UNIVERSITY OF LEEDS

The role of protein dynamics in the aggregation of biopharmaceuticals

Rhys David Thomas

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy
The University of Leeds
Astbury Centre for Structural Molecular Biology

July 2016

Acknowledgements

I am greatly indebted to my supervisors Dr. David Brockwell and Prof. Sheena Radford for their guidance and encouragement throughout my PhD as well as making me feel welcome in the lab from the start. I also wish to thank my industrial supervisor Dr. Andrew Buchanan for his encouragement and useful discussions as well as establishing the beginnings of this project.

At some point I have asked/nagged every member of the Brockwell and Radford labs for help and cannot remember ever being told to go away! Therefore, I want to thank all members of the lab for their help and providing a great atmosphere to work in. In particular, I would like to thank Dr. Theodoros Karamanos for help with both experiments and theory of NMR and data processing, plus proof-reading much of this thesis. I must also thank Nasir Khan for providing everything we need and bending over backwards to help the lab and keeping me fed! I also greatly appreciate Dr. Janet Saunders, Dr. Claire Sarrell and Dr. Kevin Tipping for making me welcome when I first arrived and helping me settle in. Also Kevin let me win at squash every time we played which was really decent of him.

Dr. Arnout Kalverda and Dr. Gary Thompson deserve thanks for their tutelage in both practical and theoretical aspects of NMR and providing a great NMR facility to work in.

Thanks are also due to Dr. James Ault for carrying out all of the mass spectrometry mentioned in this thesis and helping me interpret the results.

I also wish to thank Dr. Katie Stewart for carrying out the EM mentioned in this thesis.

I would like to thank MedImmune for funding this project.

Finally I want to thank my parents for supporting me in everything I have done.

Table of Contents

List of Figures

List of Tables

Abbreviations

α-syn	α-synuclein
β2m	β2 microglobulin
ΔG^{o}_{UN}	Free energy of unfolding under standard conditions
ε_{280}	Extinction coefficient at 280 nm
τ_c	Rotational correlation time
A_{280}	Absorbance at 280 nm
Aβ	Amyloid beta
Aβ40	Amyloid beta 40 residue peptide
Aβ42	Amyloid beta 42 residue peptide
ADA	Antidrug antibody
AF4	Asymmetric field flow fractionation
ANS	8-anilino-1-napthalenesulphonic acid
APC	Antigen-presenting cells
APS	Ammonium persulphate
AUC	Analytical ultracentrifugation
BSA	Bovine serum albumin
CD	Circular dichroism
CDR	Complementarity determining regions
CHO	Chinese hamster ovary
CIS	Cytokine-inducible SH2-domain-containing
CKD	Chronic kidney disease
CNTF	Ciliary neurotrophic factor
CPMG	Carr-Purcell-Meiboom-Gill
CRH	Cytokine receptor homologous
CSA	Chemical shift anisotropy
CSP	Chemical shift perturbation
CTFR	Cystic fibrosis transmembrane conductance regulator
CTLA4	Cytotoxic T-lymphocyte associated protein 4
CV	Column volume
dAb	Domain antibody
DLS	Dynamic light scattering
DNA	Deoxyribonucleic acid
DPPD	Recombinant purified protein derivative
DTT	Dithiothreitol

E. coli	*Escherichia coli*
EC_{50}	Half maximal effective concentration
EDTA	Ethylenediaminetetraacetic acid
EM	Electron Microscopy
EPO	Erythropoietin
ESI-MS	Electrospray ionisation mass spectrometry
EtOH	Ethanol
Fab	Antigen-binding fragment
FAP	Familial amyloidotic polyneuropathy
Fc	Crystallisable fragment
FDA	Food and Drug Administration
FID	Free induction decay
Flt3	Fms-like tyrosine kinase 3 ligand
G-CSF	Granulocyte-colony stimulating factor
G-CSFR	Granulocyte-colony stimulating factor receptor
Gammabody	Grafted amyloid-motif antibody
GFP	Green fluorescent protein
GH	Growth hormone
GM-CSF	Granulocyte macrophage-colony stimulating factor
GPCR	G-protein coupled receptor
GdnHCl	Guanidine hydrochloride
HER2	Human epidermal growth factor receptor 2
HIC	Hydrophobic interaction chromatography
HIV	Human immunodeficiency virus
HMQC	Heteronuclear multiple quantum coherence
HSQC	Heteronuclear single quantum coherence
IAPP	Islet amyloid precursor protein, or amylin
IgG	Immunoglobulin G
IFN	Interferon
IL	Interleukin
Im7	Immunity protein 7
INEPT	Insensitive nuclei enhancement by polarisation transfer
INN	International Nonproprietary Name
IPTG	Isopropyl β-D-1-thiogalactopyranoside
IR	Infrared
Jak	Janus kinase
LB	Luria-Bertani

LIF	Leukaemia inhibitory factor
M-CSF	Macrophage-colony stimulating factor
mAb	Monoclonal antibody
MCS	Multiple Cloning Site
MHC	Major histocompatibility complex
mRNA	Messenger RNA
M_{UN}	Dependence of ΔG_{UN} on denaturant concentration
MWCO	Molecular weight cut off
NAC	Nascent-chain-associated complex
NEB	New England Biolabs
NMR	Nuclear Magnetic Resonance
nOe	Nuclear Overhauser effect
ns	nanosecond
NTA	Nanoparticle tracking analysis
OD_{600}	Optical density at 600 nm
o/n	overnight
OSM	Oncostatin M
PAGE	Polyacrylamide gel electrophoresis
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PEG	Polyethylene glycol
PF	Protection factor
PIAS	Protein inhibitor of activated Stats
PIDS	Polarisation intensity differential scattering
PL	Prolactin
PMSF	Phenylmethanesulphonyl fluoride
PRCA	Pure red cell aplasia
R_1	Longitudinal relaxation rate
R_2	Transverse relaxation rate
RAC	Ribosome-associated complex
R_h	Hydrodynamic radius
Rh	Rhesus
Rh(D)	Rhesus D antigen
RMM	Resonant mass measurements
RNA	Ribonucleic acid
rpm	Revolutions per minute
RT-PCR	Reverse transcription PCR

S. cerevisiae	*Saccharomyces cerevisiae*
SAP	Spatial aggregation propensity
SASA	Solvent accessible surface area
SCF	Stem cell factor
scFv	Single-chain variable fragment
scTCR	Single-chain T-cell receptor
SDS	Sodium dodecyl sulphate
SEC	Size exclusion chromatography
SHP	SH2-domain-containing tyrosine phosphatase
SMR	Suspended microchannel resonator
SOCS	Suppressor of cytokine signalling
SOFAST	Band-selective optimised-flip-angle short-transient
srlf	Single-residue local frustration
STAT	Signal transducers and activators of transcription
TAE	Tris-acetate-EDTA
TCA	Trichloroacetic acid
TCEP	Tris(2-carboxyethyl)phosphine hydrochloride
TDA	Taylor dispersion analysis
TEMED	Tetramethylethylenediamine
TF	Trigger factor
Tris	Tris(hydroxymethyl)aminomethane
tRNA	Transfer RNA
TTR	Transthyretin
U	Units
UPS	Ubiquitin proteasome system
UV	Ultraviolet
v/v	volume:volume ratio
V_H	Variable domain from antibody heavy chain
V_L	Variable domain from antibody light chain
WHO	World Health Organisation
w/v	weight:volume ratio

Abstract

Proteins have the potential to be potent biopharmaceuticals with few side effects because of their endogenous specificity. However, the development of therapeutic proteins is frequently hindered by aggregation, a phenomenon that can occur at every phase of the manufacturing process. Aggregates can elicit immunogenic responses in patients, which along with reducing efficacy have the potential to cause life-threatening side effects. Lack of knowledge concerning the mechanism or nature of the aggregates underlying the immune response to therapeutic proteins means that any aggregates in protein drug formulations are a major concern. As a result, all aggregates must be removed during manufacturing or protein drugs reformulated to prevent aggregation, which are difficult, expensive and time consuming. Therefore, preventing therapeutic protein aggregation is paramount and a better knowledge of the conformational states involved will enable development of rational strategies to inhibit aggregation.

In this study, the folding, dynamics and aggregation of a variant of the biopharmaceutical Granulocyte-Colony Stimulating Factor (G-CSF) were characterised and a subtle conformational change of the native state was elucidated in aggregation-prone conditions. Biophysical comparison of the variant C3 relative to wild-type G-CSF suggested they possessed similar structures. In addition, characterisation of the dependence of G-CSF C3 aggregation on pH, protein concentration and the presence of urea concentrations that were too low to cause global denaturation revealed the aggregation behaviour of G-CSF C3 to be similar to wild-type G-CSF. A change in the environment of the side chain of Tyr86 on helix B of G-CSF C3 was identified, along with increased dynamics of the latter half of the AB loop that is proximal spatially to Tyr86. Tyr86 is predicted to be intrinsically aggregation-prone and is buried in the native conformation, but is hypothesised to become exposed by increased mobility of the AB loop, driving aggregation from a highly native-like state.

Chapter 1 Introduction

1.1 Protein folding

1.1.1 From pathways to funnels and energy landscapes

The information required for a protein to fold into its functional conformation is encoded entirely within its amino acid sequence, as originally demonstrated by Anfinsen *et al*.[1,2]. For proteins that adopt a three-dimensional structure in order to function, this native conformation represents the lowest free energy (and therefore most stable) structure[1,2]. However, given the vast array of different conformations that can be adopted by a polypeptide in solution, folding to the native state on a biologically relevant timescale cannot involve unbiased exploration of all possible conformations. Levinthal proposed that this paradox could be solved by proteins folding along specific pathways, driven by local interactions that progressively lead to forming the native conformation[3].

The first pathway proposed to describe general protein folding was the nucleation-growth model, which involved formation of tertiary structure from a nucleus of secondary structure[4]. From this the framework model emerged to better account for the observation of intermediate species, which proposed packing together of secondary structure elements into tertiary structures[5,6] (Figure 1.1). Analogous to this is the diffusion-collision model that suggested diffusion of secondary structure elements and their collision to form tertiary structure[7]. Whereas these models specified a hierarchical progression from primary to tertiary structures, the hydrophobic collapse model suggested initial collapse of the polypeptide chain due to burial of hydrophobic regions, sequestering them from the surrounding solvent and guiding the search for the native topology, which was followed by secondary structure formation[8,9] (Figure 1.1).

Arguably the most successful model for describing the folding of small single domain proteins, the nucleation-condensation model emerged as an amalgamation of features from the above models[10]. In this model, the polypeptide chain collapsed into a nucleus due to burial of hydrophobic surface, within which local secondary

structure was stabilised by long range tertiary interactions. This nucleus represented the transition state in the folding pathway, from which the remainder of secondary and tertiary structures formed[11-14] (Figure 1.1).

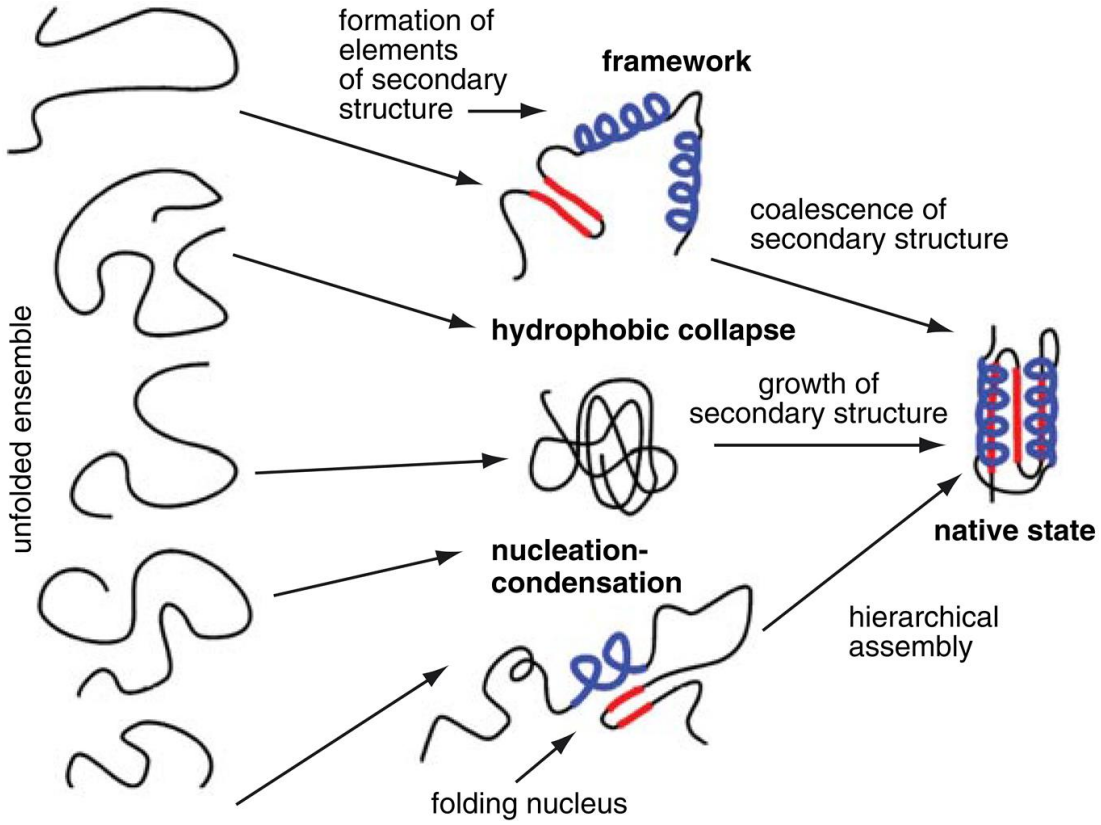


Figure 1.1 Illustrations of some of the early pathways proposed for protein folding. Schematics for the framework[5,6], hydrophobic collapse[8,9] and nucleation-condensation[11-14] model pathways are shown. The entropy of the polypeptide decreases as the structural complexity increases from left to right. This figure was reproduced from Morris and Searle (2012)[15].

Pathways such as these considered sequential transitions from denatured to native states through defined structures, constituting what is known as the classical view of protein folding[6]. In the 'new' view, protein folding is described by funnel-shaped energy landscapes, in which an ensemble of structures are populated and parallel folding trajectories are simultaneously explored[16,17]. An energy landscape is the internal free energy of each protein conformation as a function of its degrees of freedom i.e. the conformational entropy (Figure 1.2). Similar conformations are geometrically close to each other on the energy landscape. These conformations may be higher in energy than structurally similar conformations, meaning they appear as peaks in the landscape, whereas lower energy conformations appear as troughs[16].

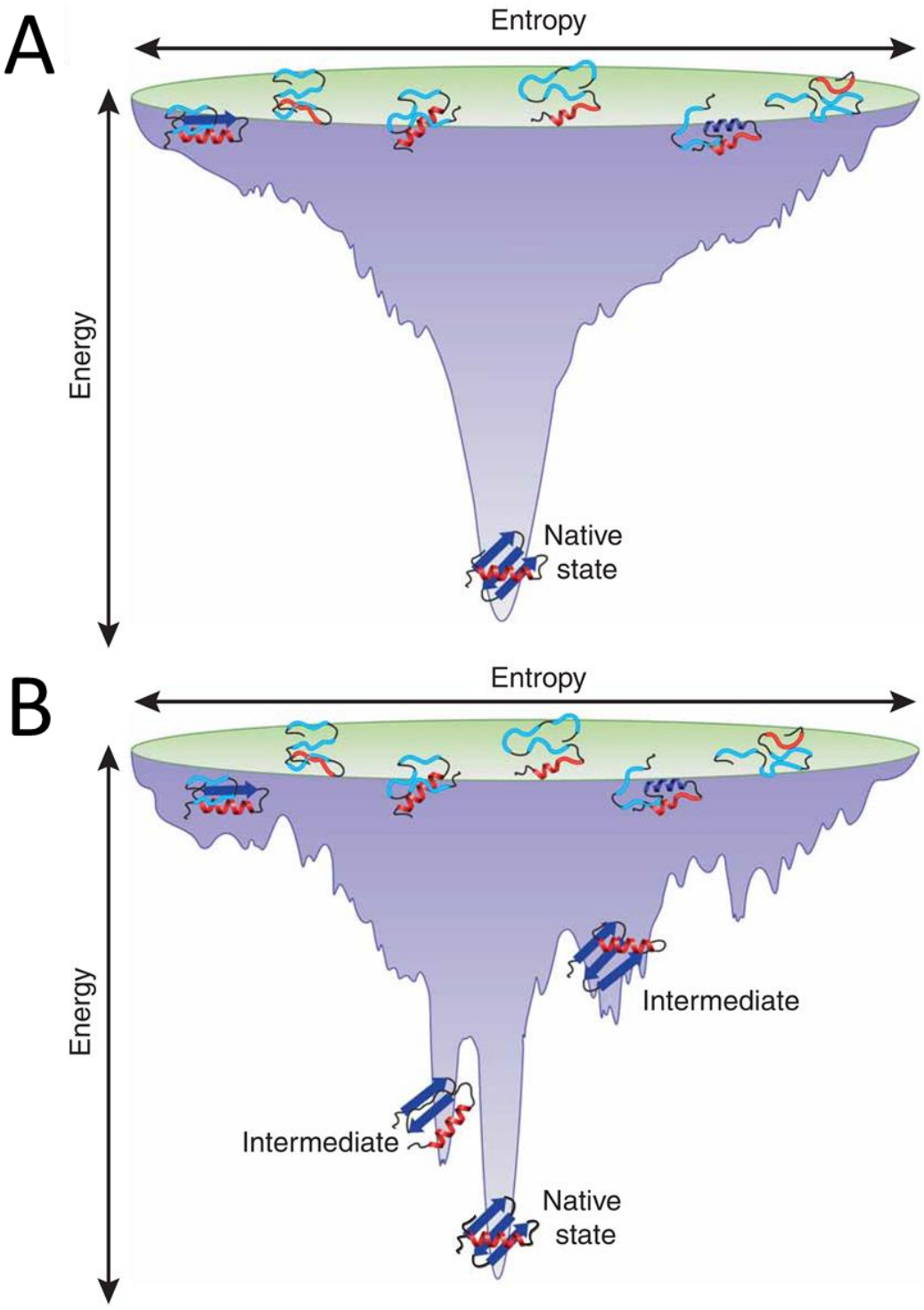


Figure 1.2 Schematic representations of different protein folding energy landscapes. The denatured state at the top of the funnel encompasses a broad ensemble of high energy structures containing both native and non-native interactions, whereas the native state at the bottom of the funnel encompasses a narrower range of lower energy structures. Native interactions are energetically favoured over non-native interactions, resulting in the protein being funnelled towards the native conformation. A smooth energy landscape is shown in A that represents two-state folding. The energy landscape shown in B is more rugged due to population of intermediates. This figure was reproduced from Bartlett and Radford (2008)[18].

The principle of minimal frustration states that the internal free energy of the protein decreases as it adopts conformations that are more native-like. This reduction in internal free energy is greater than can be accounted for by chance alone and allows the polypeptide to traverse the energy landscape and reach the native conformation, overcoming kinetic traps such as intermediates and the reduced conformational entropy of the polypeptide[19]. The reduction in internal free energy arises from the formation of native contacts e.g. hydrogen bonds, electrostatic and hydrophobic interactions, which are more stabilising than non-native contacts[17]. The final step of folding involves packing of side chains into their unique three-dimensional structure and exclusion of water from the protein core[20].

The importance of native interactions in energy landscape theory means that native state topology plays a central role in defining the folding events that occur[21]. Using models of the folding landscape where energetic frustration has been removed, structural features of the transition states of proteins have been predicted from their native topologies[22,23]. Such models have also been used to correctly predict folding rates[24] and the existence of intermediates[25] that had previously been determined experimentally. The importance of native state topology in defining the folding trajectory also helps rationalise the higher folding rates observed for proteins with higher degrees of native state complexity, measured by the contact order[26], as well as the conservation of structural features in the folding mechanisms of protein structural families, despite these proteins often exhibiting little sequence similarity[27].

Although the native state of proteins represents the minimally frustrated conformation[19], a degree of local frustration may persist, arising from a particular arrangement of amino acids that is energetically unfavourable but may be required for function e.g. in an enzyme active site[28]. As a result, proteins are only marginally stable[29,30], having evolved to harness the roughness of their energy landscape to facilitate dynamic motions between an ensemble of native-like conformations[31,32].

1.1.2 The role of intermediates in protein folding

The idea of a two-state transition between denatured and native states during folding of a protein only really applies to small globular proteins[10,33] or may not be applicable at all[34,35]. For proteins greater than 100 amino acids in length, which constitutes over 95 % of proteins in the predicted human proteome[36], folding generally occurs via one or more intermediate states[34,37] (Figure 1.2B). Ever since the first discovery of on-pathway[38] and off-pathway[39] intermediates in folding reactions it has become apparent that they can represent partially-folded stepping stones to the native conformation or misfolded kinetic traps that reduce the efficiency of folding[40]. The energy levels of folding intermediates and their positions along a folding trajectory relative to the unfolded, transition and folded states determines their ability to be detected in equilibrium and kinetic folding experiments[34,35] (Figure 1.3).

Partially folded and misfolded intermediates typically expose hydrophobic residues and regions of unstructured polypeptide backbone to the solvent that would otherwise be buried in the native conformation[37], which can promote intermolecular interactions and lead to protein aggregation. Like protein folding, aggregation is primarily driven by hydrophobic interactions, but in a concentration-dependent manner[41]. The crowded cellular environment, with protein concentrations in the cytosol in the region of 300-400 g l^{-1} [42], thus greatly increases the likelihood of aggregation, which competes with folding to determine the fate of nascent proteins.

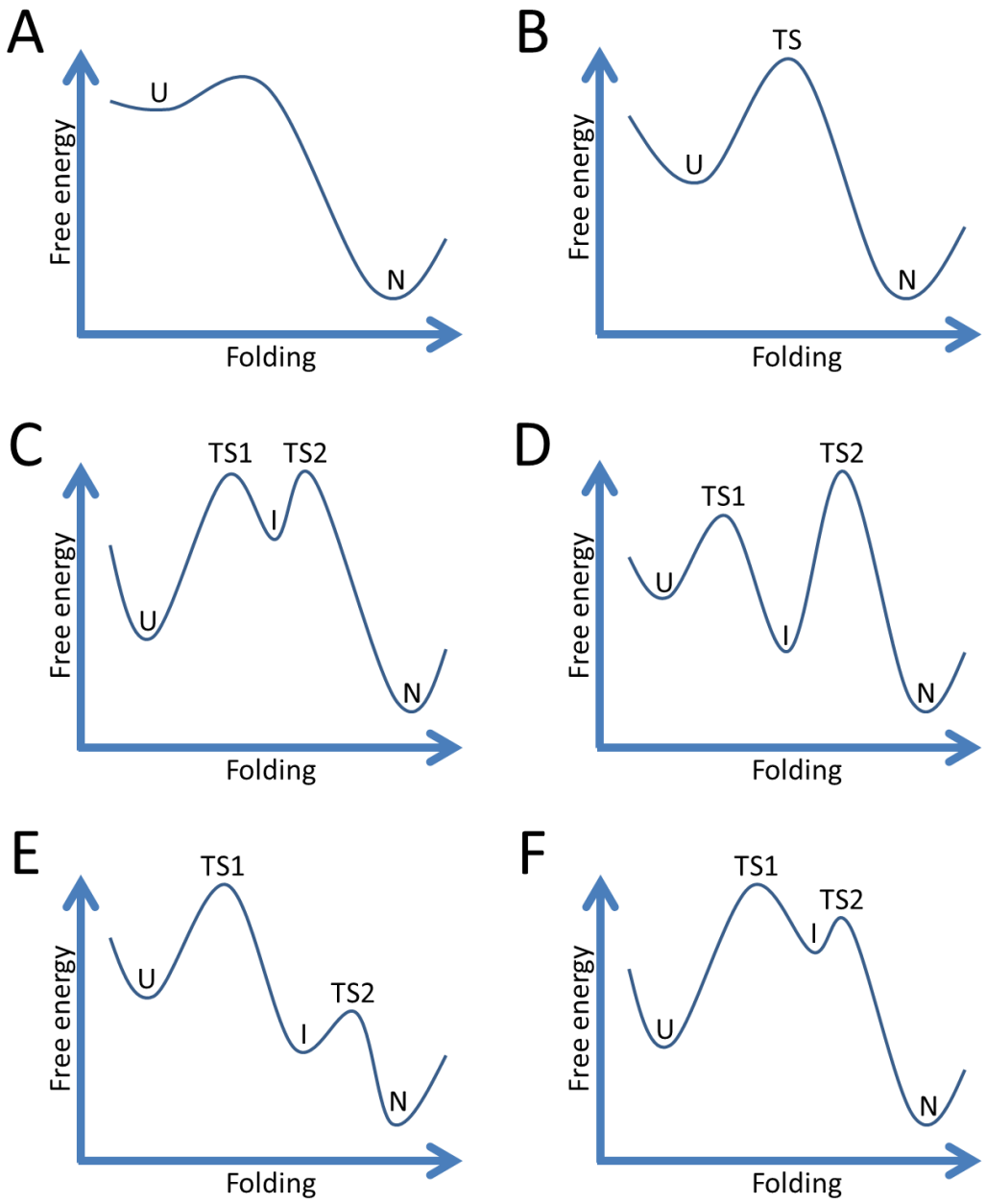


Figure 1.3 Free energy diagrams for different folding intermediates. In the absence of energetic frustration folding from the unfolded (U) to the native state (N) proceeds in a downhill manner[43] (A). A transition state (TS) barrier between the U and N states without population of intermediates (I) results in two-state folding (B). An intermediate that is high in energy relative to both the U and N states can be detected by measuring folding and unfolding kinetics but not by equilibrium folding studies (C). However, an I state that is higher in energy than the N state but lower in energy than the U state and occurs before the rate limiting TS is observable during measurements of folding kinetics but not unfolding kinetics or equilibrium folding measurements (D). Conversely, if the same I state occurs after the rate limiting TS than it is observable during measurements of unfolding kinetics but not folding kinetics[34] and can also be elucidated from equilibrium and kinetic native state hydrogen/deuterium exchange[35,44] or NMR relaxation-dispersion experiments[45] (E). An I state that is higher in energy than the U and N states but occurs after the rate limiting TS is not observable in kinetic or thermodynamic studies or equilibrium native state hydrogen/deuterium exchange experiments but can be probed using kinetic hydrogen exchange[44] or NMR relaxation-dispersion experiments[45]. This figure was redrawn from Brockwell and Radford (2007)[34].

1.1.3 Protein folding *in vivo*

In the cell, protein folding is aided by molecular chaperones, which are defined as proteins that interact with, stabilise or aid another protein in reaching its functional conformation without being present in the final structure[46]. Generally, molecular chaperones aid folding through the binding and release of exposed hydrophobic surfaces of non-native states, whereby binding of exposed hydrophobic surfaces blocks aggregation and release allows folding to occur. These cycles of binding and release allow kinetic partitioning of non-native states from folding[36]. There is also a contribution from interactions of the nascent polypeptide chain with the surface of the ribosome that helps prevent misfolding and aggregation[47,48]. In addition, reorganisation of non-native disulphide bonds and *cis-trans* isomerisation of peptide bonds involving Pro are catalysed by protein disulphide isomerases and peptidyl prolyl isomerases, respectively to aid folding of non-native states to the native conformation[49].

The action of molecular chaperones during co- and post-translational folding is highly organised[36], beginning with interactions between the polypeptide chain and ribosome-binding chaperones during translation by the ribosome[50,51], such as trigger factor (TF) in bacteria and the ribosome-associated and nascent-chain-associated complexes in eukaryotes (RAC and NAC, respectively)[51]. After release from the ribosome the polypeptide is then bound by chaperones such as the Hsp70 system[52]. If folding is not accomplished by Hsp70, proteins may be passed to chaperonins such as GroEL/ES in bacteria[53] or TriC in eukaryotes[54], which enclose a single molecule of non-native protein in their central cavity to prevent aggregation and facilitate folding. Misfolded states that cannot be recovered by the molecular chaperone network are targeted for degradation by the ubiquitin proteasome system (UPS) in eukaryotes[55].

Despite the extensive molecular chaperone network, many proteins still aggregate in the cell. Disassembly of aggregated proteins is carried out by Hsp70 in combination with Hsp110 and Hsp40 in eukaryotes[56,57], while in prokaryotes Hsp70 co-ordinates with AAA+ ATPase complexes of the Hsp100 family[58]. After disassembly, misfolded non-native states may be targeted for degradation by the UPS, while aggregates that cannot be disassembled may be cleared by lysosomal degradation or selective autophagy[55].

1.2 Protein aggregation

1.2.1 Structural features of protein aggregation

Although folding of a protein into its native conformation reduces its propensity to aggregate, owing to the burial of hydrophobic surfaces and polypeptide backbone away from solvent[37], the native conformations of proteins are only marginally stable[29,30]. Consequently, single point mutations or changes in environmental conditions e.g. pH, temperature or protein concentration can increase the population of partially unfolded intermediate states that may be aggregation-prone[59]. Partially unfolded states are more likely to aggregate than the globally unfolded state because of the proximity of hydrophobic residues that creates hydrophobic surfaces[60]. Conversely, natively unfolded proteins have a greater overall charge and reduced hydrophobicity, which reduces the likelihood of aggregation along with promoting disorder[61]. The dependency of aggregation on the charge, hydrophobicity and secondary structure propensities of an amino acid sequence have resulted in the development of many programs designed to predict the aggregation propensity of a protein sequence[62-66].

Aggregation can result in formation of insoluble aggregates that are amorphous or amyloid in nature (Figure 1.4). In contrast to amorphous aggregates, amyloid fibrils are highly ordered and incredibly stable structures with a cross β-sheet structure at their core that is perpendicular to the axis of the fibril[67]. Typically, formation of amorphous or amyloid aggregates results in increased formation of intermolecular β-sheet[68]. The ordered nature of amyloid aggregates reflects their assembly via a nucleation-growth mechanism, where fibril elongation proceeds by addition of monomer to an aggregation nucleus. This aggregation nucleus is unstable and so represents the rate-limiting step in amyloid fibril formation[59].

Amino acid sequences of native proteins appear to exhibit patterns that instil resistance to amyloid aggregation, such as avoiding long stretches of consecutive hydrophobic residues and patterns of alternating polar and non-polar residues[69]. Other intrinsic features of polypeptide sequences that appear to inhibit amyloid formation include conserved positioning of ‘gatekeeper’ Pro[70] and Gly[71] residues, which have a low propensity for β-sheet formation, as well as the inhibiting effect of individual charged residues being placed on the edges of aggregation-prone

segments[72,73]. However, polypeptides exhibiting a range of different physicochemical properties have been shown to form amyloid[74,75], which coupled to the high stability of this aggregate has led to the notion of amyloid representing a free-energy minimum accessible to all polypeptide chains[73].

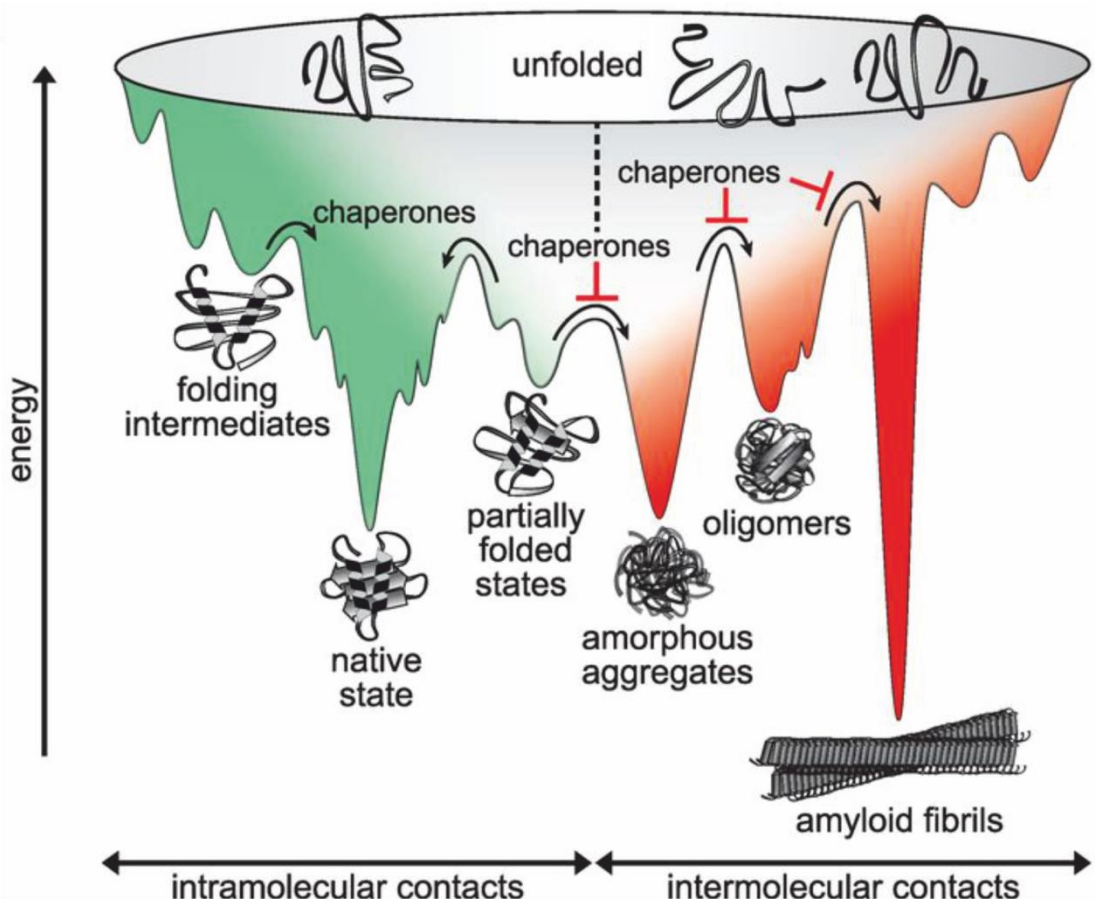


Figure 1.4 Schematic representation of protein folding and aggregation energy landscapes. Partially-folded states can possess exposed hydrophobic residues and unstructured regions of polypeptide backbone that cause aggregation. Chaperones aid the folding of partially folded and misfolded states to the native state through binding and release of exposed hydrophobic regions. Binding sequesters exposed hydrophobic residues thus inhibiting intermolecular contacts while release allows partially folded states to form intramolecular contacts. Despite the action of chaperones, proteins may still aggregate *in vivo*, resulting in formation of amorphous aggregates or highly ordered amyloid fibrils. This figure was reproduced from Balchin, Hartl and Hartl (2016)[36].

1.2.2 Protein aggregation and disease

Despite the extensive proteostasis network that exists to regulate folding[36] (see Section 1.1.3), many human pathologies are associated with protein misfolding and aggregation[74]. Some of these diseases are associated with misfolded proteins that have escaped the quality control network and exhibit reduced or abolished function,

as observed for the cystic fibrosis transmembrane conductance regulator (CTFR) in cystic fibrosis[76] and p53 in a range of cancers[77]. Such diseases tend to be hereditary as mutations affect the energy landscape of a protein and so can increase the population of misfolded states[37]. More frequently, diseases resulting from misfolded species are due to the formation of intractable aggregates[41] (Figure 1.4).

Many of the diseases caused by aggregation are characterised by deposition of intra- or extracellular deposits of amyloid fibrils, which build up to form amyloid plaques and so are called amyloidoses[78]. Many neurodegenerative diseases are amyloidoses, including Alzheimer's disease, Parkinson's disease and Huntington's disease. However, other examples of amyloidoses include type II diabetes, haemodialysis related amyloidosis and cataract formation[41,74,78]. Although amyloidoses are characterised by formation of insoluble deposits of amyloid fibrils, cellular toxicity in these diseases is believed to be caused by soluble oligomers[41,79,80], which are thought to promote amyloid formation by exposure of hydrophobic regions and unpaired β-strands[80]. The mechanisms by which cytotoxicity is exerted by soluble oligomeric species are uncertain due to the difficulties in their characterisation, which results from their instability and structural heterogeneity[79,80]. The cytotoxicity of soluble oligomers has been shown to relate to their size and surface hydrophobicity[81] and is hypothesised to be caused by common structural features[82] interacting with cellular components and membranes[79].

Chaperones have been shown to inhibit various steps of aggregate formation[83] (Figure 1.4) but a decline in the proteostasis network with aging results in reduced effectiveness of these quality control systems[84]. Consequently, amyloidoses such as neurodegenerative diseases are becoming increasingly prevalent in aging populations[74].

Protein aggregation can occur from a myriad of states on the energy landscape. The unfolded 40 and 42 residue polypeptides of amyloid beta (denoted Aβ40 and Aβ42, respectively), α-synuclein (α-syn) and islet amyloid precursor protein (IAPP or amylin) are the causative agents of Alzheimer's disease, Parkinson's disease and type II diabetes, respectively[41,74]. Amyloid aggregation of β2 microglobulin (β2m) causes haemodialysis related amyloidosis[85] and has been shown to occur from an acid unfolded state[86,87] as well as a native-like intermediate state under physiological

conditions[88]. The F9S mutant of γS-crystallin has also been shown to form amyloid aggregates via partial unfolding of the native state[89,90], resulting in cataract formation in the eye lens[91].

Transthyretin (TTR) is a protein that forms amyloid aggregates as a result of subtle fluctuations in the native conformation[92,93]. Pathogenic variants of TTR (V30M and L55P) resulting in familial amyloidotic polyneuropathy (FAP)[94] have been shown to exhibit increased rates of hydrogen exchange of residues in β-strands at the core of the native tetramer, suggesting transient partial unfolding, while for a variant of TTR associated with suppression of FAP (T119M) the rate of hydrogen exchange for these residues was reduced[92]. The varied amounts of core β-strand exposure observed for wild-type and T119M TTR occurred despite the similar structures these variants obtained by crystallography[95]. NMR relaxation-dispersion studies of a monomeric form of native TTR (F87M/L110M) at physiological pH also indicated a subtle conformational change affecting certain β-strands, which was increasingly observed from changes in the contribution of chemical exchange to measured transverse relaxation rates in aggregation-prone conditions and largely abolished for the T119M suppressor variant[93]. Overall, these results point to increased dynamics of the native conformation of TTR as causing aggregation.

1.3 Biopharmaceuticals

1.3.1 The rise of biopharmaceuticals

A biopharmaceutical is a medicinal product that is biological in nature and manufactured using biotechnology[96]. Typically, this biotechnology refers to recombinant expression in bacterial cells i.e. *E. coli* or more commonly mammalian cells, which are primarily Chinese hamster ovary (CHO) cells[97]. Like all pharmaceutical substances, biopharmaceuticals are designated an International Nonproprietary Name (INN) by the World Health Organisation (WHO) as well as their trade name given by the manufacturing company[98]. The vast majority of biopharmaceuticals are parenteral products i.e. they must be administered by injection[99].

The vast majority of biopharmaceuticals currently in clinical use are recombinantly expressed proteins[97,100]. Proteins are attractive as therapeutic agents because of their inherent specificity, which theoretically means higher efficacy and fewer adverse effects compared to small molecule drugs. Additionally, proteins often regulate multiple components of a biological process e.g. a signalling pathway, which is difficult to mimic using small molecules. Moreover, many diseases relate to reduced or eradicated function of a protein due to a genetic mutation, meaning a protein drug can provide effective treatment through replacement[101]. The importance of biopharmaceuticals is reflected in the consistent generation of new products for clinical use (Table 1.1), targeting a range of disease areas from cancers to neurodegenerative and inflammatory diseases[102].

Time period	Number of product approvals
Up to 1989	9
1990-1994	19
1995-1999	58
2000-2004	56
2005-2009	54
2010-2014	54

Table 1.1 Total numbers of biopharmaceutical products approved for use in the EU and US over time periods from 1982-2014. This data was reproduced from Walsh (2014)[97].

Due to the wide range of applications for biopharmaceuticals and the myriad of disease areas targeted, these products can be more easily categorised according to broad functional classifications[101] as shown in Table 1.2. The mechanisms of action of biopharmaceuticals by which these functions can be exerted are also shown, along with examples of products that act via these mechanisms.

Functional classification	Mechanism of action	Example (Trade name)	Function	Clinical use
Enzymatic or regulatory activity	Replacing a protein that is deficient or abnormal	Insulin[103,104] (Humulin®)	Regulates blood glucose	Diabetes mellitus
	Augmenting an existing pathway	Erythropoietin[98] (Epogen®)	Stimulates erythropoiesis	Anaemia
	Providing a novel function or activity	Human deoxyribonuclease I[105] (Pulmozyme®)	Degrades DNA in pulmonary secretions	Cystic fibrosis (decreases respiratory tract infections)
Specialised targeting activity	Interfering with a molecule or organism	Trastuzumab[106] (Herceptin®)	Binds cell surface receptor HER2 to control cancer cell growth	Breast cancer
	Delivering other compounds or proteins	Denileukin diftitox[107] (Ontak®)	Delivers cytocidal action of diphtheria toxin to cells expressing IL-2 receptor	Persistent or recurrent cutaneous T-cell lymphoma
Vaccines	Protecting against a deleterious foreign agent	Hepatitis B surface antigen[108] (Engerix®)	Non-infectious protein on surface of hepatitis B virus	Hepatitis B vaccination
	Treating an autoimmune disease	Anti-Rh immunoglobulin G[109] (Rhophylac®)	Neutralises Rh antigens that could elicit anti-Rh antibodies in Rh negative patients	Prevention of Rh(D) immunisation in Rh(D) negative women
	Treating cancer	Ipilimumab[110,111] (Yervoy®)	Binds cell surface receptor CTLA4 to upregulate activity and proliferation of T-cells	Late stage melanoma
Diagnostics	*In vivo* diagnostics	Recombinant purified protein derivative (DPPD)[112]	Non-infectious protein from *Mycobacterium tuberculosis*	Diagnosis of tuberculosis exposure
	In vitro diagnostics	HIV antigens (Enzyme immunoassay)[113]	Detects human antibodies to HIV	Diagnosis of HIV

Table 1.2 Functional classifications of biopharmaceuticals. The mechanisms of action of biopharmaceuticals within each classification are shown along with examples of products that act by each mechanism. HER2 – human epidermal growth factor receptor 2; IL – interleukin; Rh – Rhesus; Rh(D) – Rhesus D antigen; CTLA4 – cytotoxic T-lymphocyte associated protein 4; HIV – human immunodeficiency virus. This data was reproduced from Leader (2008)[101].

1.3.1.1 Insulin

Bacterially expressed human insulin was the first biopharmaceutical produced using recombinant DNA technology[103], having been approved by the US Food and Drug Administration (FDA) as Humulin® in 1982 for the treatment of diabetes mellitus[104]. This was followed by approval of recombinantly expressed human growth hormone (hGH) as Protropin® in 1985 for the treatment of dwarfism[114].

Recombinant human insulin was also the first biopharmaceutical to be engineered for improved therapeutic traits using site-directed mutagenesis[100]. For instance, introducing substitutions that made the protein more resistant to aggregation resulted in development of insulin lispro (inversion of Pro28 and Lys29 in B-chain) and later insulin aspart (P28A substitution in B-chain) and insulin glulisine (N3K and K29E substitutions in B-chain)[104], which exhibit faster modes of action as a result of these monomeric insulin variants (monomer referring to the native insulin dimer) being able to enter the bloodstream sooner after subcutaneous or intramuscular administration than wild-type insulin, which forms zinc-bound hexamers in its formulation[100,104]. Interestingly, a slower acting form of insulin (insulin glargine) has also been developed by introducing mutations that make insulin more aggregation-prone under physiological conditions (N21G substitution in A-chain and B-chain elongated by two Arg residues). Insulin glargine is soluble in its formulation at pH 4 but precipitates upon subcutaneous administration. Therefore, aggregated insulin must solubilise before it can enter the bloodstream, resulting in slower release of monomeric insulin[104,115].

1.3.1.2 Monoclonal antibodies (mAbs)

The highest number of biopharmaceutical approvals in more recent years has come from mAbs, which have constituted an increasing proportion of the total number of biopharmaceuticals approved[97] (Table 1.3) and are the most commercially successful category of biopharmaceutical (Figure 1.5). The establishment of mAbs as the most successful class of biopharmaceutical is largely due to protein engineering, which enabled production of mAbs with reduced immunogenicity.

Time period	Percentage of mAb approvals
Up to 1989	11
1990-1994	13
1995-1999	24
2000-2004	20
2005-2009	22
2010-2014	26.5

Table 1.3 Approvals of mAbs as a percentage of the total number of biopharmaceuticals approved for use in the EU and US over time periods from 1982-2014. This data was reproduced from Walsh (2014)[97].

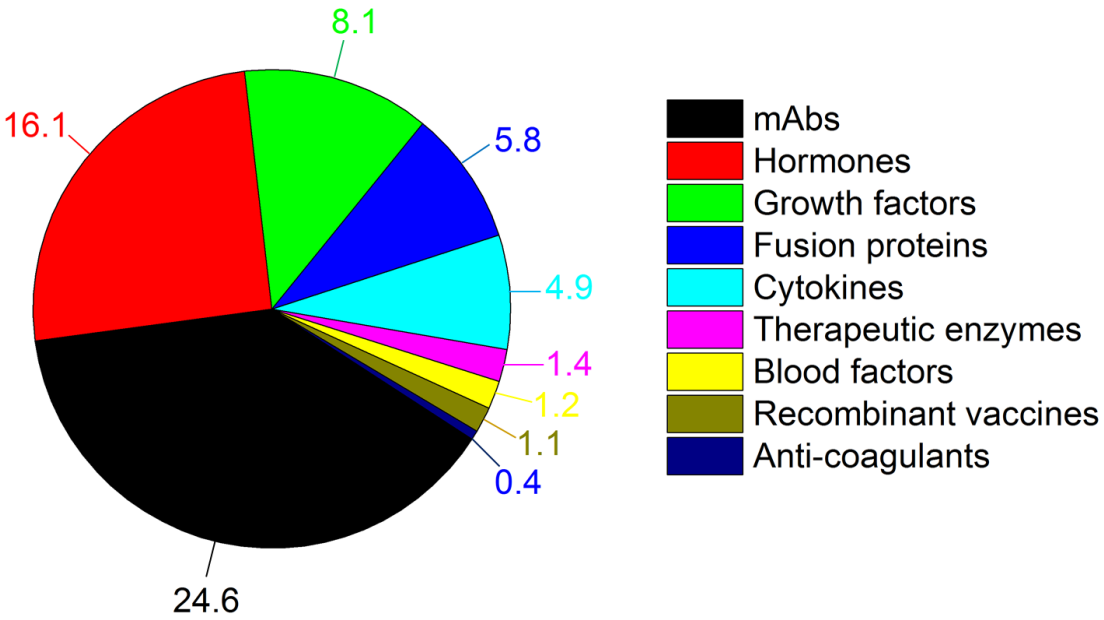


Figure 1.5 US sales of different types of biopharmaceuticals in 2012. Sales are shown as billions of US dollars. This figure was redrawn from Aggarwal (2014)[102].

Early mAbs were raised against a target protein in mice, which generated an immune response when subsequently used in humans. Protein engineering enabled generation of humanised antibodies that were less immunogenic. In humanised antibodies, the complementarity determining regions (CDRs) that make up the antigen binding site of a mAb (Figure 1.6) are replaced by CDRs from murine antibodies that have been generated against a particular target[116]. As well as reducing the immunogenicity of the antibody, molecular grafting of murine CDRs onto a human scaffold means that a human Fc region is present (Figure 1.6), which

is involved in mediation of effector functions such as the complementation cascade, a process that leads to the lysis of antibody-targeted cells by the immune system[114,117]. The Fc region also recruits immune cells for endocytosis and subsequent chemical and enzymatic breakdown of CDR-bound soluble molecules[101].

Later, the binding affinity of mAbs was significantly improved through the use of phage display, which involved cloning of genes encoding antibody variable domains into the genome of a filamentous bacteriophage. The inserted genes encode the variable (V) domains of the heavy (V_H) and light (V_L) chains of the antibody (Figure 1.6) joined by a flexible $(Gly_4\text{-}Ser)_3$ linker, which is referred to as a single-chain Fv (scFv). Upon bacterial expression, the scFv is expressed on the N-terminal end of the gene III protein of a filamentous phage, resulting in display of the scFv on the phage surface that can still bind its target antigen[118]. Further development of this technology enabled display of the entire heavy and light chains making up the Fab region of the antibody (Figure 1.6) on the surface of phage[119]. Due to the high affinity of the antibody fragment for its target antigen, antibody fragments on the surface of phage can be purified by affinity chromatography using the antigen, along with the attached phage containing the DNA encoding this antibody fragment[118].

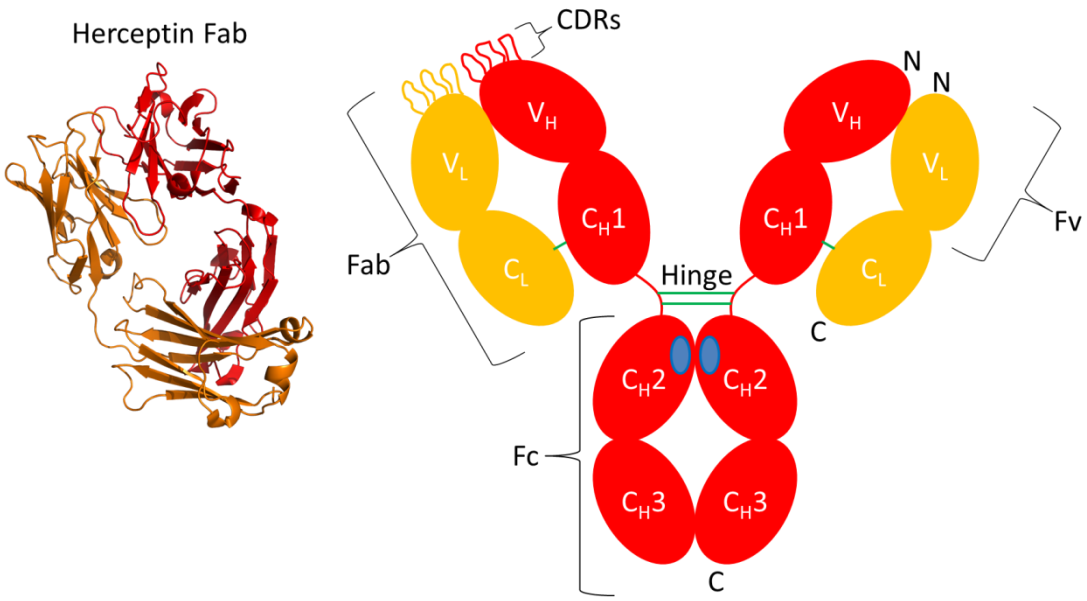


Figure 1.6 Schematic of an IgG antibody. Immunoglobulin G (IgG) is the major antibody found in serum and is made up of two 50 kDa heavy (H) polypeptide chains and two 25 kDa light (L) chains. Disulphide bonds (green) are formed between the flexible linker regions of the two heavy chains in the hinge region as well as the constant (C) domains of the heavy and light chains. The C_H2 domains are glycosylated at Asn297 (blue) and these glycans mediate effector functions[120]. The variable (V) domains contain the complementarity determining regions (CDRs), which are involved in antigen recognition[117]. CDRs are variable flexible loops that are flanked by relatively conserved framework regions made up of β-sheet[116]. The N- and C- termini and CDRs of one heavy and one light chain are shown and the Fc, Fab and Fv regions indicated[117]. Linking together the V_H and V_L domains of the Fv region produces a single-chain Fv (scFv) fragment[118]. The crystal structure of the Fab fragment of trastuzumab (Herceptin®; PDB 1N8Z[121]) is also shown using the same colour scheme and in the same orientation as the left-hand Fab region of the schematic. This figure was redrawn from Lee, Perchiacca and Tessier (2013)[120].

1.3.2 Screening technologies to improve therapeutic properties of proteins

Screening of vast DNA libraries encoding different mutants of a protein is often carried out to identify variants with improved therapeutic properties, such as increased ligand binding affinity[122], increased recombinant expression yield[123] or reduced aggregation propensity[124]. These libraries are usually constructed by random or saturation mutagenesis, using chemically synthesised DNA oligonucleotides and polymerase chain reaction (PCR) and DNA shuffling based methods[125]. In this way, desired properties can be selected for in the absence of structural information of the target or therapeutic protein candidate, with repeated rounds of selection resulting in directed evolution of a particular desired trait[123].

1.3.2.1 In-cell screening methods

Yeast cell surface display has been used to screen for proteins with increased expression level and enhanced thermal stability. In *Saccharomyces cerevisiae*, proteins with increased aggregation propensity should be removed by the protein quality control mechanisms within eukaryotic cells (see Section 1.1.3). As a result, proteins with reduced aggregation propensity are selected by the proteostasis network for secretion as well as display on the surface of yeast when fused to the protein Aga2p found on the cell wall[126]. Although this method has been used to identify variants of single-chain T-cell receptors (scTCRs) with increased secretion levels and display on the cell surface, which correlated with improved thermal stability *in vitro*[126], yeast cell surface display has also been shown to result in selection of variants that were stable as molten globules i.e. lacking a folded structure. This suggests that yeast cell surface display does not always select for the native conformation of a protein[127], which is typically the functional and least aggregation-prone state[37].

Bacteria have also been used to screen a wide range of variants for improved solubility *in vivo*. An example of this is fusion of variants at their C-terminus to Green Fluorescent Protein (GFP). In this construct, GFP only becomes soluble and fluorescent if the attached variant is soluble, providing a qualitative readout of protein solubility. This has also been shown to correlate with fluorescence intensities of fusion proteins when translated using an *in vitro* protein synthesis system and allowed directed evolution of increasingly soluble proteins[128].

A potential flaw with this system is that incorporation of GFP may improve the solubility of the attached variant, resulting in an inaccurate interpretation of variant protein solubility. This has been rectified to an extent by development of a split GFP system that involves attaching the C-terminal 15 amino acids of mutated GFP to the C-terminus of the variant proteins of interest and co-expressing the remaining 215 amino acid segment. These fragments of GFP spontaneously reassemble to form fluorescent GFP provided that the client protein is soluble, with the 15 amino acid tag having less of an effect of altering solubility than full-length GFP[129].

Screening of variants fused to GFP for improved solubility in the cytoplasm of *E. coli* relies on the protein possessing no disulphide bonds in its native structure as these

cannot form in the reducing environment of the cytoplasm[130]. Considering the importance of disulphide bonds in maintaining the structures of many biopharmaceuticals e.g. antibodies[120] (Figure 1.6) and four-helical cytokines[131] (see Section 1.5.1), screening in the cytoplasm of *E. coli* reduces the utility of such systems to biopharmaceuticals.

Development of a screening method for proteins in the periplasm of *E. coli* has enabled the aggregation-propensity of more therapeutically relevant proteins to be tested i.e. the disulphide-containing domain antibodies (dAbs) HEL4 and Dp47d, which are single V_H domains[120]. In addition, this screen has been used to investigate the aggregation of proteins linked to amyloidoses, namely the disulphide-containing protein β2m and the unfolded polypeptides Aβ40, Aβ42, α-syn and IAPP (see Section 1.2.2)[132].

This screen relies on incorporation of the protein of interest into a loop of TEM-1 β-lactamase, an enzyme which confers resistance of *E. coli* to β-lactam antibiotics. If the incorporated protein is soluble then the two segments of β-lactamase can fold to form the functional enzyme that is able to hydrolyse ampicillin. On the other hand, if the protein aggregates than the enzyme will no longer function and the antibiotic resistance of the cell is severely reduced[132,133]. This screen has also been used to identify variants of immunity protein 7 (Im7) that increase stability *in vivo*, which correlated with increases in the thermodynamic and kinetic stabilities of these Im7 variants *in vitro*[133].

