure Prediction from Sequence Variation. *Nature Biotechnology*, 30(11), 1072–1080. Retrieved from http://dx.doi.org/10.1038/nbt.2419

Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., et al. (2005). The PANTHER Database of Protein Families, Subfamilies, Functions and Pathways. *Nucleic Acids Research*, 33(Database Issue), D284–D288. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=540032&tool=pmcentrez&rendertype=abstract

Mi, H., Muruganujan, A. & Thomas, P. D. (2013). PANTHER in 2013: Modeling the Evolution of Gene Function, and Other Gene Attributes, in the Context of Phylogenetic Trees. *Nucleic acids research*, 41(Database issue), D377-86. Retrieved August 24, 2013, from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531194&tool=pmcentrez&rendertype=abstract

Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. (2016). PANTHER Version 10: Expanded Protein Families and Functions, and Analysis Tools. *Nucleic Acids Research*, 44(D1), D336–D342.

Mirarab, S., Bayzid, M. S. & Warnow, T. (2016). Evaluating Summarymethods for Multilocus Species Tree Estimation in the Presence of Incomplete Lineage Sorting. *Systematic Biology*, 65(3), 366–380.

Mitchell, A., Chang, H., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., et al. (2014). The InterPro Protein Families Database: The Classification Resource after 15 Years. *Nucleic acids research*, 43(Database issue), D213-21. Retrieved from http://nar.oxfordjournals.org/content/43/D1/D213%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4383996&tool=pmcentrez&rendertype=abstract

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T. L., Agarwala, R. & Schäffer, A. a. (2008). Database Indexing for Production MegaBLAST Searches. *Bioinformatics*, 24(16), 1757–1764.

Mottaz, A., David, F. P. A., Veuthey, A. L. & Yip, Y. L. (2010). Easy Retrieval of Single Amino-Acid Polymorphisms and Phenotype Information Using SwissVar. *Bioinformatics*, 26(6), 851–852.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. (2014). Critical Assessment of Methods of Protein Structure Prediction (CASP) - Round X. *Proteins: Structure, Function and Bioinformatics*, 82(SUPPL.2), 1–6.

Needleman, S. B. & Wunsch, C. D. (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of molecular biology*, 48(3), 443–453.

Nei, M., Saitou, N. & Nei, M. (1987). The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution*, 4(4), 406–425. Retrieved from https://academic.oup.com/mbe/article/4/4/406/1029664/The-neighborjoining-method-a-new-method-for

Ng, P. C. & Henikoff, S. (2003). SIFT: Predicting Amino Acid Changes That Affect Protein Function. *Nucleic Acids Research*, 31(13), 3812–3814.

Nguyen, T. H., Ranwez, V., Pointet, S., Chifolleau, A.-M. A., Doyon, J.-P. & Berry, V. (2013). Reconciliation and Local Gene Tree Rearrangement Can Be of Mutual Profit. *Algorithms for molecular biology : AMB*, 8(1), 12. Retrieved from

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3871789&tool=pmcentrez&rendertype=abstract

Notredame, C., Higgins, D. G. & Heringa, J. (2000). T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment. *Journal of molecular biology*, 302(1), 205–17. Retrieved May 22, 2013, from http://www.ncbi.nlm.nih.gov/pubmed/10964570

Noutahi, E., Semeria, M., Lafond, M., Seguin, J., Boussau, B., Guéguen, L., et al. (2016a). Efficient Gene Tree Correction Guided by Genome Evolution (M. Robinson-Rechavi, Ed.). *PLOS ONE*, 11(8), e0159559. Retrieved from http://dx.plos.org/10.1371/journal.pone.0159559

Noutahi, E., Semeria, M., Lafond, M., Seguin, J., Boussau, B., Guéguen, L., et al. (2016b). Efficient Gene Tree Correction Guided by Genome Evolution (M. Robinson-Rechavi, Ed.). *PLOS ONE*, 11(8), e0159559. Retrieved from https://hal.archives-ouvertes.fr/hal-01162963

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2012). Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. Retrieved from http://dl.acm.org/citation.cfm?id=2078195%5Cnhttp://arxiv.org/abs/1201.0490

Philippidis, A. (2015). The Top 25 Best-Selling Drugs of 2014. *Genetic Engineering & Biotechnology News*.