1.3.2.2 Phage display and cell-free screening methods

The *in vivo* and *in vitro* aggregation behaviours of proteins are often correlated as the cellular screening methods outlined above can be used to differentiate variants with low and high solubility *in vitro*[134]. However, the recombinant expression level of the variants being screened also relates to their transcription and translation levels[135] as well as their solubility[136]. Consequently, transcription and translation rates will influence the levels of the reporter protein and result potentially in variants appearing falsely more or less aggregation-prone. Additionally, the stresses applied to variants being screened in a cell are numerous and only qualitatively defined[36]. On the other hand, carrying out screening of DNA libraries outside of a cell allows application of more stringent and tightly controlled selection pressures that may be

harmful to cells[125], thus providing more guidance on protein drug formulations, as well as allowing investigation of folding and aggregation properties specifically rather than general solubility. Plus, variants can be selected for functional activity through the binding affinity to their target[118,137,138].

The high *in vitro* stability of the phage particle to denaturation allows a range of defined conditions to be used for selection of variants with increased stability by phage display[139], such as temperature[140-143], chemical denaturation[143,144] and proteolysis[145]. Since its initial development (see Section 1.3.1.2), phage display has been used to select for aggregation-resistant antibody V_H domains, based on transient heating of these domains when displayed on the surface of M13 bacteriophage. Heating for ten minutes at 80 ^oC resulted in thermal unfolding of V_H domains but did not affect the phage. After cooling, V_H domains were selected by binding to protein A on the basis that aggregation-resistant antibodies would refold to their native functional conformation upon cooling[140,141]. This increased resistance to heat-induced aggregation was also shown to correlate with improved recombinant expression yield of variable domains[141].

The development of cell-free based screening technologies has enabled the production of even larger libraries of protein variants (10^{13}-10^{14} proteins in libraries produced by cell-free methods compared to 10^9 for phage display and bacterial libraries[146]) by removing the requirement for transformation of DNA into host cells[125]. These cell-free screening technologies rely on transcription and translation of DNA libraries *in vitro*, resulting in complexes where the encoding mRNA is attached to the protein variant[137,138], either via the ribosome in the case of ribosome display[137] or via puromycin in the case mRNA display[138].

In ribosome display, DNA constructs encoding scFv fragments without a stop codon were transcribed and translated *in vitro*, resulting in a complex of DNA, transcribed mRNA and translated protein. The temperature was then dropped from 37 to 4 ^oC and magnesium added to stabilise these complexes. Immobilised antigen was then used to select for scFv fragments with functional activity, resulting in isolation of this scFv along with the mRNA encoding it. Reverse transcription of the mRNA then provided the cDNA encoding the scFv, which could potentially be subjected to another round of selection for improved binding affinity[137,147] and/or thermodynamic stability[148].

Ribosome display has also been used in conjunction with hydrophobic interaction chromatography (HIC) matrices[149,150] and proteolysis[149] to select for variants with improved folding i.e. a more compact native conformation with reduced exposure of hydrophobic surface. Ribosome display in the presence of dithiothreitol (DTT) and HIC matrices was also used to select for the G-CSF variant C3[150] used in this study (see Section 1.6.4).

In mRNA display, DNA constructs encoding *myc* epitope peptides of different lengths were transcribed *in vitro*, followed by attachment of the peptidyl acceptor antibiotic puromycin[151,152] at their 3' end via an A_{27} linker region. Translation of these mRNA constructs *in vitro* resulted in covalent attachment of the polypeptide to the transcribed mRNA via the puromycin-DNA linker region. The poly-A linker paused the ribosome, giving puromycin time to enter the acceptor site (A-site) before dissociation of the nascent chain and become covalently linked to the polypeptide. Selection of a particular length of amino acid *myc* peptide was carried out using an anti-*myc* mAb, resulting in isolation of this *myc* peptide in complex with the encoding mRNA, which could then be reverse transcribed as above[138].

mRNA display has been used to select for peptides with high binding affinity for the signalling protein Gαi1[153], which mediates effector functions of cell surface G-coupled protein receptors (GPCRs)[117]. Additionally, these peptides have been made more resistant to proteolysis by chymotrypsin through mRNA display-mediated selection, resulting in an approximate 100-fold increase in their serum half-life[154].

Although cell-free libraries allow exploration of a higher amount of protein sequence space, the intrinsic instability of mRNA reduces the stringency of selection conditions that can be applied to protein variants[146]. Conversely, phage are resistant to denaturation[139] and as a result phage display is currently more established as a screening technique than cell-free methods[146]. The different methods that can be used for construction and screening of variant protein libraries are summarised schematically in Figure 1.7.

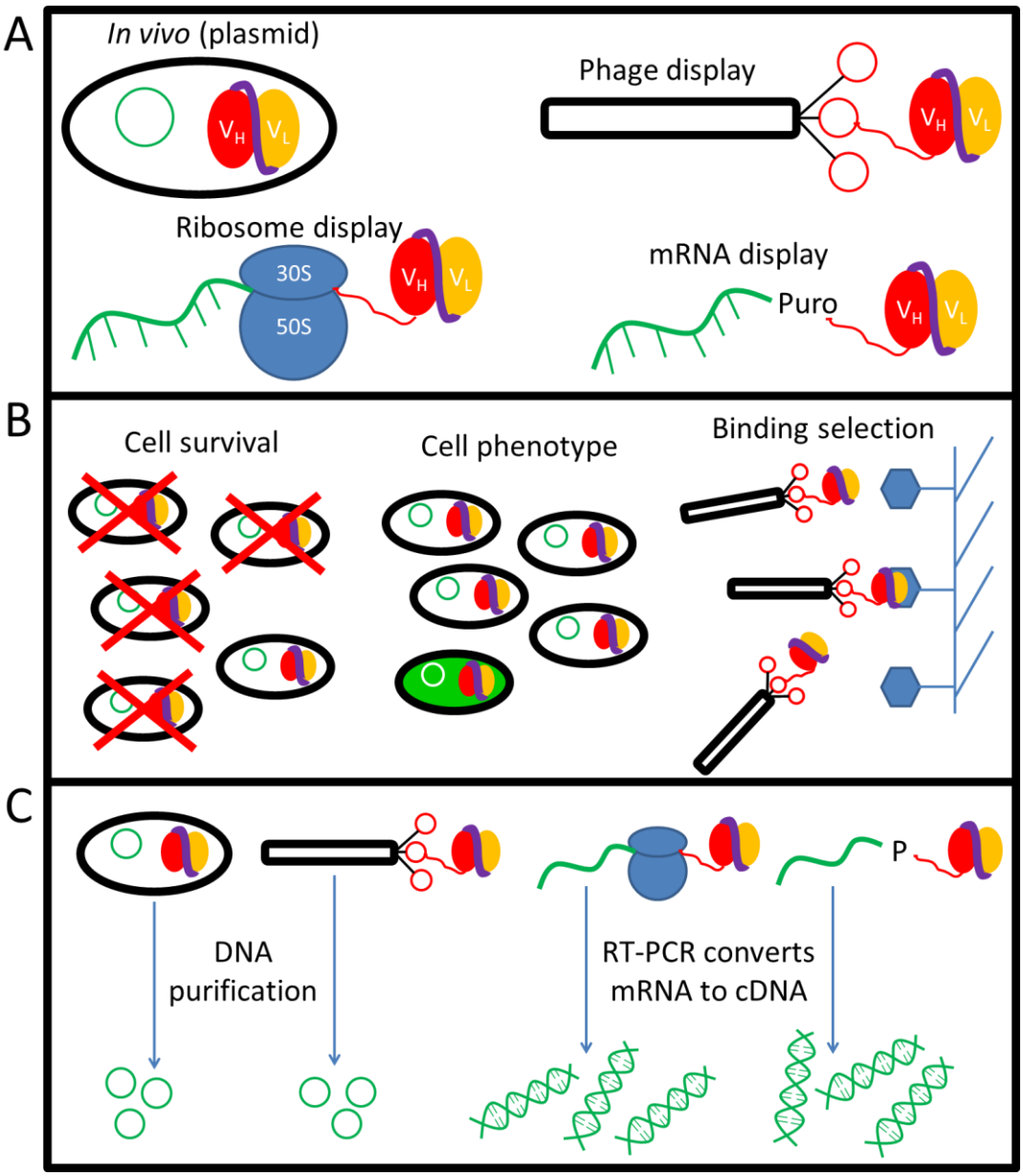


Figure 1.7 Schematic representation of methods used to construct and screen a library of protein variants. In this example, the proteins are scFv fragments. A library of DNA constructs encoding different variants of a protein is constructed using random or saturation mutagenesis[125] and must be subsequently transcribed and translated, either *in vivo* or *in vitro* (A). Protein libraries in cells may be selected for based on survival e.g. variants conferring β-lactam antibiotic resistance[132,133] or their phenotype e.g. GFP fluorescence[128,129], while protein libraries displayed on the surface of phage[118] or bound in mRNA-ribosome-protein[137] or mRNA-puromycin-protein[138] complexes are selected based on binding affinity for an immobilised target antigen (B). DNA is subsequently purified from cells or phage encoding selected protein variants or reverse transcription polymerase chain reaction (RT-PCR) is used to amplify cDNA sequences from mRNA encoding selected protein variants (C). This figure was redrawn from Magliery and Regan (2004)[125].

1.4 The problem of aggregation in biopharmaceutical development

1.4.1 Therapeutic protein aggregates cause immunogenic responses

The screening technologies outlined in Section 1.3.2 are often applied to identify variants of proteins with improved solubility and/or increased stability e.g. through interactions with HIC matrices or resistance to thermally induced aggregation. The purpose of this is to reduce the aggregation propensity of protein variants while attempting to maintain their functionality, which is inherently difficult given the relationship between the marginal stability of the native conformation[29,30] and the dynamic motions that are required for function[31,32].

There are some examples of protein drug formulations where aggregation is advantageous, such as in the case of insulin glargine[104,115] (see Section 1.3.1). However, in the vast majority of protein drug formulations aggregates must be avoided due to the immunogenic responses that they elicit in patients. Antidrug antibodies (ADAs) that are generated to therapeutic protein aggregates reduce the efficacy of the drug and can induce anaphylactic shock, but more worryingly can cause immunisation against a patient's endogenous proteins leading to life-threatening side-effects[155-157]. Although all protein drugs have the potential to be immunogenic and the factors surrounding their immunogenicity are wide ranging[155], aggregates are considered to be one of the main causes[156,157].

Lack of clarity about the aggregate species responsible for causing immunogenic responses[156,158] as well as the immunological mechanism that leads to generation of ADAs[156,157] greatly complicates the issue of aggregates in therapeutic protein formulations. Aggregates have been proposed to elicit an immune response via the T-cell independent pathway[155]. In the T-cell dependent pathway, the antigen is digested within antigen-presenting cells (APCs) and the resulting peptides are displayed on the surface of the APCs via the major histocompatibility complex (MHC) Class II molecule. These peptides are recognised by T cells and in turn result in activation of B cells, forming antibody secreting plasma B cells that can be short- or long-lived. Some B cells also become memory cells upon activation, which react rapidly to repeated presentation of the same T-cell epitopes. Conversely, in the T-cell independent pathway the antigen is endocytosed by peripheral dendritic cells,

which migrate to the spleen and present the antigen to B cells in order to activate them[159].

How aggregates elicit an immune response is unclear. One theory is that aggregates display an array of repetitive epitopes on their surfaces resembling that of a pathogen, which triggers the action of APCs[156]. An alternative theory is that insoluble aggregates are endocytosed by peripheral dendritic cells in the same way as particulate vaccines of an equivalent size (20-200 nm in diameter), resulting in digestion of aggregates and their presentation to B cells.

On the other hand, larger aggregates (0.5-5 µm) would be ingested by macrophages through phagocytosis or macropinocytosis[155,160]. After internalisation of a captured particulate, phagosomes and macropinosomes mature through fusion and fission events with early and late endosomes and lysosomes, which contain hydrolytic enzymes and vaculoar H^+-ATPases. Maturation of phagosomes and macropinosomes thus presumably facilitates digestion of aggregates for liberation of antigen, which is transferred to MHC class II molecules via fusion with immature dendritic cells. These dendritic cells subsequently mature into APCs and present this antigen on their surface to B-cells in order to activate them[160].

The resulting ADAs are classified broadly into two types; neutralising and non-neutralising. Non-neutralising ADAs are those that bind the therapeutic protein but do not inhibit its function, which may still affect the efficacy of the drug by altering its clearance from the body[161]. In addition, protein drug aggregates are usually significantly less or non-functional, which also reduces the efficacy of the drug[162].

Neutralising ADAs bind the therapeutic protein and reduce or abolish its functional activity by disturbing interactions with the target protein, meaning they have a greater effect of reducing the efficacy of the therapeutic protein[161]. Crucially, neutralising ADAs can also inhibit the function of endogenous proteins, which could cause autoimmune diseases[155-157]. An example of this is pure red cell aplasia (PRCA), an anaemia which is thought to be linked to aggregation of recombinant human erythropoietin (Eprex®)[163]. Although the incidence of autoimmune disease resulting from therapeutic protein aggregation is rare, the potential for such life-threatening conditions to result from therapeutic protein aggregates make their detection, characterisation and subsequent removal major priorities[164].

Current studies of the immunogenicity of protein aggregates suggest that insoluble aggregates are more immunogenic than soluble aggregates and that smaller insoluble aggregates are more immunogenic[155]. However, a more recent study using aggregates of therapeutic antibodies suggested that small native-like oligomeric aggregates (<100 nm) elicited a greater immune response than larger non-native aggregates (µm diameter) or native monomer, based on measurement of the ADA levels in mice after intravenous or subcutaneous administration of aggregates of known size and conformation[158].

Generally, the potential severity of aggregates is hypothesised to relate to the proportion of the native structure that is present, with a balance existing between the amount of non-native structure required to elicit an immune response and the amount of native structure required to generate epitopes for ADAs that result in their cross-reactivity with endogenous proteins[155,156].

1.4.2 Protein aggregates are difficult to detect and characterise

The problem of protein drug aggregation is compounded by the huge range of sizes that aggregates can adopt and the heterogeneous nature of aggregates in samples. Although a range of technologies exist to characterise protein aggregates[162,165], there is no single technology that can be used to characterise all possible sizes from soluble dimers to visually observable precipitates[162,166]. Examples of the existing technologies used for profiling of protein aggregates are shown in Table 1.4.

Technique	Aggregate size / μm	Parameter measured
Size exclusion chromatography (SEC)[167,168]	0.001-0.05	Molecular mass
Dynamic light scattering (DLS)[169]	0.001-5	Hydrodynamic radius (R_h)
Analytical ultracentrifugation (AUC)[169]	0.001-0.1	Molecular mass
Asymmetric flow field flow fractionation (AF4)[169]	0.001-100	Particle size
Nanoparticle tracking analysis (NTA)[170]	20-1000	Particle size
Taylor dispersion analysis (TDA)[171]	0.01-5	Hydrodynamic radius (R_h)
Resonant mass measurements (RMM)[172]	0.5-5	Particle size
Polarisation intensity differential scattering (PIDS)[173]	0.04-2	Particle size

Table 1.4 Examples of techniques used for aggregate sizing in protein solutions. Approximate aggregate size ranges that can be detected are shown for each technique along with the parameter measured. This data was reproduced from Hamrang, Rattray and Pluen (2013)[165].

The most commonly used method for aggregate profiling in protein drug formulations is SEC because of its speed, simplicity, relatively low sample requirements and ability to resolve and quantify monomeric and oligomeric species[167,168]. However, there are numerous issues with SEC, such as interaction of aggregates with the resin affecting their elution volume or preventing elution entirely, as well as changes in the amounts of aggregates resulting from non-specific interactions or dilution during the chromatography run[168]. As a result of this, aggregate profiling in biopharmaceutical formulations is often carried out using orthogonal approaches. Techniques typically used alongside SEC include DLS, AUC and AF4[169], which alone do not provide the same resolution or quantification abilities of SEC, but in combination provide validation of results and the ability to further resolve larger aggregates[162].

Aggregates in the size range of 0.1-10 μm are too small to detect by visual inspection (<100 μm) or light scattering but are too large for detection by SEC or AUC (Table 1.4). Although this size range of particle can be detected by DLS and

AF4, they are not well resolved by these techniques[165,174]. These potentially immunogenic aggregates are typically near the solubility limit and often low in concentration, making their detection by the orthogonal methods outlined above difficult[162].

One technique that may provide greater detection of particles in the 0.1-10 µm range is RMM (Table 1.4). This technique uses a suspended microchannel resonator (SMR), which consists of a microfluidic channel within a suspended cantilever to detect and resolve micron and submicron particles. Particles that flow through the channel are detected by causing a momentary change in the resonant frequency of the cantilever, which is due to their differing density to the surrounding solution. The reliance on density for particle detection in RMM, rather than optical properties or Brownian motion in the approaches above, means that this technique has the potential for future use in orthogonal approaches to characterise subvisible aggregate formation[172].

1.4.3 The manufacturing process promotes therapeutic protein aggregation

The extent of aggregation of biopharmaceuticals can be influenced by a variety of factors that can be classified as intrinsic or extrinsic. Intrinsic factors relate to the primary, secondary, tertiary and quaternary structures of a protein while extrinsic factors refer to the environmental conditions that the protein experiences[175]. Both of these factors can affect the proportions of different conformations accessible to a protein, which may have different aggregation propensities as shown in Figure 1.8.

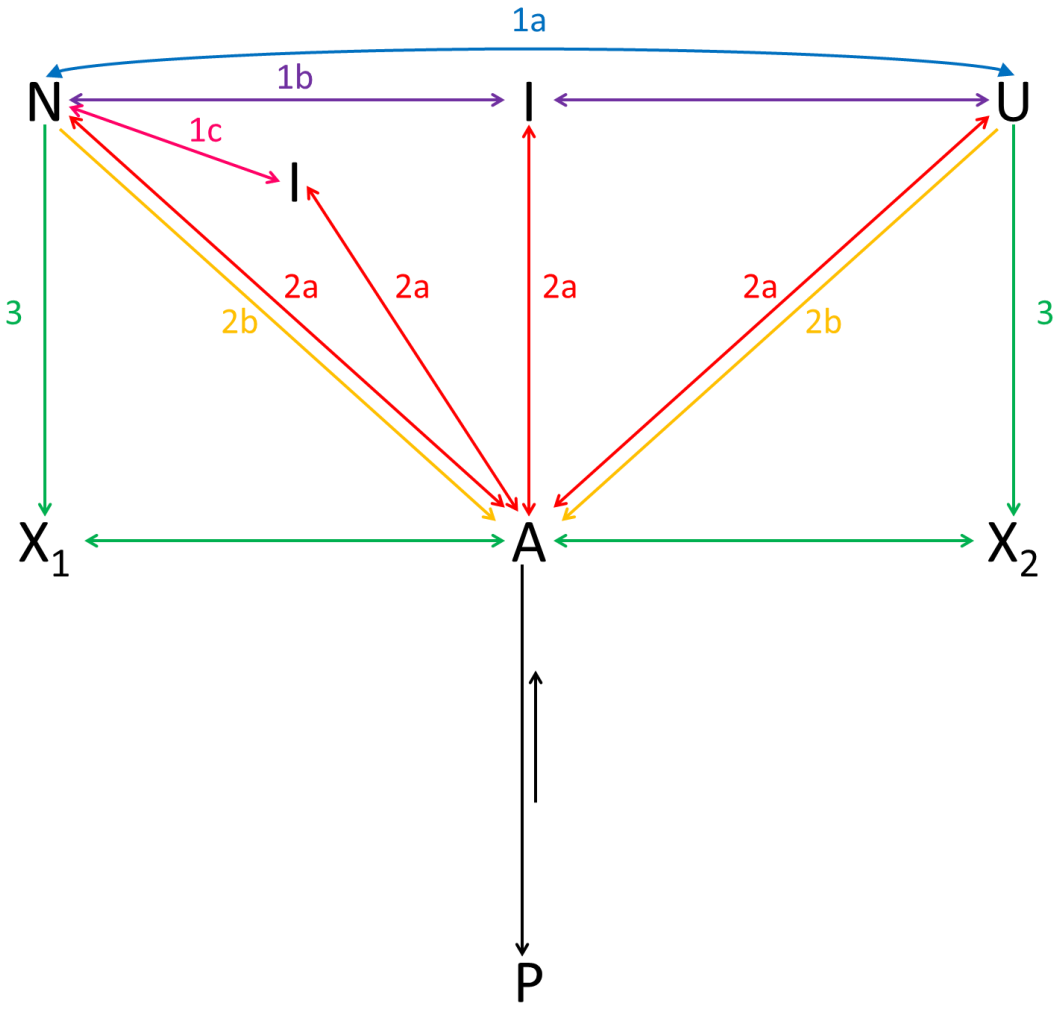


Figure 1.8 Illustrations for the major pathways of protein aggregation. Pathway 1 shows that aggregation can occur from global (1a) or partial unfolding (1b and 1c) of the native state (N), resulting in population of the unfolded (U) or intermediate (I) states, respectively. Intermediates resulting from partial unfolding may be on-pathway for folding (1b) or misfolded off-pathway (1c) states[40]. Pathway 2 shows that these states can form aggregate (A) by self-association through noncovalent (2a) or covalent interactions (2b). Pathway 3 shows that aggregation can occur through chemically modified species (X) e.g. deamidation or oxidation. For simplicity, all possible sizes and morphologies of soluble aggregates are represented by an A state that can further self-associate to form precipitate (P)[175]. This figure was redrawn from Wang, Nema and Teagarden (2010)[175].

During the manufacturing process, biopharmaceuticals are subjected to a range of extrinsic factors that can cause aggregation. Consequently, protein aggregation can occur at every phase of drug development, from recombinant expression to packaging and storage of the final product and administration[175].

Recombinant expression of therapeutic proteins may be carried out in bacteria or eukaryotic cells, depending on the requirement for disulphide bonds or

posttranslational modifications such as glycosylation[176]. Expression in bacteria often leads to formation of insoluble aggregates, termed inclusion bodies. After washing, inclusion bodies are largely made up of the recombinantly expressed protein which eases purification, but they must be solubilised and refolded to attain the native functional conformation. Whereas similar recombinant expression and purification protocols may be used for many different proteins, the optimum conditions for refolding vary considerably, with inefficient refolding potentially causing aggregation[177]. Purification of recombinantly expressed proteins is required regardless of whether they have been refolded, during which they are subjected to changes in environmental conditions such as pH, temperature and ionic strength[175].

An important step in protein manufacturing is viral inactivation, which may occur through exposure to temperatures high enough to cause thermal denaturation of the therapeutic protein (as high as 60 ^{o}C[178]) or acidic pH (as low as pH 3.8[179]). Thermal denaturation of proteins occurs through disruption of noncovalent interactions, resulting in exposure of hydrophobic surface area to potentially cause aggregation[178]. At temperatures required for viral inactivation the hydrophobic effect is increased compared to at room temperature and the diffusion of protein molecules by Brownian motion is increased, making intermolecular collisions and aggregation more likely[180]. Changes in pH can affect the type and distribution of charges on proteins and so may alter electrostatic interactions in the native conformation as well as affecting the overall charge on the native state, which may promote aggregation[181].

Proteins may be frozen and thawed after production or between different stages of the manufacturing process, which can induce aggregation through interactions of hydrophobic surfaces with ice-water interfaces. Generally, interactions with surfaces promote partial unfolding and increased exposure of hydrophobic surface, which upon desorption from the surface can cause protein aggregation[182]. Aggregation is also induced as a result of pH changes caused by crystallisation of buffer components during freezing[178].

Agitation of protein solutions during manufacturing, storage and administration can cause aggregation by exposing proteins to air-water interfaces. The hydrophobicity of air relative to water induces enrichment of hydrophobic surfaces of proteins at the air-water interface, facilitating their exposure and subsequently initiating aggregation[175]. Generally, surface adsorption can promote protein aggregation[182],

with the degree of adsorption being dependent on the material of the surface[183]. Shear stress resulting from flow can also expose hydrophobic surfaces on a potein, promoting aggregation, although the effect of shear is thought to be less significant than (and perhaps dependent on) adsorption at solid-liquid and air-liquid interfaces[184].

Additionally, due to the long shelf-life required for biopharmaceuticals (at least 18 months[172,175,190]) and preparation of multi-dose formulations, many products contain antimicrobial preservatives. One of the most commonly used is benzyl alcohol[99], which has been shown to promote aggregation of therapeutic proteins through interaction with partially unfolded states[185-188]. Even exposure to light/UV radiation can promote aggregation, either through direct photo-oxidation or indirect oxidation by free radicals[189,190].

Lyophilisation (freeze-drying) is commonly used for preparation of solid products, which makes them more resistant to the physical and chemical degradation processes that occur in solution. However, drying removes part of the hydration layer, which can potentially impact the native conformation and promote aggregation upon reconstitution[178], along with the potential impact of freezing outlined above. Additionally, pharmaceutical containers, rubber stoppers, syringes and needles used during storage and administration, respectively are often siliconized to provide lubrication and a protective coating, but trace amounts of silicone oil can initiate protein aggregation due to its hydrophobicity[191].

As well as the increased risk of therapeutic protein aggregation through noncovalent intermolecular interactions (pathway 2a in Figure 1.8) that may be promoted by manufacturing and storage, chemical modification of proteins may also occur that promote their aggregation. These may involve formation of chemical cross-links between proteins, such as intermolecular disulphide bonds via disulphide exchange[192] (pathway 2b in Figure 1.8), or could be chemical alterations of proteins that increase their aggregation propensity.

Deamidation, glycation and oxidation are examples of chemical modifications that have been suggested to promote aggregation (pathway 3 in Figure 1.8). Deamidation is the hydrolysis of amide groups on the side chains of Asn and Gln, resulting in formation of carboxylic acids and these residues becoming Asp and Glu, respectively[182]. This chemical modification has been linked to the reduced solubility

of crystallins in lenses containing cataracts[193] and increased crystallin aggregation *in vitro*[194]. Glycation is the reaction between a carbonyl group of a reducing sugar and a basic side chain[182] and has been shown to induce amyloid aggregation of albumin[195]. His, Met, Cys, Tyr and Trp side chains are potentially oxidised by reactive oxygen species[182] and Met oxidation has been shown to alter the structure and thermal stability of an IgG as well as increasing its aggregation rate[196].

1.4.4 Methods employed to prevent therapeutic protein aggregation

1.4.4.1 Additives

The requirement for a minimum shelf life of 18 months[175,178,197] means that the potential physical and chemical stresses that can cause biopharmaceutical aggregation must be counteracted. Usually, this is attempted through the use of excipients/additives that act either to affect the conformational stability of the native protein or its colloidal stability i.e. reducing intermolecular interactions that cause aggregation[178,182].

A common mechanism of action of excipients is stabilisation of the native conformation through preferential exclusion of the excipient from the protein surface[198], exerted by excipients such as sugars, polyols and amino acids[178]. Addition of excipients such as sucrose results in hydration of the protein surface as the interaction of the protein with water is increasingly favourable relative to the interaction with sucrose. Preferential exclusion of sucrose and hydration of the protein surface has a greater destabilising effect on conformations with increased exposure of hydrophobic surface area, meaning that the denatured and intermediate states are destabilised to a greater extent than the native conformation[198].

Arginine is widely used for refolding of proteins from inclusion bodies[199-203], owing to the ability of this amino acid in particular to suppress aggregation[200,201]. However, the mechanism by which this suppression is achieved is unclear. There is most likely a contribution from preferential exclusion of the amino acid from exposed hydrophobic surfaces of a protein making the native conformation more energetically favourable as outlined above[198], but this does not fully account for the ability of arginine to suppress aggregation[200]. Numerous studies have reported the

ability of arginine to interact with[203] or improve the solubility of[201,204] unfolded and partially-unfolded states. Interaction of the guanidine group of the arginine side chain with tryptophan side chains has been proposed[200,205], along with hydrogen bonding interactions with the amide backbone during the initial stages of folding[203,205].

Non-ionic surfactants are commonly used in the pharmaceutical industry to reduce surface adsorption and subsequent aggregation of proteins during storage, filtration, purification and transportation[182]. Non-ionic surfactants such as polysorbates act by aligning at interfaces to block their interactions with exposed hydrophobic patches of proteins[206], as well as binding weakly to these patches on proteins to inhibit the same interactions[207]. However, the ability of non-ionic surfactants to inhibit surface adsorption of proteins depends on the properties of both the protein solution and the surface[208]. In some cases, non-ionic surfactants have even been shown to promote aggregation through partial denaturation[209]. Ionic surfactants are less commonly used due to their ability to interact with polar and non-polar regions of proteins that can facilitate denaturation, as in the case of sodium dodecyl sulphate (SDS)[206].

The buffer conditions and excipients/additives used in a protein drug formulation are generally selected on a trial-and-error basis, often biased by prior knowledge of conditions that have worked previously[131,210]. In addition, use of multiple excipients and stabilising agents during protein formulation adds an additional burden during characterisation and shelf-life studies[197]. An alternative route to preventing aggregation is through chemical modification of the therapeutic protein.

1.4.4.2 Chemical conjugation

The most common form of chemical conjugation of therapeutic proteins is attachment of polyethylene glycol (PEG)[211]. 'PEGylation' has been shown to increase the thermal stability of the derivatised protein[212] as well as resulting in formation of smaller oligomeric species during aggregation[213,214]. PEGylation was initially designed as a method for improving the *in vivo* half-life of a protein by providing a degree of shielding of the attached protein to proteolytic degradation[215] and reducing renal clearance by making the protein larger[211]. The longer half-life of PEGylated protein drugs is beneficial because fewer doses are required, which improves patient compliance[211]. Additionally, the PEG group itself does not elicit an

immune response and provides shielding of the protein from interaction with immune response elements, reducing the potential immunogenicity of the drug[216]. However, the increased resistance of PEGylated proteins to renal clearance[211], proteolysis[215] and clearance by the immune system[216] results in increased liver uptake and accumulation, which may increase the risk of toxicity[217].

Glycosylation is a chemical modification of proteins that occurs naturally. Glycosylation has been shown to reduce the occurrence of a range of chemical and physical degradation processes[218] and is thought to hinder aggregation through steric repulsion[219,220], as well as increasing the thermodynamic stability of the protein[219,221]. In humans, glycosylation involves attachment of sugars principally on the side chains of Asn and Ser/Thr residues (N- and O-linked glycosylation, respectively). Not all of the potential glycosylation sites on a protein are occupied, which coupled to the variability in the sequence, order and branching of the sugars attached results in a huge range of different glycoforms of proteins[218].

1.4.4.3 Protein engineering

Protein engineering methods have also been used in an attempt to reduce the aggregation tendency of therapeutic proteins. Often this involves screening of vast arrays of mutational variants using methods such as those outlined in Section 1.3.2. More rational approaches have employed the use of algorithms that predict the aggregation propensity of an amino acid sequence[62-66] and/or a three dimensional structure[63]. However, these algorithms only consider the native and denatured states, whereas partially unfolded intermediates are often more aggregation-prone due to localisation of hydrophobic residues into surface exposed patches[60].

Protein engineering strategies have been applied successfully to the generation of antibodies with reduced aggregation propensities[120]. The vulnerability of antibodies to aggregation is wide-ranging, owing to the aggregation-prone motifs often being located in the hypervariable CDRs[140]. Mutation of CDRs to inhibit aggregation can result in significantly reduced binding affinity[222], but mutation of residues adjacent to CDRs has been shown to significantly reduce aggregation while allowing retained functional activity. An example of this is in grafted amyloid-motif antibodies (gammabodies), which are dAbs that display hydrophobic segments from Aβ in their CDRs. Gammabodies are able to selectively bind Aβ oligomers and fibrils over

monomer, enabling targeting of the causative agents of Alzheimer's disease (see Section 1.2.2). However, the exposed hydrophobic segments in the CDRs of gammabodies make them highly aggregation-prone. Interestingly, addition of just three negatively charged residues near the hydrophobic loops within the CDR of gammabodies was shown to suppress aggregation without affecting binding activity[223].

The variability of the primary structure of CDRs coupled to their role in antigen-binding makes it unlikely that a general set of mutations will reduce the aggregation propensity of all antibodies while allowing functional activity to be retained[120]. Mutations introduced into an antibody may also have different effects on aggregation depending on which CDR they are localised to. However, introduction of certain mutations at specific positions within CDRs of dAbs that exhibit significant sequence diversity has been shown to elicit generic reductions in aggregation propensity, which was subsequently applied to the generation of scFv variants derived from Herceptin® with improved solubility[224].

Studies of full-length IgGs have also targeted the CDRs for mutation in an attempt to inhibit aggregation. One study reported a ten-fold increase in solubility of an antibody as a result of substituting just three hydrophobic residues for charged residues in a CDR. However, mutation of these hydrophobic residues resulted in more than a thousand-fold decrease in binding affinity. The aggregation propensity of this IgG was also significantly reduced by targeted glycosylation of a neighbouring CDR, which resulted in an eight-fold increase in solubility without altering binding[225].

Other studies of IgGs have used atomistic molecular dynamics simulations to identify aggregation-prone regions, enabling the solvent exposure of every residue to be identified and a spatial aggregation propensity (SAP) to be calculated. Crucially, using molecular dynamics simulations helped to reveal regions whose solvent exposure is dependent on dynamics and so would not necessarily be revealed from analysis of static structures[222,226-229]. In some cases, the aggregation propensity of identified regions was confirmed experimentally by substitution of hydrophobic for charged residues and demonstrating reduced aggregation[222,227].

The above studies highlight the importance of the dynamics of the native conformation in exposure of aggregation-prone regions. However, as mentioned previously aggregation is arguably more likely to occur from a partially unfolded

state than the native or denatured states, with increased population of partially unfolded states potentially resulting from one of the many stresses of the manufacturing process outlined in Section 1.4.3. These intermediates may be sparsely populated under aggregation-deficient conditions, making them difficult to structurally characterise or model in molecular dynamics simulations.

1.5 Four-helical cytokines

1.5.1 The structural superfamily of four-helical cytokines

The four-α-helix bundle is a common structural motif among globular proteins, consisting of four α-helices connected by flexible linker regions. These helices are commonly connected in an up-down-up-down arrangement, where the C-terminus of one helix is connected to the N-terminus of the next helix in the bundle via a short unstructured linker region typically less than 10 amino acids (Figure 1.9A and C). Alternatively, helices may be connected via long overhand loops that span the helical bundle, which are usually longer than 10 residues, enabling connected helices to be orientated in the same direction (Figure 1.9B and D). The 'handedness' of the arrangement of α-helices in a four-helix bundle is determined based on the position of helix B relative to helix A (when helix A is pointing upwards from N- to C-termini; Figure 1.9E)[230] or helix C relative to helix A in cases where helix B is directly opposite helix A in the bundle (Figure 1.9F).

Connection of helices A and B plus C and D via long overhand loops is observed in the structural superfamily of four-helical cytokines, creating an up-up-down-down topology. Conversely, the linker connecting helices B and C is shorter, which results in neighbouring helices in the bundle being arranged antiparallel to each other (Figure 1.9B and D). The structures of many members of this superfamily are stabilised by disulphide bonds[231].

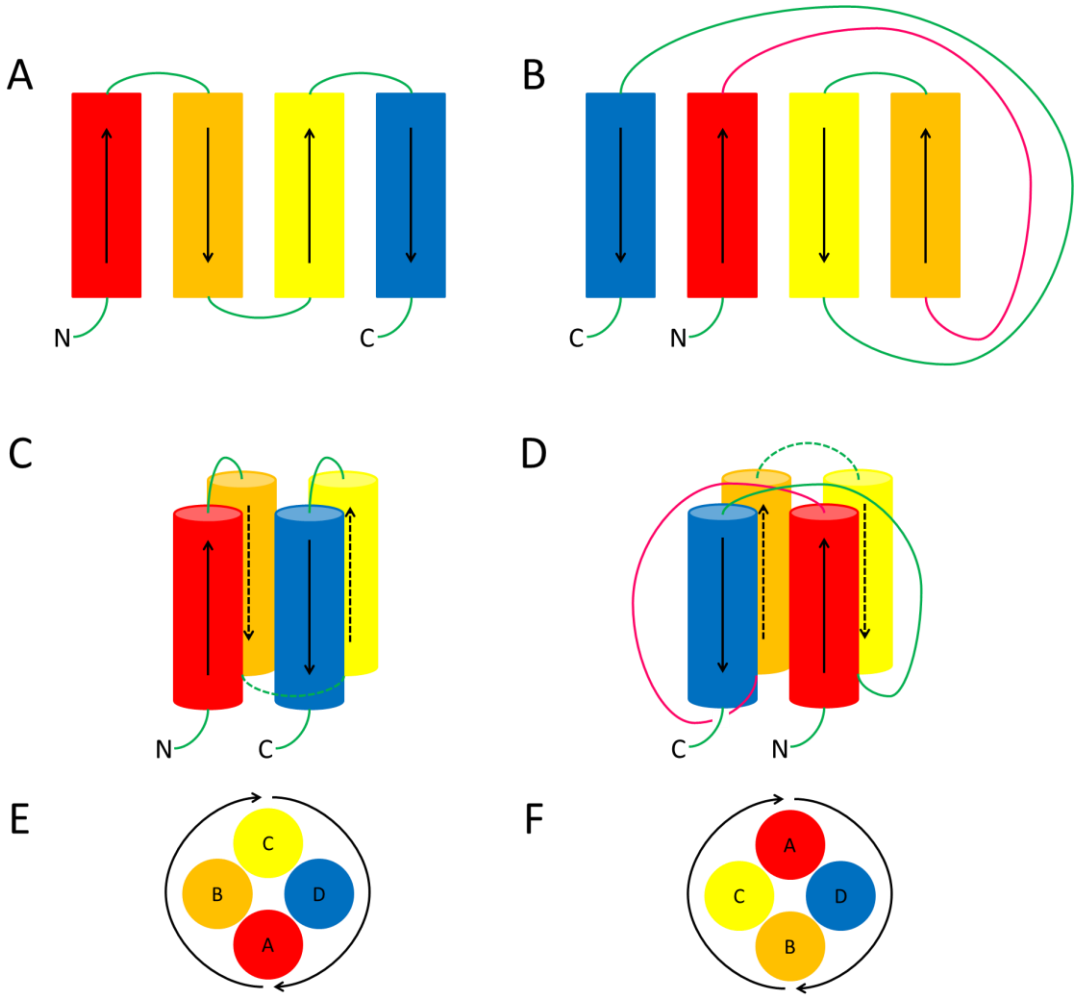


Figure 1.9 Schematic representation of different four-helix bundle topologies. Helices are denoted A, B, C and D from N- to C- termini and are coloured red, orange, yellow and blue, respectively in two- (A and B) and three-dimensional (C and D) representations of four-helix bundles of different topologies. A and C represent the conventional four-helix bundle topology while B and D represent the topology found in four-helical cytokines. The conventional four-helix bundle topology in C is shown from above in E to illustrate the arrangement of helices in a left-handed four-helix bundle. Disordered N- and C-termini and loops connecting helices are shown in green, except for the AB loop that is shown in pink. The position of this loop relative to the CD loop and helix D is indicative of a short- or long-chain four-helical cytokine[231,232]. Arrows indicate the direction of polypeptide chains from N- to C-termini and dotted lines indicate lines or arrows in the background in three-dimensional representations. This figure was redrawn from Presnell and Cohen (1988)[230].

The structures of members of the four-helical cytokine family are similar despite little or no sequence homology[232] and the superfamily can be further divided into short-chain, long-chain and interferon (IFN)/IL-10 families[231]. Short-chain four-helical cytokines comprise on average 140 amino acids, with the four main helices being 10-20 amino acids in length and may possess antiparallel β-sheet structures formed between residues in their AB and CD loops that pack against the core helices in the bundle. Long-chain four-helix bundle cytokines comprise on average 180 amino

acids, with the four main helices being 20-30 residues and may exhibit some α-helical structure in the AB loop (coloured pink in Figure 1.9B and D) that does not pack against the core helices[232,233]. Additionally, in short-chain four-helical cytokines the AB loop passes under the CD loop and behind helix D while in long-chain four-helical cytokines the AB loop passes over the CD loop and in front of helix D[231,232].

All long-chain four-helical cytokines are monomeric, as are the majority of short-chain cytokines, with known exceptions being stem cell factor (SCF)[234], macrophage-colony stimulating factor (M-CSF)[235], interleukin 5 (IL-5)[236] and fms-like tyrosine kinase 3 ligand (Flt3)[237]. Other than SCF, these dimeric proteins all contain disulphide bonds formed between subunits, but the IL-5 dimer has the added feature of intertwined helices, whereby helix D of one subunit completes the four-helix bundle of the other subunit.

Members of the IFN/IL-10 family[233] exhibit structural features of both short- and long-chain four-helical cytokines[232]. For instance, the AB loop passes in front of helix D as observed for long-chain four-helical cytokines, whereas the packing angle between the AD and BC helices in the bundle more closely resembles that observed in short-chain (around 35^{o}) rather than long-chain (around 18^{o}) four-helical cytokines[231,232]. An additional feature characteristic of the IFN/IL-10 family is a short helical section in the CD loop that packs against the core helices of the bundle[232].

Of the members of this family, IFN-α[238] and IFN-β[239] are monomeric while IFN-γ[240] and IL-10[241] are intertwined dimers. These intertwined dimers are formed by exchange of helix D between subunits as observed for IL-5. However, the subunits in IFN dimers are packed side-by-side whereas the subunits in dimers of short-chain four-helical cytokines (e.g. M-CSF and IL-5) are packed end-to-end. Example members from the short-chain, long-chain and IFN/IL10 families are shown in Table 1.5 and Table 1.6.

Family	Protein	Monomer/Dimer	Function
Short-chain	Granulocyte macrophage-colony stimulating factor (GM-CSF)	Monomer[242]	Neutrophil and macrophage production[243]
	Macrophage-colony stimulating factor (M-CSF)	Dimer[b 235]	Macrophage and monocyte production[243]
	IL-2	Monomer[244]	Immune response (primarily T-cells)[244]
	IL-4	Monomer[245]	Immune response (primarily B-cells)[246]
	IL-5	Dimer[b,c 236]	Immune response (eosinophil specific)[246]
	Fms-like tyrosine kinase 3 ligand (Flt3)	Dimer[b 237]	HSC proliferation and differentiation[247]
	Stem cell factor (SCF)	Dimer[a 234]	HSC proliferation and differentiation[247]
	Erythropoietin (EPO)*	Monomer[248,249]	Erythrocyte production[250]
Long-chain	Granulocyte-colony stimulating factor (G-CSF)	Monomer[251]	Neutrophil production[252]
	Growth hormone (GH)	Monomer[253]	Growth regulation[254]
	IL-6	Monomer[255]	Immune system[255]
	IL-11	Monomer[256]	Platelet production[257]
	Leukaemia inhibitory factor (LIF)	Monomer[258]	Differentiation and suppression of leukemic cells[259]
	Ciliary neurotrophic factor (CNTF)	Monomer[260]	Motor neurone development[260]
	Oncostatin M (OSM)	Monomer[261]	Suppression of tumour cell proliferation[262]
	Leptin	Monomer[263]	Body weight regulation[264]
	Prolactin (PL)	Monomer[265]	Mammalian reproductive biology[265]

Table 1.5 Examples of four-helical cytokines from the short- and long-chain families. The native conformations of these proteins are either monomeric or dimeric. Dimers formed without or with disulphide bonds between subunits are labelled a and b, respectively while dimers formed by intertwining of helices are labelled c. *EPO possesses structural features of both short- and long-chain families. Some of the physiological functions of these proteins are also shown. HSC – haematopoietic stem cell. This data was reproduced from Hill, Morea and Chothia (2002)[233].

Family	Protein	Monomer/Dimer	Function
IFN/IL-10	IFN-α	Monomer[238]	Antiviral, antiproliferative and immunomodulatory activities[266,267]
	IFN-β	Monomer[239]	
	IFN-γ	Dimer[c 240]	
	IL-10	Dimer[c 241]	Inhibits synthesis of other cytokines e.g. IFN-γ[241]

Table 1.6 Examples of four-helical cytokines from the IFN/IL-10 family. Proteins are labelled as monomeric or dimeric as outlined in the legend of Table 1.5. Some of the physiological functions of these proteins are also shown.

1.5.2 Functions and physiological roles of four-helical cytokines

Four-helical cytokines are small secreted proteins that mediate intercellular communication in order to help regulate the circulatory system as well as immune and inflammatory responses (Table 1.5 and Table 1.6). Cytokines exert their functions by binding to receptors on the surface of cells, which trigger receptor dimerisation and a subsequent cascade of intracellular signalling reactions. Many of these intracellular signalling reactions involve the Janus kinase family of tyrosine kinases (Jaks) and transcription factors from the signal transducers and activators of transcription (Stats) family[268,269].

The majority of four-helical cytokines signal through binding to Type I cytokine receptors, which are transmembrane proteins that possess conserved Cys residues and a WSXWS motif in their extracellular region. Type II receptors are structurally related to Type I receptors but do not possess the WSXWS motif and are only bound by four-helical cytokines of the IFN/IL-10 family[269]. The intracellular portions of cytokine receptors are more varied but contain Box1 and Box2 domains, which are conserved among Type I receptors[252], while a peptide motif within Box1 is conserved among Type II receptors[270]. Signalling by four-helical cytokines is complicated by the use of common and ligand-specific subunits in different cytokine receptors, allowing some cross-reactivity of cytokines with different receptors[269].

Jaks are constitutively associated with Box1 and Box2 domains[270,271] and become phosphorylated upon formation of a cytokine-receptor complex. Phosphorylation of Jaks causes their activation and phosphorylation of Tyr side chains on the intracellular portion of the receptor. Signalling proteins with SH2 domains e.g. Stats

then bind to these phosphotyrosine motifs and are phosphorylated by Jaks. Phosphorylation of Stats promotes their dissociation from the receptor and subsequent formation of homo- and heterodimers, which are recruited to the nucleus to influence gene expression and ultimately mediate a cellular response e.g. proliferation[269]. As well as promoting a cellular response, Stat signalling also increases expression of proteins that are involved in dampening and terminating cytokine signalling pathways. These proteins achieve this negative feedback regulation of cytokine signalling through inhibition of Jak and Stat activity[272] as shown in Figure 1.10.

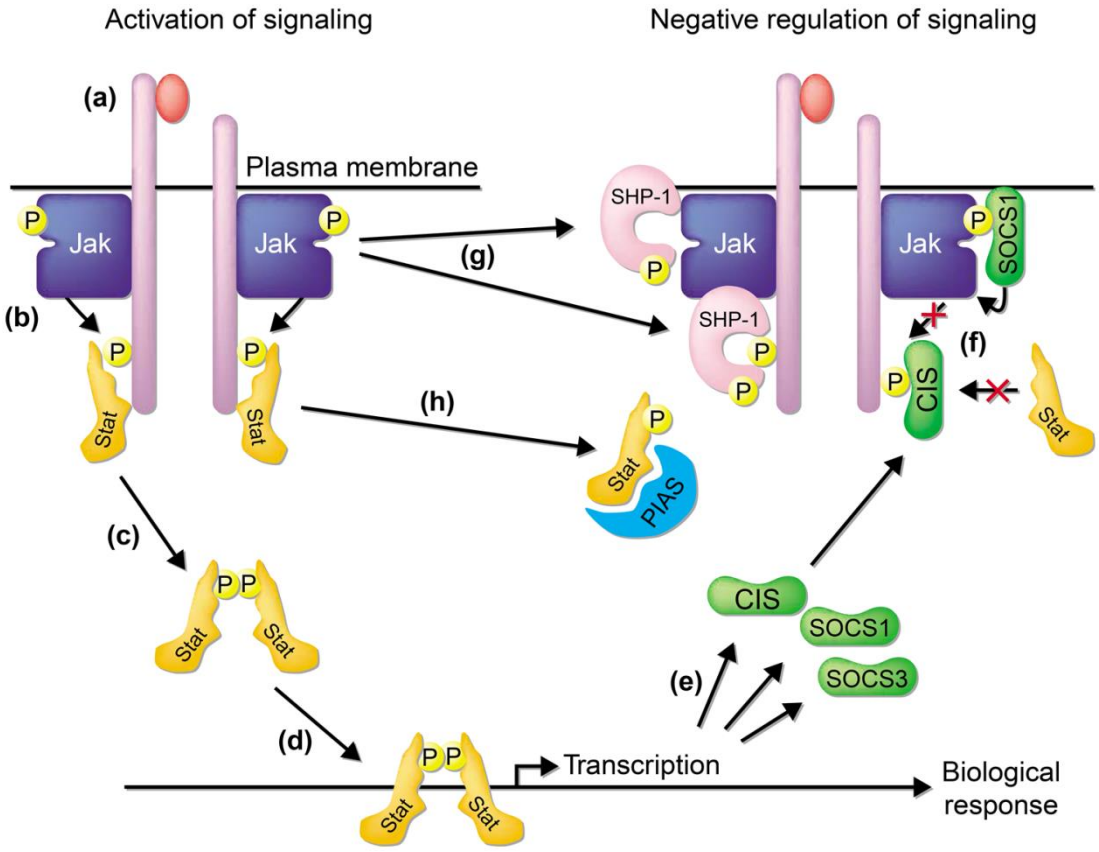


Figure 1.10 Schematic illustration of four-helical cytokine signalling pathways. Receptor dimerization is induced by binding of its four-helical cytokine (red ellipse in a), which causes phosphorylation and subsequent activation of Jaks associated with the intracellular domain. Jaks phosphorylate tyrosine side chains on the intracellular region of the receptor, resulting in recruitment of Stats that are in turn phosphorylated by Jaks (b). Phosphorylation of Stats causes their dissociation and dimerisation (c) and subsequent recruitment to the nucleus to initiate a biological response (d). Stat signalling also upregulates expression of proteins from the suppressor of cytokine signalling (SOCS) family such as SOCS1, SOCS3 and cytokine-inducible SH2-domain-containing proteins (CIS) (e). SOCS proteins suppress signalling by binding to phosphorylated receptors and Jaks to sterically hinder Stat binding (f). Upon activation, Jaks also phosphorylate and activate SH2-domain-containing tyrosine phosphatases (SHPs) that subsequently dephosphorylate Jaks and receptors (g). Constitutively expressed protein inhibitor of activated Stats (PIAS) also bind to phosphorylated Stats to inhibit their signalling. This figure was reproduced from Gadina *et al.* (2001)[269].

1.5.3 Uses of four-helical cytokines as biopharmaceuticals

The role of four-helical cytokines in regulating the proliferation, differentiation and activity of cells involved in vital processes such as haematopoiesis and the immune and inflammatory responses (Table 1.5 and Table 1.6) makes them very attractive as therapeutic agents. Examples of four-helical cytokines from the IFN/IL-10 and short- and long-chain families that are used as biopharmaceuticals are shown in Table 1.7 and Table 1.8, respectively.

Protein (Expression system)	**INN (Trade name)**	**Clinical use**
IFN-α (*E. coli*)	IFN-α-2b[266] (IntronA)	Chronic hepatitis B Chronic hepatitis C A range of lymphomas, melanomas and leukaemias
	IFN-α-2a[266] (Roferon-A)	
	PEG-IFN-α-2b[266] (PegIntron®) (ViraferonPeg®)	Chronic hepatitis C
	PEG-IFN-α-2b[131] (Sylatron®)	Melanoma
	PEG-IFN-α-2a[266] (Pegasys®)	Chronic hepatitis B Chronic hepatitis C
	Interferon-alfacon-1[266] (Infergen®)	Chronic hepatitis C
IFN-β (*E. coli*)	IFN-β-1b[266] (Betaseron®)	Multiple sclerosis
IFN-β (CHO cells)	IFN-β-1a[266] (Extavia®) (Avonex®) (Rebif®)	
IFN-γ (*E. coli*)	IFN-γ-1b[267] (Actimmune®)	Chronic granulomatous disease Osteopetrosis

Table 1.7 Recombinant human IFNs used as biopharmaceuticals. The international non-proprietary names (INN) and trade names of these drugs are shown as well as their clinical uses. The additional letters and numbers used to label IFNs indicate different naturally occurring variants that may also be glycosylated to different extents. This data was reproduced from Lipiainen *et al*. (2015)[131].

Protein (Expression system)	INN (Trade name)	Clinical use
G-CSF (*E. coli*)	Filgrastim[98] (Neupogen®)	Neutropenia HSC transplantation Chemotherapy-induced neutropenia
	Pegfilgrastim[98] (Neulasta®)	Neutropenia Chemotherapy-induced neutropenia
	Lipegfilgrastim[273] (Lonquex®)	Neutropenia
G-CSF (CHO cells)	Lenograstim[98] (Granocyte®)	Neutropenia HSC transplantation Chemotherapy-induced neutropenia
EPO (CHO cells)	Epoetin-α[98] (Eprex®) (Epogen®) (Neorecormon®)	Chronic kidney disease (CKD) and chemotherapy-induced anaemia
	Epoetin-δ[98] (Dynepo®)	
	Epoetin-β[98] (Mircera®)	
	Darbopoietin-α[274] (Aranesp®)	
GM-CSF (*E. coli*)	Molgramostim[98] (Leucomax®)	Bone marrow transplantation Chemotherapy-induced neutropenia
GM-CSF (*S. cerevisiae*)	Sargramostim[98] (Leukine®)	
IL-2 (*E. coli*)	Aldesleukin[131] (Proleukin®)	Renal cell carcinoma Melanoma
Diphtheria toxin/IL-2 fusion protein (*E. coli*)	Denileukin diftitox[107] (Ontak®)	T-cell lymphoma
IL-11 (*E. coli*)	Oprelvekin[98] (Neumega®)	Chemotherapy-induced thrombocytopenia
SCF (*E. coli*)	Ancestim[275] (Stemgen®)	Used in conjunction with Filgrastim for HSC transplantation

Table 1.8 Recombinant human four-helical cytokines used as biopharmaceuticals. The international non-proprietary names (INN) and trade names of these drugs are shown as well as their clinical uses. The different Greek letters used to label EPOs indicate different degrees of glycosylation. This data was reproduced from Lipiainen *et al.* (2015)[131].

Four-helical cytokine drug formulations tend to contain very low concentrations of the active therapeutic agent (typically <1 mg/ml[131]) in comparison to other biopharmaceutical classes such as mAbs (typically >50 mg/ml[120]). PEGylation can greatly increase the concentration at which the protein can be stored at, for instance protein drug formulations of bacterially expressed G-CSF (Filgrastim) are stored at 0.3-0.6 mg/ml[276], while PEGylated G-CSF formulations (Pegfilgrastim and Lipegfilgrastim) are stored at 10 mg/ml[131,277], with both PEGylated and non-PEGylated protein drugs being stored under similar conditions[131,276,277]. PEGylation improves the overall solubility and reduces the rate of aggregation of G-CSF, although PEGylated G-CSF still aggregates, resulting in formation of smaller soluble aggregates[214]. PEGylated G-CSF also has a longer *in vivo* half-life than the non-PEGylated version (3.5 h for Filgrastim[277] compared to 27.2 h for Pegfilgrastim and 32.4 h for Lipegfilgrastim[273]) due to the generic effects of PEGylating therapeutic proteins outlined in Section 1.4.4.

Many four-helical cytokines used as biopharmaceuticals exhibit a range of glycosylation patterns, which often confer beneficial properties to the therapeutic protein (see Section 1.4.4). For instance, glycosylated G-CSF exhibits increased thermal stability and reduced aggregation propensity relative to the non-glycosylated protein[278], enabling its formulation at neutral pH after reconstitution, as opposed to Filgrastim that is formulated at acidic pH[131]. For the majority of four-helical cytokines glycosylation is not required for function, enabling their recombinant expression in *E. coli* (Table 1.7 and Table 1.8).

An exception to this is EPO, which possesses three N-linked glycans that are essential for *in vivo* (but not *in* vitro) biological activity[279,280] and one O-linked glycan that is not[279]. The dependency of the *in vivo* but not the *in vitro* biological activity of EPO on N-linked glycosylation is most likely due to secretion, as EPO with N-linked glycosylation sites removed is secreted less effectively than fully glycosylated EPO[279]. The relationship between the *in vivo* half-life of EPO and glycosylation has enabled generation of a longer acting mutated EPO called Darbopoietin-α, which possesses two extra N-linked glycans and as a result exhibits higher potency and a three-fold increase in serum half-life compared to recombinant human EPO (Epoetin)[274].

1.6 Granulocyte-colony stimulating factor (G-CSF)

1.6.1 Structure

Human G-CSF (Figure 1.11) is a 19.6 kDa protein consisting of 174 amino acids with an O-linked glycan[281] attached to Thr133[282]. The structure of G-CSF is a monomeric long-chain four helix bundle with an up-up-down-down arrangement of helices, with each α-helix being antiparallel in relation to its neighbouring helices and an overall left-handed arrangement of helices in the bundle[251,283]. There is also a short helical segment within the AB loop consisting of a four residue 3_{10} helix followed by six residues of α-helix[251]. G-CSF contains two native disulphide bonds linking Cys36 to Cys42 and Cys64 to Cys74 as well as an unpaired Cys residue at position 17[284]. These disulphide bonds are essential for forming the native functional conformation of G-CSF[285], whereas Cys17 and O-linked glycosylation are not required for function[278,281,285].

Structures of bacterially expressed recombinant human G-CSF (rhG-CSF) have been obtained free in solution by crystallography (PDB 1RHG[251]) and NMR (PDB 1GNC[283]; Figure 1.11A) and in complex with the principal ligand-binding region of the murine G-CSF receptor (PDB 1CD9[286]) and the entire ligand-binding region of the human receptor (PDB 2D9Q[287]; Figure 1.12A). Heterologous expression of rhG-CSF in *E. coli* results in non-glycosylated G-CSF with an additional N-terminal Met residue, meaning the molecular mass of bacterially expressed rhG-CSF is 18.8 kDa[281]. Due to the mobility of the unstructured regions of rhG-CSF i.e. the N- and C-termini and the AB and CD loops[283,288] there is electron density missing in the structure of unbound rhG-CSF obtained by crystallography[251], as opposed to the solution NMR structure for which the structural arrangement of every residue has been resolved[283]. Therefore, throughout this thesis all structures of rhG-CSF free from its receptor are shown using the NMR structure PDB 1GNC[283] (Figure 1.11A).

The four main helices of the bundle are rigid while the disordered loops connecting them as well as the unstructured N- and C-termini are more flexible and mobile[283,288]. Despite the AB loop of G-CSF C3 being longer than the CD loop (residues 42-71 in the AB loop compared to residues 125-144 in the CD loop), the CD loop has been shown to be more dynamic than the AB loop. This is most likely due to the short helical section in the AB loop plus the two disulphide bonds linking

residues Cys37-Cys43 and Cys65-Cys75, which restrain the AB loop by linking residues at the N- and C-termini of this loop to helices A and B, respectively[288].

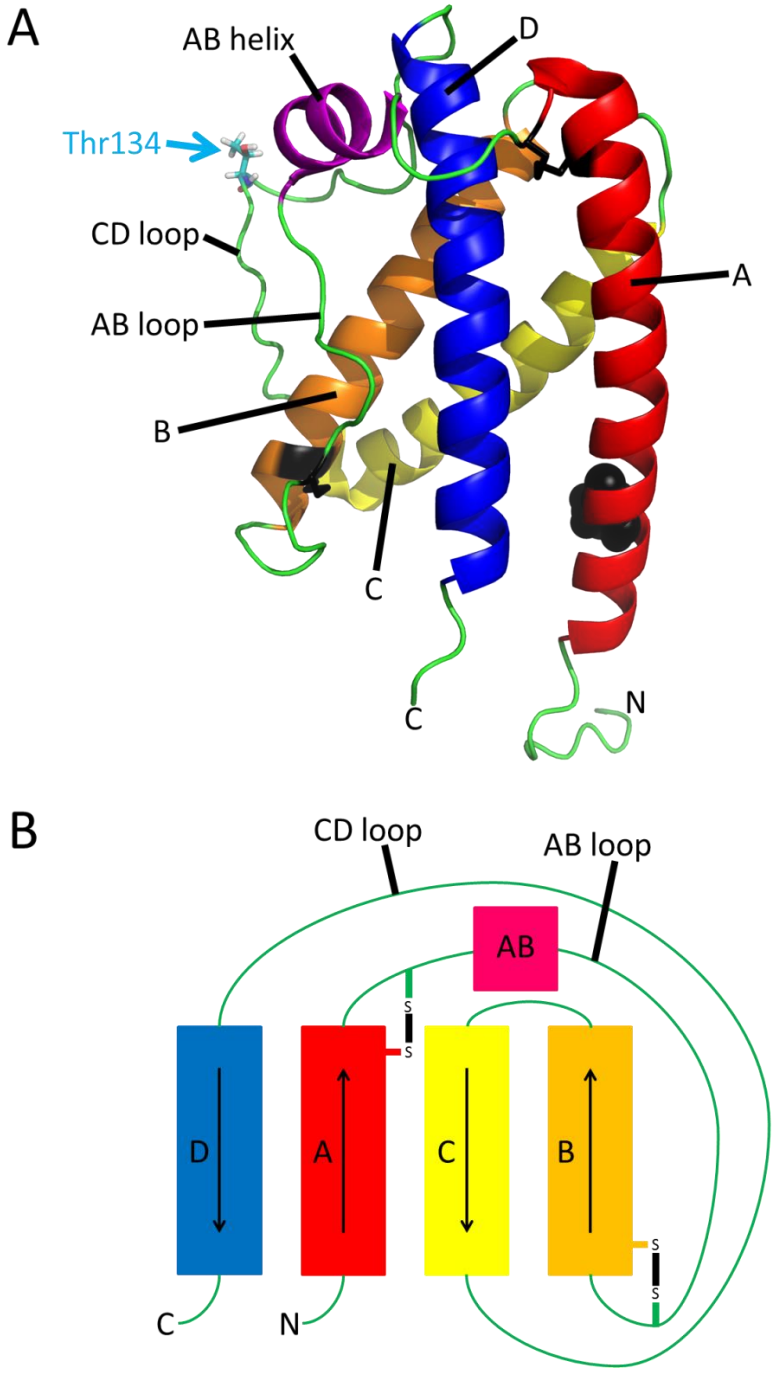


Figure 1.11 Structure of rhG-CSF. Helices A (residues 13-41), B (72-96), C (102-124) and D (145-170) are coloured red, orange, yellow and blue, respectively on the structure in A and on the schematic representation of topology in B. The short helical section in the AB loop (48-56) is shown in pink. Cys residues and disulphide bonds connecting Cys37 with Cys43 and Cys65 with Cys75 are coloured in black and unpaired Cys18 is shown as a black space-filling model in A. The unstructured N- and C-termini plus AB and CD loops are shown in green. The O-linked glycosylation site (Thr134) is shown in light blue in A. This figure was constructed using the structure of G-CSF obtained by NMR (PDB 1GNC[283]) in PyMOL[289].

1.6.2 Function and physiological roles

G-CSF is involved primarily in the response to infection[252]. Levels of G-CSF in human serum are normally undetectable or detectable at very low levels (25.3 ± 19.7 pg/ml) but show a marked increase during infection (731.8 ± 895.0 pg/ml)[290]. The main physiological role of G-CSF is in promoting neutrophil production, achieved through stimulating the proliferation of haematopoietic stem cells and their differentiation into mature neutrophils, as well as promoting survival of cells throughout the neutrophil lineage. G-CSF also enhances key activities of mature neutrophils such as superoxide production and phagocytosis[252]. The physiological role of G-CSF and the functional activity of rhG-CSF expressed in bacterial or mammalian cells[278,281] allow its use as a biopharmaceutical for treating neutropenia, often as a result of chemotherapy as well as mobilisation of haematopoietic stem cells (HSCs) for transplantation from the bone marrow[98] (Table 1.8).

G-CSF exerts its role in combating infection through binding to the G-CSF receptor (G-CSFR), which is found on the surface of all cells of the neutrophil lineage[252]. Binding of G-CSF to G-CSFR results in formation of a 2:2 complex[286,287] (Figure 1.12A) that initiates intracellular signalling cascades along multiple interconnected pathways[252], the best characterised of which are outlined below. G-CSFR lacks intrinsic kinase activity but receptor binding results in phosphorylation and activation of the Jak family kinases Jak1, Jak2 and Tyk2[291,292], which along with Src kinases Lyn[293] and Hck[294] are constitutively bound to the Box1 and Box2 domains of the membrane-proximal region of G-CSFR. Activation of Jak family kinases in turn causes phosphorylation and activation of primarily Stat3[292] but also Stat1 and Stat5 through dimerisation, resulting in formation of Stat3 homodimers plus Stat1/Stat3 and Stat3/Stat5 heterodimers[295,296]. Upon dimerisation, Stat3 is recruited to the nucleus to initiate cell proliferation. Activation of Lyn and Hck results in phosphorylation and activation of downstream PI3 kinase[297], which in turn activates Akt[298] that is subsequently recruited to the nucleus to initiate cell survival.

In addition to the Box1 and Box2 intracellular domains, phosphorylation of Tyr704 and Tyr744 at the C-terminal end of the intracellular region of G-CSFR are essential for mediating binding and subsequent activation of Stat3[299], while phosphorylation of Tyr764 is essential for Shc binding and downstream activation of Ras[300] that initiates proliferation. Tyr729 phosphorylation is also needed for initiating negative signalling

pathways to limit the intensity and duration of the neutrophil-promoting signal induced by G-CSF. Phosphorylated Tyr729 recruits SOCS3[301], which inhibits Stat3 signalling and is induced within thirty minutes of G-CSF signal initiation[302].

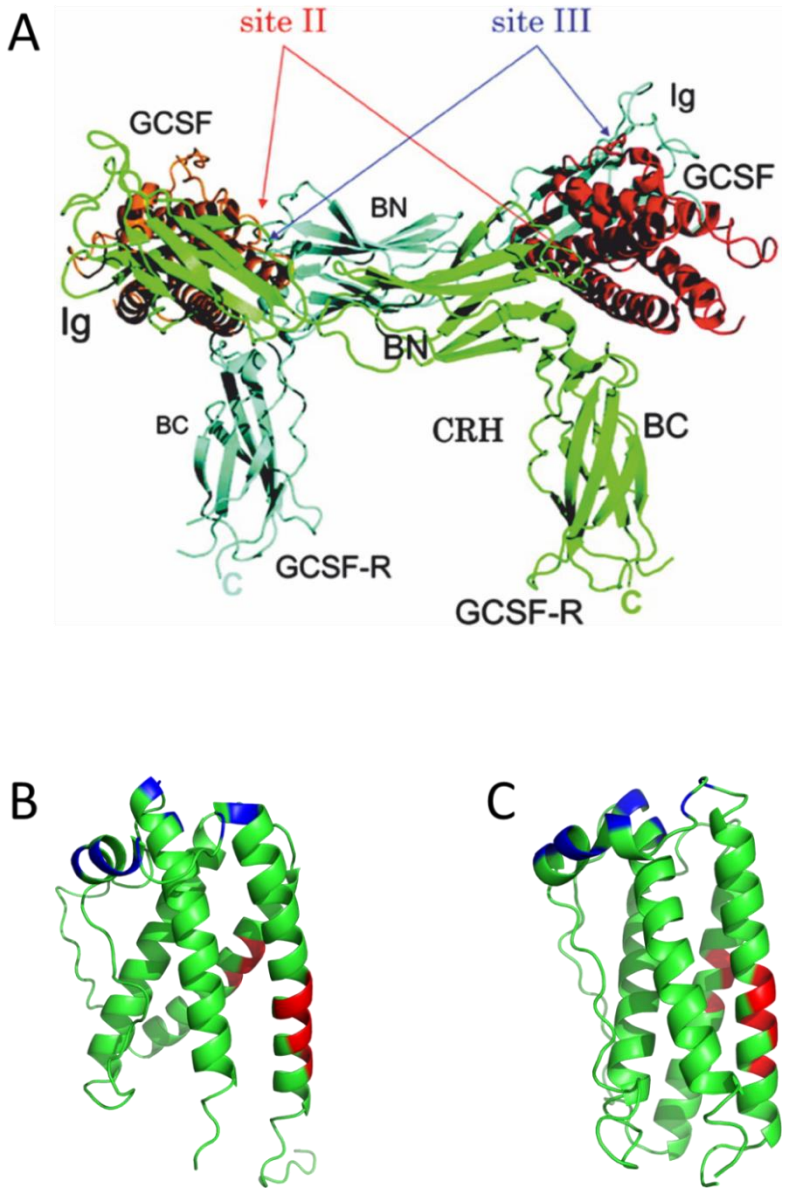


Figure 1.12 Structure of G-CSF in complex with the G-CSF receptor. A 2:2 complex is formed between G-CSF and its receptor (G-CSFR) mediated by two sites of interaction (A). The major site of interaction is at site II between residues in helices A and C of G-CSF and residues within the 'bend' region of the cytokine receptor homologous (CRH) domain of the receptor (labelled orange in A). The CRH domain is made up of N- and C-terminal domains (BN and BC, respectively). The minor site of interaction is at site III between residues in helix D and the helical section of the AB loop of G-CSF and residues in the Ig-like domain of the receptor (labelled purple in A). Interacting residues at sites II and III are listed in Table 1.9 and are also coloured red and blue, respectively on the structure of G-CSF unbound in solution in B (PDB 1GNC[283]) and on the structure of G-CSF bound to its receptor in C (PDB 2D9Q[287]), both of which were constructed using PyMOL[289]. The structure of G-CSF in 2:2 complex with G-CSFR shown in A was reproduced from Tamada *et al*. (2006)[287].

	G-CSF			G-CSF receptor		
Interaction site	**Residue**	**Atom**	**Distance / Å**	**Atom**	**Residue**	**Domain**
Major site (site II)	Lys16	Cγ	3.99	Cδ2	Leu291	Cytokine receptor homologous (CRH)
	Glu19	Oε2	2.25*	Oη	Tyr173	
		Oε2	3.47	Cγ1	Ile239	
		Oε1	2.58*	Nη2	Arg288	
	Gln20	Cδ	2.94	Cδ2	Tyr173	
	Arg22	Nε	3.42*	O	His238	
		Nε	3.40*	Oδ1	Asn240	
	Lys23	Cβ	3.13	Nδ1		
	Leu108	Cβ	3.84	Nη1	Arg167	
	Asp109	Oδ1	2.91*	Nη1		
	Asp112	Oδ1	2.86*	Nη2		
		Oδ1	3.12*	N	Leu172	
		Oδ2	3.79	Cδ1	Leu171	
Minor site (site III)	Tyr39	Cα	3.98	Cζ	Phe75	Ig-like
		O	3.16*	Nε2	Gln91	
	Leu41	Cδ2	3.48	Oε1		
	Glu46	Oε2	3.55	Cδ1	Ile88	
		O	3.03*	N		
		O	3.09*	Nε2	Gln87	
	Val48	N	3.52*	Oε1		
	Leu49	Cβ	2.54	Nε2		
		Cδ2	3.70	Cδ1	Trp82	
	Ser53	Cβ	3.65	Sγ	Cys3	
	Phe144	Cβ	3.44	Oδ1	Asp90	
		Cδ1	3.21	Cβ	Gln91	
	Arg147	Nη1	2.86*	Oε1		

Table 1.9 Amino acid residues of G-CSF and G-CSF receptor that are involved in interactions in the 2:2 ligand-receptor complex. The distances between interacting atoms on residues are listed, along with whether this interaction occurs in site II or III. The domain that the interacting residue belongs to on the receptor is also shown. Distances marked with an asterisk indicate hydrogen bond interactions. Interactions at site II are primarily hydrogen bonding with some Van der Waals, while interactions at site III are mainly hydrophobic with some hydrogen bonding and an aromatic stacking interaction between the side chains of Tyr39 of G-CSF and Phe75 of the receptor. This data was reproduced from Tamada *et al*. (2006)[287].

1.6.3 Current knowledge of the folding and aggregation of G-CSF

Studies of the folding and aggregation behaviours of G-CSF have been carried out using bacterially expressed rhG-CSF, which possesses an added N-terminal Met residue and lacks O-linked glycosylation. This thesis only concerns bacterially expressed rhG-CSF, which is referred to herein as G-CSF.

G-CSF is significantly more aggregation-prone at pH 7 than pH 4[185,276,303], despite displaying similar thermodynamic stabilities under these conditions (40 and 38 kJ mol^{-1}, respectively[142,304]). Reduced aggregation of G-CSF at pH 4 has been suggested to be the result of the side chain of unpaired Cys18 being less reactive at acidic pH[305]. However, a C18A variant has been shown to aggregate to the same extent as wild-type G-CSF at pH 7 and 37 ^{o}C, albeit at a slower rate[142]. A more likely explanation for the reduced aggregation of G-CSF observed at acidic pH values is a greater overall positive charge on the protein (pI of 6.02 predicted from the amino acid sequence[117]), which increases electrostatic repulsion between monomers[303].

Equilibrium denaturation of G-CSF at pH 4 in 0-5 M GdnHCl monitored by far-UV CD and fluorescence emission suggested a two-state transition between native and denatured states. However, studies of the folding and unfolding kinetics of G-CSF at pH 4 revealed the population of two intermediate states as shown in Equation 1.1[304].

$$U \leftrightarrow I_1 \leftrightarrow I_2 \leftrightarrow N \qquad \textbf{Equation 1.1}$$

The first intermediate state (I_1) was inferred from observation of roughly 50 % of the expected change in amplitude of the far-UV CD signal at the beginning of measurement, while the observed change in the amplitude of the fluorescence emission signal matched what was expected. Based on a dead-time of approximately 1 ms for the stopped-flow mixing apparatus used by Brems[304] in this study, it was concluded that I_1 represented formation of approximately 50 % of the α-helical structure of G-CSF without tertiary structure formation around Trp59 in the AB loop or Trp119 on helix C.

The second intermediate state (I_2) was detected based on the change in signal observed during kinetic refolding experiments having to be fitted using two

exponential functions, suggesting two folding reactions were taking place with different rate constants, which were denoted the fast and slow reactions. Conversely, the change in signal observed during unfolding experiments could be fitted using only a single exponential function. The use of W59Q and W119F single-site G-CSF mutants enabled the change in the environment of Trp119 to be attributed to the fast reaction (approximately 20-100 ms), represented by the transition from I_1 to I_2 in Equation 1.1. Completed formation of native α-helical structure was also shown to occur during this I_1 to I_2 transition. The change in the environment of Trp59 was attributed to the slow folding reaction (approximately 0.2-1 s), represented by the transition from I_2 to N in Equation 1.1, suggesting Trp59 is in a non-native environment in the I_2 state but otherwise G-CSF is native.

Interestingly, when refolding was initiated in the presence of final concentrations of 2-3.5 M GdnHCl the amplitude for the slow folding reaction was increased compared to refolding in the presence of GdnHCl concentrations lower than 2 M, suggesting increased population of the I_2 state in denaturant concentrations too low to cause global denaturation[304]. This could be indicative of partial unfolding of G-CSF.

In the presence of GdnHCl concentrations too low to cause global denaturation at pH 7 (1-2.5 M), the rate of G-CSF aggregation has been shown to increase significantly. In the presence of these low denaturant concentrations the fluorescence emission intensity of G-CSF was significantly increased, suggesting a change in the chemical environment of Trp59 and/or Trp119[142]. This change in fluorescence emission intensity in low GdnHCl concentrations was not observed at pH 4[304], although this is most likely due to the reduced emission intensity of Trp at acidic pH that results from quenching[306]. Through the use of W59Q and W119F single-site G-CSF mutants, the increased fluorescence emission intensity of wild-type G-CSF at pH 7 was revealed to be the result of a change in the environment of Trp59[142]. This indicated a possible conformational rearrangement of the AB loop in conditions that promote G-CSF aggregation and has been hypothesised to be the I_2 state observed during kinetic refolding experiments at pH 4[142,304] (Equation 1.1).

Characterisation of G-CSF aggregation at pH 7 suggests that aggregation occurs from a highly native-like conformation[142,185,303,307,308] with the only detected conformational change being an altered environment around Trp59[142]. This aggregation-prone conformation has been described as a structurally expanded species[303,308], which is supported by monitoring of native state hydrogen exchange

of G-CSF in the presence of sucrose[185,307-309] and benzyl alcohol[185,309]. G-CSF exhibits reduced aggregation at pH 7 in the presence of sucrose[185,307,308], which coincides with increased protection of residues in the four core helices from hydrogen exchange with the solvent[182,288-290]. Conversely, the aggregation of G-CSF is increased at pH 7 in the presence of benzyl alcohol[185], with a corresponding decrease in protection of the helical core from hydrogen exchange[185,309].

This could suggest that the aggregation of G-CSF is the result of a general increase in solvent exposure of the core helices, rather than a more subtle conformational change of the AB loop. However, sucrose is well known for causing compaction of the native conformation of proteins as a result of its preferential exclusion from hydrophobic surfaces[198], which is the rationale behind its frequent use as an excipient in protein drug formulations[178]. Conversely, benzyl alcohol is commonly used as an antimicrobial preservative[99] in protein drug formulations and has been shown to interact with partially unfolded forms of proteins, resulting in aggregation[186-188]. Therefore, the generally increased and decreased solvent exposure of G-CSF observed in the presence of benzyl alcohol and sucrose, respectively are to be expected and may not reflect the minimal conformational change required to promote G-CSF aggregation.

1.6.4 Generating G-CSF variant C3

To improve the soluble expression level of G-CSF, Buchanan *et al*.[150] generated a library of DNA constructs by error-prone PCR, which encoded an array of random G-CSF variants. This library was screened using *in vitro* ribosome display (see Section 1.3.2.2), with mRNA-ribosome-G-CSF complexes being incubated in the presence of DTT, HIC matrices and KCl. *In vitro* translation was also carried out in the presence of DTT to remove the stabilising effect of disulphide bonds on the native conformation.

The rationale behind the screening conditions was that variants with increasingly compact and presumably more thermodynamically stable native conformations would be able to retain their functional conformation in the presence of DTT. Additionally, these variants would exhibit increased protection of hydrophobic surface area, making them less reactive towards HIC matrices, while variants with significant affinity for HIC matrices were removed at this stage through

centrifugation. After the stability selection stage (Figure 1.13D), the remaining stabilised variants were tested for functional activity based on binding to a G-CSFR-Fc fusion (Figure 1.13E). mRNA-ribosome-G-CSF variants that survived both the stability and functional selection stages were then isolated using protein G-coated magnetic beads and the mRNA encoding these variants was purified and then amplified by RT-PCR, potentially to be used in another round of selection (Figure 1.13G). The *in vitro* ribosome display screening process is illustrated in Figure 1.13 and the selection conditions used are outlined in Table 1.10.

Round of selection	Selection conditions
1	1 mM DTT
2	5 mM DTT
3	10 mM DTT, Phenyl HIC, 0.5 M KCl
4	10 mM DTT, Octyl HIC, 1 M KCl

Table 1.10 Stability selection conditions used during screening of a G-CSF variant library by *in vitro* ribosome display. All rounds of selection were performed at room temperature. The screening process is illustrated in Figure 1.13. This data was reproduced from Buchanan *et al*. (2012)[150].

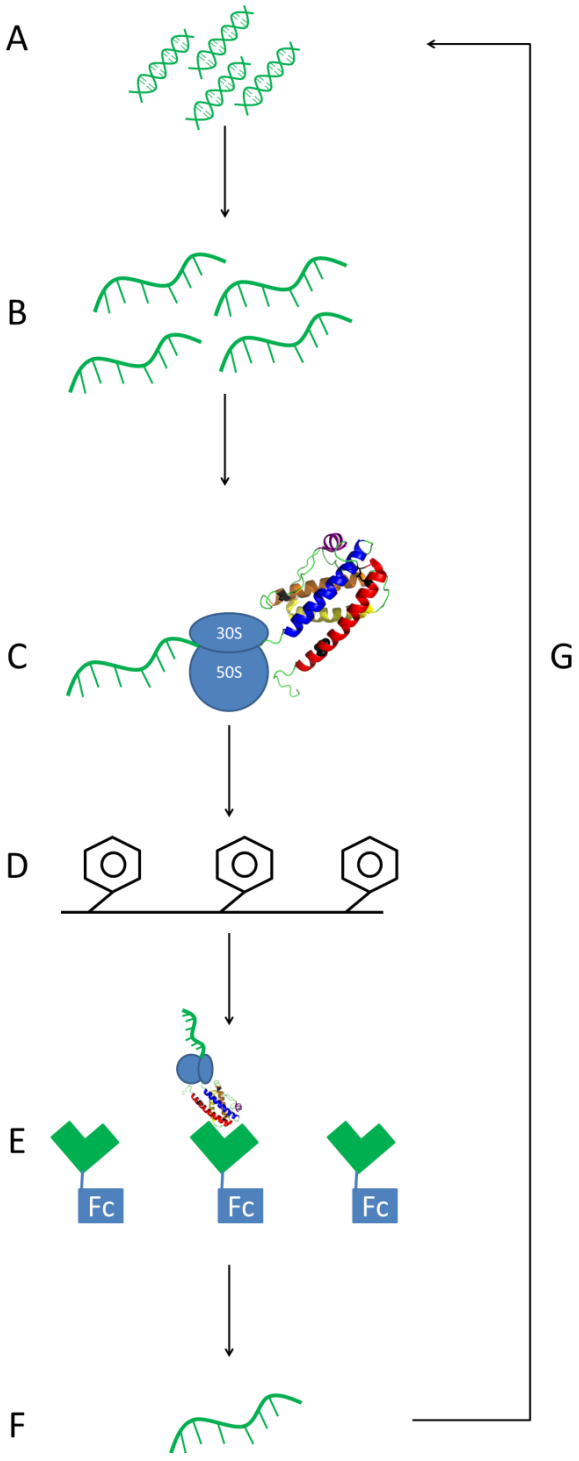


Figure 1.13 Schematic representation of screening a library of G-CSF constructs by *in vitro* ribosome display. A library of DNA constructs encoding G-CSF variants without stop codons (A) was transcribed *in vitro*. mRNA transcripts (B) were translated *in vitro* in the presence of 1 mM DTT to prevent formation of stabilising disulphide bridges. The resulting mRNA-ribosome-G-CSF complexes (C) were incubated in the presence of DTT and HIC matrices plus different KCl concentrations (selection conditions outlined in Table 1.10) (D). G-CSF variants that did not interact with HIC matrices were then tested for functional activity based on binding to a G-CSFR-Fc fusion (E). Stabilised G-CSF variants with retained functional activity were purified in their mRNA-ribosome-protein complex and the attached mRNA was purified (F) and used to amplify cDNA encoding these variants by RT-PCR. cDNA may then have been used in further rounds of selection (G). This figure was redrawn from Buchanan *et al*. (2012)[150].

After the four rounds of selection outlined in Table 1.10, 88 variants of G-CSF plus the wild-type protein were cloned into *E. coli* strain HB2151 for periplasmic expression. Quantification of expression levels by western blotting revealed two variants with significantly increased expression. One of these variants, termed G-CSF C3, was chosen for expression at 2 l scale and purified along with wild-type G-CSF by Ni affinity chromatography. Analysis of purifications by SDS-PAGE and SEC-HPLC revealed a thousand-fold increase in the yield of monomeric G-CSF C3 (8 mg per litre of 2× TY broth) relative to the wild-type protein (8 µg per litre of 2× TY broth). Moreover, SEC-HPLC revealed purified G-CSF C3 to be largely monomeric, while wild-type G-CSF was shown to contain mostly aggregated species. Crucially, G-CSF C3 showed equivalent functional activity to wild-type G-CSF in cell proliferation assays (EC_{50} values of 2.6 and 6 µM, respectively).

G-CSF variant C3 contains four substitutions in relation to the wild-type protein, these being C18G, W59R, Q71R and F84L. How these mutants cause increased expression of soluble G-CSF variant C3 in the periplasm of *E. coli* has not been determined experimentally. However, computational modelling carried out by Buchanan *et al*.[150] predicts that the substitutions W59R and Q71R confer a reduced aggregation propensity onto G-CSF, which is most likely due to the general reliance of aggregation prediction programs on hydrophobicity and charge[62-66]. The aggregation prediction software used by Buchanan *et al.*[150] is found within the Biovia (formerly Accelrys) Discovery Studio® package and calculates SAPs[222,226-229] (see Section 1.4.4.3), in this case using the structure of G-CSF obtained in complex with the CRH domain of the murine G-CSF receptor (PDB 1CD9[286]).

This suggests that the W59R substitution may be linked to the observed increase in recombinant expression of soluble protein. Interestingly, Trp59 in the AB loop of G-CSF has been revealed from fluorescence emission spectra to undergo a change in its chemical environment under aggregation-prone conditions, which is hypothesised to result from a conformational change in the AB loop. Other than this subtle conformational rearrangement, G-CSF is theorised to undergo aggregation from a native-like conformation[142,185,303,307,308]. This suggests a role for the W59R substitution found in G-CSF C3 in conferring increased soluble expression possibly via reducing aggregation propensity.

Although the C18G substitution was not predicted to affect the aggregation propensity from computational modelling, substitution of unpaired Cys18 for Ala has

previously been shown to reduce the rate of aggregation of G-CSF, although ultimately the C18A G-CSF variant aggregated to the same extent as the wild-type protein[142]. Therefore, the C18G substitution may also be responsible for the observed increase in recombinant expression level.

1.7 Objectives of this study

The outline of this study is biophysical characterisation and comparison of G-CSF wt and the variant C3 under conditions that favour and disfavour aggregation. The objective of this is to elucidate whether G-CSF C3 exhibits reduced aggregation compared to G-CSF wt and whether this is responsible for the observed increase in the expression level of soluble G-CSF C3. The other key objective is identifying and analysing the subtle conformational change that drives G-CSF aggregation.

In Chapter 3, G-CSF wt and C3 are expressed and purified and their structures compared, which revealed no significant differences in conformation between the two proteins. The work in Chapter 4 shows that G-CSF C3 exhibits similar aggregation behaviour to the wild-type protein. The observed similarities in the biophysical properties and aggregation behaviour of G-CSF wt and C3, coupled to the increased recombinant expression yield of G-CSF C3 led to a change in focus at this point to only studying G-CSF C3 under conditions where aggregation is favoured and disfavoured.

Biophysical characterisation of G-CSF C3 in Chapter 4 revealed a change in the environment of one or more Tyr residues. The conformational change in aggregation-prone conditions was further probed at amino acid resolution by NMR in Chapter 5, revealing a change in the environment of Tyr86 on helix B. A significant increase in the dynamics of the latter half of the AB loop was also revealed by NMR, which is proximal to Tyr86. As a result, increased AB loop dynamics is hypothesised to expose an intrinsically aggregation-prone region of G-CSF C3 encompassing Tyr86, which drives aggregation from a native-like state.

Chapter 2 Materials and Methods

2.1 Materials

2.1.1 General chemicals

Purite 18 MΩ distilled water was used in all protocols. 30% (*w/v*) acrylamide:0.8% (*w/v*) bis-acrylamide and 10 % (*w/v*) sodium dodecyl sulphate (SDS) were purchased from Severn Biotech Ltd., UK. Agar was purchased from Melford Laboratories, UK. Ready mixed LB-Broth (Miller) was purchased from Merck, Germany. Agarose, L-arginine, dibasic sodium phosphate (Na_2HPO_4) and ethylenediaminetetraacetic acid (EDTA) were purchased from Acros Organics, Belgium. Glucose, glycerol, hydrochloric acid (32 % (*w/v*) HCl), sodium azide, sodium chloride (NaCl), sodium hydroxide (NaOH), sucrose and tris(hydroxymethyl)aminomethane (Tris) were purchased from Fisher Scientific, UK. Carbenicillin disodium salt, dithiothreitol (DTT) and isopropyl β-D-1-thiogalactopyranoside (IPTG) were purchased from Formedium, UK. ^{15}N ammonium chloride ($^{15}NH_4Cl$) and ^{13}C-glucose were purchased from Cambridge Isotope Laboratories, USA. Deuterium oxide (D_2O) was purchased from Fluorochem, UK. 8-anilino-1-napthalenesulphonic acid (ANS), ammonium chloride (NH_4Cl), ammonium persulphate (APS), benzamidine hydrochloride hydrate, bromophenol blue, chloramphenicol, ethanol, ethidium bromide, guanidine hydrochloride (GdnHCl), magnesium chloride ($MgCl_2$), magnesium sulphate ($MgSO_4$), monobasic sodium phosphate (NaH_2PO_4), phenylmethanesulphonyl fluoride (PMSF), sodium acetate, tetramethylethylenediamine (TEMED), tris(2-carboxyethyl)phosphine hydrochloride (TCEP), triton X-100 and urea (BioXtra prep-grade) were purchased from Sigma Life Sciences, USA. Urea used for experiments was of 99 % purity or greater and was purchased from Acros Organics, Belgium or MP Biomedicals, UK.

2.1.2 Molecular biology materials

Synthetic DNA sequences encoding G-CSF wt and C3, codon optimised for *E. coli* expression, were purchased from Eurofins MWG Operon, Germany and inserted into the multiple cloning site (MCS) of expression vector pET23a (Novagen) by procedures outlined in Section 2.2, creating pET23a_G-CSF_wt and pET23a_G-CSF_C3. A plasmid map for pET23a_G-CSF is shown in Appendix 7.3 and the DNA and amino acid sequences of both G-CSF constructs used for recombinant protein expression are shown in Appendices 7.1/7.2 and 7.4/7.5, respectively. The pET23a vector used as the backbone contained a gene encoding TonB (pET23a_TonB) and was obtained from Sam Hickman, Brockwell lab, University of Leeds.

Q5 High-Fidelity DNA polymerase and Antarctic Phosphatase kits plus all restriction enzymes were purchased from New England Biolabs (NEB), USA. T4 DNA ligase, 1 kbp and 100 bp DNA ladders and 6× loading buffer were purchased from Promega, UK. QIAprep Spin Miniprep and QIAquick gel extraction kits were purchased from Qiagen, UK.

2.1.3 *E. coli* strains

E. coli strains BL21(DE3), BL21(DE3) pLysS and XL1-Blue supercompetent cells were purchased from Agilent, USA. Stocks of competent cells were prepared from these strains, using the $CaCl_2$ method[310], and stored at -80 ^{o}C. The genotypes of *E. coli* strains used are shown below:

<u>*E. coli* XL1-Blue</u>
*rec*A1 *end*A1 *gyr*A96 *thi*-1 *hsd*R17 *sup*E44 *rel*A1 *lac* [F' *pro*AB*lacI*^{q} ZΔM15 *Tn*10 (Tet^{r})]

<u>*E. coli* BL21(DE3)</u>
F^{-} *dcm ompT hsd*S_B (r_B^-, m_B^-) *gal* λ(DE3)

<u>*E. coli* BL21(DE3) pLysS</u>
F^{-} *dcm ompT hsd*S_B (r_B^-, m_B^-) *gal* λ(DE3) [pLysS (Cam^{r})]

2.1.4 Growth media

Transformed *E. coli* were cultured in Luria-Bertani (LB) medium, containing 25 g of ready mixed LB-Broth per litre of water. All media were sterilised before use by autoclaving for 20 mins at 120 ^{o}C. 100 mg/ml carbenicillin in water was sterilised by filtration through a 0.22 µm filter and added to media (after cooling) to a final concentration of 100 µg/ml. Chloramphenicol was dissolved in 100% ethanol and added to cooled media to a final concentration of 34 µg/ml. For preparation of agar plates 15 g of agar was added per litre of LB before autoclaving.

To express ^{15}N and ^{15}N/^{13}C isotopically labelled proteins, cells were grown in M9 minimal medium. 5× M9 salts were prepared as outlined below and filtered through a 0.22 µm filter:

<u>5× M9 salts</u>
67.8 g of Na_2HPO_4
30 g of KH_2PO_4
5 g of NaCl

5× M9 salts were then used to make 1 l of 1× M9 minimal medium with water after addition of 1 g of NH_4Cl (or 1 g of $^{15}NH_4Cl$ when expressing labelled protein). The following filter sterilised supplements were added per litre of M9 minimal medium after autoclaving but before adding antibiotics:

<u>Supplements</u>
2 ml of 1 M $MgSO_4$
100 µl of 1 M $CaCl_2$
20 ml of 20 % (*w/v*) glucose (or 10 ml of 20 % (*w/v*) ^{13}C-glucose when expressing labelled protein)

2.1.5 Protein chemistry materials

Superdex^{TM} 75 10/300 GL gel filtration and 5 ml HiTrap^{TM} SP HP cation exchange columns were purchased from GE Healthcare Bio-Sciences AB, Sweden and were attached to an Äkta prime plus pump system, also purchased from GE Healthcare. The Gel Filtration Low Molecular Weight Calibration Kit used to calibrate the

Superdex^{TM} 75 10/300 GL gel filtration column was also purchased from GE Healthcare. Vivaspin 2 ml and 20 ml centrifugal concentrators with respective 3 and 5 kDa molecular weight cut off (MWCO) PES membranes were purchased from Sartorius Stedim, U.K. All buffers used during protein purification were filtered by vacuum filtration using 0.45 µm membrane filters purchased from Millipore, UK. Small volumes of buffers used were syringe-filtered using 0.45 µm Minisart membrane filters purchased from Sartorius, UK. SnakeSkin Pleated Dialysis Tubing with a 3.5 kDa MWCO and Zeba^{TM} Spin Desalting Columns with a 7 kDa MWCO were purchased from Thermo Scientific, UK. Small volumes of proteins were dialysed using either 250 µl DiaEasy^{TM} Dialyzers with a 6-8 kDa MWCO or 800 µl DiaEasy^{TM} Dialyzers with a 3.5 kDa MWCO, both purchased from BioVision, USA. 10× Tris-Tricine-SDS pH 8.3 cathode running buffer was purchased from Alfa Aesar, USA. InstantBlue Coomassie stain was purchased from Expedeon, UK, while Quick Coomassie Stain was purchased from Generon, UK. Precision Plus Protein Dual Xtra Standard molecular mass marker was purchased from Bio-Rad, USA.

2.2 Molecular biology methods

2.2.1 Agarose gel electrophoresis

Agarose gels were prepared by heating 1.5 % (*w/v*) agarose in 1× Tris-acetate-EDTA (TAE) buffer in a 250 ml conical flask until the agarose dissolved. TAE buffer was prepared as a 50× stock, containing 121 g Tris, 28.55 ml glacial acetic acid and 50 ml 0.5 M EDTA (pH 8.0) in 500 ml. Once the agarose had dissolved and cooled to approximately 50 ^{o}C, ethidium bromide (10 mg/ml in water) was added to a final concentration of 1 µg/ml. The gels were then poured into a 12 cm by 15 cm gel tray, a comb was inserted, and the gel was allowed to set. DNA samples were diluted 1:5 with 6× loading buffer. Samples containing 0.1 µg of 1 kbp ladder and 0.13 µg of 100 bp ladder were also diluted 1:5 with 6× loading buffer and loaded into appropriate wells to allow size estimation of DNA within each band. Gel electrophoresis was carried out in 1× TAE buffer, running at a constant voltage of 100 V until the dye front had reached roughly two thirds down the length of the gel. The gels were imaged using a UV transilluminator (Syngene).

2.2.2 Restriction digests of plasmid DNA

Restriction digests were set up using NdeI and XhoI restriction enzymes in combination, or EcoRI in isolation. The composition of a restriction digest mixture is shown below:

- Approx. 1-5 µg of plasmid DNA
- 1× NEBuffer 4
- 1× bovine serum albumin (BSA)
- 5 U each of NdeI and XhoI
- Made up to a final volume of 50 µl with sterile water

When using EcoRI the reaction mixture was identical, apart from the absence of BSA. The samples were incubated at 37 ^{o}C for 1 h.

Where digested plasmid DNA was to be used subsequently for ligation, vector DNA was 5' dephosphorylated by adding 1× Antarctic Phosphatase reaction buffer and 5 U of Antarctic Phosphatase, followed by incubation at 37 ^{o}C for 15 mins. The phosphatase was inactivated by incubation at 65 ^{o}C for 5 mins. After digestion (or 5' dephosphorylation for vector DNA), the products were separated by agarose gel electrophoresis and purified using the QIAquick gel extraction kit (Qiagen) following the manufacturer's instructions.

2.2.3 Ligation of DNA

Ligation reactions were set up using T4 DNA ligase. The composition of a ligation reaction mixture is shown below:

- Approx. 1-2 µg of insert DNA
- Approx. 300-400 ng of vector DNA
- 1× T4 DNA ligase buffer
- 3 U T4 DNA ligase
- Made up to a final volume of 10 µl with sterile water

Negative control samples were also set up, containing all components outlined above minus insert DNA. The samples were incubated overnight at 4 ^{o}C.

2.2.4 Transformation of ligations into XL1-Blue supercompetent cells

E. coli XL1-Blue supercompetent cells were thawed on ice from a stock stored at -80 ^{o}C. 1.7 µl of 1.22 M β-mercaptoethanol was added to 100 µl of cells on ice. Cell suspensions were gently mixed and incubated on ice for 10 mins. Transformation of competent cells was carried out by adding 5 µl of ligation reaction mixture to 100 µl cell suspensions, gently mixing and incubating on ice for 30 mins. Cell/DNA suspensions were then immersed in a water bath at 42 ^{o}C for 45 s, followed by incubation on ice for a further 2 mins. 900 µl of sterile SOC medium (without antibiotic) was added to cell suspensions, which were then incubated for 1 h at 37 ^{o}C with shaking at 200 rpm. After incubation cells were centrifuged at 1000 rpm for 10 mins using a Labnet Spectrafuge 24D microcentrifuge and 900 µl of medium was discarded. Cell pellets were gently resuspended in the remaining 100 µl of SOC medium and plated onto LB agar plates containing 100 µg/ml carbenicillin.

2.2.5 Preparation of plasmid DNA

Plasmid DNA was isolated from 5 ml overnight LB cultures of XL1-Blue cells under 100 µg/ml carbenicillin selection. DNA was isolated using the QIAprep Spin Miniprep Kit (Qiagen) following the manufacturer's instructions. The concentration of the eluted DNA was calculated using an Amersham Biosciences Ultrospec 2100 Pro UV/Visible Spectrophotometer, based on A_{260} values. An A_{260} value equalling 1 corresponds to 50 µg/ml of double stranded DNA when the path length is 1 cm.

2.2.6 DNA sequencing

All DNA sequencing was carried out by Beckman Coulter Genomics. T7 promoter (5'-TAATACGACTCACTATAGGG-3') and T7 terminator (5'-GCTAGTTATTGCTCAGCGG-3') primers were used for forward and reverse sequencing, respectively of pET23a DNA.

2.2.7 Transformation of *E. coli* strains for protein expression

Competent BL21(DE3) or BL21(DE3) pLysS cells were thawed on ice from a stock stored at -80 ^{o}C. Transformation of competent cells was carried out by adding 1 µl of plasmid DNA at 100 ng/µl to 50 µl of cells on ice. The Cell/DNA suspensions were gently mixed and incubated on ice for 30 mins, followed by immersion in a water bath at 42 ^{o}C for 45 s and incubation on ice for a further 2 mins. 450 µl of sterile LB medium (without antibiotic) was added to the cell suspension, which was then incubated for 1 h at 37 ^{o}C with shaking at 200 rpm. Finally, 100 µl of the suspension was plated onto LB agar plates containing 100 µg/ml carbenicillin (and 34 µg/ml chloramphenicol for pLysS cells), which were incubated overnight at 37 ^{o}C.

2.3 General protein methods

2.3.1 Calculation of protein concentration by ultraviolet (UV) absorption at 280 nm (A_{280})

Protein concentration was determined by measuring the absorption of UV light by the protein in solution at a wavelength of 280 nm (A_{280}) and using the Beer-Lambert Law (Equation 2.1); where A represents absorbance, ε represents the molar extinction coefficient, c represents the molar concentration and l represents the path length of light in cm. Where necessary, protein solutions were diluted to produce A_{280} values below 1.

$$A = \varepsilon c l \qquad \textbf{Equation 2.1}$$

The molar extinction coefficients (ε_{280}) of G-CSF wt and C3 unfolded in GdnHCl were calculated from their amino acid sequences to be 15460 and 9770 M^{-1} cm^{-1}, respectively, using the Gill and von Hippel Law (Equation 2.2); where n is the number of that amino acid in the protein of interest. For proteins unfolded in GdnHCl εTrp is 5690 M^{-1} cm^{-1}, εTyr is 1280 M^{-1} cm^{-1} and εCystine is 120 M^{-1} cm^{-1} [311].

$$\varepsilon\text{protein} = \text{nTrp}.\,\varepsilon\text{Trp} + \text{nTyr}.\,\varepsilon\text{Tyr} + \text{nCystine}.\,\varepsilon\text{Cystine} \qquad \textbf{Equation 2.2}$$

The ε_{280} values for folded G-CSF wt and C3 were determined using 77 and 407 µM stocks, respectively. Six-fold dilutions of each G-CSF stock in 20 mM sodium phosphate, 20 mM sodium acetate pH 4 were prepared in triplicate, using the same buffer with and without 6 M GdnHCl. The A_{280} values of these solutions were then measured and average A_{280} values calculated for the three G-CSF dilutions with and without 5 M GdnHCl. Based on these average A_{280} values and the ε_{280} values of unfolded G-CSF wt and C3 calculated using the Gill and von Hippel law, ε_{280} values for folded G-CSF wt and C3 of 15981 M^{-1} cm^{-1} and 10349 M^{-1} cm^{-1} were calculated using Equation 2.3.

$$\varepsilon_{280}\ \text{folded} = \varepsilon_{280}\ \text{unfolded}\left(\frac{\text{average A}_{280}\ \text{without 5 M GdnHCl}}{\text{average A}_{280}\ \text{with 5 M GdnHCl}}\right) \qquad \textbf{Equation 2.3}$$

2.3.2 Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE)

SDS-PAGE was used to assess protein yield and purity during expression and purification. A two-layered system was used, consisting of a stacking gel on top of a resolving gel. The composition of these gels is outlined in Table 2.1. Tris-tricine buffered SDS-PAGE gels were used.

Gel component	Resolving gel volume added / ml	Stacking gel volume added / ml
30% (*w/v*) acrylamide:0.8% (*w/v*) bis-acrylamide	7.5	0.83
Gel buffer (3 M TrisHCl, 0.3% (*w/v*) SDS pH 8.45)	5.0	1.55
Water	0.44	3.72
Glycerol	2.0	0
10% (*w/v*) APS	0.10	0.10
TEMED	0.010	0.005

Table 2.1 Composition of stacking and resolving gels used to make the two layered gel system for Tris-tricine buffered SDS-PAGE. The volumes indicted are sufficient for casting two 8 cm by 10 cm mini-gels using a 1.5 mm spacer.

Two glass plates (Atto mini-gel kit) separated by a 1.5 mm spacer were assembled and clipped together according to the manufacturer's instructions. TEMED is the last component to be added to the resolving gel mixture, which is then immediately poured into the space between the glass plates, up to within 2 cm of the top of the glass plates. This is quickly followed by addition of TEMED to the stacking gel, which is immediately poured on top of the resolving gel and a comb inserted to create wells. Both are allowed to set simultaneously, as the glycerol in the resolving gel prevents mixing of the different layers.

Samples were diluted 1:1 with 2× loading buffer (50 mM TrisHCl, 100 mM DTT, 2% (*w/v*) SDS, 0.1% (*w/v*) bromophenol blue, 10% (*v/v*) glycerol pH 6.8. Gel samples were boiled for the 5 mins before loading onto the gel, with 15 µl of boiled sample being loaded per well. In addition, 5 µl of Precision Plus Protein Dual Xtra Standard molecular mass marker was loaded into another well to aid identification of molecular masses of gel bands. For gel electrophoresis, the inner reservoir of the gel tank contained cathode buffer (100 mM TrisHCl, 100 mM Tricine, 0.1% (*w/v*) SDS pH 8.3) and the outer reservoir contained anode buffer (400 mM TrisHCl pH 8.8). The gels were run at a constant current of 30 mA until the samples entered the resolving gel, at which point the current was increased to 60 mA. Electrophoresis was stopped once the dye front had reached the bottom of the gel and gels were

stained with InstantBlue stain or Quick Coomassie Stain. The gels were photographed using a UV transilluminator (Syngene).

2.3.3 Trichloroacetic acid (TCA) precipitation

Protein solutions containing 6 M GdnHCl were subjected to TCA precipitation and resuspension to allow analysis by SDS-PAGE. Unfolded protein in 6 M GdnHCl was diluted two-fold with pre-chilled 40 % (*w/v*) TCA and vortexed until white precipitate was observed, then left on ice for 20 mins. The TCA/protein suspension was centrifuged at 13300 rpm for 10 mins using a Labnet Spectrafuge 24D microcentrifuge and the supernatant discarded. The resulting pellet was resuspended in 1 ml of acetone, which had been stored at -20 ^{o}C prior to use and then centrifuged as above. After discarding the supernatant this wash step was repeated and the resulting protein pellet was dried by boiling briefly in a heating block. The dried protein pellet was resuspended in 2× loading buffer.

2.3.4 Mass spectrometry

2.3.4.1 Mass measurement

The molecular masses of expressed and purified G-CSF proteins were measured using electrospray ionisation mass spectrometry (ESI-MS). G-CSF samples in 20 mM sodium phosphate, 20 mM sodium acetate pH 4 were buffer exchanged into 66 mM ammonium acetate pH 4 using Zeba^{TM} desalting columns with 7 kDa MWCO according to the manufacturer's instructions. Samples of G-CSF were submitted at 60-80 µM for experiments and analyses carried out by Dr. James Ault in the mass spectrometry facility, University of Leeds using a Synapt HDMS instrument purchased from Waters, UK. For mass spectra of acid denatured and reduced G-CSF C3 1 % (*v/v*) formic acid and 50 % (*v/v*) acetonitrile were added to samples along with 10 mM TCEP by Dr. Ault before analysis.

2.3.4.2 Verification of disulphide bond formation

Trypsin digests of G-CSF proteins carried out by Dr. Ault were also analysed using ESI-MS, in the presence and absence of a reducing agent to confirm formation of the two native disulphide bonds. 50 µl G-CSF samples in 20 mM sodium phosphate, 20 mM sodium acetate pH 4 at 50 µM were washed twice with 200 µl of 50 mM ammonium bicarbonate pH 7, with sample volumes being reduced back to 50 µl after washes using Vivaspin 0.5 ml centrifugal concentrators with 3 kDa MWCO PES membrane. 20 ng μl^{-1} of trypsin in 50 mM ammonium bicarbonate pH 7 was added to 20 µl of G-CSF sample and incubated with shaking at 37 ^{o}C for 18 h. Digestion was quenched by addition of 5 µl of formic acid. For reduction TCEP was added to a final concentration of 5 mM and incubated at 60 ^{o}C for 30 mins before ESI-MS analysis.

Enzymatic digests of proteins in SDS-PAGE gel bands were also carried out and analysed by Dr. Ault using ESI-MS, in order to sequence and identify the proteins in these gel bands.

2.4 Recombinant expression of G-CSF proteins

A single colony of BL21(DE3) or BL21(DE3) pLysS cells freshly transformed with pET23a_G-CSF_wt or pET23a_G-CSF_C3 DNA was picked and used to inoculate a 250 ml LB culture containing 100 µg/ml carbenicillin (and 34 µg/ml chloramphenicol for pLysS cells), which was incubated overnight at 37 ^{o}C with shaking at 200 rpm.

20 ml of overnight culture was used to inoculate each litre of LB medium in a 2 l conical flask, containing 100 µg/ml carbenicillin (and 34 µg/ml chloramphenicol for pLysS cells). For recombinant expression of G-CSF proteins in M9 minimal medium, 10 ml of overnight LB culture was used to inoculate 0.5 l of M9 minimal medium, containing 100 µg/ml carbenicillin (and 34 µg/ml chloramphenicol for pLysS cells), in a 2 l conical flask. Typically, 11× 1 l of LB medium or 22× 0.5 l of M9 minimal medium were used for recombinant G-CSF expression. Cell cultures were incubated at 37 ^{o}C with shaking at 200 rpm and cell growth was monitored at regular intervals by measuring the optical density of cultures at a wavelength of 600 nm (OD_{600}). At an OD_{600} value of approximately 0.6 protein expression was induced by adding

sterile filtered IPTG (1 M in water) to a final concentration of 1 mM. After addition of IPTG all cultures were incubated for 4 h at 37 ^{o}C with shaking at 200 rpm. The cell pellets were harvested by centrifugation at 6000 rpm, 4 ^{o}C for 15 mins using a Beckman Coulter Avanti J-26 XP centrifuge with a pre-chilled JLA 8.1 rotor. Cell pellets were stored at -20 ^{o}C and the supernatant was treated with Virkon disinfectant and discarded.

For expression trials the OD_{600} was measured at the end of each hour of the 4 h post-induction incubation period as well as after overnight incubation. Pellets of 1 ml samples of cell culture were harvested by centrifugation for 5 mins at 13300 rpm in a Labnet Spectrafuge 24D microcentrifuge at room temperature. Cell pellets were stored at -20 ^{o}C and the supernatant discarded. Expression levels were determined by SDS-PAGE analysis, which involved thawing of cell pellets and subsequent lysis by resuspending in 2× loading buffer and boiling for 15 mins. The volume of 2× loading buffer used for cell lysis was varied to correct for cell density.

2.5 Purification of G-CSF proteins

2.5.1 G-CSF inclusion body isolation, washing and solubilisation

Cell pellets were thawed and resuspended in lysis buffer (50 mM TrisHCl, 5 mM EDTA, 2 mM phenylmethanesulphonyl fluoride (PMSF) and 2 mM benzamidine hydrochloride hydrate pH 8) at a ratio of 1 g of cell pellet to 10 ml of buffer and lysed by 5× 30 second periods of sonication on ice at 75 % amplitude, using a Sonics Vibra-Cell^{TM} VCX-130PB sonicator with a 6 mm diameter probe. The cytoplasmic inclusion bodies were harvested by centrifugation at 15000 rpm, 4 ^{o}C for 30 mins using a Beckman Coulter Avanti J-26 XP centrifuge with a JLA 16.250 rotor. Inclusion bodies were washed by resuspending in wash buffer 1 (50 mM TrisHCl, 5 mM EDTA, 1 % (*v/v*) triton X-100 pH 8) at a ratio of 1 g of inclusion body pellet to 20 ml of buffer and harvesting inclusion bodies by centrifugation as above. This wash step was repeated using wash buffer 2 (50 mM TrisHCl, 5 mM EDTA, 4 M urea pH 8). Washed inclusion bodies were resuspended in unfolding buffer (50 mM TrisHCl, 5 mM EDTA, 6 M GdnHCl, 10 mM DTT pH 8) at a ratio of 1 g of inclusion body pellet to 5 ml of buffer and left on a platform shaker at room temperature overnight.

Resuspended pellet was then centrifuged at 15000 rpm, 4 ^{o}C for 1 h using a Beckman Coulter Avanti J-26 XP centrifuge with a JA 25.50 rotor.

2.5.2 G-CSF refolding

G-CSF inclusion bodies solubilised in unfolding buffer were diluted ten-fold with refolding buffer (50 mM TrisHCl, 5 mM EDTA, 0.9 M L-arginine pH 8) and dialysed into 5 l of 20 mM sodium phosphate, 20 mM sodium acetate pH 4. After three dialysis changes refolded G-CSF was centrifuged at 15000 rpm for 30 mins as above and the supernatant vacuum filtered through a 0.22 µm filter to remove precipitated protein.