Podlaha, O., Riester, M., De, S. & Michor, F. (2012). Evolution of the Cancer Genome. *Trends in Genetics*, 28(4), 155–163. Retrieved from http://dx.doi.org/10.1038/nrg3317

Popp, O., Müller, D., Didzus, K., Paul, W., Lipsmeier, F., Kirchner, F., et al. (2016). A Hybrid Approach Identifies Metabolic Signatures of High-Producers for Chinese Hamster Ovary Clone Selection and Process Optimization. *Biotechnology and Bioengineering*, 113(9), 2005–2019.

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., et al. (2014). RefSeq: An Update on Mammalian Reference Sequences. *Nucleic Acids Research*, 42(D1), 756–763.

Rannala, B. & Yang, Z. (1996). Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenetic Inference. *Journal of Molecular Evolution*, 43(3), 304–311. Retrieved from papers2://publication/uuid/ED47C0A7-E7F9-4C38-B925-BCFFDE86F7EE

Rasmussen, M. D. & Kellis, M. (2011). A Bayesian Approach for Fast and Accurate Gene Tree Reconstruction. *Molecular Biology and Evolution*, 28(1), 273–290.

Reva, B., Antipin, Y. & Sander, C. (2011). Predicting the Functional Impact of Protein Mutations: Application to Cancer Genomics. *Nucleic Acids Research*, 39(17).

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., et al. (2012). Mrbayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice across a Large Model Space. *Systematic Biology*, 61(3), 539–542.

Rose, P. W., Bi, C., Bluhm, W. F., Christie, C. H., Dimitropoulos, D., Dutta, S., et al. (2013). The RCSB Protein Data Bank: New Resources for Research and Education. *Nucleic Acids Research*, 41(D1), 1–8.

Roy, A., Kucukural, A. & Zhang, Y. (2010). I-TASSER: A Unified Platform for Automated Protein Structure and Function Prediction. *Nature protocols*, 5(4), 725–738.

Roychowdhury, S. & Chinnaiyan, A. M. (2016). Translating Cancer Genomes and Transcriptomes for Precision Oncology. *CA: A Cancer Journal for Clinicians*, 66(1), 75–88. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/26528881%5Cnhttp://doi.wiley.com/10.3322/caac.21329

Rupp, O., Becker, J., Brinkrolf, K., Timmermann, C., Borth, N., Pühler, A., et al. (2014). Construction of a Public CHO Cell Line Transcript Database Using Versatile Bioinformatics Analysis Pipelines. *PloS One*, 9(1), e85568. Retrieved May 2, 2014, from http://dx.plos.org/10.1371/journal.pone.0085568

Sander, J. D. & Joung, J. K. (2014). CRISPR-Cas Systems for Editing, Regulating and Targeting Genomes. *Nature biotechnology*, 32(4), 347–55. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24584096

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., et al. (1977). Nucleotide Sequence of Bacteriophage Phi X174 DNA. *Nature*, 265, 687–695.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA Sequencing with Chain-Terminating Inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), 5463–7. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=431765&tool=pmcentrez&rendertype=abstract

Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. (2010). MutationTaster

Evaluates Disease-Causing Potential of Sequence Alterations. *Nature Methods*, 7(8), 575–576. Retrieved from http://dx.doi.org/10.1038/nmeth0810-575

Shendure, J. (2005). Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science*, 309(5741), 1728–1732. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16081699

Shendure, J. & Ji, H. (2008). Next-Generation DNA Sequencing. *Nature biotechnology*, 26(10), 1135–45. Retrieved August 6, 2013, from http://www.ncbi.nlm.nih.gov/pubmed/18846087

Shihab, H. A., Gough, J., Cooper, D. N., Day, I. N. M. & Gaunt, T. R. (2013). Predicting the Functional Consequences of Cancer-Associated Amino Acid Substitutions. *Bioinformatics*, 29(12), 1504–1510.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., et al. (2011). Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega. *Molecular Systems Biology*, 7(539).