2.5.3 Cation exchange chromatography

A 5 ml HiTrap^{TM} SP HP cation exchange column was connected to an Äkta prime plus pump system at room temperature and washed with five column volumes (CVs) of water at a flow rate 5 ml/min to remove the 20 % (*v/v*) ethanol storage solution. The column was then equilibrated with five CVs of 20 mM sodium phosphate, 20 mM sodium acetate pH 4 through lines A1 and B. Refolded G-CSF was loaded onto the column at a flow rate of 5 ml/min through line A1. A gradient of 0-100 % elution buffer (20 mM sodium phosphate, 20 mM sodium acetate, 1 M NaCl pH 4) was then run over 100 ml at 2 ml/min through line B, with elution monitored by measuring A_{280} and 2 ml fractions being collected. Elution traces did not immediately return to zero after elution of protein in the peak, resulting in the appearance of an elution peak with a long trailing ‘tail’. Fractions in both the ‘peak’ and the ‘tail’ of the elution profile were pooled separately based on the results of SDS-PAGE analysis. Both ‘peak’ and ‘tail’ fraction pools were desalted by dialysis into 5 l of 20 mM sodium phosphate, 20 mM sodium acetate pH 4 at 4 ^{o}C. After three dialysis changes G-CSF was concentrated using Vivaspin 20 ml centrifugal concentrators with 5 kDa MWCO PES membrane until the protein concentration was 2.5-5 mg/ml for G-CSF wt and 8-10 mg/ml for G-CSF C3. G-CSF proteins were snap frozen as 50-200 µl aliquots in 1.5 ml Eppendorf tubes using dry ice and ethanol and stored at -80 ^{o}C.

After assessment of protein purity by SDS-PAGE, aliquots of G-CSF C3 from both the ‘peak’ and ‘tail’ elution pools were used for experiments, whereas only aliquots of G-CSF wt from the ‘peak’ elution pool were used for experiments.

2.5.4 Size exclusion chromatography (SEC)

Analytical SEC was performed to assess whether G-CSF C3 was monomeric after refolding and purification. All buffers and cleaning solutions were filtered and degassed by vacuum filtration through 0.22 µm filters before use. A Superdex^{TM} 75 10/300 GL gel filtration column was connected to an Äkta prime plus pump system at room temperature and washed with one CV of water at a flow rate of 0.5 ml/min to remove the 20 % (*v/v*) ethanol storage solution, followed by one CV of 1 M NaOH and another CV of water at the same flow rate to remove any contaminant protein.

The column was then equilibrated with one CV of 20 mM sodium phosphate, 20 mM sodium acetate pH 4 through line A1. Aliquots of G-CSF C3 at pH 4 (see Section 2.5.3) were thawed and centrifuged for 15 mins at 13300 rpm in a Labnet Spectrafuge 24D microcentrifuge at room temperature before injection onto the column. 100 µl of G-CSF C3 was injected onto the column at 219 µM through a 500 µl loop attached to the injection port at the beginning of the run. Protein was eluted from the column at a flow rate of 0.5 ml/min and elution was monitored by measuring A_{280}, which was set to zero at the beginning of the run. G-CSF C3 was also analysed by SEC in the same pH 4 buffer with an additional 200 mM NaCl using 100 µl of 208 µM G-CSF C3.

G-CSF C3 was also analysed by SEC in 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide pH 7 with and without 200 mM NaCl. Aliquots of G-CSF C3 at pH 4 (see Section 2.5.3) were thawed and set up for dialysis into 2 l of 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide pH 7 with two changes at 4 ^{o}C. Following dialysis, G-CSF C3 was diluted with the pH 7 dialysis buffer outlined above with or without an additional 2 M NaCl to final salt concentrations of 0 or 200 mM. G-CSF C3 samples in 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide pH 7 buffer with or without 200 mM NaCl were centrifuged as above before injection onto the column. 100 µl of 406 and 443 µM G-CSF C3 were injected onto the column in the presence and absence of 200 mM NaCl, respectively. Before each experiment, the gel filtration column was

washed and cleaned, then equilibrated in the same buffer used to inject G-CSF C3 onto the column as described above.

Proteins from the Gel Filtration Low Molecular Weight Calibration Kit (GE Healthcare) were loaded onto the column in 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide, 200 mM NaCl pH 7 according to the manufacturer's instructions. A calibration curve for analytical SEC was obtained by plotting the ratio of the elution volumes of calibrant proteins (V_e) to the void volume (V_0) of the column (V_e/V_0) against $\log_{10}$ of the molecular masses of calibrant proteins in kDa and fitting to a linear function.

2.6 Steady-state spectroscopy

2.6.1 Buffers

All experiments were carried out in 20 mM sodium phosphate, 20 mM sodium acetate pH 4, referred to herein as pH 4 buffer, or in 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide pH 7, referred to herein as pH 7 buffer, unless otherwise stated. For experiments carried out in pH 4 buffer aliquots of G-CSF stocks (see Section 2.5.3) were thawed and diluted to the stated protein concentrations. For experiments carried out in pH 7 buffer, aliquots of G-CSF stocks were thawed and set up for dialysis into 2 l of pH 7 buffer with two changes at 4 ^{o}C followed by dilution to the protein concentrations stated. Buffers were made volumetrically and filtered by vacuum filtration, or syringe filtration for small volumes, through 0.22 µm filters.

The ionic strength of a buffer solution is a function of the concentration of all the ions present and can be calculated using Equation 2.4; where I represents ionic strength, c represents the concentration of ion i and z represents the charge of ion i.

$$\mathrm{I} = \frac{1}{2}\sum \mathrm{c_i z_i^2} \qquad \textbf{Equation 2.4}$$

Urea concentrations of stock solutions were checked by measuring the refractive indices (Ceti refractometer, Belgium) of pH 4 and 7 buffers with and without urea. The difference in these refractive indices was then used in Equation 2.5 to calculate urea concentration[312]; where ΔN represents the difference in refractive indices of buffer stocks with and without urea.

$$[\text{Urea}] = 117.66\Delta\text{N} + 29.753\Delta\text{N}^2 + 185.56\Delta\text{N}^3 \qquad \textbf{Equation 2.5}$$

2.6.2 Circular Dichroism (CD) spectroscopy

CD spectra were acquired on a Chirascan™ plus CD Spectrometer (Applied Photophysics, U.K.) using recording parameters of 1 nm increments, 1 s of acquisition per increment and a bandwidth of 2.5 nm at 25 ^{o}C. For measurements in the far-UV region (190-260 nm) spectra were recorded using 200 µl of sample volume in a 1 mm path length quartz cuvette (Hellma), while in the near-UV region (250-350 nm) spectra were recorded using 950 µl of sample volume in a 1 cm path length quartz cuvette (Hellma). Samples were left to equilibrate in the instrument for approximately one minute before acquisition.

Ellipticity values of buffer alone were subtracted from ellipticity values of samples containing G-CSF before calculation of molar ellipticity using Equation 2.6; where [θ] represents molar ellipticity, θ represents the ellipticity recorded in degrees, c represents molar concentration and l represents the path length of light in m. The mean residue ellipticity ($[\theta]_{MRE}$) is then calculated by dividing [θ] by the number of peptide bonds[313].

$$[\theta] = \frac{\theta}{\text{cl}} \qquad \textbf{Equation 2.6}$$

α-helices exhibit characteristic absorption minima at 208 and 222 nm[313] and $[\theta]_{MRE}$ values at these wavelengths can be used to estimate the percentage α-helical content using Equation 2.7[314] and Equation 2.8[315], respectively.

At 208 nm

$$\% \, \alpha \, \text{helix} = \left(\frac{[\theta]_{\text{MRE}} - 4000}{-33000 - 4000} \right) \times 100 \qquad \textbf{Equation 2.7}$$

At 222 nm

$$\% \, \alpha \, \text{helix} = \left(\frac{[\theta]_{\text{MRE}} - 3000}{-36000 - 3000} \right) \times 100 \qquad \textbf{Equation 2.8}$$

2.6.2.1 Characterisation of native and unfolded G-CSF proteins

Far-UV spectra of folded and unfolded G-CSF proteins in pH 4 and 7 buffers were acquired in the absence and presence of 8.1 M urea, respectively. For acquiring these far-UV spectra pH 7 buffer without azide was used. Far-UV spectra in pH 4 buffer were acquired using 5 µM G-CSF wt and 8 µM G-CSF C3, while spectra in pH 7 buffer were acquired using 4 µM G-CSF wt and 5 µM G-CSF C3. Far-UV spectra acquired in the presence of 8.1 M urea are only shown in the region of 215-260 nm because data obtained below 215 nm led to excessive dynode voltage (800-1000 V) caused by the absorbance of signal by urea.

2.6.2.2 Characterisation of G-CSF C3 in sub-denaturing urea concentrations

Far-UV spectra of G-CSF C3 in pH 4 buffer in the presence of 0-3 M urea with 1 M increments were acquired using 5-8 µM protein. Far-UV spectra of G-CSF C3 in pH 7 buffer in the presence of 0, 0.5 and 1-4 M urea with 1 M increments were acquired using 7-8 µM protein. These spectra are only shown in the region of 210-260 nm because data obtained below 210 nm led to excessive dynode voltage (800-1000 V) caused by the absorbance of signal by urea.

Near-UV CD spectra of G-CSF C3 in pH 4 buffer in the presence of 0-3 M urea with 1 M increments were acquired using 36-52 µM protein. Near-UV spectra of G-CSF C3 in pH 7 buffer in the presence of 0, 0.5 and 1-4 M urea with 1 M increments were acquired using 45-57 µM protein.

2.6.3 Fluorescence emission spectroscopy

Fluorescence emission spectra were acquired on a QM-1 Fluorimeter (Photon Technology International, U.K.) using recording parameters of 1 nm increments, 1 s of acquisition per increment at 25 ^{o}C. Samples were left to equilibrate in the instrument for approximately one minute before acquisition.

Fluorescence count values of buffer alone were subtracted from count values of samples containing G-CSF. Where necessary, fluorescence count values were adjusted to correct for different protein concentrations of samples.

2.6.3.1 Characterisation of native and unfolded G-CSF proteins

All measurements were recorded using 950 µl of sample in a 1 cm path length quartz cuvette (Hellma) using excitation and emission slit widths of 1 and 1.5 nm, respectively. After excitation at 280 nm emission spectra in the region of 290-400 nm were acquired for folded and unfolded G-CSF wt and C3 in pH 4 and 7 buffers, using the same samples as used in Section 2.6.2.1. For acquiring emission spectra of unfolded and reduced G-CSF C3 in pH 4 and 7 buffers 1 mM TCEP was added to G-CSF C3 samples in 8.1 M urea and 650 µl of sample used.

2.6.3.2 Characterisation of G-CSF C3 in sub-denaturing urea concentrations

All measurements were recorded using 950 µl of sample in a 1 cm path length quartz cuvette (Hellma). For G-CSF C3 samples in pH 4 buffer excitation and emission slit widths of 1 nm and 1.5 nm, respectively were used, while for G-CSF C3 samples in pH 7 buffer excitation and emission slit widths of 1 nm were used. After excitation at 280 nm emission spectra in the region of 290-400 nm were acquired for G-CSF C3 in pH 4 buffer in the presence of 0-3 M urea with 1 M increments, as well as in pH 7 buffer in the presence of 0, 0.5 and 1-4 M urea with 1 M increments, using the same samples as used in Section 2.6.2.2. All measurements were recorded using 950 µl of sample in a 1 cm path length quartz cuvette (Hellma).

2.6.3.3 8-anilino-1-napthalenesulphonic acid (ANS) fluorescence

1 ml samples of 250 µM ANS in pH 4 buffer with and without 5 µM G-CSF C3 were set up at urea concentrations of 0-1 M with 0.2 M increments and 1-9 M with 1 M increments, then incubated at 25 ^{o}C overnight in a water bath. 1 ml samples of 250 µM ANS in pH 7 buffer with and without 2.5 µM G-CSF C3 were also set up at the same urea concentrations and incubated at 25 ^{o}C overnight.

All measurements were recorded using 500 µl of sample in a 1 cm path length quartz cuvette (Hellma). For samples in pH 4 buffer excitation and emission slit widths of 0.5 nm and 0.75 nm, respectively were used, whereas for samples in pH 7 buffer excitation and emission slit widths of 1.5 nm and 2 nm were used. After excitation at 389 nm emission spectra in the region of 400-600 nm were acquired for samples of 250 µM ANS with and without 5 µM G-CSF C3 in pH 4 buffer in the presence of 0-1 M urea in 0.2 M increments and 1-9 M urea in 1 M increments. Spectra were also acquired for samples of 250 µM ANS in pH 7 buffer with and without 2.5 µM G-CSF C3 in pH 7 buffer in the presence of 0-1 M urea in 0.2 M increments and 1-9 M urea in 1 M increments.

2.6.4 Urea denaturation of G-CSF C3

Stocks of 7 µM G-CSF C3 in pH 4 and 7 buffers in 0, 2, 4, 6, 8 and 9 M urea were prepared by mixing appropriate amounts of pH 4 and 7 buffers without urea, pH 4 and 7 buffers with 10 M urea and concentrated G-CSF C3 protein solutions at pH 4 and 7 as outlined in Table 2.2 and Table 2.3, respectively. These stocks were mixed in the correct ratios to prepare 1 ml samples of 7 µM G-CSF C3 at urea concentrations of 0-9 M with 0.2 M increments, which were subsequently incubated at 25 ^{o}C overnight.

Urea concentration of stock / M	Buffer volume added / ml	10 M urea volume added / ml	Volume of 521 µM G-CSF C3 added / ml	Total volume of stock / ml
0	7.9	0	0.1	8
2	9.45	2.4	0.15	12
4	7.05	4.8	0.15	12
6	4.65	7.2	0.15	12
8	1.875	8	0.125	10
9	0.35	3.6	0.05	4

Table 2.2 Volumes of pH 4 buffer solution with and without 10 M urea and concentrated protein solution required to make 7 µM G-CSF C3 pH 4 stocks for equilibrium denaturation.

Urea concentration of stock / M	Buffer volume added / ml	10 M urea volume added / ml	Volume of 341 µM G-CSF C3 protein added / ml	Total volume of stock / ml
0	7.84	0	0.16	8
2	9.36	2.4	0.24	12
4	6.96	4.8	0.24	12
6	4.56	7.2	0.24	12
8	1.8	8	0.2	10
9	0.4	4.5	0.1	5

Table 2.3 Volumes of pH 7 buffer solution with and without 10 M urea and concentrated protein solution required to make 7 µM G-CSF C3 pH 7 stocks for equilibrium denaturation.

2.6.4.1 Urea denaturation monitored by fluorescence emission

For each G-CSF C3 sample fluorescence emission intensity was measured using 950 µl of protein solution in a 1 cm path length quartz cuvette (Hellma) with five repeats of 10 s acquisition at 25 ^{o}C. Samples were left to equilibrate in the instrument for approximately one minute before acquisition. For G-CSF C3 samples in pH 4 buffer emission intensity was measured at 301 and 348 nm after excitation at 280 nm. For G-CSF C3 samples in pH 7 buffer emission intensity was measured at 320 nm after excitation at 280 nm. Emission intensities at 301 nm were measured using excitation and emission slit widths of 1 and 2 nm, respectively. Emission intensities at 320 and 348 nm were measured using excitation and emission slit widths of 1 nm. After measurement of emission intensities samples were recovered and incubated at 25 ^{o}C overnight.

2.6.4.2 Urea denaturation monitored by far-UV CD

Using the same G-CSF C3 samples as described in Section 2.6.4.1 ellipticity intensity at 222 nm (θ_{222}) was measured using 200 µl of sample in a 1 mm path length quartz cuvette (Hellma) with five repeats of 10 s acquisition and a bandwidth of 2.5 nm at 25 ^{o}C. Samples were left to equilibrate in the instrument for approximately one minute before acquisition.

2.6.4.3 Calculating the free energy of unfolding (ΔG^{o}_{UN}) from urea denaturation measurements

Fluorescence emission and θ_{222} intensities (see Sections 2.6.4.1 and 2.6.4.2, respectively) were plotted against urea concentration and fitted to Equation 2.9, which assumes a two-state transition between folded and unfolded states at equilibrium; where S_{obs} is the observed signal and represents either measured fluorescence emission or θ_{222} intensity, a and c are the signals of the denatured and native states in the absence of denaturant, respectively, b and d are the dependencies of the signal of the denatured and native states, respectively on the denaturant concentration, [D] is the denaturant concentration, ΔG_{UN} is the free energy of unfolding in the absence of denaturant and M_{UN} is the dependence of ΔG_{UN} on the denaturant concentration[312].

$$S_{obs} = \frac{\left((a[D] + b)e^{\left(\frac{\Delta G_{UN} - M_{UN}[D]}{RT}\right)} + (c[D] + d)\right)}{\left(1 + e^{\left(\frac{\Delta G_{UN} - M_{UN}[D]}{RT}\right)}\right)} \quad \textbf{Equation 2.9}$$

Using parameters a, b, c and d derived from Equation 2.9 fluorescence emission and θ_{222} intensities were normalised using Equation 2.10, converting S_{obs} values into fractions of folded protein (frac(F))[312].

$$\text{frac}(F) = \frac{S_{obs} - (a[D] + b)}{(c[D] + d) - (a[D] + b)} \quad \textbf{Equation 2.10}$$

2.7 Aggregation assays

2.7.1 G-CSF C3 aggregation at different protein concentrations

100 µl samples of G-CSF C3 at concentrations of 4.5 and 8.6 mg/ml in pH 4 buffer, plus G-CSF C3 at concentrations of 1.7 and 5.5 mg/ml in pH 7 buffer were prepared from the stocks described in Section 2.5.3 in triplicate in 1.5 ml Eppendorf tubes. These samples were centrifuged for 15 mins at 13300 rpm in a Labnet Spectrafuge 24D microcentrifuge at room temperature. 10 µl of supernatant was removed from each sample and diluted with 50 µl of pH 4 or 7 buffer to allow measurement of A_{280}. The measured A_{280} values of triplicate samples were averaged and standard deviations calculated for each protein concentration. An absorbance scan across wavelengths 190-600 nm was acquired for at least one of each set of triplicate samples to ensure measured A_{280} values were not being affected by light scattering from particulates. The solutions were then incubated at 37 ^{o}C in a water bath but samples were removed at regular intervals for centrifugation and measurement of A_{280} as described above. These periods of centrifugation and measurement of A_{280} lasted for approximately 30 mins and were not included in the time course of the experiment.

Proportions of aggregated G-CSF C3 remaining at these time intervals were calculated by subtracting the average A_{280} measured at a time interval from the average A_{280} measured before incubation at 37 ^{o}C, then dividing this value by the average A_{280} measured before incubation at 37 ^{o}C. Changes in the proportion of aggregated G-CSF C3 over time for G-CSF C3 in pH 7 buffer with a starting protein concentration of 5.5 mg/ml were fitted to a single exponential decay with a y-axis offset, while changes in the proportion of aggregated G-CSF C3 over time for other samples were fitted to a linear function.

2.7.2 G-CSF C3 aggregation at different urea concentrations

100 µl samples of 87 µM G-CSF C3 in pH 4 buffer, plus 70 µM G-CSF C3 in pH 7 buffer at urea concentrations of 0-7 M with 1 M increments were prepared in triplicate in 1.5 ml Eppendorf tubes. These samples were centrifuged for 15 mins at 13300 rpm in a Labnet Spectrafuge 24D microcentrifuge at room temperature. 10 µl of supernatant was removed from each sample and diluted with 50 µl of pH 4 or 7 buffer containing the same urea concentration as the G-CSF C3 sample to allow measurement of A_{280}. Measured A_{280} values were averaged and standard deviations calculated for triplicate samples of each protein concentration. An absorbance scan across wavelengths 190-600 nm was acquired for at least one of each set of triplicate samples to ensure measured A_{280} values were not being affected by light scattering from particulates. Samples were then incubated at 37 ^{o}C in a water bath but removed at regular time intervals for centrifugation and measurement of A_{280} as described above. These periods of centrifugation and measurement of A_{280} lasted for approximately 30 mins and were not included in the time course of the experiment.

Changes in the A_{280} over time for G-CSF C3 in pH 7 buffer in urea concentrations of 5-7 M urea with 1 M increments were fitted to a single exponential decay with a y-axis offset, while changes in the A_{280} over time for other G-CSF C3 samples were fitted to a linear function. Bar graphs showing the fractions of soluble G-CSF C3 remaining after certain time intervals were calculated by dividing the A_{280} measured at that time interval by the A_{280} measured before incubation at 37 ^{o}C. For G-CSF C3 samples in pH 7 buffer at urea concentrations of 0-7 M with 1 M increments, fractions of soluble G-CSF C3 were calculated after incubation for 24 h at 37 ^{o}C. For

G-CSF C3 samples in pH 4 buffer at urea concentrations of 0-7 M urea with 1 M increments, fractions of soluble G-CSF C3 were calculated after 27 h at 37 ^{o}C.

2.7.3 G-CSF C3 aggregation at different urea concentrations monitored by Dynamic Light Scattering (DLS)

Light scattering was measured using a WyattQELS^{TM} module integrated within a miniDawn^{TM} TREOS$^{®}$ instrument (Wyatt Technology, USA). All buffers and cleaning solutions were filtered and degassed by vacuum filtration through 0.22 µm filter membranes before use, as well as being injected into the 0.7 µl flow cell through a 0.22 µm filter. The flow cell was initially flushed with 10 ml of air followed by 5 ml of water by injection through the 'IN' port with waste tubing connected to the 'OUT' port. The flow cell was then washed with 1 ml of 1 M nitric acid followed by 1 ml of water by injection through the 'OUT' port with waste tubing connected to the 'IN' port. These wash steps were repeated until the light scattering trace reached a baseline, which corresponded to rescaling of the light scattering values to four decimal places for at least 5 mins. Before injection of each G-CSF C3 sample the flow cell was equilibrated by injection of 1 ml of sample buffer through the 'OUT' port followed by a five minute delay to ensure the baseline for the light scattering trace was maintained.

600 µl samples of 83 and 85 µM G-CSF C3 in pH 4 buffer in the presence and absence of 3 M urea, respectively, or 67 and 66 µM G-CSF C3 in pH 7 buffer in the presence and absence of 4 M urea, respectively, were prepared in 1.5 ml Eppendorf tubes. These samples were centrifuged for 15 mins at 13300 rpm in a Labnet Spectrafuge 24D microcentrifuge at room temperature. 10 µl of supernatant was removed from each sample and diluted with 50 µl of pH 4 or 7 buffer containing the same urea concentration as the G-CSF C3 sample to allow measurement of absorbance across wavelengths 190-600 nm. This procedure ensured that measurements were not affected by light scattering. At this point, recording of the light scattering trace by ASTRA version 6 (Wyatt Technology, USA) was started and allowed to proceed for 5 mins before injection of G-CSF C3 samples in order to provide a pre-injection baseline. 290 µl of supernatant was injected into the flow cell through the 'OUT' port for light scattering measurement. Sample injection was followed by a five minute delay for recording of light scattering, after which time 1 ml

of sample buffer was injected into the flow cell through the 'OUT' port. Recording was then allowed to proceed for a further 5 mins to provide a post-injection baseline.

The remaining 300 µl of supernatants were incubated at 37 ^{o}C in a water bath. After 24 h, these samples were centrifuged and the supernatants analysed by DLS as outlined above.

A three minute interval within each five minute recording of light scattering data was chosen for analysis by the ASTRA version 6 software in order to avoid 'spikes' in the light scattering trace resulting from injections. Light scattering over these time intervals was fitted to a second order autocorrelation function (Equation 2.11) in ASTRA version 6; where $g^2(\tau)$ represents the second order autocorrelation function, I(t) represents the intensity of scattered light at time t, I(t + τ) represents the intensity of scattered light after time delay τ and the brackets ($\langle\ \rangle$) indicate averaging over all time[316].

$$g^2(\tau) = \frac{\langle I(t)I(t+\tau)\rangle}{\langle I(t)\rangle^2} \qquad \textbf{Equation 2.11}$$

The value of $g^2(\tau)$ obtained from Equation 2.11 was then used in Equation 2.12, which assumes the population of the sample is monodisperse in size to obtain the decay rate of the autocorrelation function (Γ); where B is the baseline of the autocorrelation function at infinite delay and β is the amplitude of the autocorrelation function at zero delay[316].

$$g^2(\tau) = B + \beta e^{-2\Gamma\tau} \qquad \textbf{Equation 2.12}$$

The value of Γ obtained from Equation 2.12 was then used in Equation 2.13 to obtain the translational diffusion coefficient (D_t) of the species within the sample; where q is the magnitude of the scattering vector and is calculated using Equation 2.14; where N is the refractive index of the sample, λ is the wavelength of the incident light and θ is the angle of the detector relative to the flow cell[316].

$$D_t = \frac{\Gamma}{q^2}$$ **Equation 2.13**

$$q = \frac{4\pi N}{\lambda} \sin\left(\frac{\theta}{2}\right)$$ **Equation 2.14**

The value of D_t obtained from Equation 2.13 was then used in the Stokes-Einstein equation (Equation 2.15) to obtain the hydrodynamic radius (R_h) of the species within the sample; where k_B is Boltzmann's constant, T is temperature and η is the viscosity of the sample[316].

$$R_h = \frac{k_B T}{6\pi\eta D_t}$$ **Equation 2.15**

Attempting to fit light scattering data for G-CSF C3 samples using Equation 2.12 in ASTRA version 6 assumes a monodisperse solution. If there are multiple sizes of species present i.e. the sample is polydisperse then each species of a particular size will decay at a different rate[316]. For polydisperse samples, the $g^2(\tau)$ value obtained from fitting of light scattering intensities to Equation 2.11 was used in Equation 2.16 to obtain a first order autocorrelation function ($g^1(\tau)$); where $\zeta(\tau)$ is the experimental noise after time delay τ. The $g^1(\tau)$ value is subsequently used in Equation 2.17, which is the weighted sum of the exponential decays of each species of a different R_h in the population; where $G(\Gamma)$ is the distribution function of the species with different decay rates (Γ) and $G(\Gamma)d\Gamma$ is the fraction of the total intensity of scattering by each species[317].

$$g^2(\tau) = (g^1(\tau))^2 + 1 + \zeta(\tau)$$ **Equation 2.16**

$$g^1(\tau) = \int G(\Gamma) e^{-\Gamma\tau}\, d\Gamma$$ **Equation 2.17**

A regularisation analysis[318] was then performed by the DYNALS algorithm in ASTRA version 6 using the distribution function G(Γ) calculated in Equation 2.17. The resulting plot shows the fractional intensity of species with different R_h values[318]. This regularisation analysis was carried out on all G-CSF C3 DLS samples to give an idea of their polydispersity.

R_h values calculated from regularisation analysis of DLS data were used in Equation 2.18 to estimate the molecular masses of species, although this assumes these species are spherical; where M_r is the molecular mass of the species, ρ is the average density of a protein, which is assumed to be 1.37 g cm^{-3}, N_A is Avogadro's number and r_w is the radius of the hydration layer surrounding the protein, which is assumed to be 3.2 Å for one hydration shell[319].

$$R_h = \sqrt[3]{\frac{3M_r}{(4\pi\rho N_A)} + r_w} \qquad \textbf{Equation 2.18}$$

2.8 Electron Microscopy

Transmission electron microscope images of aggregates were acquired by Dr. Katie Stewart, University of Leeds on a JEOL JEM-1400 transmission electron microscope (JEOL Ltd, Japan). Aggregates were prepared as described in Section 2.9.4 and thawed, then 10 µl of aggregated samples were placed on formvar/carbon-coated copper grids (provided by Martin Fuller, Electron Microscopy facility, University of Leeds). After 30 s incubation, excess sample was removed by blotting with filter paper followed by negative staining with 10 µl of 2% (*w/v*) aqueous uranyl acetate. Grids were imaged after a 30 second drying period.

2.9 Nuclear Magnetic Resonance (NMR) spectroscopy

NMR experiments were carried out in partnership with Dr. Gary Thompson and Dr. Arnout Kalverda in the NMR facility, University of Leeds. In particular, transverse relaxation experiments, ^{13}C-correlated spectra used for backbone assignment and ^{1}H-^{15}N SOFAST HMQC spectra used to monitor hydrogen exchange were optimised and carried out under the supervision of Dr. Kalverda.

NMR spectra were acquired using a Varian Inova spectrometer (Agilent) operating at a proton resonance frequency of 600 MHz and equipped with a room temperature probe unless otherwise stated. For NMR experiments 10 % (*v/v*) D_2O was added to G-CSF samples, resulting in the buffer components in the pH 4 and 7 buffers being diluted by 10 %. Spectra were acquired using 300 µl of sample in a Shigemi microtube matched to D_2O (Sigma Life Sciences, USA) at 25 ^{o}C unless otherwise stated. All NMR spectra were processed using NMRPipe[320] and analysed using CCPN Analysis[321].

2.9.1 Backbone assignments by double- and triple-resonance NMR

Multi-dimensional NMR relies on the transfer of magnetisation between nuclei, specifically from nuclei with a high gyromagnetic ratio (e.g. ^{1}H) to nuclei with a low gyromagnetic ratio such as ^{15}N and ^{13}C. The gyromagnetic ratio of a nucleus is the ratio of its magnetic dipole moment to its angular momentum and this ratio governs the precession frequency of a nucleus in a magnetic field according to Equation 2.19; where ω_0 represents the precession frequency of the nucleus i.e. the Larmor frequency, γ represents the gyromagnetic ratio and B_0 represents the magnetic field strength. Consequently, nuclei with a higher γ are more sensitive to detection by NMR. Shielding of nuclei from the magnetic field (B_0) by local electron density alters ω, giving rise to the different chemical shifts observed in NMR spectra.

$$\omega_0 = -\gamma B_0 \qquad \textbf{Equation 2.19}$$

In two- and three-dimensional NMR, transfer of magnetisation from more to less sensitive nuclei is achieved through insensitive nuclei enhancement by polarisation transfer (INEPT) radiofrquency pulse sequences[322], examples of which can be seen within the pulse sequence of the heteronuclear single quantum coherence (HSQC) experiment[319,323] (Figure 2.1).

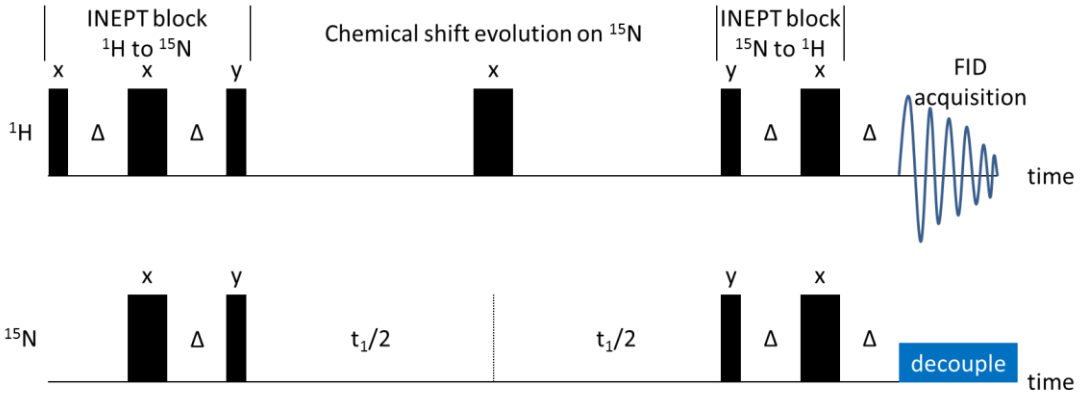


Figure 2.1 A generic pulse scheme for the ^{1}H-^{15}N HSQC experiment. Narrow and wide bars represent 90° and 180° pulses, respectively, along x- or y-axes. Δ represents the time delay between pulses and is equal to $(4J)^{-1}$ for the J-coupling of the nuclei being studied, in this case ^{1}H and ^{15}N linked by an amine bond. The first INEPT block transfers magnetisation from ^{1}H to ^{15}N nuclei. The chemical shift is then allowed to evolve on the ^{15}N nucleus for a time period t_1, during which time a 180° pulse decouples ^{1}H and ^{15}N. A second INEPT block transfers magnetisation back from ^{15}N to ^{1}H nuclei for detection by acquisition of a Free Induction Decay (FID). During acquisition there is ^{15}N broadband decoupling to ensure only one peak is observed in the ^{1}H-^{15}N HSQC spectrum per amine bond[319,323].

Introduction of nuclei with spin quantum number I = ½ into a magnetic field results in alignment of their spins parallel to the magnetic field, either in the same direction along the z-axis (corresponding to the low energy +½ or α state) or in the opposite direction along the z-axis (corresponding to the high energy –½ or β state). The J-coupling between two nuclei connected by a covalent bond is the difference in the resonance frequency of one nucleus that results from the other nucleus in the bond being in the α or β spin state, which is caused by differing interactions of these spin states with the local electron density of the bond[324]. INEPT blocks exploit the J-couplings of nuclei connected by covalent bonds. Figure 2.2 shows J-couplings for covalent bonds commonly used in triple-resonance NMR for backbone assignment.

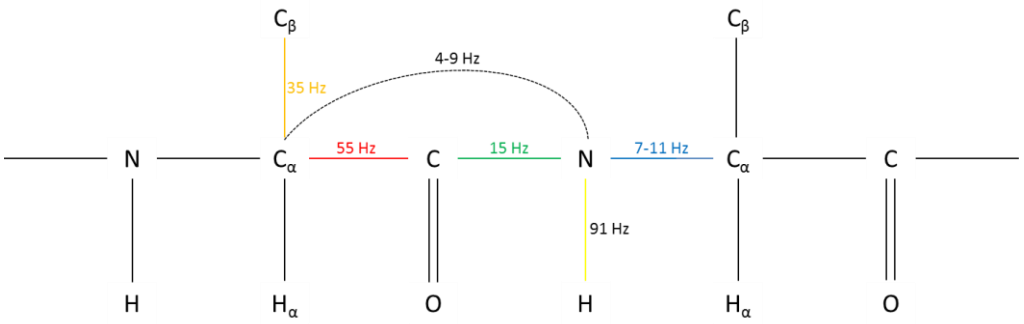


Figure 2.2 J-coupling constants for nuclei linked by covalent bonds commonly used for assignment of the polypeptide backbone. Frequencies of couplings J_{NH} ~ 91 Hz (yellow), J_{NCO} ~ 15 Hz (green), $J_{C\alpha CO}$ ~ 55 Hz (red), $J_{C\alpha C\beta}$ ~ 35 Hz and $J_{NC\alpha}$ ~ 7-11 Hz (blue) are depicted[319].

Acquisition of triple-resonance spectra involves the transfer of magnetisation between ^{1}H, ^{15}N and ^{13}C nuclei, resulting in an HSQC spectrum with a third ^{13}C dimension perpendicular to the ^{1}H-^{15}N plane. For example, an HNCA spectrum results in two $^{13}C_{\alpha}$ peaks for every backbone amide proton as shown in Figure 2.3. Due to the stronger coupling of the amide nitrogen nucleus to the alpha carbon nucleus of the same amino acid residue, denoted the i residue, compared to the preceding residue, denoted the i-1 residue, an HNCA spectrum contains higher and lower intensity peaks for the alpha carbon nuclei of the i and i-1 residues, respectively[325-327]. On the other hand, an HN(CO)CA spectrum only gives a single $^{13}C_{\alpha}$ peak for every backbone amide proton corresponding to the i-1 residue through utilisation of the J-coupling between the amide nitrogen and the connecting carbonyl group[326,328]. Thus, HNCA and HN(CO)CA spectra can be used in conjunction to distinguish $^{13}C_{\alpha}$ peaks belonging to i and i-1 residues as well as link together residues based on the amino acid sequence.

The same principle applies to the use of HNCO[325,326,329] and HN(CA)CO[330] spectra, which produce one and two ^{13}CO peaks, respectively with overlapping peaks for the carbonyl carbon nucleus of the i-1 residue. Finally, a triple-resonance spectrum that produces $^{13}C_{\alpha}$ and $^{13}C_{\beta}$ peaks such as an CBCA(CO)NH[331] spectrum (Figure 2.3) can help provide the identity of amino acid residues, based on the characteristic chemical shift values of $^{13}C_{\alpha}$ and $^{13}C_{\beta}$ peaks for particular residues.

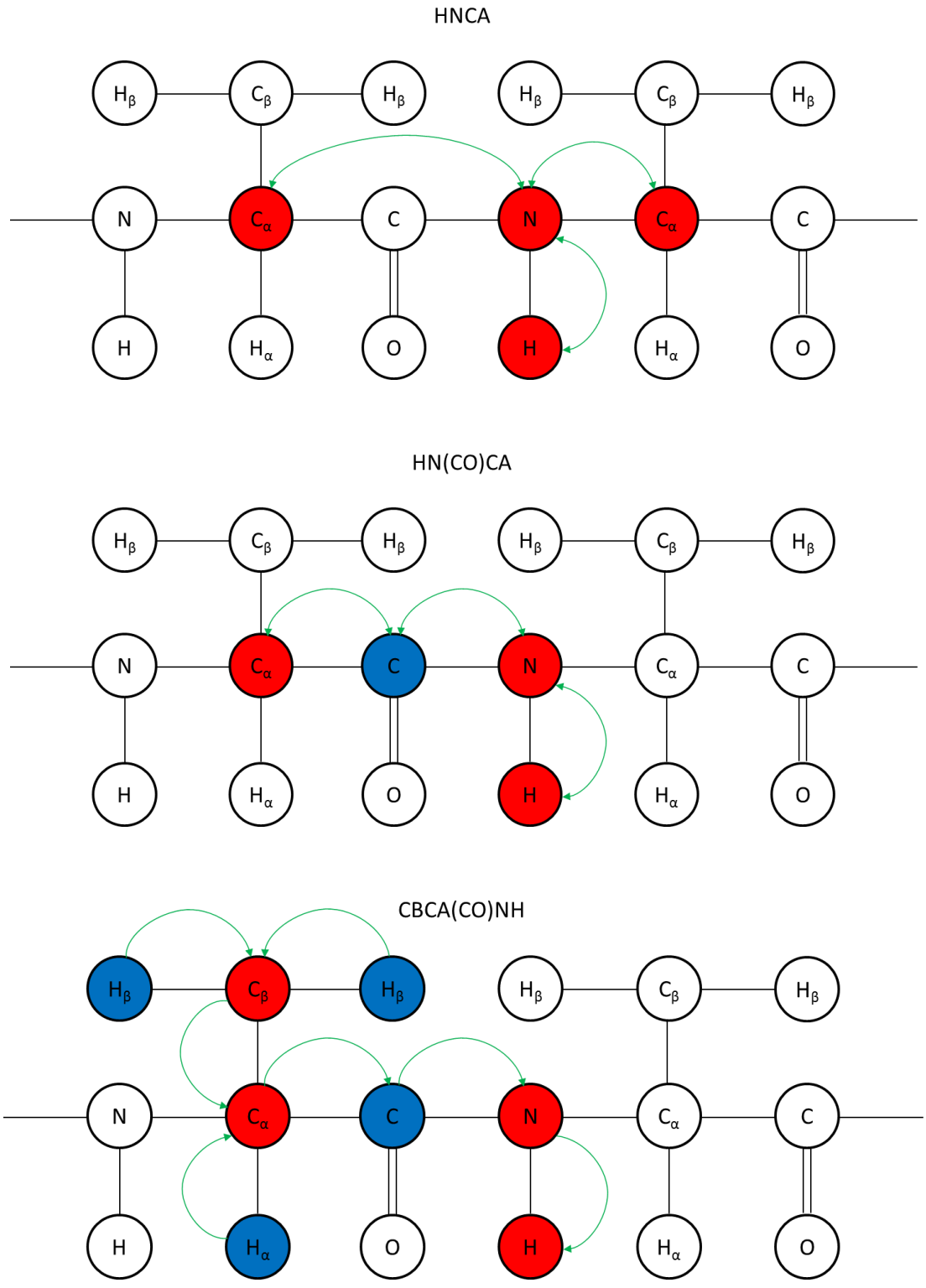


Figure 2.3 Magnetisation pathways for HNCA, HN(CO)CA and CBCA(CO)NH triple-resonance NMR spectra. Chemical shifts are evolved on nuclei in red producing peaks in these spectra, while nuclei in blue are used to transfer magnetisation via J-couplings but the chemical shift is not evolved for these nuclei. Double-headed arrows in the HNCA[325-327] and HN(CO)CA[330] schematics indicate initial excitation of the amide proton, which is subsequently transferred to nuclei in red and blue as indicated. Magnetisation is then transferred back to the amide proton for detection. Single-headed arrows in the CBCA(CO)NH[331] schematic indicate excitation of H_α and H_β protons, which is then transferred to nuclei in red and blue as indicated finishing at the amide proton for detection.

Backbone assignments of 473 and 176 µM G-CSF C3 were carried out in pH 4 buffer in the absence and presence of 3 M urea, respectively using the NMR spectra and acquisition parameters outlined in Table 2.4 and Table 2.5. Backbone assignments of 458 and 208 µM G-CSF C3 were carried out in pH 7 buffer in the absence and presence of 4 M urea, respectively using the NMR spectra and acquisition parameters outlined in Table 2.6 and Table 2.7.

Parameter	**HSQC**	**HNCA**	**HN(CO)CA**	**HNCO**	**HN(CA)CO**	**CBCA(CO)NH**
np	2048	2048	2048	2048	2048	2048
nt	48	16	16	16	32	28
sw / Hz	7225	9615	9615	9615	9615	9615
sw (^{15}N) / Hz	1700	1600	1600	2500	1600	1600
sw (^{13}C) / Hz	N/A	4000	4000	3000	3000	10500
ni (^{15}N)	256	32	32	32	32	32
ni (^{13}C)	N/A	32	32	32	32	32
d1 / s	1	0.97	1	0.95	1	1.04
Acquisition time / h	8.5	21	22	21	43	39.5

Table 2.4 NMR spectra and acquisition parameters used for assignment of 473 µM G-CSF C3 in pH 4 buffer without urea. np refers to the number of data points (real+imaginary) acquired along the FID in the directly detected dimension; nt refers to the number of transients, meaning the number of scans of each data point along the FID; sw refers to the spectral width for detection in the dimension shown; ni refers to the number of increments of the evolution time t_1 (see Figure 2.1), corresponding to the number of FIDs (complex points) recorded in the indirect dimensions shown; d1 refers to the time delay after acquisition of the FID to allow recovery of equilibrium magnetisation before the pulse sequence is applied.

Parameter	HSQC	HNCA	CBCA(CO)NH
np	2048	2048	2048
nt	48	16	32
sw / Hz	7225	9615	9615
sw (^{15}N) / Hz	1600	1600	1600
sw (^{13}C) / Hz	N/A	4000	10500
ni (^{15}N)	256	32	32
ni (^{13}C)	N/A	32	32
d1 / s	1	0.97	1.04
Acquisition time / h	8.5	21	45

Table 2.5 NMR spectra and acquisition parameters used for assignment of 176 µM G-CSF C3 in pH 4 buffer with 3 M urea. See the legend of Table 2.4 for a description of each parameter.

Parameter	HSQC	HNCA	HN(CO)CA	HNCO	HN(CA)CO	CBCA(CO)NH
np	2048	2048	2048	2048	2048	2048
nt	64	32	32	8	64	64
sw / Hz	7225	9615	9615	9615	9615	9615
sw (^{15}N) / Hz	1700	1600	1600	2500	1600	1600
sw (^{13}C) / Hz	N/A	4000	4000	3000	3000	10500
ni (^{15}N)	232	32	32	32	28	32
ni (^{13}C)	N/A	32	32	36	30	32
d1 / s	1	0.97	1	1	1	1.04
Acquisition time / h	10	42	44	12	70.5	90

Table 2.6 NMR spectra and acquisition parameters used for assignment of 458 µM G-CSF C3 in pH 7 buffer without urea. See the legend of Table 2.4 for a description of each parameter.

Parameter	HSQC	HNCA	CBCA(CO)NH
np	2048	2048	2048
nt	64	32	32
sw / Hz	7225	9615	9615
sw (^{15}N) / Hz	1600	1600	1600
sw (^{13}C) / Hz	N/A	4000	10500
ni (^{15}N)	232	32	32
ni (^{13}C)	N/A	32	32
d1 / s	1	0.97	1.04
Acquisition time / h	10	42	90

Table 2.7 NMR spectra and acquisition parameters used for assignment of 208 µM G-CSF C3 in pH 7 buffer with 4 M urea. See the legend of Table 2.4 for a description of each parameter.

Backbone assignments of amide proton and nitrogen atoms plus alpha, beta and carbonyl carbon atoms of G-CSF C3 in pH 4 buffer were aided by previous backbone assignments of G-CSF wt by Zink *et al*.[283] and Werner *et al*.[288], which in turn assisted assignment of the same atoms of G-CSF C3 in pH 7 buffer. Backbone assignments of amide proton and nitrogen atoms of G-CSF C3 in pH 4 and 7 buffers in the presence of urea were aided by urea titrations monitored by ^{1}H-^{15}N HSQC spectra outlined in Section 2.9.3.

Assignments of backbone amide proton and nitrogen atoms of G-CSF wt in pH 4 buffer were achieved using previous backbone assignments of G-CSF wt by Zink *et al*.[283] and Werner *et al*.[288]. 111 µM G-CSF wt in pH 4 buffer was used to acquire an ^{1}H-^{15}N HSQC spectrum for backbone assignment using the parameters outlined in Table 2.8.

Parameter	HSQC
np	1024
nt	40
sw / Hz	8621
sw (^{15}N) / Hz	1600
ni (^{15}N)	256
d1 / s	0.9
Acquisition time / h	6

Table 2.8 Acquisition parameters for ^{1}H-^{15}N HSQC spectrum used for assignment of 111 µM G-CSF wt in pH 4 buffer. See the legend of Table 2.4 for a description of each parameter.

2.9.2 Comparison of G-CSF C3 from 'peak' and 'tail' elution pools of cation exchange chromatography by ^{1}H-^{15}N HSQC spectra

^{1}H-^{15}N HSQC spectra of 173 and 177 µM G-CSF C3 in pH 4 buffer obtained from 'peak' and 'tail' elution pools of cation exchange chromatography, respectively (see Section 2.5.3) were acquired using the parameters outlined in Table 2.9, in order to elucidate any structural differences between G-CSF C3 obtained from these different elution pools.

Parameter	HSQC
np	1024
nt	16
sw / Hz	8621
sw (^{15}N) / Hz	1600
ni (^{15}N)	256
d1 / s	0.9
Acquisition time / h	2.5

Table 2.9 Acquisition parameters for ^{1}H-^{15}N HSQC spectra used for assignment of 173 and 177 µM G-CSF C3 in pH 4 buffer obtained from 'peak' and 'tail' elution pools of cation exchange chromatography. See the legend of Table 2.4 for a description of each parameter.

2.9.3 Urea titrations of G-CSF C3 monitored by ^{1}H-^{15}N HSQC spectra

^{1}H-^{15}N HSQC spectra of 103 and 81 µM G-CSF C3 in pH 4 and 7 buffers, respectively were acquired in the presence of 0-7 M urea with 0.5 M increments. Spectra were acquired using the parameters outlined in Table 2.10.

Parameter	HSQC
np	2048
nt	4
sw / Hz	9615
sw (^{15}N) / Hz	1700
ni (^{15}N)	128
d1 / s	1
Acquisition time / mins	20

Table 2.10 Acquisition parameters for ^{1}H-^{15}N HSQC spectra used to monitor urea titration of G-CSF C3. 103 and 81 µM G-CSF C3 were used in pH 4 and 7 buffers, respectively in the presence of 0-7 M urea with 0.5 M increments. See the legend of Table 2.4 for a description of each parameter.

2.9.4 G-CSF C3 aggregation at different urea concentrations monitored by ^{1}H-^{15}N HSQC spectra

^{1}H-^{15}N HSQC spectra were acquired using a Varian Inova spectrometer (Agilent) at 37 ^{o}C operating at a proton resonance frequency of 750 MHz and equipped with a cryogenic probe. Assignments for backbone amide proton and nitrogen nuclei were transferred from 25 to 37 ^{o}C by acquisition of ^{1}H-^{15}N HSQC spectra at 3 ^{o}C temperature intervals within this temperature range. Spectra of 78 and 85 µM G-CSF C3 in pH 4 buffer in the presence and absence of 3 M urea, respectively, plus 63 and 65 µM G-CSF C3 in pH 7 buffer in the presence and absence of 4 M urea, respectively were acquired using the parameters outlined in Table 2.11. Samples were left to equilibrate at each temperature interval for at least 1 h before starting acquisition.

Parameter	HSQC
np	2048
nt	8
sw / Hz	11962
sw (^{15}N) / Hz	1971
ni (^{15}N)	128
d1 / s	1
Acquisition time / mins	40

Table 2.11 Acquisition parameters for ^{1}H-^{15}N HSQC spectra used to monitor temperature titration of G-CSF C3. 78 and 85 µM G-CSF C3 were used in pH 4 buffer in the presence and absence of 3 M urea, respectively, while in pH 7 buffer 63 and 65 µM G-CSF C3 were used in the presence and absence of 4 M urea, respectively. See the legend of Table 2.4 for a description of each parameter.

For monitoring of G-CSF C3 aggregation, spectra of 83 and 85 µM G-CSF C3 in pH 4 buffer in the presence and absence of 3 M urea, respectively, or of 67 and 66 µM G-CSF C3 in pH 7 buffer in the presence and absence of 4 M urea, respectively were acquired using the parameters outlined in Table 2.12. After an overall acquisition period of 24 h at 37 ^{o}C samples were snap frozen using dry ice and ethanol and stored at -80 ^{o}C. The resulting losses of amide proton signals for G-

CSF C3 aggregation assay samples in pH 7 buffer with and without 4 M urea were fitted to a single exponential decay with a y-axis offset using in-house software written in Python by Dr. Theodoros Karamanos. Error calculation was performed based on the experimental noise determined by CCPN Analysis[321].

The amide proton peak intensities from spectra of G-CSF C3 aggregation assays carried out in pH 4 buffer with or without 3 M urea could not be fitted to a single exponential decay so these data were not included in analysis. However, insoluble aggregates from these samples were analysed by electron microscopy (EM), along with insoluble aggregates from pH 7 aggregation assay samples (see Section 2.8).

Parameter	HSQC
np	2118
nt	6
sw / Hz	11962
sw (^{15}N) / Hz	1971
ni (^{15}N)	130
d1 / s	1
Acquisition time / mins	30

Table 2.12 Acquisition parameters for ^{1}H-^{15}N HSQC spectra used to monitor aggregation of G-CSF C3 at 37 ^{o}C. 83 and 85 µM G-CSF C3 were used in pH 4 buffer in the presence and absence of 3 M urea, respectively, while in pH 7 buffer 67 and 66 µM G-CSF C3 were used in the presence and absence of 4 M urea, respectively. See the legend of Table 2.4 for a description of each parameter.

2.9.5 ANS binding to G-CSF C3 monitored by ^{1}H-^{15}N HSQC spectra

^{1}H-^{15}N HSQC spectra were acquired using a Varian Inova spectrometer (Agilent) at 25 ^{o}C operating at a proton resonance frequency of 750 MHz and equipped with a cryogenic probe. Spectra of 85 and 65 µM G-CSF C3 in pH 4 and 7 buffers respectively, in the presence of 700 µM ANS were acquired using the parameters outlined in Table 2.13.

Parameter	HSQC
np	2048
nt	12
sw / Hz	11962
sw (^{15}N) / Hz	1971
ni (^{15}N)	256
d1 / s	1
Acquisition time / h	2

Table 2.13 Acquisition parameters for ^{1}H-^{15}N HSQC spectra used to monitor ANS binding of G-CSF C3 at 25 ^{o}C. 85 and 65 µM G-CSF C3 in pH 4 and 7 buffers, respectively in the presence of 700 µM ANS were used. See the legend of Table 2.4 for a description of each parameter.

2.9.6 Calculating chemical shift perturbations (CSPs) from amide proton chemical shifts

CSPs were mainly calculated in order to probe for conformational changes of G-CSF C3 in pH 4 and 7 buffer upon addition of sub-denaturing urea concentrations, as well as assessing structural differences between G-CSF C3 at pH 4 and 7 in the absence of urea. ^{1}H and ^{15}N chemical shifts of backbone amide proton and nitrogen nuclei for G-CSF C3 at pH 4 in 0, 0.5 and 3 M urea, as well as G-CSF C3 at pH 7 in 0, 0.5 and 4 M urea were obtained from backbone assignments (see Section 2.9.1) and urea titrations (see Section 2.9.3). CSPs were also calculated to probe for conformational changes of G-CSF C3 during temperature titration of ^{1}H-^{15}N HSQC spectra from 25-37 ^{o}C (see Section 2.9.4). In addition, CSPs were calculated to determine regions of G-CSF C3 that may be involved in binding to ANS. Changes in the chemical shifts of amide protons and nitrogens for G-CSF C3 due to the presence of ANS were assessed by acquiring ^{1}H-^{15}N HSQC spectra (see Section 2.9.5).

The structural similarity between recombinantly expressed G-CSF wt used in this study to previous studies of G-CSF wt by NMR was also assessed by calculating CSPs. For assessing the structural similarity of G-CSF wt produced in this study to protein produced in previous studies of G-CSF wt by NMR, the centres of the

nitrogen and proton dimensions of the ^{1}H-^{15}N HSQC spectrum of G-CSF wt were adjusted until chemical shifts for the backbone amide nitrogen and proton of Thr2 matched the chemical shifts reported for this residue by Zink *et al*.[283] and Werner *et al*.[288]. Thr2 was chosen because this residue is within the unstructured twelve residue N-terminal segment, meaning any structural differences between G-CSF wt used in this study and wild-type protein used in previous studies would be unlikely to affect this residue. CSPs were also used to determine the structural similarity between G-CSF wt and C3 at pH 4 prepared in this study.

CSPs were calculated using Equation 2.20; where the scaling factor of 0.14 for differences in ^{15}N chemical shifts represents the ratio of the range of ^{1}H chemical shifts (in ppm) occupied to the range of ^{15}N chemical shifts (in ppm) occupied[332].

$$\mathrm{CSP} = \sqrt{(\Delta\delta^{1}\mathrm{H})^{2} + (0.14\Delta\delta^{15}\mathrm{N})^{2}} \qquad \textbf{Equation 2.20}$$

2.9.7 Hydrogen/deuterium exchange monitored by ^{1}H-^{15}N HSQC spectra

In natively folded proteins exchangeable protons are involved in hydrogen bonding, either with the solvent or within the secondary and tertiary structures of proteins. Consequently, exchange of these protons with solvent deuterons relies on a structural isomerisation represented by Equation 2.21[333-335].

$$\mathrm{NH_{closed}} \underset{k_{cl}}{\overset{k_{op}}{\rightleftharpoons}} \mathrm{NH_{open}} \overset{k_{ch}}{\leftrightarrow} \mathrm{ND} \qquad \textbf{Equation 2.21}$$

The observed rate constant (k_{obs}) for hydrogen exchange depends on the rates of the opening (k_{op}) and closing (k_{cl}) reactions as well as the intrinsic rate of hydrogen exchange (k_{ch}) according to Equation 2.22[333-335]. k_{ch} represents the rate of hydrogen exchange of the amino acid sequence of the fully unfolded and solvent exposed protein. k_{ch} is heavily dependent on pH and temperature[335] as well as amino acid sequence[336], the effects of which have all been quantified precisely allowing calculation of k_{ch} using freely available computational tools such as CIntX[337].

$$k_{obs} = \frac{k_{op}k_{ch}}{(k_{op} + k_{cl} + k_{ch})}$$ **Equation 2.22**

For a natively folded protein k_{op} is assumed to be much lower than k_{cl}, meaning Equation 2.22 can be simplified to Equation 2.23.

$$k_{obs} = \frac{k_{op}k_{ch}}{(k_{cl} + k_{ch})}$$ **Equation 2.23**

Under conditions where the intrinsic rate of hydrogen exchange is low i.e. $k_{ch} << k_{cl}$ the value of k_{obs} becomes dependent on the proportion of exchangeable protons that are exposed to solvent, which is represented by the opening equilibrium constant (K_{op}) in Equation 2.24. Conversely, when the intrinsic rate is high i.e. $k_{ch} > k_{cl}$ hydrogen exchange occurs as soon as exchangeable protons become exposed to solvent, meaning k_{obs} is only limited by k_{op} (Equation 2.25). Due to the bimolecular and unimolecular nature of the exchange mechanisms depicted in Equation 2.24 and Equation 2.25 they have been termed EX2 and EX1, respectively[334].

EX2 $$k_{obs} = K_{op}k_{ch}$$ **Equation 2.24**

EX1 $$k_{obs} = k_{op}$$ **Equation 2.25**

Buffer components required to prepare 10 ml of pH 4 and 7 buffer were dissolved in 4 ml of D_2O and lyophilised. Buffer components required to prepare 10 ml of pH 4 buffer with 3.33 M urea and pH 7 buffer with 4.44 M urea were dissolved in 9 ml of D_2O and lyophilised. After freeze-drying buffer components were resuspended and made up to a final volume of 10 ml using D_2O and the pD was adjusted to 4.8 or 7.8 for pH 4 and 7 buffer components, respectively. pD 4.8 and pD 7.8 buffers exhibited pH meter readings of 4.4 and 7.4, respectively[338].

A concentrated G-CSF C3 pH 4 stock was diluted to a final concentration of 52 µM in 500 µl of pD 4.8 deuterated buffer with and without a final urea concentration of 3

M, followed by brief manual mixing to initiate hydrogen/deuterium exchange reactions. A concentrated G-CSF C3 pH 7 stock was also diluted to a final concentration of 52 µM in 300-350 µl of pD 7.8 deuterated buffer with and without a final urea concentration of 4 M, followed by brief manual mixing to initiate exchange.

After mixing, hydrogen exchange reaction samples were placed into a Shigemi microtube and inserted into the NMR spectrometer. Hydrogen exchange was monitored based on measuring peak intensities of amide protons using ^{1}H-^{15}N band-selective optimised-flip-angle short-transient (SOFAST) HMQC spectra[339], the parameters of which are outlined in Table 2.14. The dead-time between initiating exchange and beginning acquisition was 7-8 mins. In total, spectra of G-CSF C3 samples in pD 4.8 buffer were acquired over 24 h, while spectra of G-CSF C3 samples in pD 7.8 buffer were acquired over 15 h.

Parameter	pD 4.8 samples	pD 7.8 samples
flip angle	120^{o}	120^{o}
np	512	512
nt	24	8
sw / Hz	8503	8503
sw (^{15}N) / Hz	1600	1600
ni	68	60
d1 / s	0.302	0.248
Acquisition time / mins	20	5

Table 2.14 Acquisition parameters for ^{1}H-^{15}N SOFAST-HMQC spectra used to monitor amide proton hydrogen exchange reactions of G-CSF C3. 52 µM G-CSF C3 was used in pD 4.8 buffer in the presence and absence of 3 M urea as well as in pD 7.8 buffer in the presence and absence of 4 M urea. Flip angle refers to the degree of inversion of amide protons relative to the z-axis. See the legend of Table 2.4 for a description of other parameters.

The resulting loss of amide proton signal was fitted to a single exponential decay with a y-axis offset using in-house software written in Python by Dr. Karamanos. Error calculation was performed based on the experimental noise determined by CCPN Analysis[321]. Measured rate constants (k_{obs}) were used with intrinsic rate constants (k_{ch}) calculated using CIntX[337] in Equation 2.26 to calculate protection

factors (PFs)[335]. The ratio of ln(PF) for G-CSF C3 in pD 4.8 and 7.8 buffers to ln(PF) in the urea concentrations specified was then calculated to elucidate residues with altered solvent exposure upon addition of urea.

$$PF = \frac{k_{ch}}{k_{obs}} \qquad \textbf{Equation 2.26}$$

2.9.8 Amide nitrogen transverse relaxation (R_2)

Introduction of nuclei with spin quantum number I = ½ into a magnetic field causes the energy levels represented by the α and β spin states to be populated according to the Boltzmann distribution, resulting in a slightly higher population of spins in the α state compared to the β state and causing net magnetisation along the z-axis. Excitation of nuclei by application of a radiofrequency pulse disturbs the Boltzmann distribution, giving rise to precession of spins perpendicular to the axis of the magnetic field in the xy plane. Relaxation encompasses the processes by which the Boltzmann distribution of spins is restored, meaning net magnetisation is lost in the xy plane and restored along the z-axis.

The three processes by which relaxation occurs are longitudinal and transverse relaxation, the rates of which are denoted R_1 and R_2, respectively and the nuclear Overhauser effect (nOe). Longitudinal relaxation is the return of magnetisation to the z-axis, while transverse relaxation is the loss of coherence in the xy plane. Consequently, R_1 determines the time period before nuclei can be excited again by a radiofrequency pulse and R_2 determines the width of peaks. The nOe is the exchange of magnetisation between nuclei through space and is proportional to the inverse of the sixth power of the distance between nuclei.

For spin I = ½ nuclei, the two mechanisms by which relaxation occurs are dipolar relaxation and chemical shift anisotropy (CSA). Dipolar relaxation arises from the interaction of the magnetic field of one spin with another spin. The strength of this dipolar coupling is proportional to the gyromagnetic ratios of the interacting nuclei and the inverse cube of the distance between nuclei, while it is also affected by the direction of the vector connecting the two nuclei relative to the magnetic field. The relaxation induced by dipolar coupling is proportional to the square of the dipolar

coupling, i.e. to the inverse sixth power of the distance between nuclei. Fluctuations of the local magnetic field caused by either molecular tumbling of nuclei in solution or protein motions are the cause of dipolar relaxation in proteins. Dipolar relaxation also causes transition of spins between α and β states, giving rise to the nOe.

The electronic density around a nucleus, which gives rise to the chemical shift, is not spherical (isotropic), but varies as the orientation of the nucleus changes with respect to the magnetic field. This phenomenon gives rise to CSA. In solution, fast molecular tumbling causes CSA to be rotationally averaged across all possible orientations, resulting in a single chemical shift. However, changes in the strength and direction of the magnetic field experienced by nuclei caused by molecular tumbling cause fluctuations in the local magnetic fields of spins, resulting in relaxation.

Dipolar relaxation and CSA are the mechanisms by which longitudinal and transverse relaxation occur. However, transverse relaxation is also contributed to by spins precessing at different frequencies in the xy plane. This additional contribution results in longitudinal and transverse relaxation rates having different dependencies on correlation time as depicted in Figure 2.4[324].

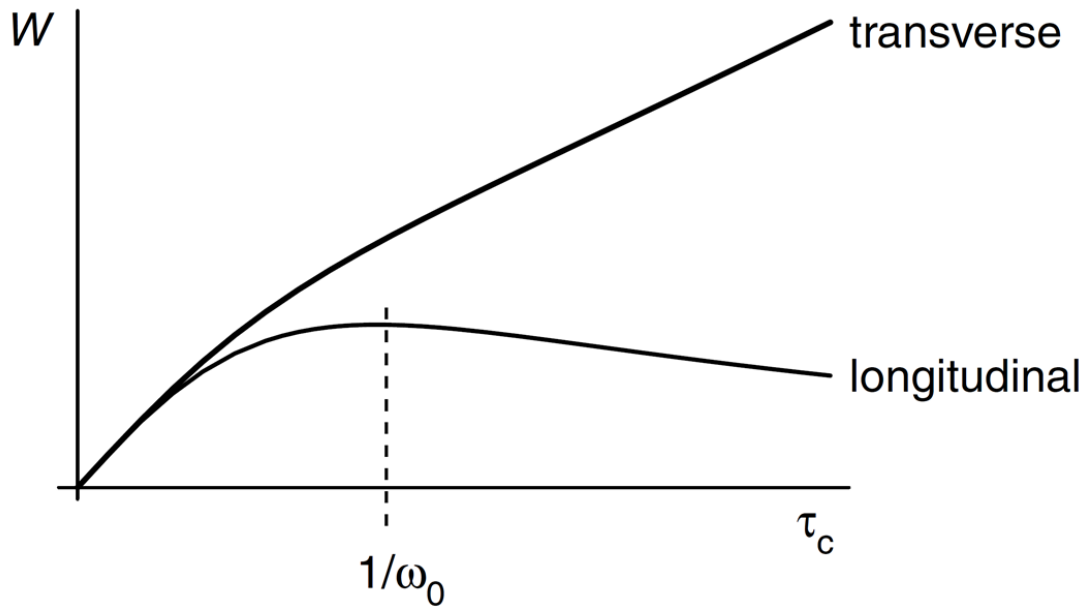


Figure 2.4 Variation of the relaxation rate constant (W) with correlation time (τ_c) for longitudinal and transverse relaxation of nuclei. τ_c is the time taken for the molecule to rotate through one radian. At the point where τ_c is equal to the inverse of the Larmor frequency ($1/\omega_0$) the longitudinal rate reaches a maximum but the transverse rate continues to increase. This figure is reproduced from Keeler (2005)[324].

Measurement of relaxation properties of proteins typically uses backbone amide nitrogen nuclei. The low gyromagnetic ratio of nitrogen ($\gamma = -2.71 \times 10^7$ T s^{-1})[319], coupled with the high gyromagnetic ratio of proton ($\gamma = 2.67 \times 10^8$ T s^{-1})[319], means that R_1 and R_2 relaxation of backbone amide nitrogen nuclei is predominantly due to dipole-dipole interactions with the attached proton[319,340] and so can be used to investigate local fluctuations of the polypeptide backbone. Measured R_2 rates may also be affected by chemical exchange resulting from larger amplitude protein motions[341]. In order to reduce the contributions of R_1 relaxation and chemical exchange to measured R_2 rates a series of 180^o pulses with set time delays before and after each pulse is applied in the indirect dimension, referred to as a Carr-Purcell-Meiboom-Gill (CPMG) pulse train[342,343]. The CPMG sequence is implemented between the first INEPT block and the t_1 evolution period in the indirect dimension of the HSQC pulse sequence (see Figure 2.1), keeping spins of nitrogen nuclei in the xy plane while decoherence occurs. By varying the number of repetitions of the CPMG block time delays of different lengths are created and ^{1}H-^{15}N HSQC spectra with different amide proton peak intensities are acquired, which can be used to calculate R_2 rates for each amide nitrogen[341].

162, 151 and 163 µM G-CSF C3 in pH 4 buffer in the presence of 0, 0.5 and 3 M urea, respectively and 130, 135 and 134 µM G-CSF C3 in pH 7 buffer in the presence of 0, 0.5 and 4 M urea, respectively were used for measuring backbone ^{15}N R_2 rates. Time delays of 16.1, 32.2, 48.4, 64.5, 80.6, 96.7, 112.8 and 129.0 ms were used in R_2 experiments. Relaxation occuring during these mixing times resulted in ^{1}H-^{15}N HSQC spectra with varying amide proton peak intensities, the parameters of which are outlined in Table 2.15. Amide proton peak intensities at each relaxation delay were fitted to a single exponential decay using in-house software written in Python by Dr. Thompson and Dr. Karamanos. Errors were calculated as the difference of duplicate peak intensities from the mean intensity acquired using relaxation delays of 32.2 and 96.7 ms. Monte Carlo error analysis with one thousand steps was performed to calculate the error of the parameters obtained from fitting of peak intensities to a single exponential decay. The ratio of R_2 for G-CSF C3 in pH 4 and 7 buffers to R_2 in the urea concentrations specified was then calculated to elucidate residues with altered dynamics upon addition of urea.

Parameter	HSQC
np	2048
nt	8
sw / Hz	12019
sw (^{15}N) / Hz	1600
ni (^{15}N) / Hz	160
d1 / s	2.5
Acquisition time / h	20

Table 2.15 Acquisition parameters for ^{1}H-^{15}N HSQC spectra used to measure backbone ^{15}N R_2 rates of G-CSF C3. 162, 151 and 163 µM G-CSF C3 were used in pH 4 buffer in the presence of 0, 0.5 and 3 M urea, respectively, while in pH 7 buffer 130, 135 and 134 µM G-CSF C3 were used in the presence of 0, 0.5 and 4 M urea, respectively. See the legend of Table 2.4 for a description of each parameter.

Chapter 3 Biophysical comparison of the native states of G-CSF wt and C3

3.1 Introduction

The variant of G-CSF identified by Buchanan *et al*.[150], termed G-CSF C3, was identified by screening a library of G-CSF DNA constructs generated randomly by error-prone PCR. Screening consisted of four rounds of *in vitro* ribosome display, with increasingly stringent selection pressures being employed during each round (see Section 1.6.4). G-CSF variants remaining after these four rounds of selection were subsequently cloned and expressed in the periplasm of *E. coli* strain HB2151 along with wild-type G-CSF. One of the variants identified as showing increased soluble expression of monomeric protein was G-CSF C3, which was corroborated by SDS-PAGE and SEC-HPLC analysis of G-CSF C3 purified from 2 l periplasmic expression in parallel with the wild-type protein. G-CSF C3 contains four amino acid substitutions relative to wild-type, these being C18G, W59R, Q71R and F84L. Despite these substitutions, G-CSF C3 was shown to have equivalent functional activity to wild-type G-CSF from cell proliferation assays. Although the substitutions W59R and Q71R are predicted to reduce aggregation propensity, based on computational modelling of G-CSF with and without the substitutions found in the variant C3[150], reduced aggregation was not confirmed experimentally. Likewise, there have been no attempts to investigate the structure of G-CSF C3 and make comparisons of its aggregation behaviour to wild-type G-CSF.

G-CSF proteins used in this study were expressed in the cytoplasm of *E. coli*, resulting in formation of insoluble inclusion bodies. Consequently, after refolding of G-CSF proteins from cytoplasmic inclusion bodies and subsequent purification it was critical to establish that the functional native conformation of monomeric G-CSF had been reached by comparison with previous biophysical studies of wild-type G-CSF. In this chapter, the expression of G-CSF wt and C3 are described, along with their purification and characterisation using a range of biophysical techniques to confirm they were correctly folded, as well as providing more structural information about the variant C3. In addition, isotope labelling of G-CSF wt (^{15}N) and C3 (^{15}N/^{13}C) proteins are described and their purification to allow assignment of ^{1}H-^{15}N HSQC spectra. Acquisition and assignment of ^{1}H-^{15}N HSQC spectra allowed higher

resolution structural comparison of G-CSF wt and C3 and provided a more reliable measure of these proteins having reached their native conformation.

In the following chapter all experiments carried out in pH 4 and 7 buffers were performed in 20 mM sodium phosphate, 20 mM sodium acetate pH 4 buffer or 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide pH 7 buffer unless otherwise stated.

3.2 Results

3.2.1 Recombinant expression of G-CSF proteins

Expression trials of G-CSF C3 in BL21(DE3) and BL21(DE3) pLysS cells were carried out to determine which of these *E. coli* strains would provide the highest yield of G-CSF proteins. Competent BL21(DE3) and BL21(DE3) pLysS cells (see Section 2.1.3) were transformed with pET23a_G-CSF_C3 (Appendix 7.3) and grown on LB and agar plates containing 100 µg/ml carbenicillin (and 34 µg/ml chloramphenicol for pLysS cells; see Section 2.2.7). Single colonies were used to inoculate 250 ml overnight LB cultures containing 100 µg/ml carbenicillin (and 34 µg/ml chloramphenicol for pLysS cells), which were incubated at 37 ^{o}C with shaking at 200 rpm. 20 ml of overnight cultures were used to inoculate 1 l LB expression cultures that were subsequently incubated at 37 ^{o}C with shaking at 200 rpm. At an OD_{600} of 0.6 protein expression was induced in half of the expression cultures by addition of 1 mM IPTG, followed by incubation of both induced and uninduced expression cultures for a further 4 h at 37 ^{o}C with shaking at 200 rpm. At the end of each hour the OD_{600} was measured and pellets harvested from 1 ml samples of expression culture by centrifugation at 13300 rpm for 5 mins at room temperature. Cell pellets were lysed by resuspending in appropriate amounts of 2× loading buffer to correct for cell density and expression levels analysed by SDS-PAGE (Figure 3.1).

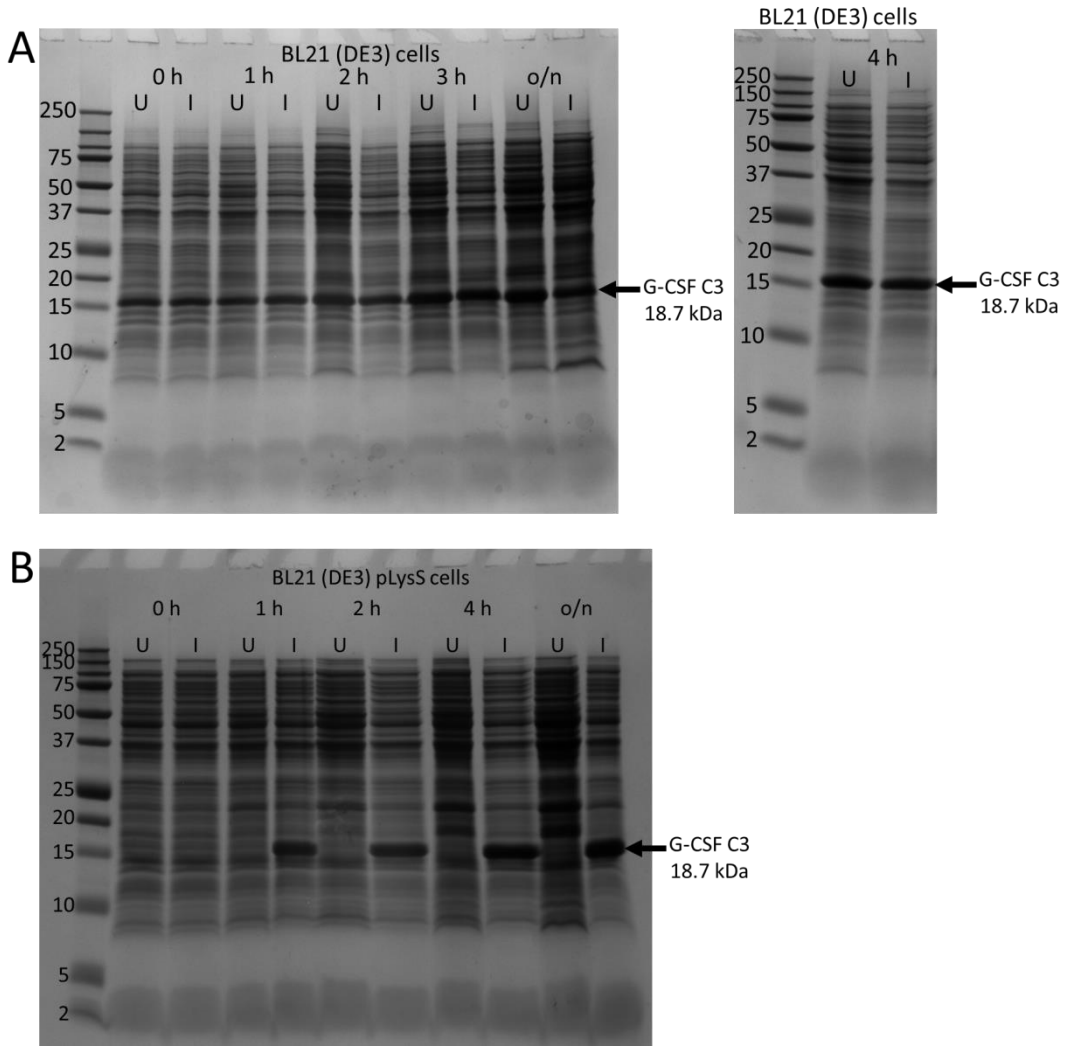


Figure 3.1 Recombinant expression trials of G-CSF C3 in *E. coli* strains BL21(DE3) and BL21(DE3) pLysS. At an OD_{600} of 0.6 protein expression was induced in half of the expression cultures (I) and the other half were left uninduced (U). After induction samples of both induced and uninduced expression cultures of BL21(DE3) (A) and BL21(DE3) pLysS (B) cells were harvested at the time periods shown, which included samples taken from cultures that had been left overnight (o/n). Samples of cultures from expression trials of G-CSF C3 in BL21(DE3) pLysS cells were not taken 3 h after induction. The sizes of protein molecular mass markers are shown on the left in kDa.

Expression trials showed that the highest expression of G-CSF C3 is achieved using *E. coli* strain BL21(DE3) pLysS with induction by addition of 1 mM IPTG, followed by incubation for 4 h or overnight at 37 ^{o}C with shaking at 200 rpm. Interestingly, recombinant protein expression appears to occur regardless of the addition of IPTG when *E. coli* strain BL21(DE3) is used, but when using BL21(DE3) pLysS cells expression only occurs after induction with IPTG.

pET vectors are commonly used for recombinant expression due to the high levels of mRNA synthesised, typically leading to recombinant protein expression

encompassing 40-50% of total cellular protein[176,344]. High levels of transcription are due to incorporation of the target DNA sequence downstream of the promoter for the highly processive bacteriophage T7 RNA polymerase. T7 RNA polymerase is not endogenous to bacterial chromosomes, meaning it must be supplied during recombinant expression. In the case of BL21(DE3) cells, DNA encoding viral T7 RNA polymerase is integrated into the bacterial chromosome by prophage λDE3, under control of the IPTG-inducible *lac*UV5 promoter[344].

Positioning of the T7 polymerase DNA sequence downstream of the *lac*UV5 promoter should provide stringent regulation of recombinant protein expression. However, high levels of recombinant expression using the pET system without inducer have been shown previously for a number of soluble and insoluble proteins expressed in LB medium[345-348], including G-CSF[348]. This phenomenon is thought to occur when cells cultured in complex media reach the stationary phase and is caused by a lack of carbon sources. The *lac* operon plays an endogenous role in carbohydrate scavenging, thus being activated under such nutrient-limited conditions in a cAMP-dependent process[349].

Presumably, recombinant expression of G-CSF wt and C3 stimulates this nutrient limited response and uninduced recombinant protein expression in *E. coli* BL21(DE3) cells. The additional pLysS plasmid found in BL21(DE3) pLysS cells encodes bacteriophage T7 lysozyme, which binds T7 RNA polymerase and inhibits transcription[350]. T7 lysozyme thus inhibits transcription by small amounts of T7 RNA polymerase that may be expressed by cells, especially in carbon-starved conditions, but after induction the expression level of T7 RNA polymerase is greatly increased and T7 lysozyme is no longer able to inhibit transcription. This additional regulation restores the ability to induce protein expression with IPTG[176], as well as improving the recombinant expression level of G-CSF C3 (Figure 3.1).

Expression trials of G-CSF wt and C3 using *E. coli* strain BL21(DE3) pLysS were also carried out to assess whether the cytoplasmic expression level of G-CSF C3 is higher than wild-type in line with the periplasmic expression levels of these proteins[150]. The results of these expression trials show that the cytoplasmic expression level of G-CSF C3 is significantly higher than that of the wild-type protein (Figure 3.2).

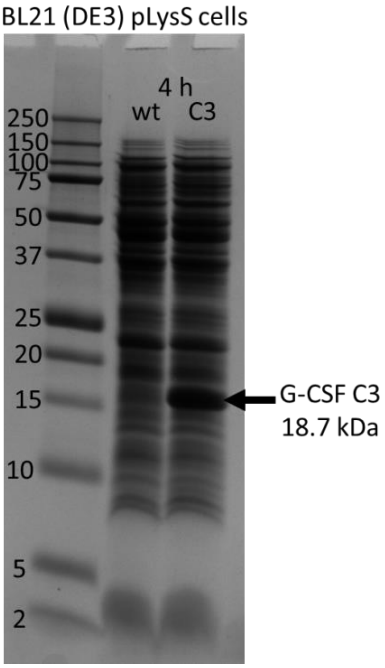


Figure 3.2 Recombinant expression trials of G-CSF wt and C3 in *E. coli* strain BL21(DE3) pLysS. At an OD_{600} of 0.6 protein expression was induced and samples of expression cultures were harvested 4 h after induction. The sizes of protein molecular mass markers are shown on the left in kDa.

3.2.2 Purification of G-CSF proteins

3.2.2.1 G-CSF inclusion body preparation and solubilisation

G-CSF proteins expressed in the cytoplasm of *E. coli* form insoluble inclusion bodies, evidenced from the appearance of G-CSF in the insoluble fraction of cell lysate (Figure 3.3). This is most likely due to the reducing environment of the *E. coli* cytoplasm[130], which prevents the formation of the two disulphide bonds needed for G-CSF to refold to its native biologically active conformation[285]. After recombinant expression in BL21(DE3) pLysS cells and cell lysis by sonication, G-CSF inclusion bodies were harvested by centrifugation at 15000 rpm for 30 mins at 4 ^{o}C. Inclusion bodes were washed by resuspending in 50 mM TrisHCl, 5 mM EDTA, 1 % (*v/v*) triton X-100 pH 8 (wash buffer 1) to remove cell membranes and harvested by centrifugation as above. This wash step was repeated using 50 mM TrisHCl, 5 mM EDTA, 4 M urea pH 8 (wash buffer 2) to solubilise contaminant proteins. Finally, G-CSF inclusion bodies were solubilised by resuspending in 50 mM TrisHCl, 5 mM EDTA, 6 M GdnHCl, 10 mM DTT pH 8 (unfolding buffer) and centrifuged as above to remove any remaining insoluble material (Figure 3.3).

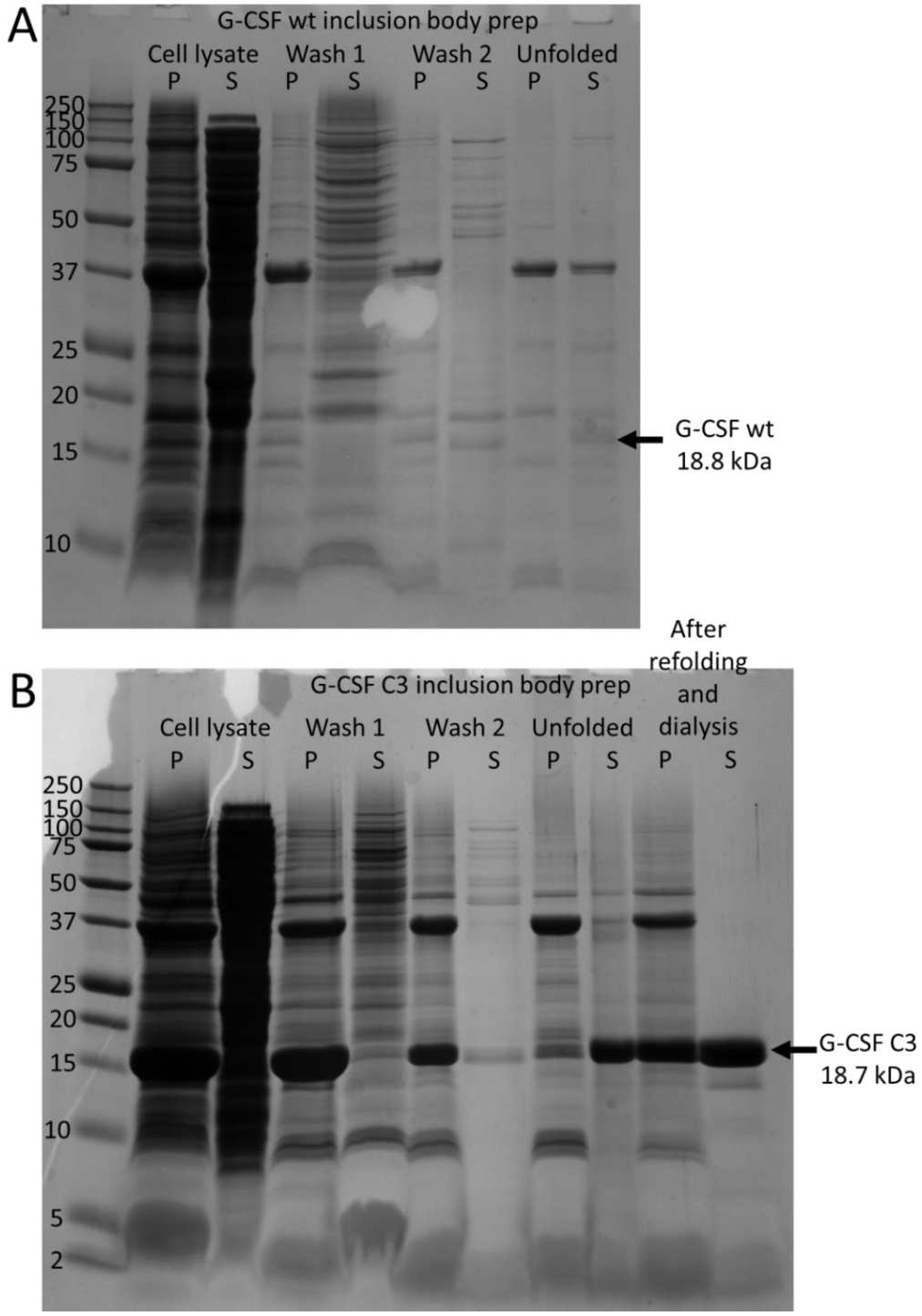


Figure 3.3 Inclusion body extraction, washing and solubilisation for G-CSF wt and C3. Cell lysates were centrifuged in order to separate pellet (P) and supernatant (S). G-CSF wt (A) and C3 (B) inclusion bodies were subsequently washed by resuspending in wash buffers containing 1 % (*v/v*) triton X-100 and 4 M urea (washes 1 and 2, respectively), with G-CSF inclusion bodies being harvested after each wash step. After washing, G-CSF inclusion bodies were solubilised and unfolded by resuspending in buffer containing 6 M GdnHCl and 10 mM DTT (unfolded) and centrifuged to remove any remaining insoluble material. Dialysis of refolded G-CSF into pH 4 buffer for cation exchange resulted in large amounts of aggregate that were removed by centrifugation and filtration. Samples of pellet and supernatant from G-CSF C3 after refolding and dialysis are shown in B. The sizes of protein molecular mass markers are shown on the left in kDa.

3.2.2.2 G-CSF refolding and purification by cation exchange chromatography

After solubilising and unfolding G-CSF proteins from inclusion bodies, refolding was achieved by ten-fold dilution in 50 mM TrisHCl, 5 mM EDTA, 0.9 M arginine pH 8 buffer. Arginine has been frequently used for refolding of proteins from inclusion bodies[199-203], including G-CSF[202,203], due its ability to suppress aggregation[200,201] (see Section 1.4.4.1).

Refolded G-CSF proteins were dialysed into pH 4 buffer for purification by cation exchange chromatography, a commonly used final purification step during preparation of G-CSF[203,348,351,352]. However, during dialysis aggregation was observed. SDS-PAGE analysis of pellet and supernatant from G-CSF C3 after refolding and dialysis revealed this aggregate to be a mixture of G-CSF C3 and contaminant proteins (Figure 3.3B). Insoluble aggregates were removed by centrifugation at 15000 rpm for 30 mins at 4 ^{o}C followed by filtration using a 0.22 µm filter before loading onto an SP HP cation exchange column for purification. Ten-fold dilution of G-CSF proteins during refolding resulted in large volumes of dilute protein solutions, which required loading onto the column through line A1 of an Äkta prime plus pump system. The bound protein was eluted from the column over an elution gradient of 0-1 M NaCl over 100 ml with 2 ml fractions being collected. The purity of eluted G-CSF proteins was determined by SDS-PAGE (Figure 3.4).

Unusually, the elution traces observed for G-CSF proteins during cation exchange chromatography did not immediately return to baseline after elution of the main peak (Figure 3.4A and B), giving these elution traces the appearance of a ‘peak’ with a succeeding ‘tail’. SDS-PAGE analysis shows that G-CSF proteins are the only component of elution fractions collected from the peak or tail (Figure 3.4C and D). The reason for this tail may be the appearance of G-CSF proteins with different degrees of deamidation of glutamine residues[276], which would affect the overall charge of the protein. Consequently, elution fractions from the peak and tail of cation exchange chromatography were pooled separately and desalted by dialysis into pH 4 buffer.

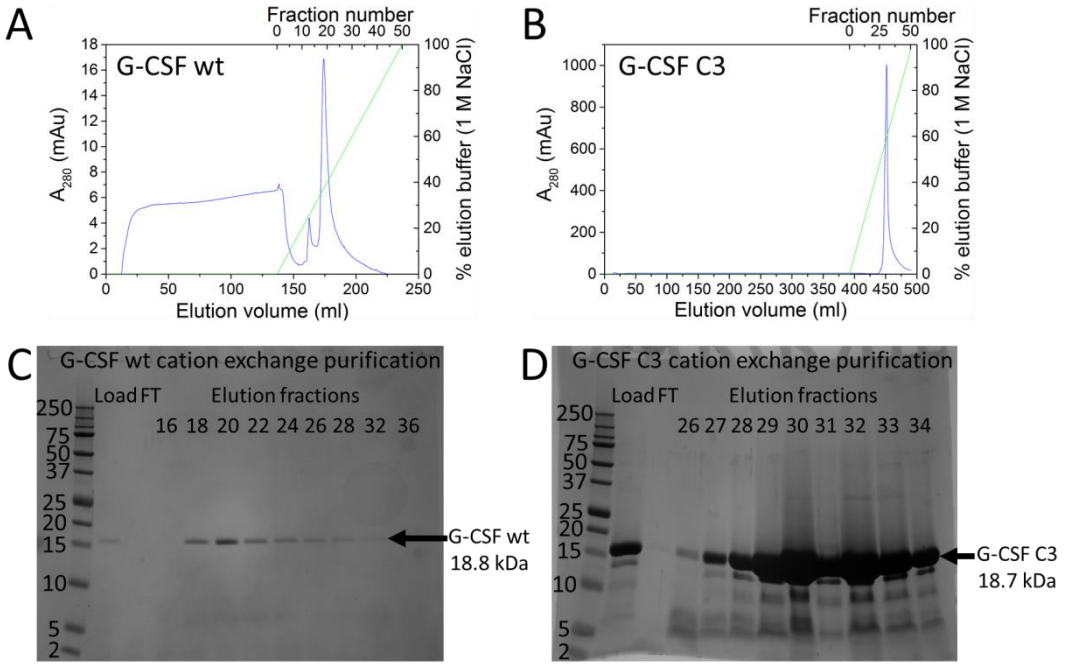


Figure 3.4 Cation exchange purification of refolded G-CSF wt and C3. G-CSF proteins were loaded onto an SP HP cation exchange column in pH 4 buffer. Approximately 130 ml of G-CSF wt was loaded, although the protein concentration was too dilute to measure accurately (A). Approximately 390 ml of G-CSF C3 was loaded at a protein concentration of 0.57 mg/ml (B). Loading and elution of protein was followed by measurement of A_{280} (blue line and left hand y-axis of elution traces). Bound protein was eluted from the column using a gradient of 0-100 % elution buffer (green line and right hand y-axis of elution traces) applied over 100 ml with 2 ml fractions being collected. Elution buffer used was pH 4 buffer with 1 M NaCl. Numbers of elution fractions collected are shown on the top x-axis and the purity of elution fractions collected in or around the elution peak were assessed by SDS-PAGE for G-CSF wt (C) and C3 (D). The sizes of protein molecular mass markers are shown on the left in kDa.

Peak pools constituted fractions 16-23 and 27-32 from purifications of G-CSF wt and C3, respectively, while tail pools constituted fractions 24-40 and 33-45 from purifications of G-CSF wt and C3, respectively (Figure 3.4). After desalting, these elution pools were concentrated to 2.5-5 mg/ml for G-CSF wt pools and 8-10 mg/ml for G-CSF C3 pools then snap frozen as aliquots of 50-200 µl. The purity of these pools was assessed by SDS-PAGE (Figure 3.5).

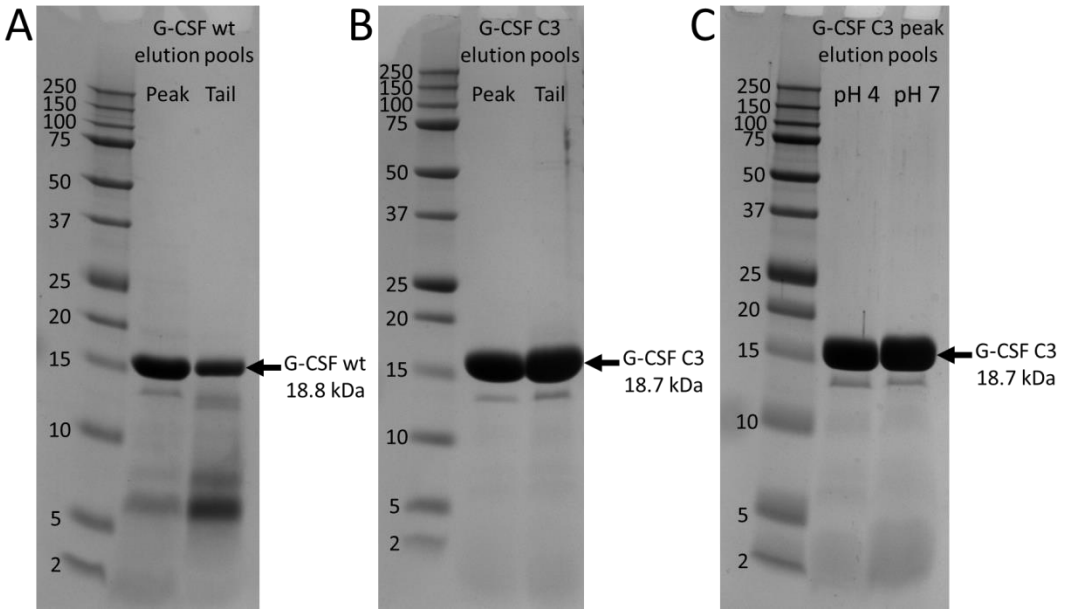


Figure 3.5 Purified G-CSF wt and C3 obtained from the 'peak' and 'tail' portions of cation exchange chromatography elution traces. G-CSF wt samples from peak and tail elution pools were analysed at 27 and 25 µM, respectively (A). G-CSF C3 samples from peak and tail elution pools were analysed at 36 and 42 µM, respectively (B). G-CSF C3 samples from the peak elution pool at pH 4 and 7 were analysed at 50 µM (C). The sizes of protein molecular mass markers are shown on the left in kDa.

In all elution pools there appears to be a faint band below the most intense gel band corresponding to G-CSF (Figure 3.5). Amino acid sequencing of the protein from both faint and strong gel bands by enzymatic digest and electrospray ionisation tandem mass spectrometry (ESI-MS/MS), carried out by Dr. James Ault, identified them both as G-CSF (results not shown). Dialysis of a sample of G-CSF C3 in pH 4 buffer into pH 7 buffer and subsequent analysis by SDS-PAGE did not remove this additional faint gel band (Figure 3.5C), suggesting that this band is not caused by inefficient reduction of the two disulphide bonds of G-CSF by DTT at pH 4. Therefore, this faint extra band observed in SDS-PAGE is most likely the result of a small amount of proteolytic degradation[276] reducing the mass of the protein. G-CSF wt obtained from the tail portion of cation exchange was not pure enough for experimental use and so was discarded, while all other aliquots of G-CSF wt and C3 were sufficiently pure for use in experiments.

The total yields of pure G-CSF wt and C3 were approximately 0.1 and 6.5 mg, respectively per litre of LB medium used for expression. Unusually, the total yields of pure isotopically labelled G-CSF wt and C3 from expression in M9 minimal medium (see Section 2.1.4) were slightly higher than those observed from expression in LB

medium, with yields of approximately 0.2 and 9.1 mg per litre of M9 medium used for expression of G-CSF wt and C3, respectively.

3.2.3 Mass spectrometry of G-CSF proteins

3.2.3.1 Confirmation of the molecular masses of G-CSF proteins

The molecular masses of purified G-CSF wt and C3 were confirmed by electrospray ionisation mass spectrometry (ESI-MS), carried out by Dr. Ault (see Section 2.3.4.1). Before analysis by ESI-MS, G-CSF samples were buffer exchanged into 66 mM ammonium acetate pH 4, which was calculated to match the ionic strength (I) of the pH 4 buffer used in this study (43 mM). G-CSF C3 from the peak and tail elution pools of cation exchange were analysed to determine whether there were any differences in the molecular masses of G-CSF C3 from these pools. The differences between the measured molecular masses of G-CSF wt and C3 obtained from the most intense peaks in mass spectra and the molecular masses predicted based on amino acid sequence were within the 0.01 % error of mass measurement by ESI-MS. There appear to be no significant differences between mass spectra obtained for G-CSF C3 from the peak or tail of cation exchange chromatography (Figure 3.6B-E).

An additional peak corresponding to a species with a molecular mass of 18566 Da was observed in the mass spectrum of native G-CSF wt (Figure 3.6A), corresponding to removal of the N-terminal Met and Thr residues. This proteolytic degradation is also suspected to be responsible for the faint additional gel band observed below the main G-CSF gel band in SDS-PAGE analysis (Figure 3.5). Mass spectra of native G-CSF C3 produced broad mass distributions, most likely the result of ion adducts. The vast majority of these peaks were removed by acid denaturation of G-CSF C3 and reduction of disulphide bonds, which was achieved by addition of 1 % (*v/v*) formic acid/50 % (*v/v*) acetonitrile and 10 mM TCEP, respectively before analysis. Mass spectra of acid denatured and reduced G-CSF C3 and native G-CSF wt showed four additional peaks that were higher in molecular mass than the main G-CSF peak, each separated by approximately 18 Da. These peaks were most likely the result of oxidation of the side chains of the four Met residues in G-CSF[276].

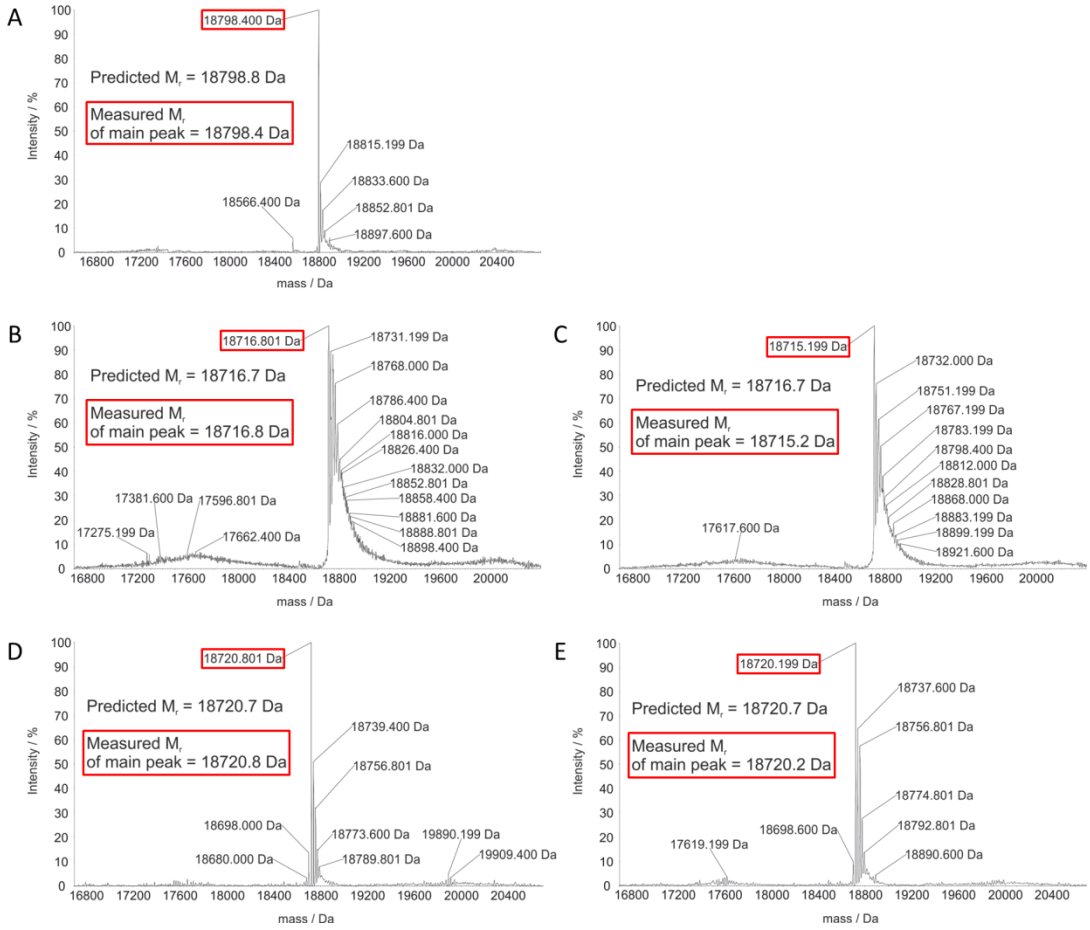
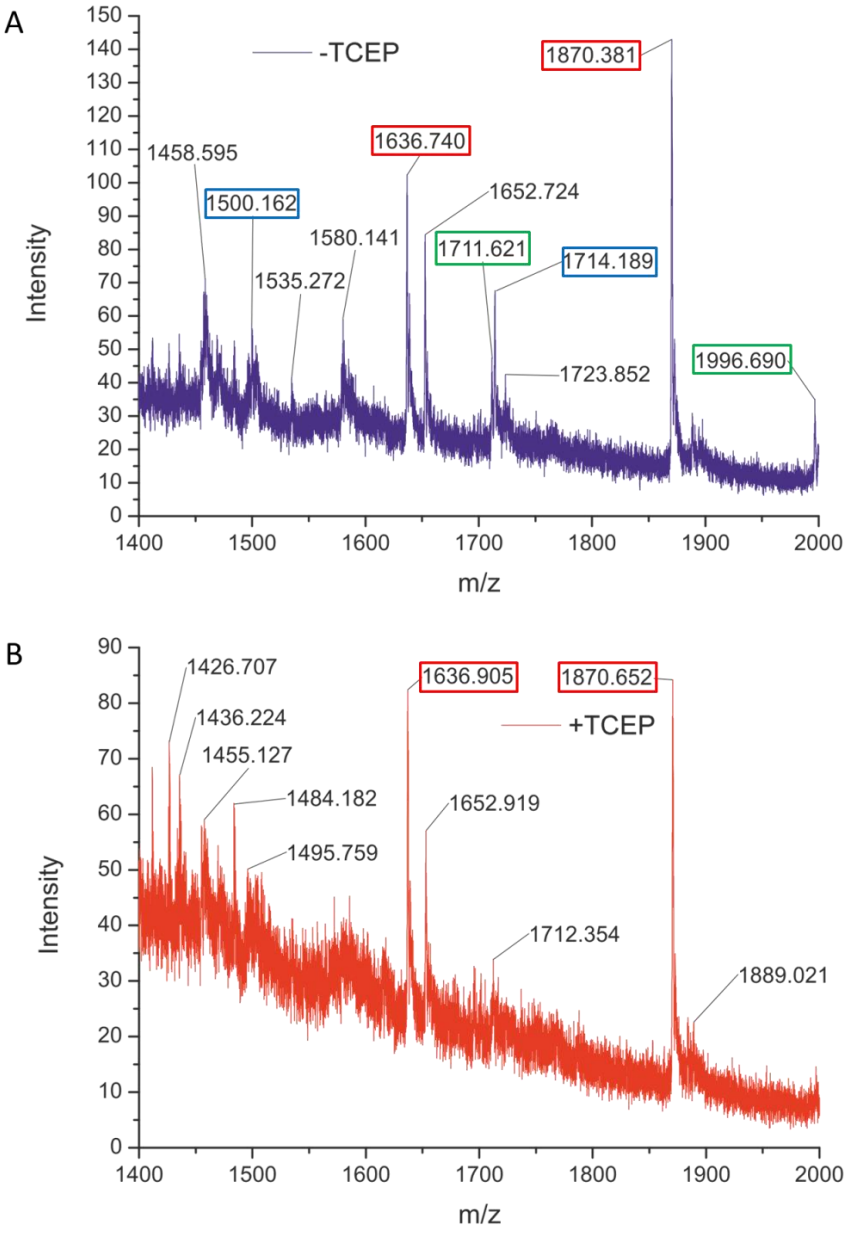


Figure 3.6 Confirmation of the molecular masses of G-CSF wt and C3 by ESI-MS. Mass spectra are shown as the mass of species in Da plotted against percentage intensity, where the most intense peak is normalised to 100 %. Mass spectrometry of native G-CSF wt was carried out at 79 µM (A), while native G-CSF C3 was analysed at 69 and 83 µM for peak (B) and tail (C) samples, respectively in 66 mM ammonium acetate pH 4 buffer (I = 43 mM) by Dr. Ault. Mass spectra of acid denatured and reduced G-CSF C3 peak and tail samples are shown in D and E, respectively. Peaks with masses closest to the predicted mass of G-CSF wt and C3 are highlighted in red.

3.2.3.2 Confirmation of native disulphide bond formation in G-CSF proteins

ESI-MS analysis of G-CSF C3 revealed a 4 Da increase in molecular mass upon reduction (Figure 3.6B-E), corresponding to the reduction of the two disulphide bonds found in native G-CSF C3. The measured molecular mass of native G-CSF wt was also consistent with the formation of these two disulphide bonds (Figure 3.6A). To confirm the two disulphide bonds had formed between Cys residues 37-43 and 65-75 found in native G-CSF, trypsin digests were carried out by Dr. Ault (see Section 2.3.4.2) and the results analysed by ESI-MS in the presence and absence of 5 mM TCEP (Figure 3.7 and Figure 3.8). Before digestion, G-CSF samples were buffer exchanged into 50 mM ammonium bicarbonate pH 7.

The theoretical m/z values of peptides containing disulphide bonds predicted to result from digestion of G-CSF wt with trypsin were used to identify peaks in mass spectra caused by disulphide linked peptides. These predicted m/z values were calculated from isotopically averaged molecular masses of peptides and their charge states (Figure 3.7C). In the absence of a reducing agent, the mass spectrum of G-CSF wt contained peaks with m/z values corresponding to disulphide containing peptides, which are highlighted in Figure 3.7A. Upon addition of TCEP four of these peaks were no longer observed, suggesting that they represent species containing disulphide bonds. However, two of these peaks remained after addition of TCEP, highlighted in red on Figure 3.7B. The m/z values of these peaks suggest reduction of one of the two disulphide bonds in this peptide (Figure 3.7C). Peaks for fully reduced peptides of digested G-CSF wt were not observed in mass spectra.



C

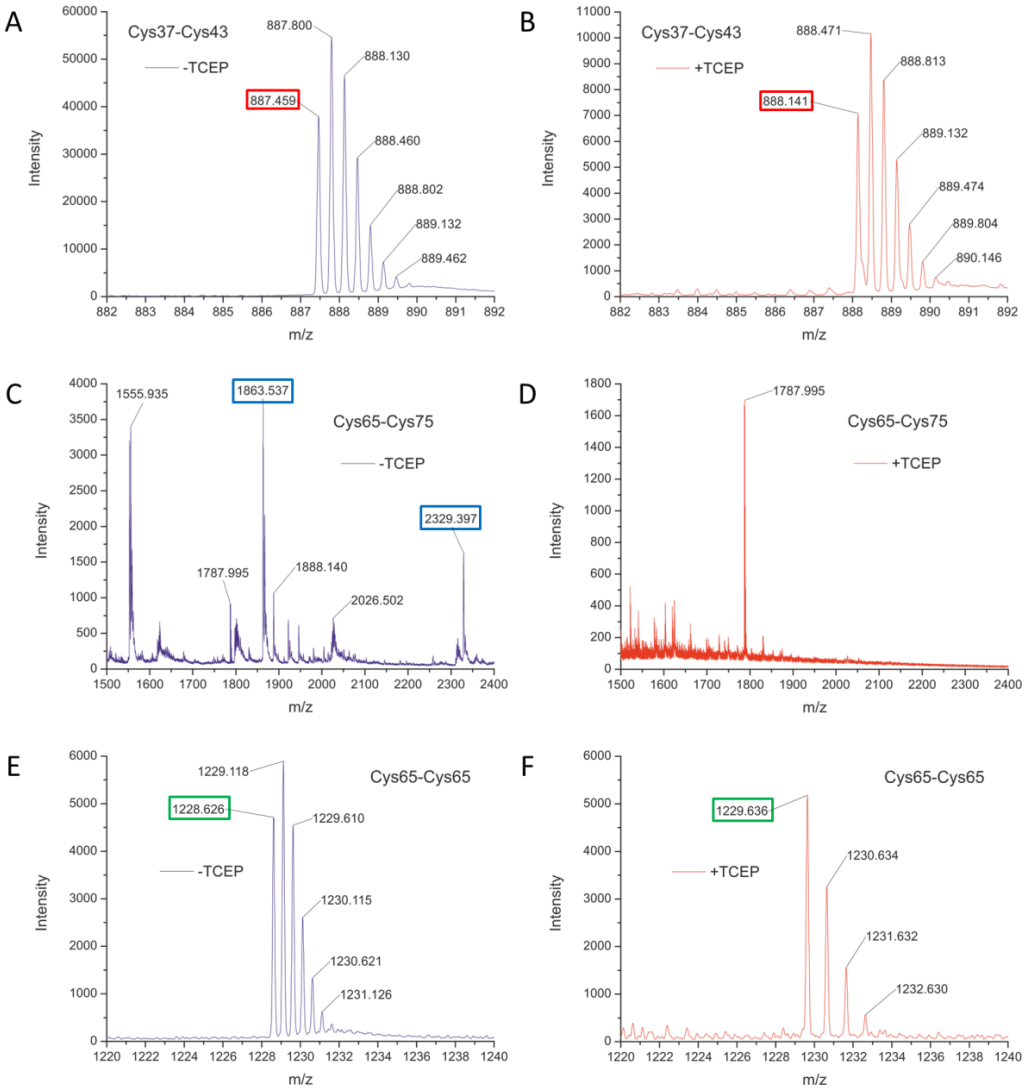
LCATYK LCHPEELVLLGHSLGIPWAPLSSCPSQALQLAGCLS QLHSGLFLYQGLLQALEGISPELGPTLDTLQLDVAD FATTIWQQMEELGMAPALQPTQGAMPAFASAFQR	Predicted $[M+8H]^{8+}$ = 1500.106 Predicted $[M+7H]^{7+}$ = 1714.264	Measured $[M+8H]^{8+}$ = 1500.162 Measured $[M+7H]^{7+}$ = 1714.189	Highlighted in blue in A
IQGDGAALQEKLCATYKLCHPEELVLLGHSLGIPWAPLSSCPSQALQLAGCLS QLHSGLFLYQGLLQALEGISPELGPTLDTLQLDVADFATTIWQQMEELGMAPA LQPTQGAMPAFASAFQR	Predicted $[M+8H]^{8+}$ = 1636.756 Predicted $[M+7H]^{7+}$ = 1870.436	Measured $[M+8H]^{8+}$ = 1636.740 Measured $[M+7H]^{7+}$ = 1870.381	Highlighted in red in A
LCATYKLCHPEELVLLGHSLGIPWAPLSSCPSQALQLAGCLS QLHSGLFLYQGLLQALEGISPELGPTLDTLQLDVADFATTIW QQMEELGMAPALQPTQGAMPAFASAFQR	Predicted $[M+7H]^{7+}$ = 1711.690 Predicted $[M+6H]^{6+}$ = 1996.805	Measured $[M+7H]^{7+}$ = 1711.621 Measured $[M+6H]^{6+}$ = 1996.960	Highlighted in green in A
IQGDGAALQEKLCATYKLCHPEELVLLGHSLGIPWAPLSSCPSQALQLAGCLS QLHSGLFLYQGLLQALEGISPELGPTLDTLQLDVADFATTIWQQMEELGMAPA LQPTQGAMPAFASAFQR	Predicted $[M+8H]^{8+}$ = 1637.006 Predicted $[M+7H]^{7+}$ = 1870.721	Measured $[M+8H]^{8+}$ = 1636.905 Measured $[M+7H]^{7+}$ = 1870.652	Highlighted in red in B

Figure 3.7 Trypsin digest of G-CSF wt analysed by ESI-MS to confirm the presence of native disulphide bonds Cys37-Cys43 and Cys65-Cys75. Mass spectra are shown as the mass-to-charge ratio (m/z) plotted against intensity in arbitrary units. Digested G-CSF wt samples were analysed at 50 µM in 50 mM ammonium bicarbonate pH 7 buffer (I = 65 mM) in the absence (A) or presence (B) of 5 mM TCEP (blue and red, respectively) by Dr. Ault. Peaks with measured m/z values corresponding to peptides containing disulphide bonds are highlighted as outlined in C. The predicted m/z values of these peptides are also shown in C, along with their charge states and amino acid sequences. Peaks for fully reduced peptides were not observed in mass spectra and so are not listed in C.