Simmons, M. P. & Gatesy, J. (2015). Coalescence vs. Concatenation: Sophisticated Analyses vs. First Principles Applied to Rooting the Angiosperms. *Molecular Phylogenetics and Evolution*, 91, 98–122. Retrieved from http://dx.doi.org/10.1016/j.ympev.2015.05.011

Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J. & Timp, W. (2017). Detecting DNA Cytosine Methylation Using Nanopore Sequencing. *Nature Methods*, 14(4), 407–410. Retrieved from http://dx.doi.org/10.1038/nmeth.4184

Smith, T. & Waterman, M. (1981). Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, 147(1), 195–197. Retrieved from http://dx.doi.org/10.1016/0022-2836(81)90087-5

Sneath, P. H. A. & Sokal, R. R. (1973). *Numerical Taxonomy. The Principles and Practice of Numerical Classification.* San Francisco: W.H. Freeman.

Söding, J. (2005). Protein Homology Detection by HMM-HMM Comparison. *Bioinformatics*, 21(7), 951–960.

Sonnhammer, E. L. L., Eddy, S. R. & Durbin, R. (1997). Pfam: A Comprehensive Database of Protein Domain Families Based on Seed Alignments. *Proteins: Structure, Function and Genetics*, 28(3), 405–420.

Stamatakis, A. (2014). RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. *Bioinformatics*, 30(9), 1312–1313.

Stamatakis, A., Hoover, P. & Rougemont, J. (2008). A Rapid Bootstrap Algorithm for the RAxML Web Servers. *Systematic Biology*, 57(5), 758–771.

Stamatakis, A., Ludwig, T. & Meier, H. (2005). RAxML-III: A Fast Program for Maximum Likelihood-Based Inference of Large Phylogenetic Trees. *Bioinformatics*, 21(4), 456–463.

Stenson, P. D., Mort, M., Ball, E. V., Howells, K., Phillips, A. D., Cooper, D. N., et al. (2009). The Human Gene Mutation Database: 2008 Update. *Genome Medicine*, 1(1), 1–6.

Stoler, D. L., Chen, N., Basik, M., Kahlenberg, M. S., Rodriguez-Bigas, M. A., Petrelli, N. J., et al. (1999). The Onset and Extent of Genomic Instability in Sporadic Colorectal Tumor Progression. *Proceedings of the National Academy of Sciences of the United States of America*, 96(26), 15121–6. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=10611348%5Cnhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC24783/pdf/pq015121.pdf%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/10611348%5Cnhttp://www.pubmedcentral.nih.gov/articl

Sun, T., Li, C., Han, L., Jiang, H., Xie, Y., Zhang, B., et al. (2015). Functional Knockout of FUT8 in Chinese Hamster Ovary Cells Using CRISPR/Cas9 to Produce a Defucosylated Antibody. *Engineering in Life Sciences*, 15(6), 660–666.

Swiech, K., Picanço-Castro, V. & Covas, D. T. (2012). Human Cells: New Platform for Recombinant Therapeutic Protein Production. *Protein Expression and Purification*, 84(1), 147–153. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S1046592812001313

Szöllosi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E. & Daubin, V. (2013). Efficient Exploration of the Space of Reconciled Gene Trees. *Systematic Biology*, 62(6), 901–912.

Tai, C. H., Bai, H., Taylor, T. J. & Lee, B. (2014). Assessment of Template-Free Modeling in CASP10 and ROLL. *Proteins: Structure, Function and Bioinformatics*, 82(SUPPL.2), 57–83.

Tang, H. & Thomas, P. D. (2016). PANTHER-PSEP: Predicting Disease-Causing Genetic Variants Using Position-Specific Evolutionary Preservation. *Bioinformatics*, 32(14), 2230–2232.

Tannock, I. & Guttman, P. (1981). Response of Chinese Hamster Ovary Cells to Anticancer Drugs under Aerobic and Hypoxic Conditions. *British Journal of Cancer*, 43(2), 245–248.