G-CSF C3 contains two extra Arg residues compared with wild-type G-CSF, owing to the amino acid substitutions W59R and Q71R. Consequently, there are two more trypsin cleavage sites in G-CSF C3 than wild-type. The resulting disulphide containing peptides of G-CSF C3 were smaller than those found in G-CSF wt, which appeared to make them more amenable to analysis by ESI-MS based on the intensities of peaks observed in Figure 3.8 compared to Figure 3.7.

The smaller peptides of G-CSF C3 containing disulphide bonds were visible both in oxidised and reduced forms (Figure 3.8A, B, E and F). The narrow m/z ranges over which these smaller peptides are shown in Figure 3.8A, B, E and F means that peaks resulting from incorporation of ^{13}C at natural abundance were observed. Predicted m/z values of peptides containing oxidised and reduced Cys residues were calculated using monoisotopic molecular masses when peaks resulting from ^{13}C isotope distribution could be observed, whereas for peptides detected over wider m/z ranges (where isotope distributions could not be resolved) predicted m/z values were calculated using isotopically averaged molecular masses (Figure 3.8G).

The larger of the two peptides linked by a disulphide bond between Cys65 and Cys75 was no longer detectable when this disulphide linkage was broken as shown in Figure 3.8C and D, which were acquired before and after treatment with 5 mM TCEP, respectively. Interestingly, peaks with masses corresponding to the smaller peptide in the Cys65-Cys75 disulphide linked species were observed both in the absence and presence of TCEP (Figure 3.8E and F, respectively). This is most likely due to the formation of Cys65-Cys65 disulphide bonds during the trypsin digest that were reduced by TCEP.



G

LCATYKLCHPEELVLLGHSLGIPR	Predicted $[M+3H]^{3+}$ = 887.470	Measured $[M+3H]^{3+}$ = 887.459	Highlighted in red in A
LCATYKLCHPEELVLLGHSLGIPR	Predicted $[M+3H]^{3+}$ = 888.136	Measured $[M+3H]^{3+}$ = 888.141	Highlighted in red in B
APLSSCPSQALR LAGCLSQLHSGLLLYQGLLQALEGISPELGPTLDTL QLDVADFATTIWQQMEELGMAPALQPTQGAMPAF ASAFQR	Predicted $[M+5H]^{5+}$ = 1863.738 Predicted $[M+4H]^{4+}$ = 2329.423	Measured $[M+5H]^{5+}$ = 1863.537 Measured $[M+4H]^{4+}$ = 2329.397	Highlighted in blue in C
APLSSCPSQALR APLSSCPSQALR	Predicted $[M+2H]^{2+}$ = 1228.622	Measured $[M+2H]^{2+}$ = 1228.626	Highlighted in green in E
APLSSCPSQALR	Predicted $[M+H]^{+}$ = 1229.622	Measured $[M+H]^{+}$ = 1229.636	Highlighted in green in F

Figure 3.8 Trypsin digest of G-CSF C3 analysed by ESI-MS to confirm the presence of native disulphide bonds Cys37-Cys43 and Cys65-Cys75. Mass spectra of peptides containing disulphide linked and reduced Cys37 and Cys43 residues are shown in A and B, respectively. The mass spectrum of peptides linked by a disulphide bond between Cys65 and Cys75 is shown in C. After reduction with TCEP the larger of these peptides is no longer observed (D) but the smaller peptide is still detected (F). The mass spectrum of peptides linked by a disulphide bond between Cys65 residues is shown in E. Mass spectra are shown as the mass-to-charge ratio (m/z) plotted against intensity in arbitrary units. Digested G-CSF C3 samples were analysed at 50 µM in 50 mM ammonium bicarbonate pH 7 buffer (I = 65 mM) in the absence or presence of 5 mM TCEP (blue and red, respectively) by Dr. Ault. Peaks with measured m/z values corresponding to peptides containing disulphide bonds are highlighted as outlined in G. The predicted m/z values of these peptides are also shown in G, along with their charge states and amino acid sequences. Peaks for the peptide containing reduced Cys75 were not observed in mass spectra and so are not listed in G.

3.2.4 Analytical size exclusion chromatography (SEC) of G-CSF C3

Analytical SEC was used to ensure purified G-CSF C3 was monomeric under the buffer conditions used for experiments in this study. The low yield from expression and purification of G-CSF wt, coupled with the high protein concentrations required for analytical SEC (at least 1 mg/ml) meant that only G-CSF C3 was analysed.

G-CSF C3 was loaded onto a Superdex^{TM} 75 10/300 GL gel filtration column via a 500 µl loop attached to the injection port of an Äkta prime plus pump system and elution monitored over roughly one column volume. G-CSF C3 chromatography runs were carried out in pH 4 and 7 buffers in the presence and absence of 200 mM NaCl, but calibrants were only run at pH 7 in the presence of 200 mM NaCl in accordance with the manufacturer's instructions (Figure 3.9).

Elution traces of G-CSF C3 at pH 7 show two peaks, the elution volumes of which correspond to monomeric and dimeric G-CSF C3 as revealed by the calibration plot. Observations of dimeric species have also been reported for G-CSF wt[307,308], although the role of dimeric G-CSF in aggregation is uncertain. Dimeric G-CSF has been proposed as an on-pathway species to aggregation[307], but also as an indistinguishable mixture of on-pathway dimer that leads to irreversible aggregation and off-pathway dimer that exists in equilibrium with native monomer[303,308].

Unusually, G-CSF C3 at pH 4 elutes from the gel filtration column after a larger elution volume than predicted from its molecular mass. Addition of 200 mM NaCl decreases the observed elution volume of G-CSF C3, suggesting that there may be nonspecific adhesion of G-CSF C3 to the gel filtration column at pH 4. However, even in the presence of 200 mM NaCl the elution volume of G-CSF C3 is significantly larger at pH 4 than monomeric G-CSF C3 at pH 7. Considering the calibrant proteins have not been analysed by SEC at pH 4, it is not clear whether the different elution volumes observed for G-CSF C3 at pH 4 and 7 are due to the biophysical properties of the protein or the gel filtration column.

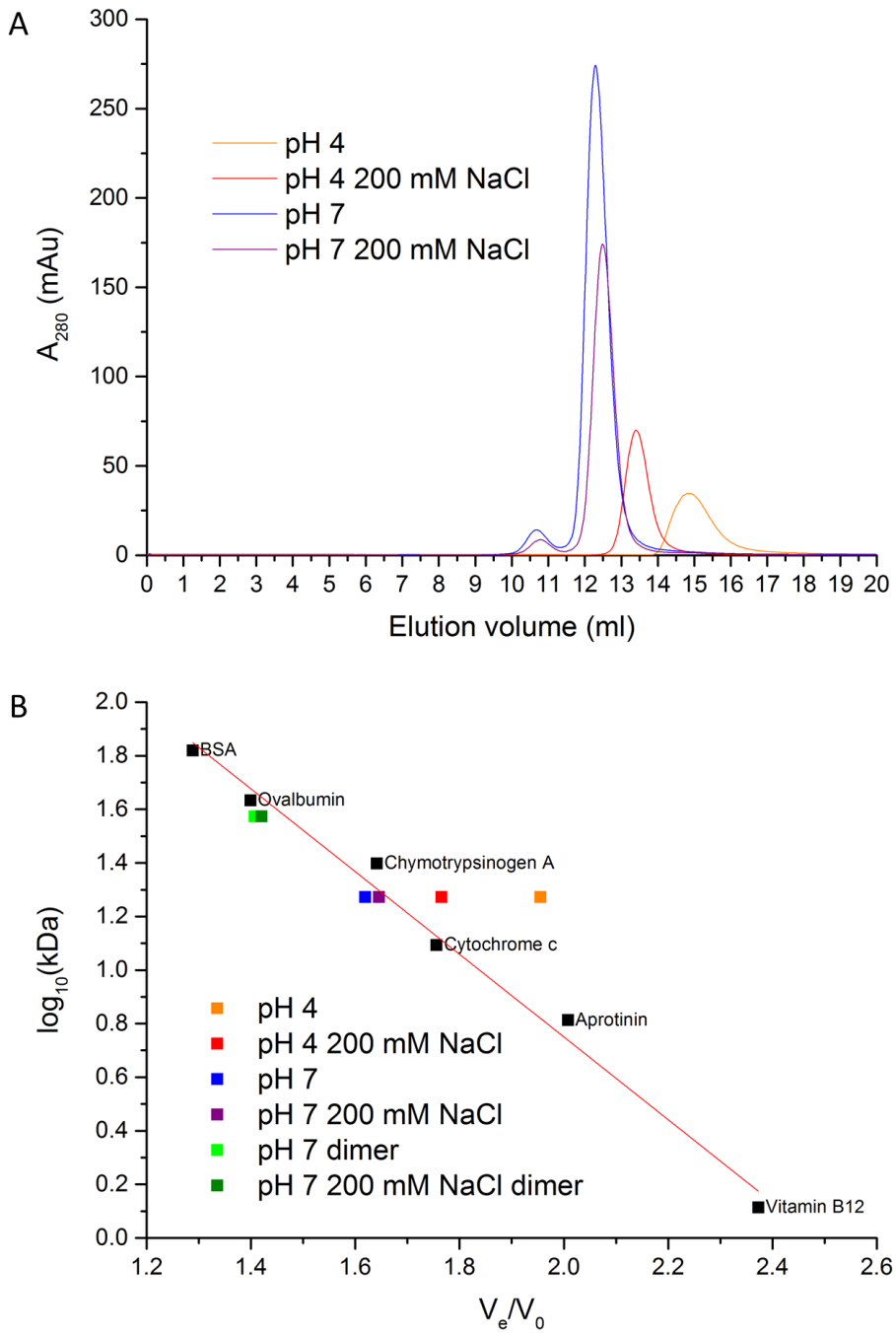


Figure 3.9 Analytical SEC of G-CSF C3. 100 µl of each G-CSF C3 sample was injected onto the column via a 500 µl loop attached to the injection port of an Äkta prime plus pump system. 406 and 443 µM G-CSF C3 were used in 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide pH 7 buffer (I = 80 mM) with or without 200 mM NaCl, respectively (purple and blue, respectively) , while 208 and 219 µM G-CSF C3 were used in 20 mM sodium phosphate, 20 mM sodium acetate pH 4 buffer (I = 43 mM) with or without 200 mM NaCl, respectively (red and orange, respectively). Chromatograms show A_{280} over roughly one column volume after injection of protein onto the column (A). The elution volumes of calibrant proteins (traces not shown in A) were used to produce a calibration plot, showing the ratio of the elution volumes of calibrant proteins (V_e) to the void volume (V_0) of the column (V_e/V_0) against log_{10} of the molecular masses of calibrant proteins in kDa fitted to a linear function (B). Calibration values for major peaks in the elution traces are coloured in the same way as the elution traces, while values for minor peaks in the elution traces of G-CSF C3 in pH 7 buffer in the presence and absence of 200 mM NaCl are shown in dark and light green, respectively.

3.2.5 Characterising and comparing the secondary structures of G-CSF wt and C3 using far-UV circular dichroism (CD)

Far-UV CD was used to characterise and compare the secondary structures of G-CSF wt and C3. CD measures the differential absorption of left and right circularly polarised components of plane-polarised light by chiral groups. In proteins, one such chiral group are peptide bonds arranged in secondary structures, which exhibit distinct far-UV CD spectra. For example, α-helical structures exhibit characteristic absorption minima at 208 and 222 nm[313].

Far-UV CD spectra of G-CSF wt and C3 at pH 4 and 7 exhibit absorption minima at 209 and 221 nm, confirming they are both predominantly α-helical. Upon addition of 8.1 M urea these absorption minima are lost due to unfolding of secondary structure (Figure 3.10).

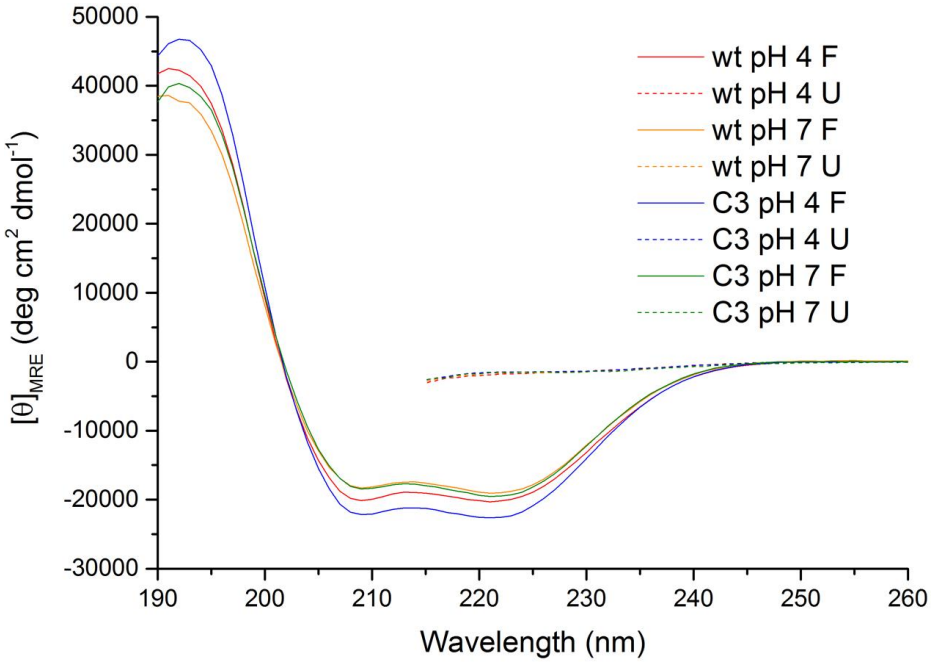


Figure 3.10 Far-UV CD spectra of G-CSF wt and C3 folded and unfolded. Spectra of G-CSF wt were acquired using protein concentrations of 5 and 4 µM in 20 mM sodium phosphate, 20 mM sodium acetate pH 4 buffer (I = 43 mM) and 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide pH 7 buffer (I = 80 mM), respectively (red and orange, respectively), while spectra of G-CSF C3 were acquired using protein concentrations of 8 and 5 µM in pH 4 and 7 buffers, respectively (blue and green, respectively). pH 7 buffer without azide was used for acquiring spectra to avoid noise caused by absorbance of signal by azide. Spectra of folded (F) and unfolded (U) proteins are represented by solid and dashed lines, respectively, with spectra of unfolded proteins being acquired in the presence of 8.1 M urea. Mean residue ellipticity ($[\theta]_{MRE}$) values acquired at wavelengths below 215 nm are not shown for unfolded samples due to excessive dynode voltage (800-1000 V) caused by the absorbance of signal by urea.

The mean residue ellipticity ($[\theta]_{MRE}$) at 208[314] and 222[315] nm can be used to estimate the percentage of secondary structure that is α-helical in folded proteins. $[\theta]_{MRE}$ values at 208 and 222 nm for G-CSF wt at pH 7 give estimations of 59 and 56 % α-helical structure, respectively. The percentage α-helical content of G-CSF wt at pH 7 has previously been reported as 59 % based on the $[\theta]_{MRE}$ value at 208 nm[353]. G-CSF C3 at pH 7 is estimated to possess 60 and 58 % α-helical structure based on $[\theta]_{MRE}$ values at 208 and 222 nm, respectively, suggesting that the secondary structures of G-CSF wt and C3 are similar at pH 7. G-CSF wt at pH 4 has estimated α-helical content of 64 and 59 % based on $[\theta]_{MRE}$ values at 208 and 222 nm, respectively, which is comparable to the 61 % α-helical content reported previously[353]. Interestingly, G-CSF C3 at pH 4 possesses estimated 70 and 66 % α-helical content based on calculations using $[\theta]_{MRE}$ values at 208 and 222 nm, respectively, which is notably higher than the estimated α-helical content of the wild-type protein. However, other previous studies of G-CSF wt have estimated α-helical percentages of 75[354] and 69[276] % at pH 4.5, based on the $[\theta]_{MRE}$ value at 208 nm, making it difficult to determine whether the difference between estimated α-helical percentages of G-CSF wt and C3 at pH 4 observed in this study is significant.

3.2.6 Characterising and comparing the tertiary structures of G-CSF wt and C3 using fluorescence

Fluorescence emission spectra of G-CSF wt and C3 were studied to allow comparison of their tertiary structures. The amino acids Trp, Tyr and Phe are intrinsically fluorescent, but the higher quantum yield and extinction coefficient of Trp, coupled with the sensitivity of Trp fluorescence to the chemical environment make this residue the most useful reporter of changes in tertiary structure[355]. Due to the large dipole moment of Trp[356] the fluorescence emission energy is sensitive to the polarity of the environment. When buried within the hydrophobic core of the protein interior the wavelength at which the maximum intensity of fluorescence emission is observed is typically 335 nm[357], but when exposed to solvent this wavelength is red-shifted to 355 nm and a decrease in intensity is usually observed[355].

The intensity of Trp fluorescence emission may be reduced by quenching from other proximal amino acid side chains. Lys and Tyr side chains quench by excited-state proton transfer and Gln, Asn, Glu, Asp, His and cysteine/cystine side chains quench

by excited-state electron transfer. Additionally, Asp, Glu and His side chains are better able to quench tryptophan fluorescence when protonated[358-360]. Excited-state proton transfer over an intermediate pH range (4-11) involves transfer of a proton from the donor to the C2, C4 or C7 positions of the indole ring[361-364] of Trp and is assumed to be dependent on the orientation of donor and acceptor groups[359]. Conversely, excited-state electron transfer occurs from the indole ring to an acceptor[358,360,365,366] and can theoretically occur over distances as great as 10 Å and is less dependent on orientation[359].

The fluorescence emission spectra shown in Figure 3.11 have been corrected for protein concentration, such that any differences in the intensities of spectra are due to fluorescence properties of G-CSF proteins. Generally, spectra of G-CSF wt are more intense than G-CSF C3 due to the extra Trp residue found in the wild-type protein.

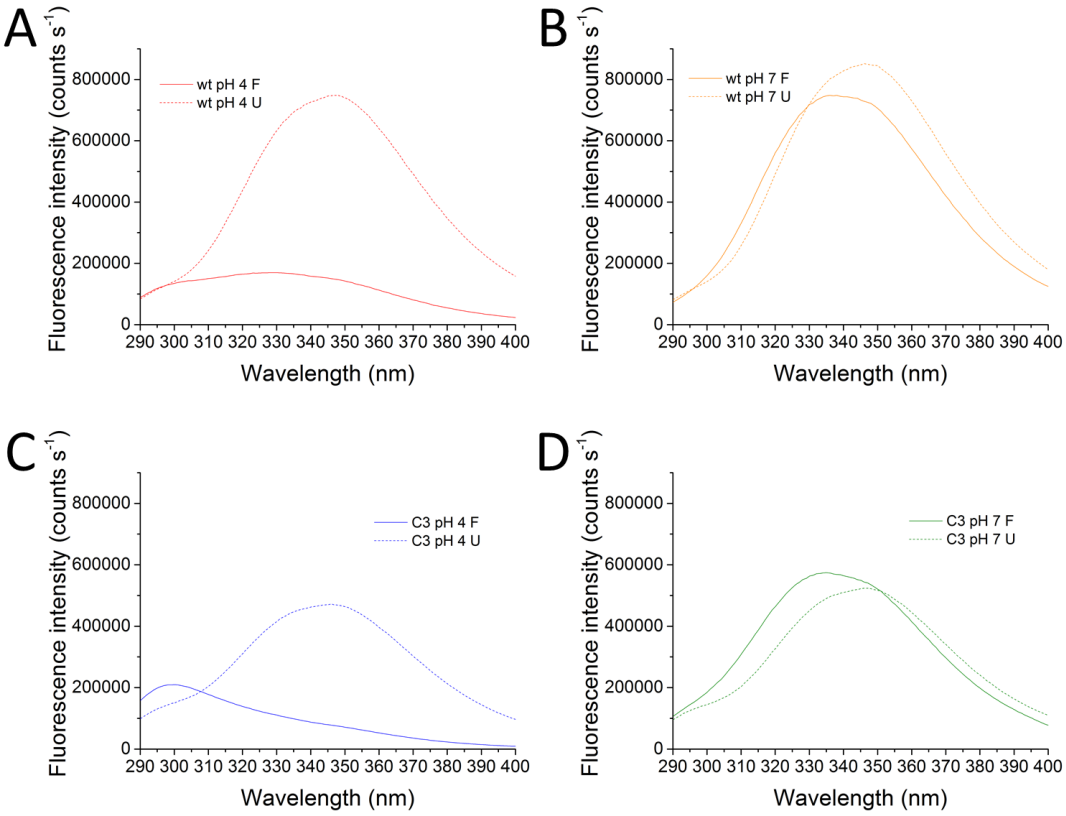


Figure 3.11 Fluorescence emission spectra of G-CSF wt and C3 folded and unfolded. Spectra of G-CSF wt were acquired using protein concentrations of 5 and 4 µM in 20 mM sodium phosphate, 20 mM sodium acetate pH 4 buffer (I = 43 mM) and 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide pH 7 buffer (I = 80 mM), respectively (A/red and B/orange, respectively), while spectra of G-CSF C3 were acquired using protein concentrations of 8 and 5 µM in pH 4 (C) and 7 (D) buffers, respectively (blue and green, respectively). Fluorescence emission intensities have been corrected for the different protein concentrations used. pH 7 buffer without azide was used for acquiring spectra because the same samples were also used for acquiring far-UV CD spectra (see the legend of Figure 3.10). Spectra of folded (F) and unfolded (U) proteins are represented by solid and dashed lines, respectively, with spectra of unfolded proteins being acquired in the presence of 8.1 M urea.

Spectra of G-CSF wt and C3 at pH 7 show emission maxima at 336 and 335 nm, respectively (Figure 3.11B and D). Upon addition of 8.1 M urea a red-shift in the wavelength of maximum emission is observed to 346 nm, indicating solvent exposure of Trp due to unfolding of tertiary structure. The spectra of G-CSF wt and C3 at pH 4 (Figure 3.11A and C) exhibit emission intensities around 335 nm far lower than the intensities observed for emission of G-CSF wt and C3 around this same wavelength at pH 7. This is indicative of Trp fluorescence quenching and has been reported previously for G-CSF wt at pH 4[306].

The emission intensities of spectra of G-CSF wt and C3 at pH 4 are highest at wavelengths of 329 and 300 nm, respectively (Figure 3.11A and C), which is most

likely due to the W59R substitution found in G-CSF C3. As well as removal of a Trp residue contributing to reduced emission intensity around 335 nm, Trp59 has been proposed to be involved in quenching Tyr fluorescence[306]. Consequently, removal of Trp59 brings about a shift in the wavelength of maximum emission observed in the spectrum of G-CSF C3 compared to the wild-type protein at pH 4 to 300 nm, which is characteristic of Tyr fluorescence emission[367]. Addition of 8.1 M urea results in a red-shift in the wavelength of maximum emission observed to 347 and 346 nm for G-CSF wt and C3, respectively at pH 4. This red-shift is accompanied by an increase in emission intensity owing to the removal of Trp fluorescence quenching caused by G-CSF unfolding.

Intriguingly, an increase in the overall emission intensity is observed in the unfolded spectrum of G-CSF wt at pH 7 (Figure 3.11B), while a decrease in the emission intensity is observed in the unfolded spectrum of G-CSF C3 at pH 7 (Figure 3.11D). This would suggest quenching of Trp59 occurs at both pH 4 and 7 in G-CSF wt, which is removed upon unfolding. Again, this may be due to energy transfer between Trp59 and a proximal Tyr side chain in the wild-type protein[306]. The structure of G-CSF wt suggests the pH independent quenching of Trp59 fluorescence is most likely due to the proximal side chain of Tyr86 (Figure 3.12A). On the other hand, quenching of Trp119 fluorescence only occurs at acidic pH and is most likely due to the proximal side chain of His80 (Figure 3.12B).

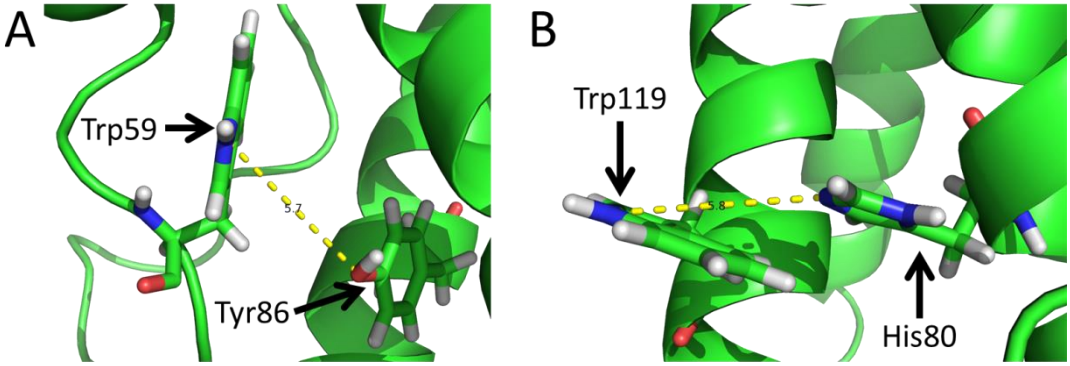


Figure 3.12 Proximity of Trp side chains in G-CSF wt to amino acid side chains suspected to be involved in Trp fluorescence quenching. The indole nitrogen atom of Trp59 is 5.7 Å away from the oxygen atom in the phenol side chain of Tyr86 (A). The indole nitrogen atom of Trp119 is 5.8 Å away from the deprotonated nitrogen atom of the imidazole side chain of His80 that becomes protonated at acidic pH (B). Distances in Å are shown for atoms connected by yellow dotted lines. This figure was constructed using the structure of G-CSF wt (PDB 1GNC[283]) in PyMOL[289].

Unusually, the spectra of unfolded G-CSF wt at pH 4 and 7 exhibit different intensities (Figure 3.11A and B), a phenomenon which is also observed for G-CSF C3 (Figure 3.11C and D). To determine if this is due to disulphide bonds still being present in urea denatured G-CSF proteins, 1 mM TCEP was added to G-CSF C3 samples unfolded in 8.1 M urea at pH 4 and 7 and fluorescence emission spectra acquired (Figure 3.13). Reduction of disulphide bonds in G-CSF C3 causes these spectra to overlay around wavelengths associated with Trp fluorescence (around 345 nm), suggesting that the difference in intensities of spectra of unfolded and oxidised G-CSF wt and C3 seen in Figure 3.11 is likely due to disulphide bonds, which are effective quenchers of Trp fluorescence[359].

Reduction of disulphide bonds also causes a difference in intensity around wavelengths associated with Tyr fluorescence (around 300 nm) as shown in Figure 3.13, which may be due to the pH dependence of Tyr fluorescence. In the ground state, the pK_a of the phenolic hydroxyl group of Tyr is 10, but in the excited state the pK_a of this group decreases to approximately 4-5. Ionisation of the phenolic side chain of Tyr causes tyrosinate formation and a red-shift of the wavelength at which maximum emission is observed to 340 nm[367]. Therefore, a higher population of tyrosinate side chains compared to tyrosine side chains in the unfolded and reduced G-CSF C3 sample at higher pH could be responsible for the reduced intensity observed around 300 nm. The expected increase in fluorescence emission at 340 nm as a result of tyrosinate formation in the higher pH sample is most likely obscured by Trp fluorescence.

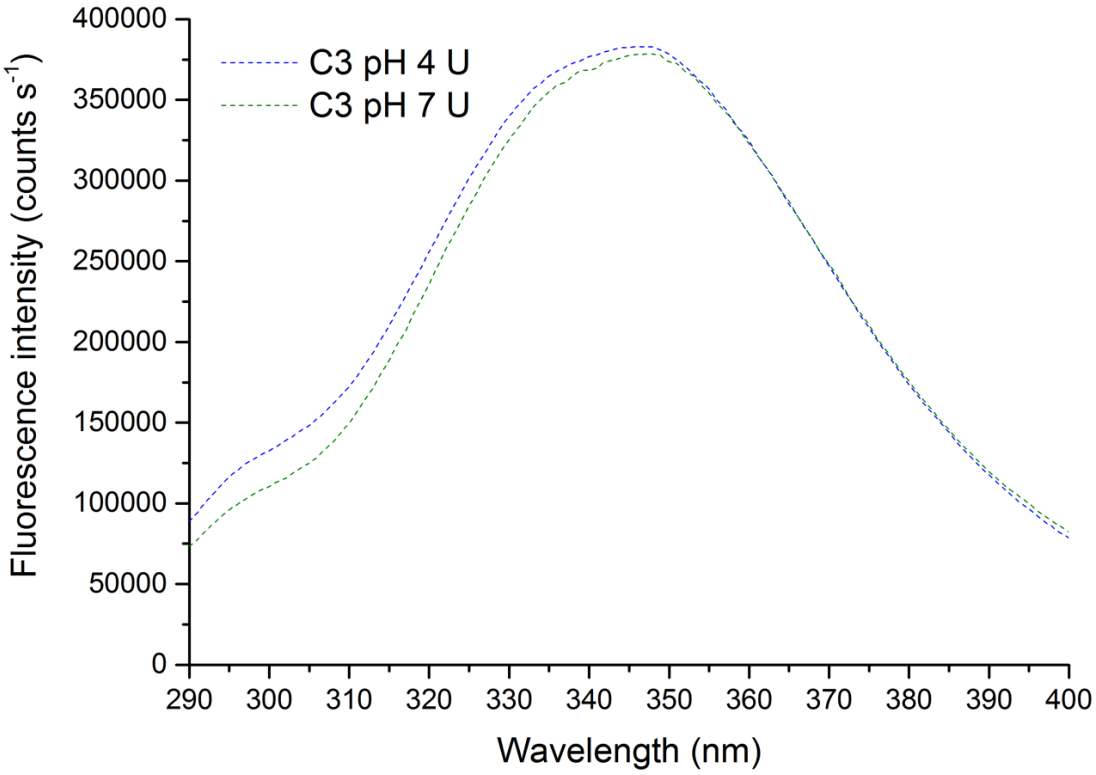


Figure 3.13 Fluorescence emission spectra of unfolded and reduced G-CSF C3. Spectra of G-CSF C3 were acquired using protein concentrations of 8 and 5 µM in 20 mM sodium phosphate, 20 mM sodium acetate pH 4 buffer (I = 43 mM) and 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide pH 7 buffer (I = 80 mM), respectively (blue and green, respectively) in the presence of 8.1 M urea and 1 mM TCEP. Spectra were acquired after addition of TCEP without readjustment of pH. Fluorescence emission intensities have been corrected for the different protein concentrations used. pH 7 buffer without azide was used for acquiring spectra because the same samples were also used for acquiring far-UV CD spectra (see the legend of Figure 3.10).

3.2.7 Nuclear Magnetic Resonance (NMR) spectroscopy of G-CSF proteins

The ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) spectrum provides a peak for every amide proton of a protein, allowing high resolution structural comparison of G-CSF wt and C3. For NMR experiments 10 % (*v/v*) D_2O was added to G-CSF samples, resulting in 10 % dilution of pH 4 and 7 buffer components. Consequently, all NMR experiments were carried out in 18 mM sodium phosphate, 18 mM sodium acetate, 10 % (*v/v*) D_2O pH 4 buffer or 18 mM sodium phosphate, 18 mM sodium acetate, 0.018 % (*w/v*) sodium azide, 10 % (*v/v*) D_2O pH 7 buffer unless otherwise stated.

3.2.7.1 Comparing G-CSF C3 from the peak and tail of cation exchange chromatography using ^{1}H-^{15}N HSQC spectra

^{1}H-^{15}N HSQC spectra were used to probe for differences in G-CSF C3 obtained from the peak or tail portions of the elution trace from cation exchange chromatography. Spectra acquired using the same parameters, with very similar concentrations of G-CSF C3 from peak and tail samples in the same buffer conditions revealed no differences between protein acquired from the peak or tail of cation exchange chromatography (Figure 3.14).

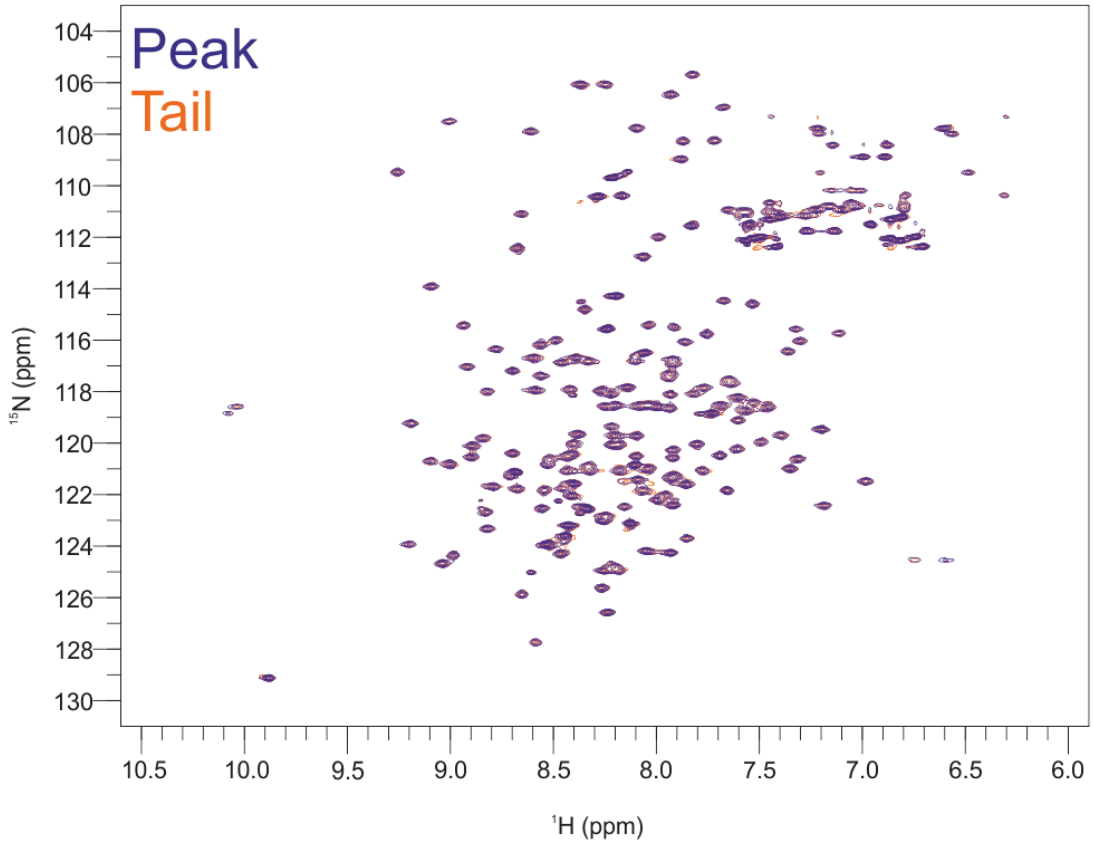


Figure 3.14 ^{1}H-^{15}N HSQC spectra of G-CSF C3 from peak and tail samples. Spectra were acquired using 173 and 177 µM ^{15}N G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 10 % (*v/v*) D_2O pH 4 buffer (I = 39 mM) from the peak (blue) and tail (orange) of cation exchange chromatography, respectively. Spectra were acquired at 25 ^{o}C using a spectrometer operating at a proton resonance frequency of 600 MHz.

3.2.7.2 Assignment of ^{1}H-^{15}N HSQC spectra

From assignment of the ^{1}H-^{15}N HSQC spectra of G-CSF proteins at pH 4 the positions of 160 non-proline residues were identified, but peaks for Ser97 and the N-terminal Met residue could not be assigned. The N-terminal residue is often not observable in ^{1}H-^{15}N HSQC spectra due to rapid exchange with the solvent, while the peak for Ser97 is presumed to be too broad for detection or overlapping with other peaks. The ^{1}H-^{15}N HSQC spectrum of G-CSF wt at pH 4 was assigned based on previous assignments of wild-type protein (Figure 3.15), in which assignments of Ser97 and the N-terminal Met residue were also not reported[283,288].

Despite G-CSF C3 only containing four substitutions relative to the wild-type protein, the ^{1}H-^{15}N HSQC spectrum of G-CSF C3 was too different from the spectrum of wild-type G-CSF to allow assignment from reported assignments of wild-type protein alone[283,288]. Therefore, ^{1}H-^{15}N HSQC spectra for G-CSF C3 at pH 4 and 7 were assigned using HNCA, HN(CO)CA, CBCA(CO)NH, HNCO and HN(CA)CO triple-resonance spectra. Example strips from HNCA, HN(CO)CA and CBCA(CO)NH spectra are shown in Figure 3.16. At pH 4, all non-proline residues in G-CSF C3 were assigned apart from Ser97 and Met1 (Figure 3.17), while at pH 7 Thr2, His44 and Ala144 also could not be assigned. These peaks may be lost at higher pH due to a higher rate of exchange with the solvent causing broadening. Alternatively, these peaks may be broadened as a result of changes in local polypeptide dynamics or conformational exchange brought about by this pH change.

^{1}H-^{15}N HSQC spectra of G-CSF wt and C3 at pH 4 both possess some residues that exhibit two or more resonances. These satellite peaks are observed for residues in unstructured regions of G-CSF around Pro residues and have been attributed to slow conformational exchange resulting from cis-trans isomerisation of peptide bonds with Pro residues[288]. In addition, extra peaks are observed around some Gln residues in G-CSF wt and C3 e.g. Gln33 and Gln174, which most likely correspond to deamidation of Gln residues. Tables of assignments for amide proton and nitrogen atoms of G-CSF wt at pH 4 and G-CSF C3 at pH 4 and 7, as well as alpha, beta and carbonyl carbon atoms of G-CSF C3 at pH 4 and 7 are shown in Appendices 7.6-7.8.

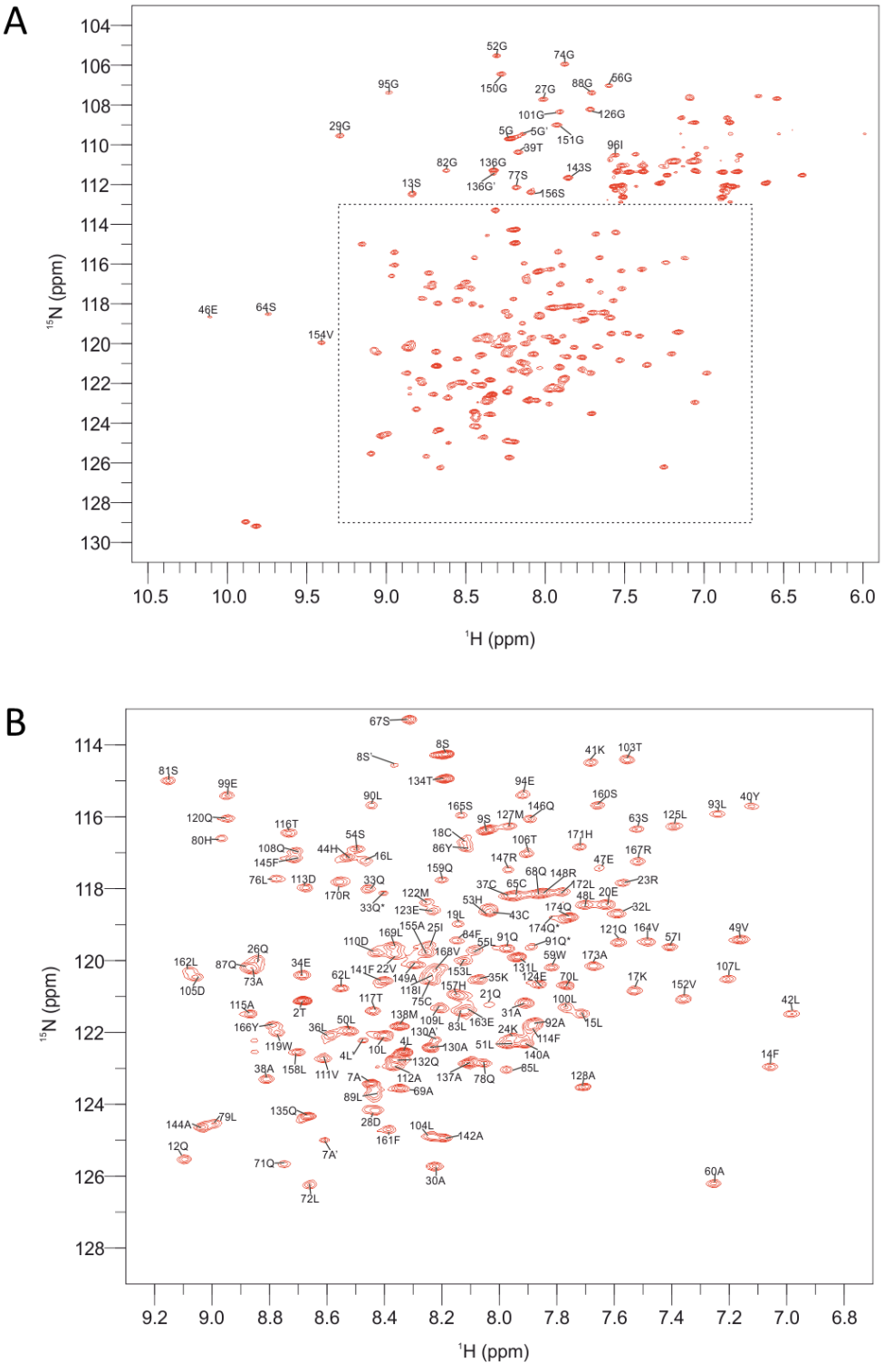


Figure 3.15 ^{1}H-^{15}N HSQC spectrum of wild-type G-CSF assigned at pH 4. The spectrum was acquired using 111 µM ^{15}N G-CSF wt in 18 mM sodium phosphate, 18 mM sodium acetate, 10 % (*v/v*) D_2O pH 4 buffer (I = 39 mM) and assigned using previous assignments[283,288]. Peaks labelled with an apostrophe indicate satellite peaks, while peaks labelled with an asterisk indicate peaks suspected to result from deamidation of Gln residues. The region highlighted in the dotted square in A is shown magnified in B. Unlabelled peaks correspond to side chains or noise. The spectrum was acquired at 25 ^{o}C using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

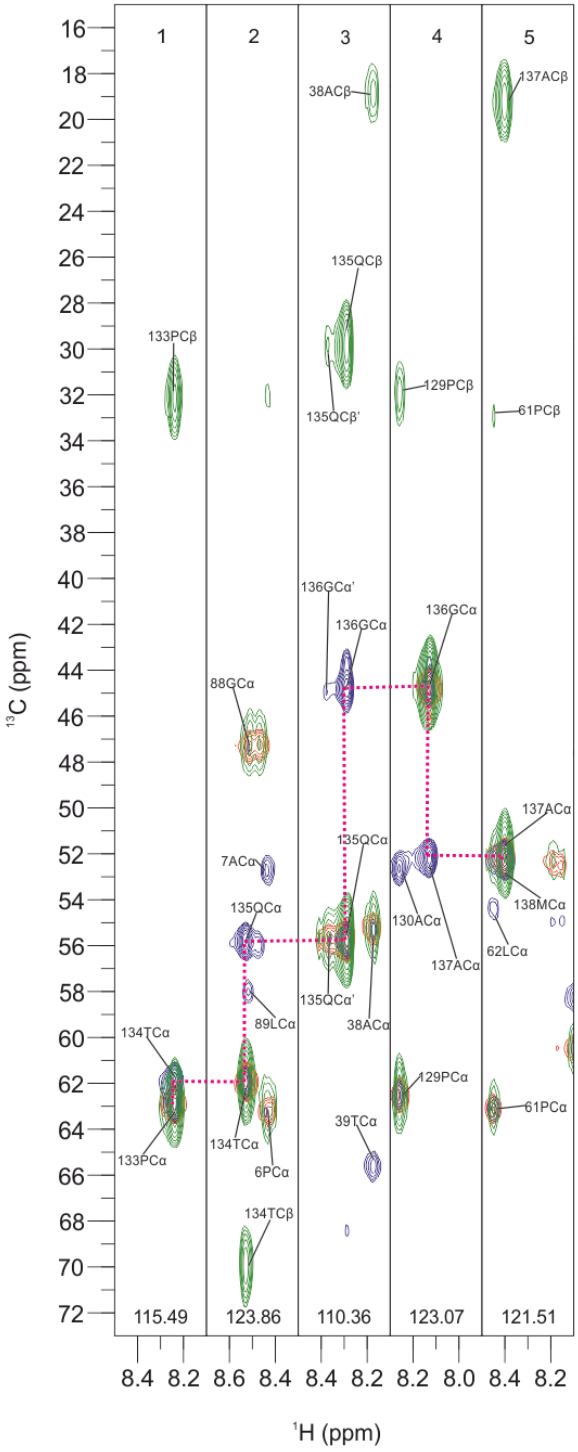


Figure 3.16 Example strips of HNCA, HN(CO)CA and CBCA(CO)NH spectra used for assignment of the ^{1}H-^{15}N HSQC spectrum of G-CSF C3. HNCA (blue), HN(CO)CA (red) and CBCA(CO)NH (green) spectra were acquired using 473 µM ^{13}C ^{15}N G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 10 % (*v/v*) D_2O pH 4 buffer (I = 39 mM). Peaks are shown for assignment of residues T134-M138 of G-CSF C3 and a pink dotted line is drawn to indicate linking of C_α atoms of sequential residues along the backbone. Numbers at the top of each strip indicate the strip number and numbers at the bottom of each strip (within each strip) indicate the nitrogen chemical shift in ppm. The nitrogen chemical shift axis is perpendicular to the proton and carbon chemical shift axes. Peaks labelled with an apostrophe are satellite peaks. Spectra were acquired at 25 ^{o}C using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

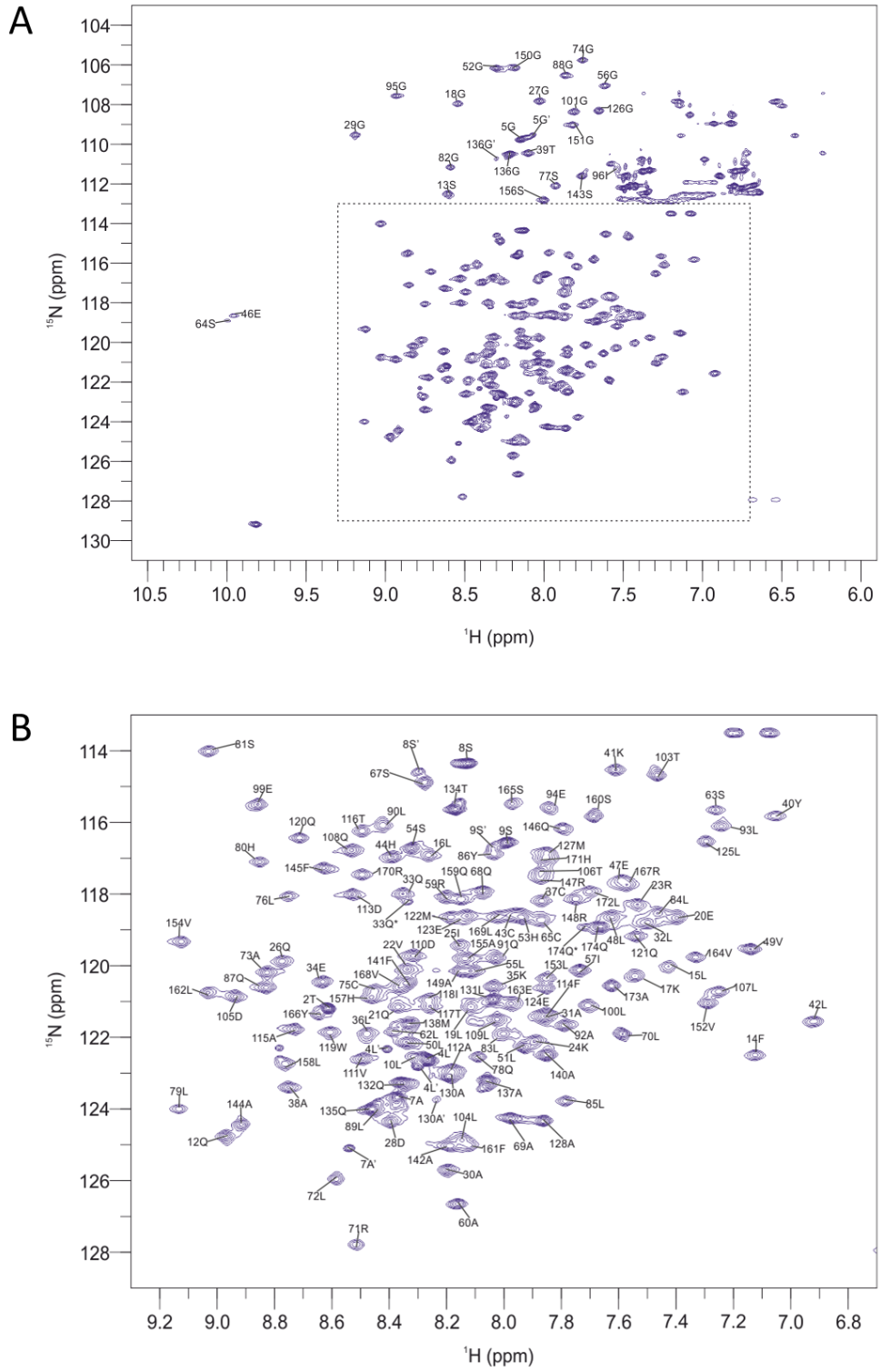


Figure 3.17 ^{1}H-^{15}N HSQC spectrum of G-CSF C3 assigned at pH 4. The spectrum was acquired using 473 µM ^{13}C ^{15}N G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 10 % (*v/v*) D_2O pH 4 buffer (I = 39 mM) and assigned using HNCA, HN(CO)CA, CBCA(CO)NH, HNCO and HN(CA)CO triple-resonance spectra. Peaks labelled with an apostrophe are satellite peaks, while peaks labelled with an asterisk indicate peaks suspected to result from deamidation of Gln residues. The region highlighted in the dotted square in A is shown magnified in B. Unlabelled peaks correspond to side chains or noise. The spectrum was acquired at 25 °C using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

3.2.7.3 Comparing the structures of G-CSF wt and C3 using chemical shift perturbations (CSPs)

To determine the structural similarity of G-CSF wt prepared in this study to wild-type protein used in previous NMR studies CSPs were calculated from amide proton and nitrogen chemical shifts (see Section 2.9.6). A CSP is the Euclidean distance between two peaks and so can provide a measure of how different the chemical environments of a particular residue are under different conditions. CSPs were calculated using the assignments of G-CSF wt at pH 4 from this study and previous assignments of the wild-type protein from studies by Zink *et al*.[283] and Werner *et al*.[288] (Figure 3.18A and B, respectively).

^{1}H-^{15}N HSQC spectra of G-CSF wt in this study were acquired in 18 mM sodium phosphate, 18 mM sodium acetate pH 4 buffer at 25 ^{o}C, while spectra in the study by Zink *et al*.[283] were acquired in 50-60 mM sodium phosphate pH 3.5 buffer at 27 ^{o}C. CSPs for the majority of residues of G-CSF wt used in this study in relation to G-CSF wt used by Zink *et al*.[283] are equal to or lower than 0.1 (Figure 3.18A), suggesting that there are only small differences between G-CSF wt prepared in this study with previous preparations. Additionally, the residues that exhibit greatest CSPs are distributed throughout the protein rather than being clustered in certain structural segments, suggesting that these differences are likely the result of the different buffer conditions and temperature used in this study rather than differences in the conformations of G-CSF wt used in these studies.

Spectra in the study by Werner *et al*.[288] were acquired in 50 mM deuterioacetic acid pH 3.9 buffer at 30-32 ^{o}C. Again, these are different buffer conditions and temperatures to those used for acquiring ^{1}H-^{15}N HSQC spectra of G-CSF wt in this study and this is reflected in the small CSPs observed for most residues of G-CSF wt used in these studies (Figure 3.18B). However, large CSPs are observed for residues Leu15, Lys17, Leu19, Glu20 and Gln21, which is probably due to the C18S substitution found in the G-CSF protein used by Werner *et al*.[288]. Significantly large CSPs are also observed for Ile118 (coloured red in Figure 3.18B) and other proximal residues in the amino acid sequence, which is most likely due to the proximity of these residues to Cys18 in the tertiary structure of G-CSF.

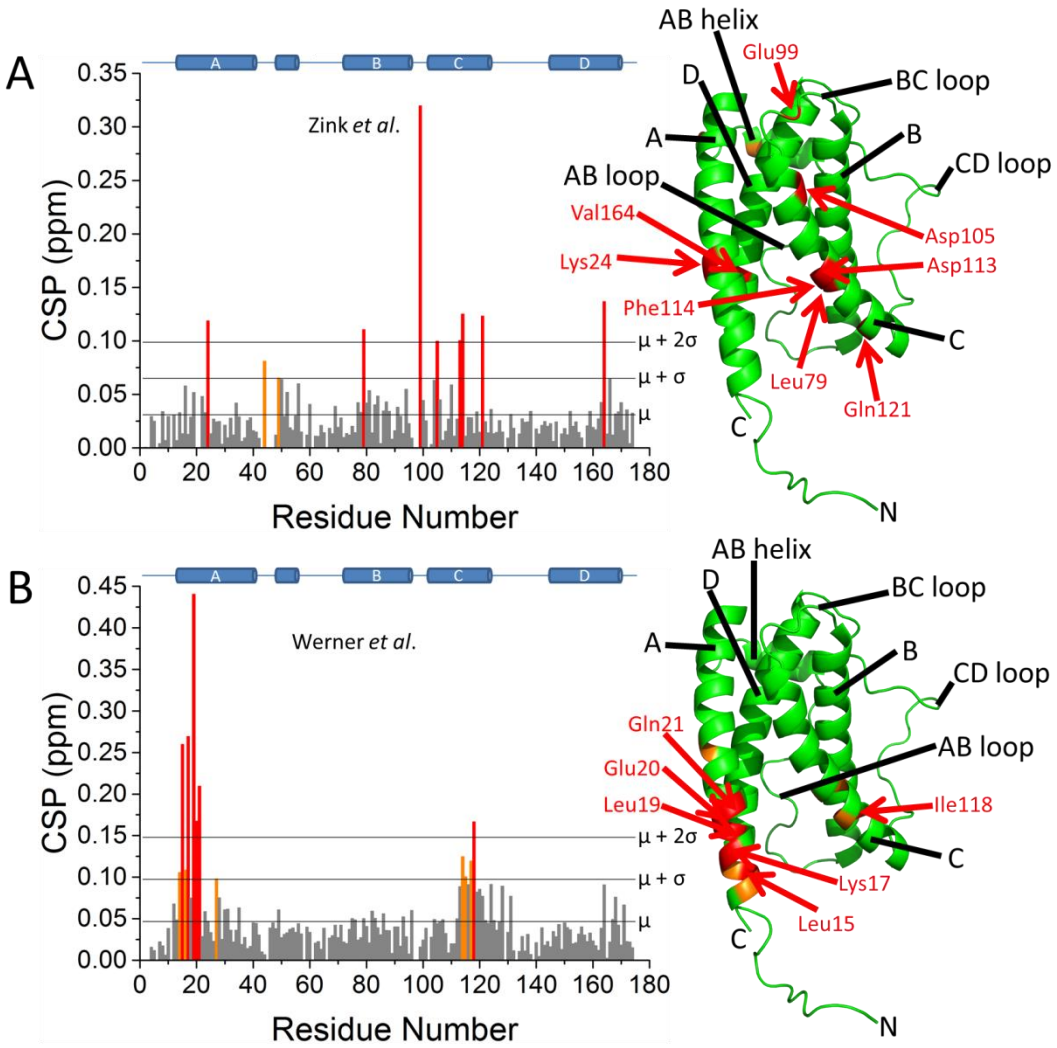


Figure 3.18 CSPs for previous assignments of G-CSF wt relative to assignments of G-CSF wt used in this study. Previous assignments used were from Zink *et al.*[283] (A) and Werner *et al.*[288] (B). Horizontal lines show the mean, mean + σ and mean + 2σ CSP values. Orange and red bars correspond to residues exhibiting CSPs greater than the mean + σ and mean + 2σ, respectively. A secondary structure schematic is shown above for reference. Right: Residues with significant CSPs are shown mapped onto the structure of G-CSF wt (PDB 1GNC[283]) using the same colour scheme, constructed using PyMOL[289]. Red residues are also labelled.

CSPs were also used to compare the structures of G-CSF wt and C3 at pH 4. Unsurprisingly, the largest CSPs for G-CSF C3 relative to wild-type are the four substitutions C18G, W59R, Q71R and F84L, along with some neighbouring residues of these substitutions (Figure 3.19). In particular, substitution of Cys for Gly brings about a large CSP, due to the lower characteristic chemical shift value of amide nitrogen atoms of Gly residues. Significant CSPs are also observed for Val154 and His157, which is most likely due to the proximity of these residues to the W59R substitution site in the tertiary structure of G-CSF. Therefore, an effect of the W59R substitution on the local conformation of G-CSF in the region of the latter half of the

AB loop and the section of helix D containing Val154 and His157 cannot be ruled out. A large CSP is also observed for Leu169, which is proximal to the C18G substitution. Apart from around the sites of these four substitutions, CSPs throughout the protein appear to be of a similarly low amplitude, suggesting no conformational differences between G-CSF wt and C3 beyond the local environments of these four substitutions.

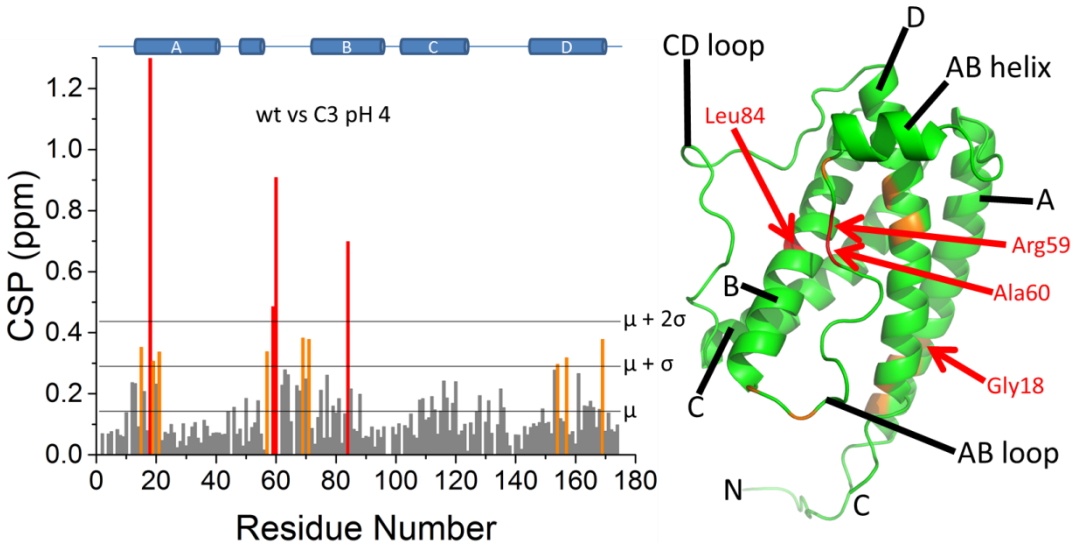


Figure 3.19 CSPs for G-CSF wt relative to the variant C3 at pH 4. Horizontal lines show the mean, mean + σ and mean + 2σ CSP values. Orange and red bars correspond to residues exhibiting CSPs greater than the mean + σ and mean + 2σ, respectively. A secondary structure schematic is shown above for reference. Right: Residues with significant CSPs are shown mapped onto the structure of G-CSF wt (PDB 1GNC[283]) using the same colour scheme, constructed using PyMOL[289]. Red residues are also labelled.

Assignment of the ^{1}H-^{15}N HSQC spectra of G-CSF C3 at pH 4 and 7 enabled CSPs to be calculated for G-CSF C3 under these different conditions. Many of the largest CSPs are observed for His residues and their neighbouring residues (Figure 3.20). This is due to the protonation of His side chains that occurs at pH 4 but not at pH 7. G-CSF possess five His residues, which are located at positions 44, 53, 80, 157 and 171 in the amino acid sequence. Unfortunately, His44 could not be assigned in the ^{1}H-^{15}N HSQC spectrum of G-CSF C3 at pH 7, along with Thr2 and Ala144, presumably because these peaks are broadened at higher pH or overlap with other peaks in the spectrum.

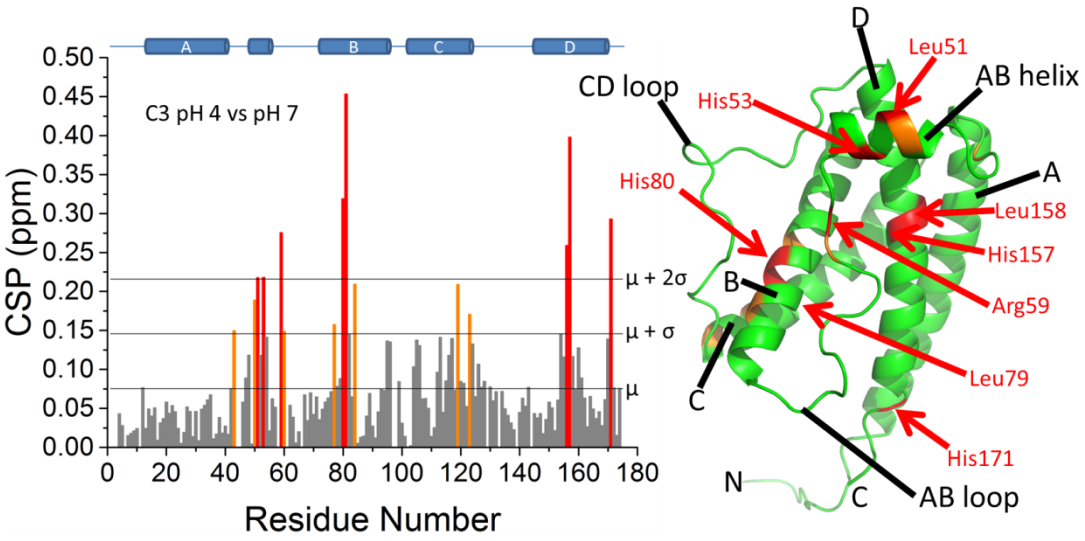


Figure 3.20 CSPs for G-CSF C3 at pH 4 relative to G-CSF C3 at pH 7. Horizontal lines show the mean, mean + σ and mean + 2σ CSP values. Orange and red bars correspond to residues exhibiting CSPs greater than the mean + σ and mean + 2σ, respectively. A secondary structure schematic is shown above for reference. Right: Residues with significant CSPs are shown mapped onto the structure of G-CSF wt (PDB 1GNC[283]) using the same colour scheme, constructed using PyMOL[289]. Red residues are also labelled.

Interestingly, Arg59 exhibits a large CSP, despite there being no nearby His residues in the amino acid sequence. The structure of G-CSF wt shows that the side chains of His53 and His157 are in close proximity to each other and the side chain of Arg59 (Figure 3.21). Therefore, protonation of His53 and His157 side chains most likely causes electrostatic repulsion between these residues and Arg59 around the short helical section of the AB loop, possibly affecting the local structure in this region. Other than in this region and around other His residues, CSPs throughout the protein sequence are of a similarly low magnitude, suggesting no significant differences in the conformation of G-CSF C3 beyond the local environment of His residues at pH 4 and 7.

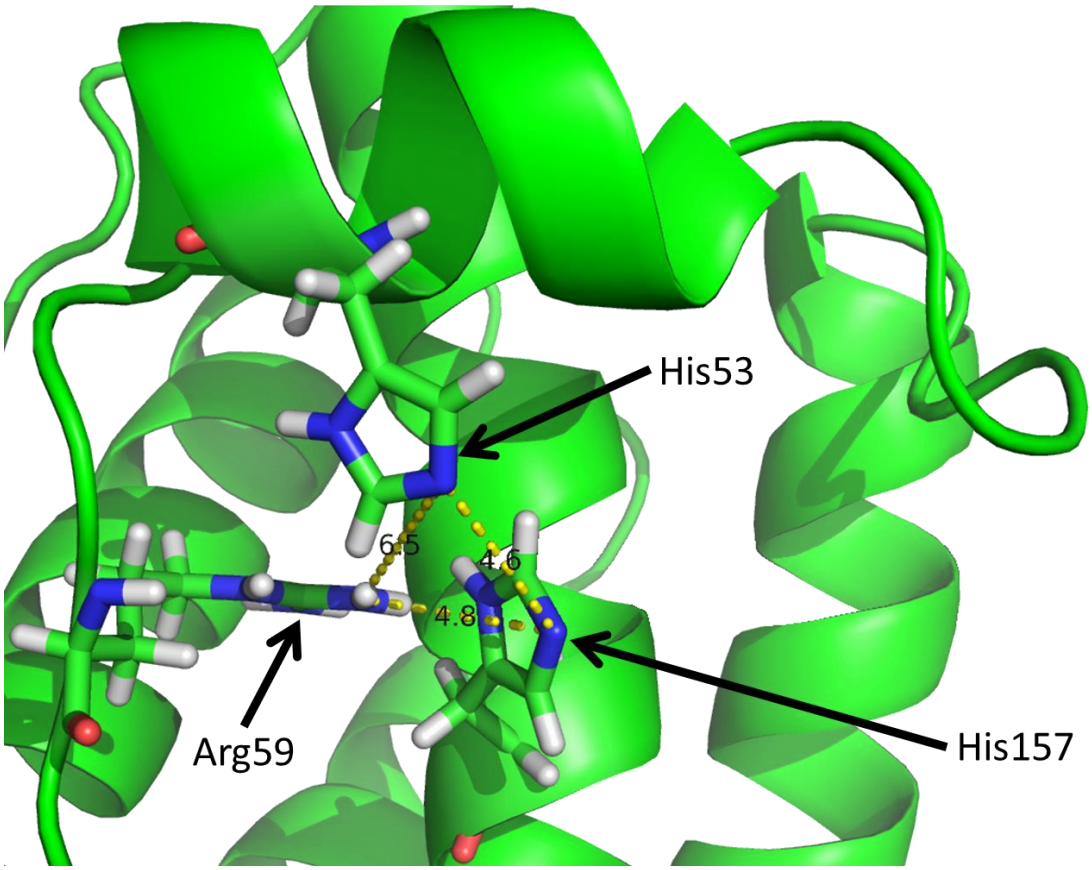


Figure 3.21 Proximity of the side chains of His53, His157 and Arg59 in G-CSF C3. The deprotonated imidazole nitrogen atom from the side chain of His53 is 4.6 Å away from the equivalent atom of His157 and 6.5 Å away from the closest nitrogen atom of the guanidine group in the Arg59 side chain. The deprotonated imidazole nitrogen atom from the side chain of His157 is 4.8 Å away from the closest guanidine nitrogen atom of the Arg59 side chain. Although shown deprotonated in this figure, these imidazole nitrogen atoms are protonated at acidic pH. Distances in Å are shown for atoms connected by yellow dotted lines. This figure was constructed using the structure of G-CSF wt (PDB 1GNC[283]) in PyMOL[289]. To create this figure Trp59 in the G-CSF wt structure has been mutated to an Arg residue.

3.3 Discussion

Recombinantly expressed and purified human G-CSF must reach its native conformation in order to function as a biopharmaceutical. Since the purpose of this study is to investigate the aggregation of G-CSF as a biopharmaceutical it was imperative to establish that the native conformation had been reached under the conditions that would be used to study aggregation. Additionally, aggregation of G-CSF is thought to occur from a native-like state[142,303,307,308], meaning biophysical characterisation of the native conformation was essential to probe for subtle changes that may be linked to aggregation. Investigation into the structure of G-CSF C3 and subsequent comparison to wild-type may also help to rationalise the vast

increase in expression level of soluble G-CSF C3 when expressed in the periplasm of *E. coli*[150].

As well as confirming the correct molecular masses of G-CSF proteins prepared in this study (Figure 3.6), ESI-MS was used to show that disulphide bonds linking residues Cys37-Cys43 and Cys65-Cys75 had formed (Figure 3.7 and Figure 3.8). These disulphide bonds are required for G-CSF wt to form its native conformation[285].

Far-UV CD spectra of G-CSF wt and C3 showed that these proteins have an estimated 59 % and 60 % α-helical structure, respectively at pH 7 (Figure 3.10), which matches the percentage of α-helical structure reported previously for G-CSF wt[353]. The estimated percentages of α-helical structure for G-CSF wt and C3 at pH 4 are 64 % and 70 %, respectively. Previously, G-CSF wt had been reported to have an estimated 61 % α-helical structure at pH 4[353], which is similar to G-CSF wt used in this study but not the variant C3, suggesting G-CSF C3 may be slightly more α-helical than wild-type. However, other previous studies have reported estimated percentages of 75[354] and 69[276] % α-helical structure for G-CSF wt at pH 4.5, making it unclear whether the differences in percentage α-helical structure estimated for G-CSF wt and C3 at pH 4 are significant.

Fluorescence emission spectra of G-CSF wt and C3 showed Trp fluorescence quenching at pH 4 (Figure 3.13A and C). This has been reported previously for G-CSF wt[306] and shows that Trp119 is in a similar chemical environment in G-CSF wt and C3, with quenching most likely the result of protonation of the proximal side chain of His80 (Figure 3.12B). In addition, there is additional Trp fluorescence quenching observed for G-CSF wt at pH 7 that is not observed for G-CSF C3 (Figure 3.11B and D). This is evident from the increase in fluorescence intensity of emission spectra of G-CSF wt at pH 7 upon unfolding, plus the intensity of emission spectra of G-CSF C3 at pH 7 decreasing upon unfolding. This quenching has been attributed to Tyr residues[306] and must involve Trp59 as this Trp residue is substituted to Arg in G-CSF C3. Substitution of Trp59 for Arg also causes an increase in Tyr fluorescence intensity for G-CSF C3 compared to wild-type protein at pH 4, suggesting quenching of Tyr fluorescence by Trp side chains. The structure of G-CSF wt suggests that quenching of Trp59 fluorescence is caused by the side chain of Tyr86 (Figure 3.12A). Overall, fluorescence emission spectra suggest that the G-CSF wt used in this study possesses similar tertiary structure around Trp

residues to G-CSF wt studied previously by fluorescence[306] and that G-CSF C3 possesses a similar tertiary structure around Trp119.

Assignments of amide proton and nitrogen atoms of G-CSF wt from ^{1}H-^{15}N HSQC spectra allowed calculation of CSPs, providing high resolution structural comparison of G-CSF wt prepared in this study to previous studies of G-CSF wt by NMR (Figure 3.18). The results suggest that there is little difference between G-CSF wt prepared in this study and G-CSF wt prepared in previous studies, with differences suspected to be the result of the different buffer conditions and temperatures used for acquisition of spectra, as well as the C18S substitution found in the G-CSF protein used by Werner *et al*.[288].

CSPs of G-CSF C3 relative to the wild-type protein at pH 4 suggest that there is no observable difference in conformation beyond the four substitutions found in the variant C3. However, residues Val154 and His157 around the W59R substitution exhibit higher CSPs than residues throughout the rest of G-CSF C3, despite not being close in amino acid sequence to the site of this substitution, meaning a local structural difference around this region in G-CSF C3 compared to the wild-type protein cannot be excluded (Figure 3.19). Additionally, CSPs calculated for G-CSF C3 at pH 4 relative to G-CSF C3 at pH 7 imply a possible difference in local structure around the short helical section of the AB loop (Figure 3.20), likely due to electrostatic repulsion between the Arg59 side chain and the side chains of His53 and His157 that are protonated at pH 4 but not at pH 7 (Figure 3.21).

The W59R substitution is located in the loop connecting helices A and B in G-CSF. This AB loop has been hypothesised to undergo rearrangement under aggregation-prone conditions, promoted by the addition of GdnHCl concentrations that are too low to cause global denaturation of the protein at pH 7[142]. Rearrangement of the AB loop was identified from a change in the fluorescence intensity of Trp59 in low GdnHCl concentrations, while the remainder of G-CSF was shown to be native-like[142]. Under less aggregation-prone conditions at pH 4, a conformation of G-CSF wt with fully formed native α-helical structure but non-native fluorescence of Trp59 has also been identified as a late stage kinetic folding intermediate[304]. Therefore, substitution of Trp59 in the AB loop in G-CSF C3 could alter the local structure of this loop without affecting the global conformation, instilling a reduced propensity for aggregation. Alternatively, substitution of uncharged residues for charged residues in this region may disturb a hydrophobic aggregation interface or simply increase

electrostatic repulsion between G-CSF molecules, thus hindering aggregation without affecting the structure of the AB loop.

Substitution of the unpaired Cys18 residue for Gly in G-CSF C3 may also contribute to reduced aggregation propensity by preventing intermolecular disulphide bond formation, without affecting the overall conformation of G-CSF. Substitution of Cys18 for Ala has been shown to slow the rate of G-CSF aggregation and prevent formation of disulphide cross-linked aggregates, although ultimately wild-type and C18A proteins were reported to aggregate to the same extent[142]. Considering the variant C3 was identified from screening of a randomly generated library of G-CSF DNA constructs, it is possible that not all of the four substitutions contribute to the observed increase in soluble expression level of G-CSF C3 relative to the wild-type protein. This could account for the incorporation of the F84L substitution in the variant C3 as there is no reported involvement of this residue in G-CSF aggregation.

Although it has been suggested that the observed increase in expression of soluble G-CSF C3 is due to reduced aggregation propensity relative to wild-type, this has not been confirmed experimentally. This assumption is supported by the reduced aggregation propensity predicted around substitutions W59R and Q71R, based on computational modelling of G-CSF wt and C3[150]. However, given that aggregation is commonly driven by hydrophobic interactions, it is unsurprising that these substitutions are predicted to reduce aggregation propensity, particularly W59R. Additionally, the cytoplasmic expression level of G-CSF C3 is higher than wild-type (Figure 3.2), even though expression of G-CSF in the cytoplasm yields insoluble inclusion bodies due to the inability to form disulphide bonds[130]. In turn, this suggests that the heterologous expression levels of G-CSF wt and C3 in *E. coli* contribute significantly to the improved expression of soluble G-CSF C3 in the periplasm relative to wild-type protein[150], as opposed to only being the result of the hypothesised reduced aggregation propensity of G-CSF C3.

In the early stages of this study, recombinant expression of G-CSF wt and C3 was carried out in the periplasm of *E. coli* using pCANTAB6, a MedImmune in-house vector provided by Dr. Andrew Buchanan (MedImmune, UK). The realtive yields of G-CSF wt and C3 from periplasmic expression reflected the yields reported by Buchanan *et al*[150] (results not shown). However, analysis of the molecular mass of G-CSF C3 purified from the periplam by ESI-MS (carried out by Dr. Ault) revealed cleavage of the N-terminal Thr residue. More importantly though, expression in

pCANTAB6 resulted in the addition of 26 extra residues at the C-terminus of G-CSF proteins, encoding Myc and hexahistidine tags as well as linker residues between and on either side of these detection motifs that are encoded by pCANTAB6. These tags are not present in commercially available formulations of G-CSF and could have affected the folding, dynamics and aggregation of G-CSF proteins. Therefore, untagged G-CSF was chosen for this study in order to reflect more closely G-CSF as it is found in biopharmaceutical formulations used clinically.

The recombinant expression level of G-CSF C3 is dramatically higher than the wild-type protein, regardless of whether these proteins are expressed in their soluble native forms in the periplasm[150] or as insoluble aggregates in the cytoplasm of *E. coli* (Figure 3.2). This would suggest that the higher expression yield of G-CSF C3 is related to the transcription and/or translation rates of these proteins. Although degradation by proteases also affects the recombinant expression level of soluble proteins, proteins in inclusion bodies are generally protected from this[176]. Theoretically, the transcription rates of G-CSF wt and C3 DNA constructs should be comparable due to their incorporation into the same vector.

The translation rates of mRNA transcripts can be affected by differential usage of codons in prokaryotic and eukaryotic cells. Rarely used codons have fewer cognate tRNAs, which slow the rate of translation[368]. Stalling is amplified in regions where rare codons are clustered and can cause amino acid substitutions and deletions, as well as premature termination of translation due to ribosome dissociation[369]. However, clustering of rare codons at 5' ends of mRNA transcripts has been suggested to destabilise mRNA secondary structure, which helps to expose the ribosome binding site and translation start codon[370].

mRNA secondary structures are related to the proportions of purine and pyrimidine bases in DNA sequences. A decrease in GC content at the 5' end of DNA encoding G-CSF has been shown to improve recombinant expression in *E. coli*, which has been hypothesised to be the result of reduced secondary structure[371]. However, A/U rich regions are recognised by the protein bS1, found in the 30S ribosomal subunit[372]. bS1 is thought to facilitate unfolding of secondary structure and incorporation of mRNA in the mRNA channel of the 30S subunit[373,374], which may also account for a general increase in recombinant expression of DNA sequences with reduced GC content at their 5' ends in *E. coli*[375]. Previous studies have rationalised improved recombinant expression of G-CSF in *E. coli* from replacement

of rare codons at the 5' end of DNA sequences with more commonly used codons, although this also resulted in a reduction in GC content[376,377].

G-CSF wt and C3 DNA constructs used by Buchanan *et al*.[150] have 18 % and 21 % rare codons, respectively, while the DNA constructs used in this study are made up of 16 % and 17 % rare codons, respectively. The DNA sequences of the constructs used in this study show similar distributions of rare codons (Figure 3.22), while the DNA sequences used by Buchanan *et al*.[150] are 99 % identical (Figure 3.23), suggesting that the notable difference in expression levels of these G-CSF proteins is not due to rare codons. Additionally, DNA constructs used by Buchanan *et al*.[150] and in this study have GC contents of 66 % and 58 %, respectively. Alignments of these DNA sequences reveal no obvious differences in the distribution of purines and pyrimidine bases between G-CSF wt and C3 DNA constructs, suggesting that the translation efficiencies of these mRNA transcripts would be similar.

```
1    ATGACTCCTC TCGGTCCTGC ATCTAGCCTT CCGCAAAGCT TCCTGCTGAA        wt
1    ATGACCCCGT TAGGCCCTGC CTCCTCTCTT CCGCAATCCT TCCTGCTGAA        C3

51   GTGCCTTGAA CAGGTCCGCA AAATCCAGGG TGATGGCGCT GCCCTGCAGG        wt
51   AGGCCTGGAA CAGGTGCGCA AAATCCAGGG TGATGGTGCA GCACTGCAAG        C3

101  AGAAACTGTG TGCGACCTAC AAACTGTGCC ATCCGGAAGA ACTGGTGCTG        wt
101  AGAAGCTGTG TGCGACCTAC AAACTGTGCC ATCCGGAAGA GCTGGTGTTG        C3

151  TTGGGCCACT CTCTGGGCAT TCCGTGGGCA CCCCTGAGTA GCTGTCCATC        wt
151  CTCGGCCATA GCCTGGGTAT TCCGCGTGCA CCCCTCAGCA GCTGTCCGTC        C3

201  CCAAGCCCTG CAATTAGCCG GCTGCTTATC GCAGCTGCAC TCAGGACTCT        wt
201  ACAGGCGTTA CGTCTGGCCG GTTGCCTTAG TCAGCTGCAC TCTGGACTGT        C3

251  TCCTGTATCA GGGCCTGTTA CAGGCGTTAG AGGGGATTTC CCCGGAACTG        wt
251  TGCTGTATCA GGGGTTACTC CAAGCCCTGG AAGGCATTTC GCCGGAACTG        C3

301  GGTCCCACAC TCGATACGCT TCAGTTGGAT GTTGCGGACT TTGCAACCAC        wt
301  GGTCCAACTC TGGATACGCT GCAGTTGGAC GTTGCGGATT TTGCCACGAC        C3

351  CATCTGGCAG CAGATGGAGG AACTGGGGAT GGCACCAGCG TTGCAACCGA        wt
351  CATCTGGCAG CAGATGGAGG AACTTGGCAT GGCTCCAGCG CTGCAACCCA        C3

401  CGCAAGGTGC GATGCCGGCC TTTGCGAGTG CCTTTCAGCG TCGTGCTGGA        wt
401  CACAAGGGGC AATGCCTGCG TTTGCTTCAG CGTTTCAACG GCGTGCGGGA        C3

451  GGTGTACTCG TGGCTTCACA TCTGCAGTCG TTCCTGGAAG TGAGCTATCG        wt
451  GGCGTACTGG TTGCCAGTCA CTTGCAGAGC TTCCTCGAAG TCTCGTATCG        C3

501  CGTTCTGCGC CATTTGGCGC AACCG                                   wt
501  CGTGTTACGC CATCTTGCTC AGCCG                                   C3
```

Figure 3.22 Alignment of DNA sequences of G-CSF wt and C3 used for cytoplasmic expression. Rare codons were identified using *E. coli* Codon Usage Analyzer 2.1[378] and are highlighted in cyan. Codons encoding different amino acids in G-CSF wt and C3 are highlighted in red. Rare codons encoding different amino acids in G-CSF wt and C3 are highlighted in green. All other nucleotide variations between G-CSF wt and C3 DNA sequences are silent and not highlighted. Vertical lines indicate sites where purine bases have been swapped for pyrimidine bases or vice versa. The nucleotide at the beginning of each line is numbered on the left.

Figure 3.23 Alignment of DNA sequences of G-CSF wt and C3 used for periplasmic expression. DNA sequences of G-CSF wt and C3 were provided by Dr. Andrew Buchanan (MedImmune, UK). DNA encoding the signal sequence that is cleaved after protein translocation into the periplasm is not shown. Rare codons were identified using *E. coli* Codon Usage Analyzer 2.1[378] and are highlighted in cyan. Codons encoding different amino acids in G-CSF wt and C3 are highlighted in red. Rare codons encoding different amino acids in G-CSF wt and C3 are highlighted in green. All other nucleotide variations between G-CSF wt and C3 DNA sequences are silent and not highlighted. Vertical lines indicate sites where purine bases have been swapped for pyrimidine bases or vice versa. The nucleotide at the beginning of each line is numbered on the left.

The translation efficiency of mRNA transcripts can also be affected by the physicochemical properties of amino acids on the peptidyl-tRNA and aminoacyl-tRNA[135]. During catalysis of peptide bond formation the aminoacyl-tRNA and peptidyl-tRNA are incorporated in the A- and P- sites of the ribosome, respectively. Incorporation of the aminoacyl-tRNA in the A-site is achieved through base-pairing of the mRNA codon encoding a particular amino acid in the A-site with the antisense codon of the tRNA that is covalently attached to that amino acid. After peptide bond formation the deacylated tRNA is transferred from the P-site to the E-site for release from the ribosome[117,135].

The rate of the peptidyl transfer step of translation is most often limited by incorporation of the aminoacyl-tRNA in the A-site of the ribosome and is not usually affected by the rate of the peptide bond forming reaction itself[379]. However, Pro is both a poor acceptor group to peptide bond formation when incorporated in the A-site[380,381] and a poor donor group when incorporated in the P-site[382,383], meaning this residue greatly reduces the rate of peptide bond formation. The reason for this is unclear, but is thought to be due to the steric properties of Pro and its orientation in the peptidyl transfer centre of the ribosome[384]. Translation is also often slowed at clusters of Arg and Lys residues, most likely owing to the interaction of positively charged segments of nascent peptide with the negatively charged walls of the ribosome exit tunnel[385]. Gly residues can also slow the rate of peptide bond formation, although the reason for this observed effect is unclear. Rather than being due to the physiochemical properties of Gly, the effect of this residue has been hypothesised to be the result of base-pairing of codons encoding Gly with the anti-Shine-Dalgrano sequence found in the 16S rRNA of the ribosome[386], which normally binds the Shine-Dalgrano sequence of mRNA transcripts for initiation of translation[117].

Considering that G-CSF C3 has two more Arg residues and one more Gly residue than wild-type protein, the effect of these substitutions may be to slow the translation rate. Additionally, the W59R substitution found in the variant C3 results in the amino acid sequence GIPRAP, which contains many residues that could stall translation. Consequently, this section of the amino acid sequence may stall translation of G-CSF C3 to a greater extent than wild-type protein, although intuitively reducing the rate of translation should not increase the recombinant expression level of G-CSF C3.

The effects of local translation kinetics on protein folding and aggregation are complex and wide ranging and may have implications on recombinant protein expression[135]. Translation kinetics have been shown to affect the folding of newly synthesised proteins through interaction with the ribosome surface[47] and exit tunnel[387], as well as modulating the proportions of co- and post-translational folding[388]. Co-translational folding can result in stabilisation of intermediates through interactions with the ribosome that would otherwise be less populated during post-translational folding[47,387], which can even affect the energy landscape of proteins after dissociation from the ribosome[389]. Stabilisation of such intermediates by the ribosome can also act to protect polypeptides from misfolding and aggregation[47,48].

Presumably, the effect of translation kinetics on the folding of G-CSF wt and C3 cannot account for the observed increase in heterologous expression of G-CSF C3 compared to the wild-type protein, as this effect is observed whether these proteins are expressed as insoluble aggregates in the cytoplasm (Figure 3.2) or soluble native protein in the periplasm of *E. coli*[150]. Therefore, no obvious link can be drawn between the translation rates of G-CSF wt and C3 and their vastly different recombinant expression levels in *E. coli*.

A detailed investigation of the many factors surrounding the recombinant expression of G-CSF wt and C3 in *E. coli* is beyond the scope of this thesis. Next, the focus of the work moved to analysis of the aggregation behaviour of G-CSF C3 and its comparison with the wild-type protein in Chapter 4. Furthermore, investigation of the structure and dynamics of G-CSF C3 at amino acid resolution by NMR in Chapter 5 helps to elucidate subtle changes to the native conformation that are likely to be linked to aggregation. Due to the poor yield from expression and purification of G-CSF wt (maximum yields of 0.1 and 0.2 mg/l were obtained from recombinant expression in LB and M9 media, respectively), experiments in these chapters were carried out using G-CSF C3 only and compared to published studies of G-CSF wt aggregation when possible. For these experiments, G-CSF C3 obtained from the peak and tail of cation exchange chromatography were used as no significant differences in the purities or structures of these samples could be detected from analysis by SDS-PAGE, ESI-MS or ^{1}H-^{15}N HSQC spectra. Overall, the work in this chapter has shown that G-CSF wt prepared in this study is structurally similar to wild-type G-CSF prepared in previous studies and there are no observable differences between the native conformations of G-CSF wt and C3.

Chapter 4 Characterising the aggregation properties of G-CSF C3

4.1 Introduction

The reason for the thousand-fold increase in the yield of soluble G-CSF C3 from the *E. coli* periplasm in relation to the wild-type protein is unclear, but it has been suggested to be the result of the reduced aggregation of G-CSF C3 relative to the wild-type protein. This hypothesis arose from computational modelling of G-CSF with and without the substitutions found in the variant C3, which suggests reduced aggregation propensity around substitutions W59R and Q71R[150]. This hypothesis was tested further here by investigating the relative aggregation properties of G-CSF wt and C3 using other protein aggregation prediction programs (Figure 4.1). Since G-CSF has been reported to form amorphous aggregates[308] and aggregates that bind the fluorescent dye Thioflavin-T[307], a long used indicator of amyloid formation[390], a range of prediction programs were utilised to cover both amorphous and amyloid aggregation.

Although G-CSF C3 is predicted to be less aggregation-prone than wild-type G-CSF, this had not previously been investigated experimentally. In this chapter, the *in vitro* aggregation of G-CSF C3 is characterised in terms of its dependence on protein concentration, pH and urea concentration. In its drug formulation as Filgrastim, G-CSF wt is stored at pH 4 with excipients because wild-type G-CSF aggregates less readily at acidic pH values compared to pH values above 5[185,276,303], whereas at physiological pH at 37 ^{o}C G-CSF wt has been shown to aggregate readily within hours[142,185,214,307,308,391]. Therefore, all aggregation assays are carried out at pH 4 and 7. Aggregation assays of G-CSF C3 at pH 7 are carried out at similar protein concentrations to assays carried out previously for wild-type G-CSF to allow comparison[142,185,214,308]. Aggregation assays of G-CSF C3 are also carried out using different urea concentrations to establish whether the rate of aggregation would accelerate in the presence of denaturant concentrations too low to cause global denaturation of the protein, as observed for aggregation of wild-type G-CSF in the presence of different GdnHCl concentrations[142].

Equilibrium denaturation analyses of G-CSF C3 are also carried out at pH 4 and 7 in urea to measure the thermodynamic stability of the native state under conditions used to study aggregation, as well as to probe for additional conformations that may be populated at urea concentrations at which aggregation is accelerated. From these analyses urea concentrations are identified at pH 4 and 7 that accelerate aggregation but do not cause global unfolding. In addition, curvature in the pre-transition baselines of some of the equilibrium curves measured by intrinsic fluorescence reveal a change in the fluorescence properties of G-CSF C3 in low urea, which are further investigated by near-UV CD in this chapter and NMR in Chapter 5. Additionally, the ANS binding properties of G-CSF C3 are investigated in different urea concentrations to probe for exposure of hydrophobic surfaces in aggregation-prone conditions.

In the following chapter all experiments carried out in pH 4 and 7 buffers are performed in 20 mM sodium phosphate, 20 mM sodium acetate pH 4 buffer or 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide pH 7 buffer unless otherwise stated. For NMR experiments 10 % (*v/v*) D_2O was added to G-CSF samples, resulting in 10 % dilution of pH 4 and 7 buffer components.

4.2 Results

4.2.1 *In silico* prediction of the aggregation of G-CSF wt and C3

The protein aggregation prediction program Aggrescan[62] predicts generic aggregation propensity from an amino acid sequence. The scores attributed by Aggrescan[62] represent the experimental aggregation propensity of each amino acid when used to replace Phe19 of the 42 residue amyloid-β peptide (Aβ42). Phe19 is located within the aggregation-prone central hydrophobic cluster. The aggregation propensity of each amino acid was determined from the fluorescence emission intensity of variants of Aβ42 fused to Green Fluorescent Protein (GFP) when expressed in the cytoplasm of *E. coli*, relative to the emission intensity of wild-type Aβ42 fused to GFP. Scores for G-CSF are averaged over seven residues (three either side of the residue being scored) to produce an ‘a4v’ score. Residues with positive scores are predicted to promote aggregation while residues with negative scores are predicted to promote solubility[62].

CamSol[63] can be used to predict the intrinsic solubility of an amino acid sequence, also providing a measure of generic aggregation propensity. Predictions are based on the charge, hydrophobicity and secondary structure propensity of amino acids. These scores are then averaged over a seven residue window and adjusted for patterns of alternating hydrophobic and hydrophilic residues, as well as the gatekeeping effect exerted by individual charges. Regions with scores below -1 are predicted to be poorly soluble while regions with scores above 1 are predicted to be significantly soluble[63].

CamSol[63] can also be used to predict the surface solubility of a protein structure, based on projection of intrinsic solubility scores onto a protein surface. The intrinsic solubility score of a residue is averaged over residues within a radius of 8 Å that are not contiguous in sequence. As well as considering the proximity of residues on the protein surface, this projection also takes into account solvent accessible surface area. This means that hydrophobic residues buried in the protein interior for example will produce significantly negative intrinsic solubility scores, but will likely appear neutral in solubility scores corrected for protein structure[63].

Zyggregator[64] is very similar to CamSol[63] in the sense that amyloid aggregation propensity is predicted from an amino acid sequence based on the charge, hydrophobicity and secondary structure propensity of amino acids and that scores are adjusted for patterns of alternating hydrophobic and hydrophilic residues and the gatekeeping effect of individual charges[64]. The key difference between the Zyggregator[64] and CamSol[63] algorithms is that different secondary structure propensities are employed for Gly and Pro residues, which discourage β-sheet formation and so help prevent amyloid formation, but have little affect on general solubility[63,392]. Therefore, the different secondary structure propensities used for Gly and Pro in the Zyggregator[64] algorithm help specify prediction of amyloid formation. Scores are inverted in the Zyggregator[64] program relative to CamSol[63] because the former predicts aggregation propensity while the latter predicts solubility, meaning Z_{agg} scores above 1 indicate regions that are predicted to promote amyloid aggregation[64].

TANGO[65] calculates the propensities for formation of α-helix, β-sheet, β-turn, α-helical aggregate and β-sheet aggregate from an amino acid sequence, based on a Boltzmann distribution of these states. The TANGO[65] method assumes that the main secondary structure of β-sheet aggregate is β-strand, that the regions involved

in aggregation are fully buried away from solvent and that residues with complementary charges in a particular window will promote aggregation, while net charge on the polypeptide or net charges within two residues of the prediction window disfavour aggregation[65]. Intermolecular β-sheet structure is a common feature of both amorphous and amyloid aggregates[68], meaning TANGO[65] predicts the generic β-sheet aggregation propensity of an amino acid sequence. Amino acid sequence lengths of five residues or more with a β-sheet aggregation propensity above 5 % are predicted to be involved in forming β-sheet aggregates[65].

Both Aggrescan[62] and CamSol[63] predict reduced aggregation propensity as a result of the W59R substitution found in G-CSF C3 (Figure 4.1A and B/C, respectively), although the predicted effect of this mutation is lessened in structurally corrected solubility scores (Figure 4.1C), suggesting Trp59 in G-CSF wt may be at least partially protected from solvent. The Q71R substitution found in the variant C3 is also predicted by CamSol[63] to reduce the aggregation propensity of G-CSF, based on intrinsic and structurally corrected solubility scores (Figure 4.1B and C, respectively). These predictions are in agreement with the computational modelling carried out previously for G-CSF with and without the substitutions found in the variant C3[150]. However, the changes in solubility predicted by CamSol[63] for these substitutions using the structure of G-CSF wt are within the confidence limits of -1 and 1 for prediction of residues that promote aggregation and solubility, respectively (Figure 4.1C), suggesting that these substitutions may not be significant in reducing aggregation of the native conformation of G-CSF.

Interestingly, the C18G mutation is predicted to reduce the aggregation propensity of G-CSF by Aggrescan[62] and CamSol[63] (Figure 4.1A and B/C, respectively), even though these programs do not take into consideration the propensity for intermolecular disulphide bond formation between unpaired Cys residues. Furthermore, the effect of this substitution spans the -1 confidence limit of CamSol[63] predictions both in intrinsic and structurally corrected solubility score profiles (Figure 4.1B and C, respectively). The C18G substitution is also predicted to reduce the propensity for amyloid aggregation of G-CSF, based on the Z_{agg} score predicted by Zyggregator[64] in this region dropping below the confidence limit of 1 (Figure 4.1D). This suggests that the reduced secondary structure propensity of Gly compared to Cys results in a reduced predicted propensity for both amorphous and amyloid aggregation.

The F84L substitution appears to have a more significant impact of reducing the predicted propensities for forming amyloid and β-sheet aggregates, predicted by Zyggregator[64] and TANGO[65] in Figure 4.1D and E, respectively, than reducing the general aggregation propensity of G-CSF, as predicted by Aggrescan[62] and CamSol[63] in Figure 4.1A and B/C, respectively. Although reductions in aggregation propensity are predicted by Aggrescan[62] and CamSol[63] as a result of the F84L substitution, Leu84 is still predicted to be a residue that promotes aggregation of G-CSF, based on the scores attributed to this residue and the thresholds for aggregation prediction of these programs. Conversely, the F84L substitution results in Z_{agg} scores for residues in this region dropping closer to the confidence limit of 1, suggesting that this reduction may be more significant in terms of the predicted propensity for amyloid aggregation (Figure 4.1D). Additionally, TANGO[65] predicts a large reduction in the β-sheet aggregation propensity as a result of the F84L substitution, although the propensity for forming β-sheet aggregate in the region around Leu84 is still above the 5 % threshold (Figure 4.1E).

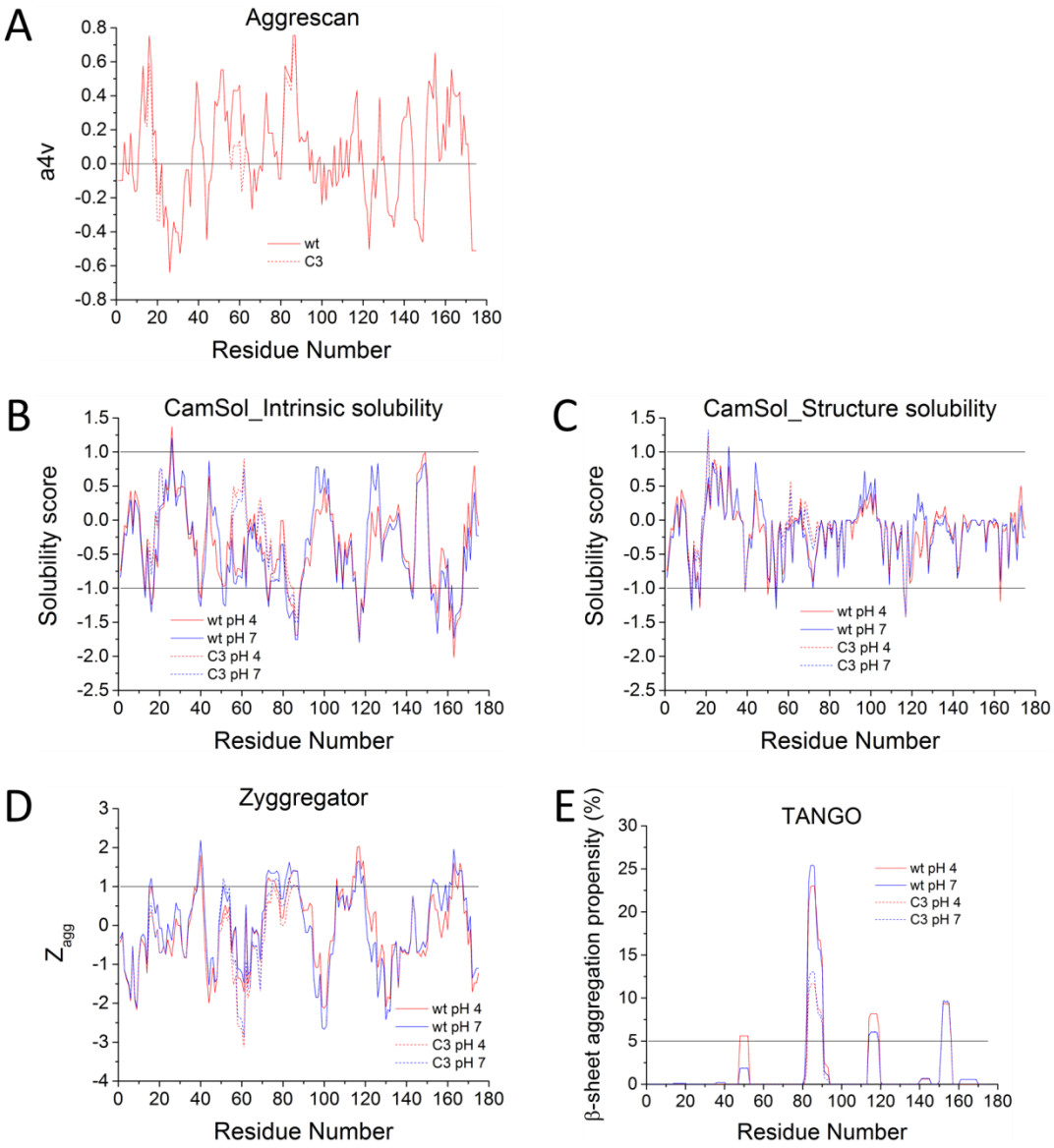


Figure 4.1 Aggregation prediction scores for G-CSF wt and C3. Scores for G-CSF wt are shown as solid lines, while scores for G-CSF C3 are shown as dashed lines. Where applicable, scores have been calculated for G-CSF wt and C3 at pH 4 (red) and 7 (blue). a4v scores calculated by Aggrescan[62] to be above and below 0 indicate residues predicted to promote aggregation and solubility, respectively *in vivo*[62] (A). CamSol[63] solubility scores above 1 indicate residues predicted to increase solubility, while scores below -1 indicate residues predicted to decrease solubility and so promote aggregation[63] (B and C). CamSol[63] solubility scores corrected for structure were calculated using the G-CSF wt structure (PDB 1GNC[283]), with residues being mutated in PyMOL[289] for calculation of G-CSF C3 solubility scores (C). Z_{agg} scores calculated by Zyggregator[64] to be above 1 indicate residues predicted to promote amyloid formation[64] (D). TANGO[65] predicts the percentage propensity for β-sheet aggregation, whereby sequences of five residues or more exhibiting percentages above 5 % are predicted to be involved in forming β-sheet aggregates[65] (E).

4.2.2 Characterising the dependence of G-CSF C3 aggregation on pH and protein concentration

The dependence of G-CSF C3 aggregation on pH and protein concentration was characterised using 100 µl samples of G-CSF C3 set up in triplicate in 1.5 ml Eppendorf tubes and incubated quiescently at 37 ^{o}C. G-CSF C3 samples were incubated in pH 4 buffer at protein concentrations of 4.5 and 8.6 mg/ml, while G-CSF C3 samples in pH 7 buffer were set up using protein concentrations of 1.7 and 5.5 mg/ml. At regular time intervals, samples were centrifuged at 13300 rpm for 15 mins at room temperature and the A_{280} value of the supernatant measured. A_{280} values were averaged across triplicate samples and standard deviations of mean A_{280} values used as errors. Mean A_{280} values measured after time intervals of incubation at 37 ^{o}C were normalised to the mean A_{280} value measured before incubation at 37 ^{o}C to produce proportions of aggregated G-CSF C3 (Figure 4.2).

Protein aggregation assays show lower proportions of aggregated G-CSF C3 in pH 4 buffer compared to pH 7 buffer, suggesting G-CSF C3 is less aggregation-prone at pH 4. This is emphasised by the aggregation profiles of 8.6 mg/ml G-CSF C3 in pH 4 buffer and 1.7 mg/ml G-CSF C3 in pH 7 buffer, which appear similar even though the G-CSF C3 sample at pH 4 has a five-fold higher protein concentration (Figure 4.2). Reduced aggregation of G-CSF C3 at acidic compared to neutral pH agrees with previously reported aggregation behaviour for the wild-type protein at acidic and neutral pH[185,276,303].

Reduced aggregation of G-CSF wt at acidic pH has been suggested to be the result of increased electrostatic repulsion between G-CSF molecules, caused by a greater overall positive charge of G-CSF at acidic compared to neutral pH[303]. Considering the pI values of G-CSF wt and C3, which are predicted from their amino acid sequences to be 6.02 and 6.67, respectively[117], it is likely that the reduced aggregation of G-CSF C3 at pH 4 compared to pH 7 is also due to an overall greater positive charge on G-CSF C3 molecules at the lower pH causing electrostatic repulsion.

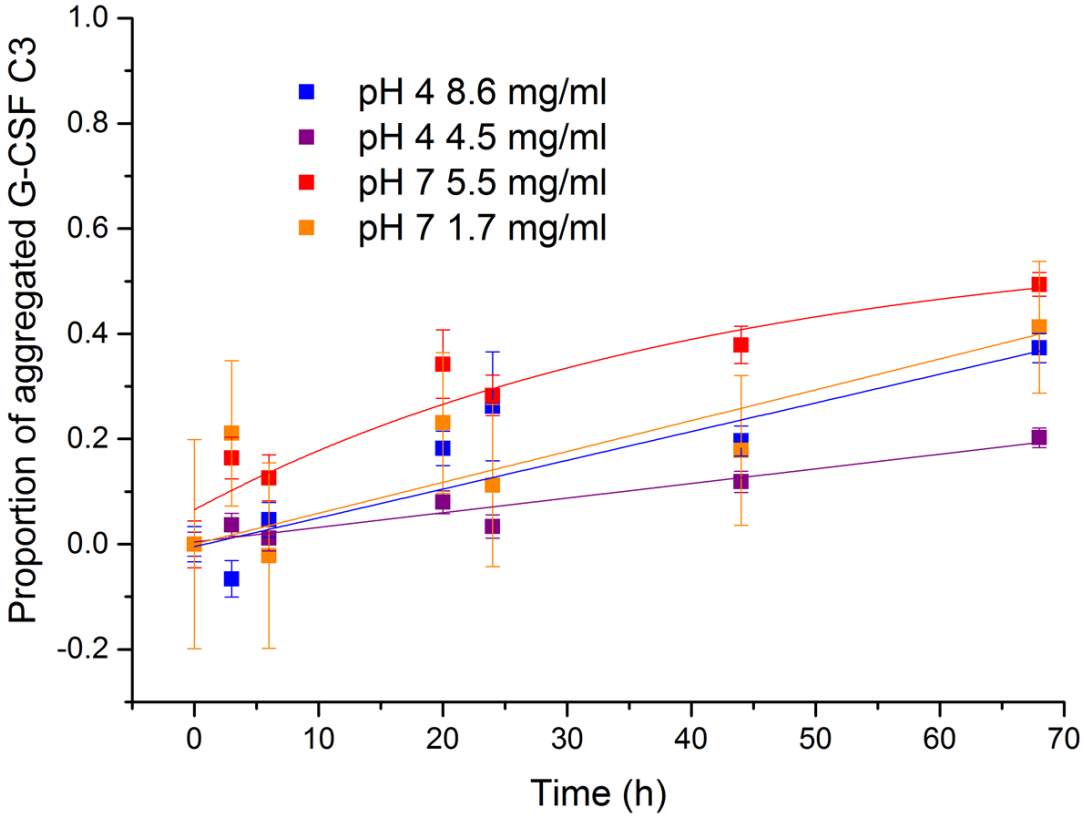


Figure 4.2 Aggregation assays for G-CSF C3 at different protein concentrations. Assays were set up using G-CSF C3 at protein concentrations of 1.7 and 5.5 mg/ml (orange and red, respectively) in 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide pH 7 buffer (I = 80 mM), while protein concentrations of 4.5 and 8.6 mg/ml (purple and blue, respectively) were used in 20 mM sodium phosphate, 20 mM sodium acetate pH 4 buffer (I = 43 mM). During centrifugation and measurement timing was suspended, meaning the x-axis in this figure represents the time samples were incubated at 37 ^{o}C. Error bars represent the standard deviation of the mean A_{280} value measured across triplicate samples. Proportions of aggregated G-CSF C3 from the 5.5 mg/ml sample were fitted to a single exponential decay function with a y-axis offset, while proportions of aggregated G-CSF C3 for all other samples were fitted to a linear function.

Surprisingly, the time courses of aggregation of G-CSF C3 samples in pH 7 buffer at 1.7 and 5.5 mg/ml shown in Figure 4.2 appear comparable to those observed for wild-type G-CSF in 100 mM MOPS pH 7 buffer at 1 and 5 mg/ml[142]. The ionic strengths of 100 mM MOPS at pH 7 and the pH 7 buffer used in this study are calculated to be 39 and 80 mM, respectively. These different buffer conditions used may affect G-CSF aggregation, for instance the aggregation of G-CSF at acidic pH has been reported to be dependent on ionic strength[303]. However, the percentage of aggregated protein for 1.7 mg/ml G-CSF C3 in pH 7 buffer after 68 h of incubation at 37 ^{o}C was 41 %, which is comparable to the approximate 35-40% aggregated protein observed after 3 days incubation at 37 ^{o}C for 1.5 mg/ml G-CSF wt in 10 mM sodium phosphate, 150 mM NaCl pH 7 buffer[185,308], even though this buffer has a higher ionic strength than the pH 7 buffer used in this study.

The percentage of aggregated protein for 5.5 mg/ml G-CSF C3 in pH 7 buffer after 44 h incubation at 37 ^{o}C was 38 %, which is significantly lower than the approximate 50 % aggregated protein observed after 2 days incubation at 37 ^{o}C for 5 mg/ml G-CSF wt in 100 mM MOPS at pH 7[142], as well as the 53 % aggregated protein reported after 2 days incubation at 37 ^{o}C for 5 mg/ml G-CSF wt in 100 mM sodium phosphate pH 6.9 buffer[214]. However, after 68 h incubation at 37 ^{o}C the percentage of aggregated protein for 5.5 mg/ml G-CSF C3 in pH 7 buffer was 49 %, suggesting that G-CSF C3 ultimately aggregates to the same extent as the wild-type protein at pH 7, albeit at a slower rate. This effect could be attributed to substitution of the unpaired Cys residue in G-CSF, which has been shown previously to slow the rate of aggregation at pH 7, although ultimately aggregation was reported to occur to the same extent as wild-type G-CSF[142].

Ultimately, G-CSF C3 appears to aggregate at a slower rate at higher protein concentrations than wild-type G-CSF but aggregates to the same extent (i.e. similar concentration of soluble protein at the end-point) as the wild-type protein and more generally, it appears that G-CSF aggregation at neutral pH does not depend on ionic strength.

4.2.3 Characterising the dependence of G-CSF C3 aggregation on urea concentration

Protein aggregation assays to characterise the denaturant dependence of G-CSF C3 aggregation were set up in the same way as assays used to characterise the dependence of G-CSF C3 aggregation on pH and protein concentration (see Section 4.2.2), although in this case mean A_{280} values were plotted against time rather than being converted into proportions of aggregated protein, to allow comparison with aggregation assays carried out by Raso *et al*.[142] in the presence of denaturant. 70 and 87 µM G-CSF C3 in pH 7 and 4 buffers, respectively were incubated quiescently at 37 ^{o}C in the presence of 0-7 M urea with 1 M increments, with triplicate samples being set up at each urea concentration (Figure 4.3).

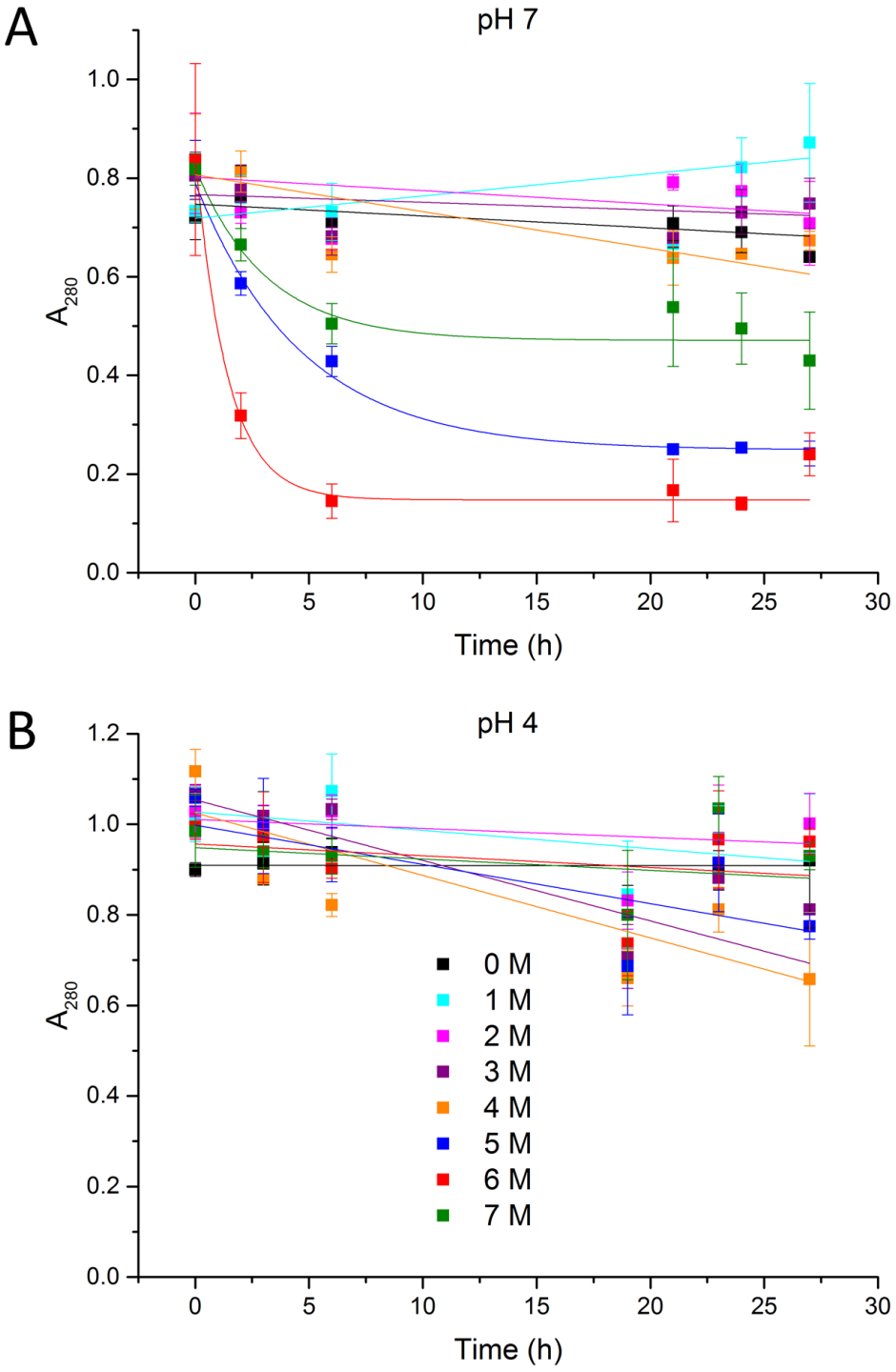


Figure 4.3 Aggregation assays for G-CSF C3 at different urea concentrations. Assays were set up using 70 and 87 µM G-CSF C3 in 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide pH 7 buffer (I = 80 mM) and 20 mM sodium phosphate, 20 mM sodium acetate pH 4 buffer (I = 43 mM), respectively (A and B, respectively) in the presence of 0 (black), 1 (cyan), 2 (pink), 3 (purple), 4 (orange), 5 (blue), 6 (red) and 7 M (green) urea. During centrifugation and measurement timing was suspended, meaning the x-axis in this figure represents the time samples were incubated at 37 °C. Error bars represent the standard deviation of the mean A_{280} value measured across triplicate samples. A_{280} values for G-CSF C3 pH 7 samples in the presence of 5, 6 and 7 M urea were fitted to a single exponential decay function with a y-axis offset as shown in A, while A_{280} values for all other samples were fitted to a linear function.