The Gene Ontology Consortium. (2013). Gene Ontology Annotations and Resources. *Nucleic Acids Research*, 41(November 2012), D530-535. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531070&tool=pmcentrez&rendertype=abstract

The UniProt Consortium. (2014). UniProt: A Hub for Protein Information. *Nucleic Acids Research*, 43(Database issue), D204-12. Retrieved from http://nar.oxfordjournals.org/content/43/D1/D204%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4384041&tool=pmcentrez&rendertype=abstract

Thomas, P. D. (2010). GIGA: A Simple, Efficient Algorithm for Gene Tree Inference in the Genomic Age. *BMC bioinformatics*, 11(1), 312. Retrieved from http://www.biomedcentral.com/1471-2105/11/312%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/20534164

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., et al. (2003). PANTHER : A Library of Protein Families and Subfamilies Indexed by Function PANTHER : A Library of Protein Families and Subfamilies Indexed by Function. , 2129–2141.

Thomas, P. D. & Kejariwal, A. (2004). Coding Single-Nucleotide Polymorphisms Associated with Complex vs. Mendelian Disease: Evolutionary Evidence for Differences in Molecular Effects. *Proceedings of the National Academy of Sciences*, 101(43), 15398–15403. Retrieved from http://www.pnas.org/cgi/doi/10.1073/pnas.0404380101

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. *Nucleic acids research*, 22(22), 4673–80. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=308517&tool=pmcentrez&rendertype=abstract

Thompson, J. D., Linard, B., Lecompte, O. & Poch, O. (2011). A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. *PloS one*, 6(3), e18093. Retrieved February 13, 2013, from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3069049&tool=pmcentrez&rendertype=abstract

Thompson, J. D., Plewniak, F. & Poch, O. (1999). BAliBASE : A Benchmark Alignment

Database for the Evaluation of Multiple Alignment Programs. , 15(1), 87–88. Retrieved from http://www-igbmc.u-strasbg.fr/BioInfo/ BAliBASE/index.html

Tonini, J., Moore, A., Stern, D., Shcheglovitova, M. & Ort??, G. (2015). Concatenation and Species Tree Methods Exhibit Statistically Indistinguishable Accuracy under a Range of Simulated Conditions. *PLoS Currents*, 7(TREEOFLIFE), 1–14.

Turcatti, G., Romieu, A., Fedurco, M. & Tairi, A.-P. (2008). A New Class of Cleavable Fluorescent Nucleotides: Synthesis and Optimization as Reversible Terminators for DNA Sequencing by Synthesis. *Nucleic Acids Research*, 36, e25.

Umaña, P., Jean-Mairet, J., Moudry, R., Amstutz, H. & Bailey, J. E. (1999). Engineered Glycoforms of an Antineuroblastoma IgG1 with Optimized Antibody-Dependent Cellular Cytotoxic Activity. *Nature biotechnology*, 17(2), 176–180.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The Sequence of the Human Genome. *Science (New York, N.Y.)*, 291(5507), 1304–51. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11181995

Vishwanathan, N., Bandyopadhyay, A., Fu, H. Y., Johnson, K. C., Springer, N. M. & Hu, W. S. (2017). A Comparative Genomic Hybridization Approach to Study Gene Copy Number Variations among Chinese Hamster Cell Lines. *Biotechnology and Bioengineering*, 114(8), 1903–1908.

Vishwanathan, N., Le, H., Jacob, N. M., Tsao, Y.-S., Ng, S.-W., Loo, B., et al. (2014). Transcriptome Dynamics of Transgene Amplification in Chinese Hamster Ovary Cells. *Biotechnology and bioengineering*, 111(3), 518–28. Retrieved March 20, 2014, from http://www.ncbi.nlm.nih.gov/pubmed/24108600

Walsh, G. (2014). Biopharmaceutical Benchmarks 2014. *Nature biotechnology*, 32(7), 992–1000.

Warnow, T. (2015). Concatenation Analyses in the Presence of Incomplete Lineage Sorting. *PLoS Currents*, 1–10. Retrieved from http://currents.plos.org/treeoflife/?p=7414

Weinstein, J. N. (2012). Drug Discovery: Cell Lines Battle Cancer. *Nature*, 483(7391), 544–545.

Weitzel, J. N., Blazer, K. R. & Macdonald, D. J. (2011). Genetics , Genomics , and Cancer Risk Assessment State of the Art and Future Directions in the Era of Personalized Medicine. *Ca Cancer J Clin*, 00(0), 1–33.

van Wijk, X. M., Döhrmann, S., Hallström, B. M., Li, S., Voldborg, B. G., Meng, B. X., et al.

(2017). Whole-Genome Sequencing of Invasion-Resistant Cells Identifies Laminin A2 as a Host Factor for Bacterial Invasion (P. Park and G. B. Pier, Eds.). *mBio*, 8(1), e02128-16. Retrieved from http://mbio.asm.org/lookup/doi/10.1128/mBio.02128-16

Wlaschin, K. F. & Yap, M. G. S. (2007). Recombinant Protein Therapeutics from CHO Cells — 20 Years and Counting. *Chemical Engineering*, 40–47.

Wong, K. C. & Zhang, Z. (2014). SNPdryad: Predicting Deleterious Non-Synonymous Human SNPs Using Only Orthologous Protein Sequences. *Bioinformatics*, 30(8), 1112–1119.

Wu, J. & Jiang, R. (2013). Prediction of Deleterious Nonsynonymous Single-Nucleotide Polymorphism for Human Diseases. *The Scientific World Journal*, 2013.

Wu, Y.-C., Rasmussen, M. D., Bansal, M. S. & Kellis, M. (2013). TreeFix: Statistically Informed Gene Tree Error Correction Using Species Trees. *Systematic Biology*, 62(1), 110–120.

Xu, D. & Zhang, Y. (2012). Ab Initio Protein Structure Assembly Using Continuous Structure Fragments and Optimized Knowledge-Based Force Field. *Proteins: Structure, Function and Bioinformatics*, 80(7), 1715–1735.

Xu, D. & Zhang, Y. (2013). Toward Optimal Fragment Generations for Ab Initio Protein Structure Assembly. *Proteins: Structure, Function and Bioinformatics*, 81(2), 229–239.

Xu, X., Nagarajan, H., Lewis, N. E., Pan, S., Cai, Z., Liu, X., et al. (2011). The Genomic Sequence of the Chinese Hamster Ovary (CHO)-K1 Cell Line. *Nature biotechnology*, 29(8), 735–41. Retrieved January 28, 2013, from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3164356&tool=pmcentrez&rendertype=abstract

Yamane-Ohnuki, N., Kinoshita, S., Inoue-Urakubo, M., Kusunoki, M., Iida, S., Nakano, R., et al. (2004). Establishment of FUT8 Knockout Chinese Hamster Ovary Cells: An Ideal Host Cell Line for Producing Completely Defucosylated Antibodies with Enhanced Antibody-Dependent Cellular Cytotoxicity. *Biotechnology and bioengineering*, 87(5), 614–22. Retrieved July 10, 2014, from http://www.ncbi.nlm.nih.gov/pubmed/15352059

Yang, Z. & Rannala, B. (1997). Bayesian Phylogenetic Inference Using DNA Sequences: A Markov Chain Monte Carlo Method. *Molecular Biology and Evolution*, 14(7), 717–724. Retrieved from http://mbe.oxfordjournals.org/cgi/content/abstract/14/7/717

Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., et al. (2015). Ensembl 2016. *Nucleic Acids Research*, 44(December 2015), gkv1157. Retrieved from http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv1157

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S. & Madden, T. L. (2012). Primer-BLAST: A Tool to Design Target-Specific Primers for Polymerase Chain Reaction. *BMC Bioinformatics*, 13(1), 134.

Zeng, S., Yang, J., Chung, B. H.-Y., Lau, Y. L. & Yang, W. (2014). EFIN: Predicting the Functional Impact of Nonsynonymous Single Nucleotide Polymorphisms in Human Genome. *BMC Genomics*, 15(1), 455. Retrieved from http://www.biomedcentral.com/1471-2164/15/455