The greatest amounts of G-CSF C3 aggregation in pH 7 buffer are observed in the presence of 5, 6 and 7 M urea. Interestingly, the amount of G-CSF C3 aggregation is highest in 6 M urea (Figure 4.3A). This is analogous to the effects of 2, 2.5 and 3.5 M GdnHCl on the amount of G-CSF wt aggregation at pH 7 reported previously, in which the highest amount of aggregation is observed in 2.5 M GdnHCl[142]. This suggests that the effect of denaturant on the aggregation of G-CSF C3 is similar to the effect on the wild-type protein, possibly indicating a common aggregation mechanism.

The relative rates of G-CSF C3 aggregation in these different urea concentrations are difficult to interpret because the curves plateau at different A_{280} values (Figure 4.3A), which is an effect not observed for G-CSF wt aggregation in the presence of GdnHCl over a similar timescale[142]. The reason for the seemingly varied endpoints of G-CSF C3 aggregation in different urea concentrations is unclear, but may be due to the overall charge of GdnHCl that is absent for urea affecting aggregation. Alternatively, removal of the unpaired Cys in G-CSF C3 may affect the aggregates formed, as aggregates of wild-type G-CSF have previously been shown to be disulphide cross-linked, whereas aggregates of a C18A G-CSF variant are not disulphide cross-linked[142].

Interestingly, addition of 3, 4 and 5 M urea also appears to increase the amount of G-CSF C3 aggregation observed in pH 4 buffer (Figure 4.3B), although to a lesser extent than observed in pH 7 buffer (Figure 4.3A). Quantification of the aggregation of G-CSF wt in the presence of denaturant at pH 4 has not previously been reported, meaning it is unknown whether the effect of denaturant on G-CSF C3 aggregation at pH 4 is also exhibited by the wild-type protein. The greatest amount of G-CSF C3 aggregation in pH 4 buffer appears to be in the presence of 4 M urea, while at higher urea concentrations less aggregation is observed (Figure 4.3B). This is similar to the effect of 6 M urea on G-CSF C3 aggregation in pH 7 buffer relative to 7 M urea (Figure 4.3A), suggesting that there may be a common structural change occurring at pH 4 and 7 in certain urea concentrations that promotes aggregation, while at higher urea concentrations global unfolding occurs and the unfolded state is solubilised. Interestingly, a subtle conformational change has been suggested to cause aggregation of G-CSF wt in 1.5-2.5 M GdnHCl, revealed by a change in the fluorescence properties of Trp59 in the AB loop[142].

4.2.4 Measuring the thermodynamic stability of G-CSF C3

Equilibrium denaturation analyses of G-CSF C3 were carried out to measure the thermodynamic stability of the protein at pH 4 and 7. Additionally, these analyses provided a measure of the thermodynamic stability in the presence of urea concentrations that have been shown to promote protein aggregation (Figure 4.3). Equilibrium denaturation of G-CSF C3 was carried out at 25 ^{o}C to allow better comparison of the free energies of unfolding (ΔG^{o}_{UN}) obtained for G-CSF C3 to those reported for wild-type G-CSF at 20 ^{o}C at pH 4[304] and at 25 ^{o}C at pH 7[142]. In addition, aggregation of G-CSF wt has been shown to occur less readily at 25 ^{o}C compared to at 37 ^{o}C[185,307], meaning ΔG^{o}_{UN} values obtained for G-CSF C3 at 25 ^{o}C were less likely to be affected by aggregation.

Equilibrium denaturation of G-CSF C3 involved setting up samples of 7 µM G-CSF C3 in pH 4 and 7 buffers in the presence of 0, 2, 4, 6, 8 and 9 M urea, which were subsequently mixed in the correct ratios to produce 1 ml samples of 7 µM G-CSF C3 in 0-9 M urea with 0.2 M increments in 1.5 ml Eppendorf tubes (see Section 2.6.4). These 1 ml samples were incubated at 25 ^{o}C overnight and used the next day to measure fluorescence emission intensity at 25 ^{o}C. The fluorescence emission intensity of the pH 7 samples was measured at 320 nm, while the emission intensity of pH 4 samples was measured at 301 and 348 nm. Intensities at different wavelengths were monitored during equilibrium denaturation of G-CSF C3 at pH 4 and 7 due to the different fluorescence emission spectra observed for folded G-CSF C3 at pH 4 and 7 (see Section 3.2.6). After measurement samples were returned to incubation at 25 ^{o}C and left overnight again. The next day samples were used to measure ellipticity at 222 nm at 25 ^{o}C by far-UV CD.

The resulting observed signal intensities were fitted to Equation 2.9 and the parameters from these fits were used in Equation 2.10 to normalise the observed signal intensities, by converting them into the fraction of folded G-CSF C3 (see Section 2.6.4.3), according to the linear-extrapolation method of analysing equilibrium denaturation data[312].

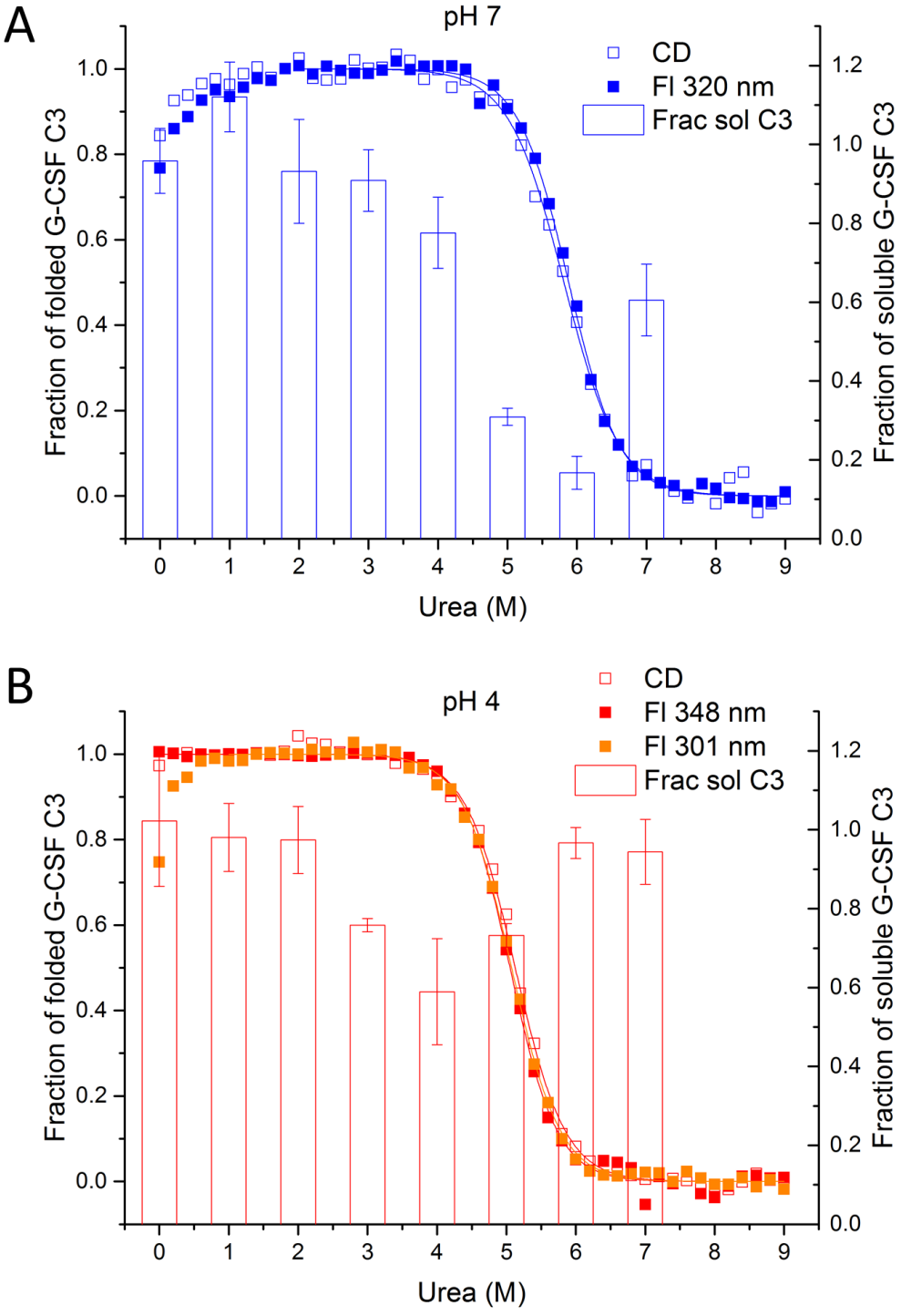


Figure 4.4 Equilibrium denaturation analyses of G-CSF C3 overlaid with fractions of soluble G-CSF C3 calculated from aggregation assays. Equilibrium denaturation of G-CSF C3 in 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide pH 7 buffer (I = 80 mM; blue data points in A) and 20 mM sodium phosphate, 20 mM sodium acetate pH 4 buffer (I = 43 mM; red and orange data points in B) is shown on the left hand y-axis and was monitored by measuring ellipticity at 222 nm (open data points) and intrinsic fluorescence emission intensity at 320 nm (blue closed data points), 301 nm (orange closed data points) and 348 nm (red closed data points) at 25 ^{o}C. Fractions of soluble G-CSF C3 remaining after incubation of 70 µM G-CSF C3 in pH 7 buffer for 24 h at 37 ^{o}C in the presence of 0-7 M urea with 1 M increments (blue bars in A), as well as after incubation of 87 µM G-CSF C3 in pH 4 buffer for 27 h at 37 ^{o}C at the same urea concentrations (red bars in B) are shown on the right hand y-axis.

Equilibrium denaturation curves for G-CSF C3 in pH 4 and 7 buffers showed concurrent loss of secondary and tertiary structure by urea that fitted well to a reversible co-operative transition, although only in urea concentrations of 2-9 M in pH 7 buffer (Figure 4.4A) and 1-9 M in pH 4 buffer (Figure 4.4B). Significant increases in ellipticity and fluorescence emission intensity at 320 nm were observed for G-CSF C3 in pH 7 buffer upon increasing the urea concentration from 0-2 M (Figure 4.4A), as well as in fluorescence emission intensity at 301 nm observed for G-CSF C3 in pH 4 buffer (Figure 4.4B). Consequently, these data obtained for 0-2 M urea at pH 7 and 0-1 M urea at pH 4 were not included in fitting of the observed signal intensities to Equation 2.9 in order to calculate ΔG^{o}_{UN} values[312].

The increases in ellipticity and fluorescence emission intensities observed for G-CSF C3 suggest a conformational change of the protein at urea concentrations that are too low to cause global unfolding, which could be linked to aggregation. Therefore, the secondary and tertiary structures of G-CSF C3 in pH 7 buffer in the presence of urea concentrations up to 4 M, as well as in pH 4 buffer in the presence of urea concentrations up to 3 M were further investigated by far- and near-UV CD (see Sections 4.2.7 and 4.2.9, respectively), intrinsic fluorescence emission (see Section 4.2.8) and later by NMR (see Chapter 5).

ΔG^{o}_{UN} and M_{UN} values, representing the free energy of unfolding and the dependence of this free energy on denaturant concentration, respectively[312], obtained for G-CSF C3 in pH 4 and 7 buffers are shown in Table 4.1. The ΔG^{o}_{UN} value calculated for G-CSF C3 in pH 7 buffer from fluorescence measurements in 2-9 M urea of 40.0 kJ mol^{-1} is similar to the ΔG^{o}_{UN} value reported for wild-type G-CSF in 100 mM MOPS pH 7 buffer with 0-5 M GdnHCl at 25 ^{o}C (40.2 kJ mol^{-1})[142]. Although the ΔG^{o}_{UN} value calculated for G-CSF C3 in pH 7 buffer from far-UV CD measurements is only 36.1 kJ mol^{-1}, the relatively large error on this value makes it unclear whether this is significantly different to the value reported for the wild-type protein.

The ΔG^{o}_{UN} value calculated for G-CSF C3 in pH 4 buffer from fluorescence measurements at 348 nm in 0-9 M urea of 37.0 kJ mol^{-1} is also similar to the ΔG_{UN} value reported for wild-type G-CSF in 20 mM sodium acetate, 20 mM MES, 50 mM Tris pH 4 buffer with 0-5 M GdnHCl at 20 ^{o}C (37.7 kJ mol^{-1})[304]. ΔG^{o}_{UN} values calculated for G-CSF C3 in pH 4 buffer from fluorescence measurements at 301 nm in 1-9 M urea and far-UV CD measurements in 0-9 M urea are 35.7 and 35.4 kJ mol^{-}

[1], respectively, which are lower than the value reported for wild-type G-CSF but are within the relatively small error margins of these calculations, suggesting that they are not significantly different from the value reported for the wild-type protein. Overall, these results suggest that the thermodynamic stability of G-CSF C3 is similar to wild-type G-CSF at pH 4 and 7.

	pH 7		**pH 4**		
	Ellipticity 222 nm	**Fluorescence 320 nm**	**Ellipticity 222 nm**	**Fluorescence 301 nm**	**Fluorescence 348 nm**
ΔG^{o}_{UN} / kJ mol^{-1}	36.12 ± 5.7	40.00 ± 4.9	35.36 ± 2.8	35.70 ± 2.5	36.97 ± 3.1
M_{UN} / kJ mol^{-1} M^{-1}	6.21 ± 1.0	6.80 ± 0.8	6.90 ± 0.6	7.04 ± 0.5	7.32 ± 0.6

Table 4.1 Free energies of unfolding (ΔG^{o}_{UN}) and the dependencies of these free energies on denaturant concentration (M_{UN}) calculated for G-CSF C3. Values were calculated from equilibrium denaturation analyses of 7 µM G-CSF C3 carried out in 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide pH 7 buffer (I = 80 mM) and 20 mM sodium phosphate, 20 mM sodium acetate pH 4 buffer (I = 43 mM), monitoring ellipticity and fluorescence at the wavelengths shown. ΔG^{o}_{UN} and M_{UN} values and errors were obtained from fitting of observed signal intensities to Equation 2.9 (see Section 2.6.4.3) shown in Figure 4.4.

Equilibrium denaturation analyses of G-CSF C3 in pH 7 and 4 buffers are shown in Figure 4.4A and B, respectively overlaid with fractions of soluble G-CSF C3, which were calculated using the mean A_{280} values shown in Figure 4.3. Mean A_{280} values measured after incubation of 70 µM G-CSF C3 in pH 7 buffer in the presence of 0-7 M urea for 24 h at 37 ^{o}C (Figure 4.3A), as well as incubation of 87 µM G-CSF C3 in pH 4 buffer in the presence of 0-7 M urea for 27 h at 37 ^{o}C (Figure 4.3B) were divided by mean A_{280} values measured for the same samples before incubation at 37 ^{o}C. Overlaying the fractions of soluble and folded G-CSF C3 reveals urea concentrations at pH 4 and 7 at which aggregation occurs but global unfolding does not, these being 3 and 4 M urea at pH 4 and 7, respectively.

As well as enabling investigation of the thermodynamic stability of G-CSF C3, equilibrium denaturation analysis has allowed a more precise definition of the

conditions in which aggregation is favoured and disfavoured. Therefore, many of the experiments performed in the remainder of this study are carried out in 0-3 M urea in pH 4 buffer and 0-4 M urea in pH 7 buffer.

4.2.5 Characterising insoluble G-CSF C3 aggregates using Electron Microscopy (EM)

EM images of G-CSF C3 aggregates were collected to determine whether they were amorphous or amyloid in nature (Figure 4.5). For aggregation assays in pH 7 buffer G-CSF C3 was used at 67 and 66 µM in the presence and absence of 4 M urea, respectively, while for assays in pH 4 buffer G-CSF C3 was used at 83 and 85 µM in the presence and absence of 3 M urea, respectively. 300 µl aggregation assays were set up in a 1.5 ml Eppendorf tube and then transferred to a Shigemi tube, which was subsequently incubated quiescently at 37 ^{o}C for 24 h in an NMR spectrometer to monitor HSQC peak intensities (see Section 2.9.4). After 24 h, each sample was transferred to 1.5 ml Eppendorf tubes and snap frozen, followed by storage at -80 ^{o}C. Samples were later thawed and 10 µl transferred to formvar/carbon-coated copper grids for imaging by Dr. Katie Stewart (see Section 2.8).

Surprisingly, G-CSF C3 aggregates formed at pH 4 and 7 appear to have different morphologies (Figure 4.5). Generally, aggregates formed at pH 7 appear amorphous in nature (Figure 4.5A-D), whereas aggregates formed at pH 4 more closely resemble protofibrils (Figure 4.5E-H), species that are considered to be late-stage intermediates in amyloid fibril formation[79]. In particular, G-CSF C3 aggregates formed at pH 4 in the presence of 3 M urea appear to be largely protofibrillar in nature (Figure 4.5G and H). EM images also suggest that there are mixtures of amorphous and protofibrillar species in the same aggregation assay samples at pH 7 (Figure 4.5D) and 4 (Figure 4.5E and F), which may account for the observation of amorphous aggregates by EM[308] and aggregates that bind Thioflavin-T formed by wild-type G-CSF at pH 7[79,307].

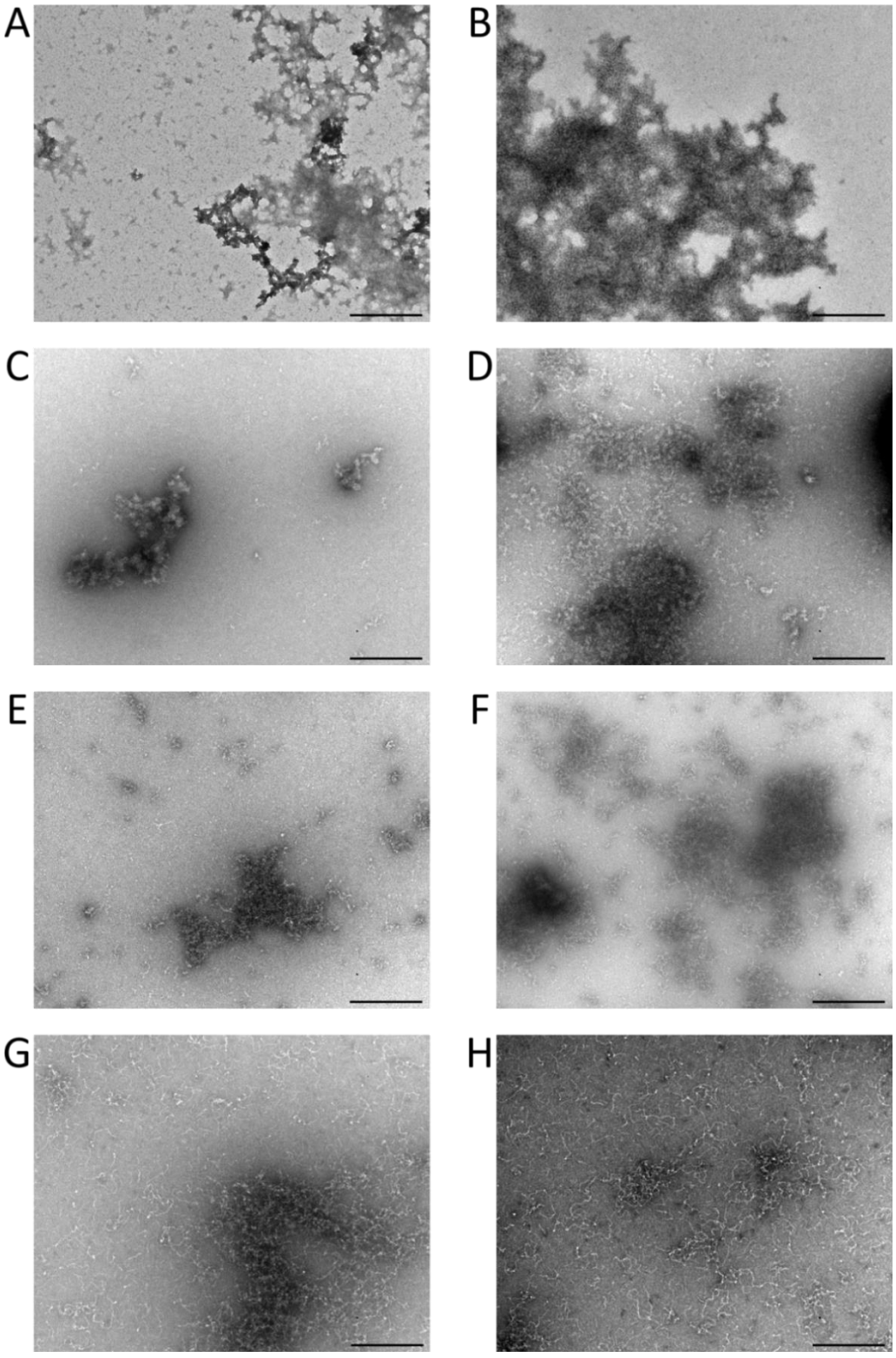


Figure 4.5 EM images of G-CSF C3 aggregates. Two images are shown for each aggregation assay sample. Images A/B and C/D show aggregates formed in 18 mM sodium phosphate, 18 mM sodium acetate, 0.018 % (*w/v*) sodium azide, 10 % (*v/v*) D_2O pH 7 buffer (I = 72 mM) in the absence or presence of 4 M urea, respectively and images E/F and G/H show aggregates formed in 18 mM sodium phosphate, 18 mM sodium acetate, 10 % (*v/v*) D_2O pH 4 buffer (I = 39 mM) in the absence or presence of 3 M urea, respectively. EM grids for imaging were prepared and imaged by Dr. Stewart. Scale bars represent 500 nm.

4.2.6 Characterising soluble G-CSF C3 aggregate formation using Dynamic Light Scattering (DLS)

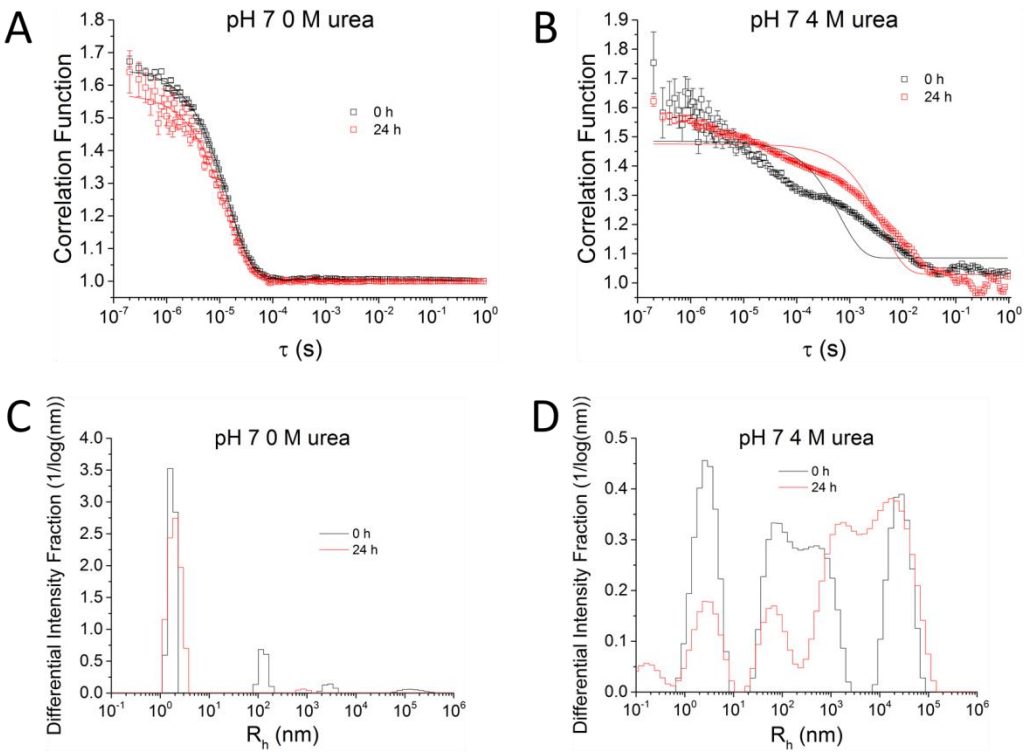
DLS involves measurements of light scattering intensity by particles over a range of time periods. The intensity of scattered light is constantly fluctuating due to Brownian motion of the particles in solution, provided a monochromatic light source is used and the wavelength of light is larger than the particle size. Light scattering intensities over time are then fitted to a second order auto-correlation function shown in Equation 2.11 (see Section 2.7.3), in order to calculate a hydrodynamic radius (R_h) for the particles. This correlation function relates the scattering intensity at the beginning of measurement to the scattering intensity after a time delay (τ). As τ becomes greater there is less correlation between the scattering intensities observed at the start of measurement and the scattering intensities after the time delay τ, meaning the correlation function decreases. Smaller particles move faster and so cause the correlation function to decay at a higher rate, whereas larger particles move slower and so cause a slower rate of decay[316].

For a population of species with a monodisperse size the decay rate of the correlation function can be calculated using a single exponential decay function i.e. Equation 2.12 (see Section 2.7.3), whereas for populations of species where a range of particle sizes is present each species has its own exponential decay rate, meaning attempting to calculate the decay rate of the correlation function using a single exponential decay function becomes less accurate[316]. For polydisperse samples, decay rates for each species of a different size are calculated using a summation of the single exponential decay functions of each of the species based on the weighting of these differently sized species i.e. Equation 2.17[317]. The decay rates of autocorrelation functions are then used in Equations 2.13-2.15 to calculate R_h values[316].

To provide an indication of the polydispersity of the sample a regularisation analysis can be carried out, based on the distribution function calculated from Equation 2.17. The resulting plot shows the fractional scattering intensity by each species of a particular R_h value[318]. All of the above calculations are performed by ASTRA 6 software. The R_h values for the different species identified can then be used to estimate their molecular mass from Equation 2.18 (see Section 2.7.3), although this assumes the particles are spherical[319].

For DLS measurements, 600 µl samples of G-CSF C3 were set up in 1.5 ml Eppendorf tubes, using 66 and 67 µM G-CSF C3 in pH 7 buffer in the presence and absence of 4 M urea, respectively and 83 and 85 µM G-CSF C3 in pH 4 buffer in the presence and absence of 3 M urea, respectively. After centrifugation at 13300 rpm for 15 mins at room temperature 290 µl of supernatant was used for DLS measurement (see Section 2.7.3). The remainder of the supernatant was incubated quiescently at 37 ^{o}C for 24 h, after which time the sample was centrifuged and DLS characterisation repeated.

The correlation functions and regularisation analyses are shown for G-CSF C3 aggregation assays at pH 7 and 4 monitored by DLS (Figure 4.6 and Figure 4.7, respectively). The correlation function plots show the results of attempting to fit the autocorrelation values to a single exponential decay function (Figure 4.6A/B and Figure 4.7A/B), while regularisation analyses show the results of fitting the autocorrelation values to a summation of the exponential decay functions of each species detected with a different R_h. The output of the regularisation analysis is the fractional intensity of each species with a distinguishable R_h (Figure 4.6C/D and Figure 4.7C/D). This cannot be used to quantify the amount of each species present due to the intensity of scattering being proportional to the sixth power of the radius of the particle. Moreover, in order for two species of different sizes to be separated their R_h values must vary by at least a factor of five[316].



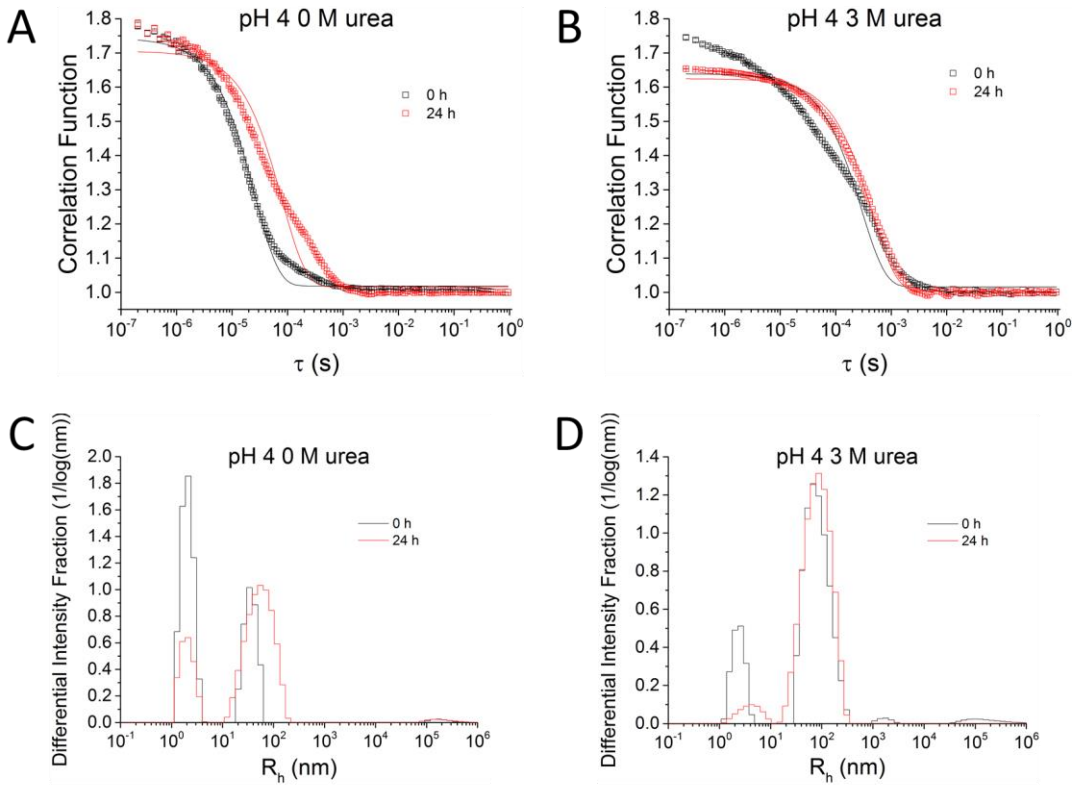
E

Peak Number	R_h of peak / nm	Percentage area under peak
Peak 1	1.61 ± 0.30	75.96 %
Peak 2	115.38 ± 26.67	17 %
Peak 1	2.06 ± 0.62	98.30 %

F

Peak Number	R_h of peak / nm	Percentage area under peak
Peak 1	2.60 ± 1.50	28.21 %
Peak 2	68.28 ± 427.46	48 %
Peak 3	28402.71 ± 14525.53	24 %
Peak 1	2.62 ± 2.08	16.20 %
Peak 2	17263.75 ± 20266.69	84 %

Figure 4.6 Aggregation assays of G-CSF C3 at pH 7 with or without 4 M urea monitored by DLS. Assays were set up in 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide pH 7 buffer (I = 80 mM) using 67 and 66 µM G-CSF C3 in the absence (A, C and E) or presence (B, D and F) of 4 M urea, respectively and then incubated quiescently at 37 ^{o}C for 24 h. DLS analysis of samples before and after incubation at 37 ^{o}C for 24 h are shown in black and red. The change in the autocorrelation functions over time in the absence or presence of 4 M urea are shown fitted to single exponential decay functions in A and B, respectively, while regularisation analyses of the changes in the autocorrelation functions over time in the absence or presence of 4 M urea are shown in C and D, respectively. The R_h values of peaks and peak widths in the regularisation analysis in the absence or presence of 4 M urea are shown in E and F, respectively, along with the percentage of area under each peak. These details are not shown for peaks that account for less than 5 % of the area occupied.



E

Peak Number	R_h of peak / nm	Percentage area under peak
Peak 1	2.14 ± 0.60	62.55 %
Peak 2	33.88 ± 10.71	36 %
Peak 1	2.12 ± 0.67	24.48 %
Peak 2	55.60 ± 35.94	74 %

F

Peak Number	R_h of peak / nm	Percentage area under peak
Peak 1	2.61 ± 0.70	18.14 %
Peak 2	68.44 ± 54.16	79 %
Peak 1	4.31 ± 1.98	5.69 %
Peak 2	87.90 ± 56.37	94 %

Figure 4.7 Aggregation assays of G-CSF C3 at pH 4 with or without 3 M urea monitored by DLS. Assays were set up in 20 mM sodium phosphate, 20 mM sodium acetate pH 4 buffer (I = 43 mM) using 85 and 83 µM G-CSF C3 in the absence (A, C and E) or presence (B, D and F) of 3 M urea, respectively and then incubated quiescently at 37 ^{o}C for 24 h. DLS analysis of samples before and after incubation at 37 ^{o}C for 24 h are shown in black and red. The change in the autocorrelation functions over time in the absence or presence of 3 M urea are shown fitted to single exponential decay functions in A and B, respectively, while regularisation analyses of the changes in the autocorrelation functions over time in the absence or presence of 3 M urea are shown in C and D, respectively. The R_h values of peaks and peak widths in the regularisation analysis in the absence or presence of 4 M urea are shown in E and F, respectively, along with the percentage of area under each peak. These details are not shown for peaks that account for less than 5 % of the area occupied.

The variation in the correlation function over time can be fitted more accurately to a single exponential decay function when fewer species of different R_h values are present in solution, as shown from DLS measurements of G-CSF C3 at pH 7 in the absence or presence of 4 M urea (Figure 4.6A and B, respectively). In the absence of urea, the correlation function for G-CSF C3 at pH 7 decays more rapidly and

reaches a baseline after a shorter time period (Figure 4.6A) compared to G-CSF C3 in the presence of 4 M urea (Figure 4.6B), indicating that the species formed in the absence of urea are smaller and so cause less light scattering.

After incubation at 37 ^{o}C the correlation functions for G-CSF C3 at pH 7 without urea show slightly increased deviation from the fit at early time periods (Figure 4.6A), suggesting formation of species of a wider range of R_h values, which is also reflected in the increased width of the main peak in the regularisation analysis (Figure 4.6C and E). R_h values and ranges of the most intense peak equate to estimated molecular masses of 3.4-13.9 and 4.9-45.4 kDa before and after incubation at 37 ^{o}C, respectively (Figure 4.6E), suggesting that within this peak is monomeric G-CSF C3, as well as a greater proportion of oilgomeric forms of G-CSF C3 after incubation at 37 ^{o}C.

In the presence of 4 M urea at pH 7, regularisation analysis shows that species with larger R_h values are observed and a wider range of R_h values is occupied (Figure 4.6D and F), suggesting formation of larger soluble G-CSF C3 aggregates as a result of adding 4 M urea. Interestingly, even before incubation at 37 ^{o}C there appear to be large soluble G-CSF C3 aggregates present in 4 M urea, but after incubation at 37 ^{o}C the distribution is shifted towards species with higher R_h values (Figure 4.6D and F), implying formation of larger soluble aggregates. Due to the wide ranges of R_h values occupied by peaks in the presence of 4 M urea (Figure 4.6F), estimated molecular masses of species have not been calculated.

Surprisingly, incubation of G-CSF C3 at pH 4 for 24 h at 37 ^{o}C in the absence of urea appears to result in formation of aggregates with larger R_h values compared to incubation of G-CSF C3 at pH 7 in the absence of urea (Figure 4.7C and Figure 4.6C, respectively), implying formation of larger soluble G-CSF C3 aggregates at pH 4. This is in contrast to the increased formation of insoluble aggregates observed for G-CSF C3 at pH 7 compared to at pH 4, which is apparent even when higher protein concentrations are used at pH 4 (Figure 4.2).

The formation of soluble G-CSF C3 aggregates at pH 4 without urea after incubation at 37 ^{o}C reduces the accuracy of a fit of the change in correlation function over time to a single exponential decay and results in a slower rate of decay (Figure 4.7A). This is also reflected in the regularisation analysis, which shows an increase in the width and area of the second peak after incubation at 37 ^{o}C (Figure 4.7C). There is

also a corresponding decrease in the area of the first peak due to incubation at 37 ^{o}C, although the width of this peak remains largely unchanged (Figure 4.7E) and corresponds to estimated molecular mass ranges of 6.3-49.0 and 5.0-52.1 kDa.

In the presence of 3 M urea at pH 4, the change in correlation function over time appears to fit better to a single exponential decay function after incubation of G-CSF C3 for 24 h at 37 ^{o}C (Figure 4.7B), suggesting that the population of species in solution becomes less disperse in size. Additionally, the decay in the correlation function occurs after a longer time period, implying larger species are present.

Regularisation analysis shows that, after incubation of G-CSF C3 at pH 4 in the presence of 3 M urea for 24 h at 37 ^{o}C, there is a shift in the distribution from population of the lower to the higher R_h peak (Figure 4.7D and F), implying increased formation of soluble aggregates. The peak of higher R_h value is also present before incubation at 37 ^{o}C (Figure 4.7D and F), suggesting soluble aggregates of G-CSF C3 are present at pH 4 in the presence of 3 M urea and incubation at 37 ^{o}C results in continued formation of these aggregates. This is similar to the effect of 4 M urea on G-CSF C3 aggregation at pH 7 observed from DLS measurements (Figure 4.6D and F). The estimated molecular masses of species in the lower R_h peak before incubation at 37 ^{o}C are 13.9-92.4 kDa, suggesting monomeric G-CSF C3 is within this peak. However, the wide ranges of R_h values occupied by other peaks in the presence of 3 M urea (Figure 4.7F) means that estimated molecular masses of species have not been calculated.

4.2.7 Probing for changes in secondary structure of G-CSF C3 in aggregation-prone conditions using far-UV CD

Far-UV CD was used to probe for changes in the secondary structure of G-CSF C3 in the presence of urea concentrations that are too low to cause global denaturation at 25 ^{o}C. The spectra of G-CSF C3 at pH 7 in the presence of 0-4 M urea exhibit absorption minima at 221-222 nm (Figure 4.8A), indicative of predominantly α-helical structure[313]. In addition, the intensities of these spectra are similar, suggesting no significant change in the α-helical content of G-CSF C3 upon addition of urea.

Spectra of G-CSF C3 at pH 4 in the presence of 0-3 M urea are also similar in terms of their shapes and intensities, showing absorption minima at 221 nm (Figure 4.8B). The spectrum of G-CSF C3 at pH 4 in the absence of urea is more negative at this absorption minimum than in the presence of all concentrations of urea, but is also more negative across all wavelengths at which ellipticity was measured. This would suggest unusually high intensities exhibited by the buffer blank that were subtracted from the intensities measured in this spectrum, possibly due to light scattering. Overall, these spectra suggest that there is no significant change in the secondary structure of G-CSF C3 in urea concentrations that favour or disfavour aggregation.

Equilibrium denaturation analysis of G-CSF C3 carried out at pH 7 and monitored by far-UV CD spectroscopy suggested an increase in the ellipticity measured at 222 nm between urea concentrations of 0-2 M (Figure 4.4A). However, acquisition of far-UV CD spectra of G-CSF C3 over wavelengths of 210-260 nm suggests no significant change in ellipticity at 222 nm in the presence of 0-4 M urea (Figure 4.8A). The increase in signal intensity observed over 0-2 M urea in the equilibrium curve obtained using far-UV CD for G-CSF C3 at pH 7 is most likely being accentuated by linear extrapolation of the pre-transition baseline, which has been carried out during fitting of intensities measured over 2-9 M urea in order to calculate ΔG^{o}_{UN} and M_{UN} values[312] (see Section 2.6.4.3). Consequently, linear extrapolation of the pre-transition baseline may be giving a false impression of the extent of the change in ellipticity at 222 nm observed in 0-2 M urea in Figure 4.4A.

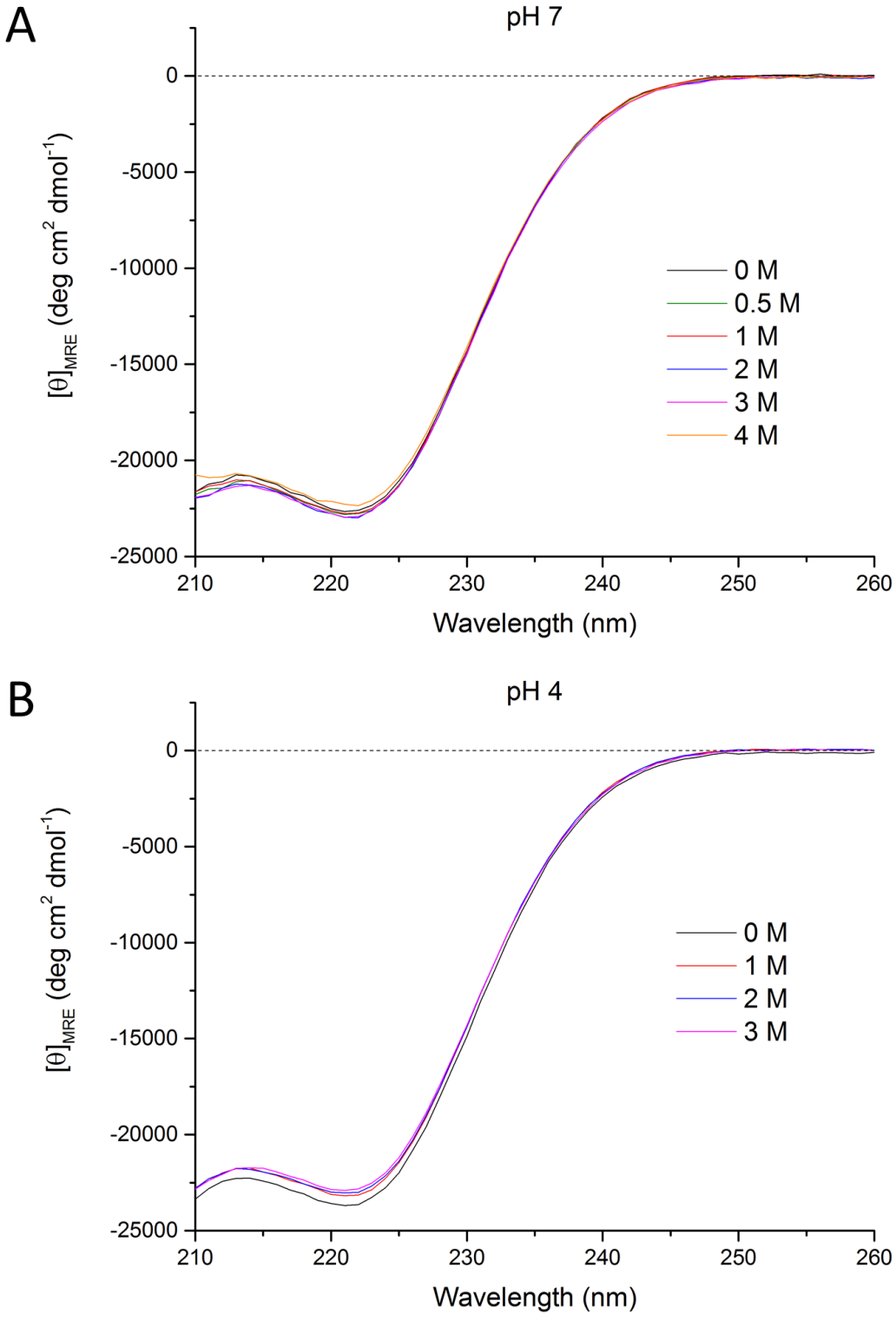


Figure 4.8 Far-UV CD spectra of G-CSF C3 at different urea concentrations. Spectra of G-CSF C3 in 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide pH 7 buffer (I = 80 mM) were acquired in the presence of 0 (black), 0.5 (green), 1 (red), 2 (blue), 3 (pink) and 4 M (orange) urea using 7-8 µM G-CSF C3 (A), while spectra of G-CSF C3 in 20 mM sodium phosphate, 20 mM sodium acetate pH 4 buffer (I = 43 mM) were acquired in the presence of the same urea concentrations, excluding 0.5 and 4 M urea, using 5-8 µM G-CSF C3 (B). A horizontal dotted line is drawn at zero ellipticity. $[\theta]_{MRE}$ values acquired at wavelengths below 210 nm are not shown due to excessive dynode voltage (800-1000 V) caused by the absorbance of signal by urea.

4.2.8 Probing for changes in the tertiary structure of G-CSF C3 in aggregation-prone conditions using fluorescence

Intrinsic fluorescence emission spectra of G-CSF C3 were used to probe for changes in the environment of Trp and Tyr residues upon addition of urea concentrations that are too low to cause global denaturation at 25 ^{o}C. The emission spectra of G-CSF C3 at pH 7 show increases in intensity across all wavelengths as the urea concentration is raised from 0 to 2 M and to a lesser extent from 2 to 4 M (Figure 4.9A). The wavelengths at which maximum emission is observed remain at 334-335 nm in the presence of 0-4 M urea. This emission wavelength is characteristic for a Trp side chain that is buried within a hydrophobic environment[357]. This suggests that the increased intensity in the fluorescence emission spectra of G-CSF C3 at pH 7 in 0-4 M urea could be due to a change in the environment of Trp side chains that is not accompanied by an increase in solvent exposure. Alternatively, the increased emission intensity observed could be caused by a change in the environment of one or more of the three Tyr side chains in G-CSF C3, as the wavelength at which maximum emission intensity is observed for Tyr side chains does not vary with solvent exposure[367].

Emission spectra of G-CSF C3 at pH 4 show a significant increase in intensity after adding 1 M urea, but at higher urea concentrations smaller increases are observed (Figure 4.9B). Again, the increase in emission intensity is not accompanied by a change in the emission maxima, which remain at 299-301 nm. An emission maximum at 303 nm is characteristic of Tyr fluorescence, although the wavelength of maximum emission of Tyr is not affected by the degree of solvent exposure[367]. The blue-shift in the fluorescence emission maximum of G-CSF C3 at pH 4 compared to pH 7 is due to quenching of Trp119 fluorescence, most likely caused by ionisation of the proximal side chain of His80 (Figure 3.12B). The increase in fluorescence emission intensity of G-CSF C3 at pH 4 around 300 nm suggests a change in the environment of one or more of the three Tyr side chains, although the reduced intensity of Trp at pH 4 due to quenching makes it unclear whether a change in the environment of Trp side chains is also occurring upon addition of urea.

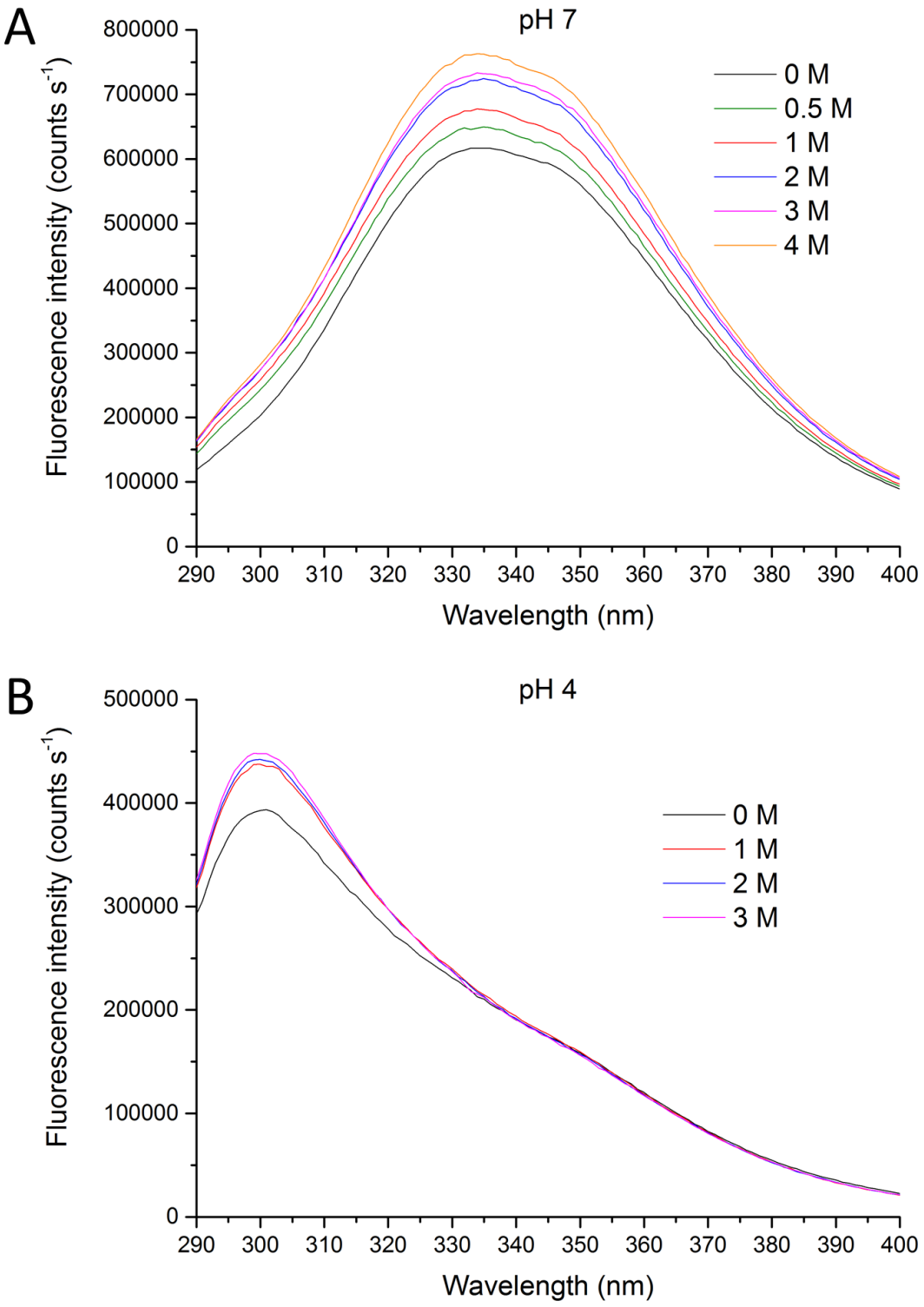


Figure 4.9 Fluorescence emission spectra of G-CSF C3 at different urea concentrations. Spectra of G-CSF C3 in 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide pH 7 buffer (I = 80 mM) were acquired in the presence of 0 (black), 0.5 (green), 1 (red), 2 (blue), 3 (pink) and 4 M (orange) urea using 7-8 µM G-CSF C3 (A), while spectra of G-CSF C3 in 20 mM sodium phosphate, 20 mM sodium acetate pH 4 buffer (I = 43 mM) were acquired in the presence of the same urea concentrations, excluding 0.5 and 4 M urea, using 5-8 µM G-CSF C3 (B). Fluorescence emission intensities have been corrected for the different protein concentrations used.

Equilibrium denaturation analysis of G-CSF C3 at pH 7 showed an increase in the fluorescence emission intensity measured at 320 nm between urea concentrations of 0-2 M (Figure 4.4A). Acquisition of fluorescence emission spectra of G-CSF C3 over wavelengths of 290-400 nm revealed a dramatic increase in emission intensity observed across all wavelengths as a result of increasing the urea concentration from 0 to 2 M, which is larger than the increase observed from 2 to 4 M (Figure 4.9A). In turn, this suggests that the inability to fit fluorescence emission intensities recorded in 0-2 M urea to a two-state transition between native and unfolded states is due to population of one or several additional conformations.

Equilibrium denaturation analysis of G-CSF C3 at pH 4 showed an increase in the fluorescence emission intensity measured at 301 nm between urea concentrations of 0-1 M, but not when measured at 348 nm (Figure 4.4B). Fluorescence emission spectra of G-CSF C3 acquired over wavelengths of 290-400 nm revealed a large increase in emission intensity around 300 nm upon addition of 1 M urea, but little change around 330-340 nm (Figure 4.9B). Again, this suggests that the inability to fit fluorescence emission intensities recorded at 301 nm in 0-1 M urea to a two-state transition is due to population of one or multiple additional conformations, which may possess an altered conformation around one or more of the three Tyr side chains of G-CSF C3. Due to Trp fluorescence quenching of G-CSF C3 at pH 4, it is unclear whether there may also be an altered conformation around Trp119 in this state.

4.2.9 Probing for changes in the tertiary structure of G-CSF C3 in aggregation-prone conditions using near-UV CD

CD can be used to characterise the tertiary structure of a protein as well as its secondary structure by measuring in the near-UV region of the spectrum. The chirality of aromatic side chains that arises from their arrangement in asymmetric environments within a folded protein results in the differential absorption of left and right circularly polarised components of plane-polarised light, which can be measured by CD. The intensity of this CD signal is reduced for more mobile aromatic side chains[313]. Typically, Phe side chains exhibit their absorbance maximum between 255-270 nm, Trp side chains exhibit a peak at 290 nm and often some absorbance at 290-305 nm and Tyr side chains are observed as an absorption minimum around 275-282 nm[393]. The CD signal observed in the near-UV region is far less intense than that observed in the far-UV region because of the

relative amounts of aromatic side chains and peptide bonds, meaning higher protein concentrations are needed for recording near-UV CD spectra[313].

To probe for tertiary structure changes in G-CSF C3 in urea concentrations that are too low to globally denature the protein at 25 ^{o}C, near-UV CD spectra were acquired in pH 7 buffer using 45-57 µM protein in the presence of 0-4 M urea, while in pH 4 buffer spectra were acquired using 36-52 µM protein in the presence of 0-3 M urea (Figure 4.10A and B, respectively).

The near-UV CD spectra of G-CSF C3 at pH 7 and 4 show an increased absorption signal in the region around 275 nm upon addition of urea (Figure 4.10A and B, respectively), suggesting a change in the environment of one or more of the three Tyr side chains. Although there are slight increases in negativity for absorption around 255 and 290 nm upon adding urea, indicative of Phe and Trp side-chains, respectively, these changes are significantly smaller in magnitude than the increased negativity observed around 275 nm, suggesting that they are largely due to the change observed around 275 nm.

A change in the environment of Tyr side chains most likely also accounts for the increased fluorescence emission intensity observed for G-CSF C3 at pH 7 and 4 upon adding urea (Figure 4.9A and B, respectively). Interestingly, the absorption minima at 275 nm observed in the near-UV CD spectra become more negative in the presence of urea (Figure 4.10), implying Tyr side chains are becoming less mobile as a result of the conformational change occurring. This would suggest that Tyr side chains are not becoming more solvent exposed as a result of adding urea.

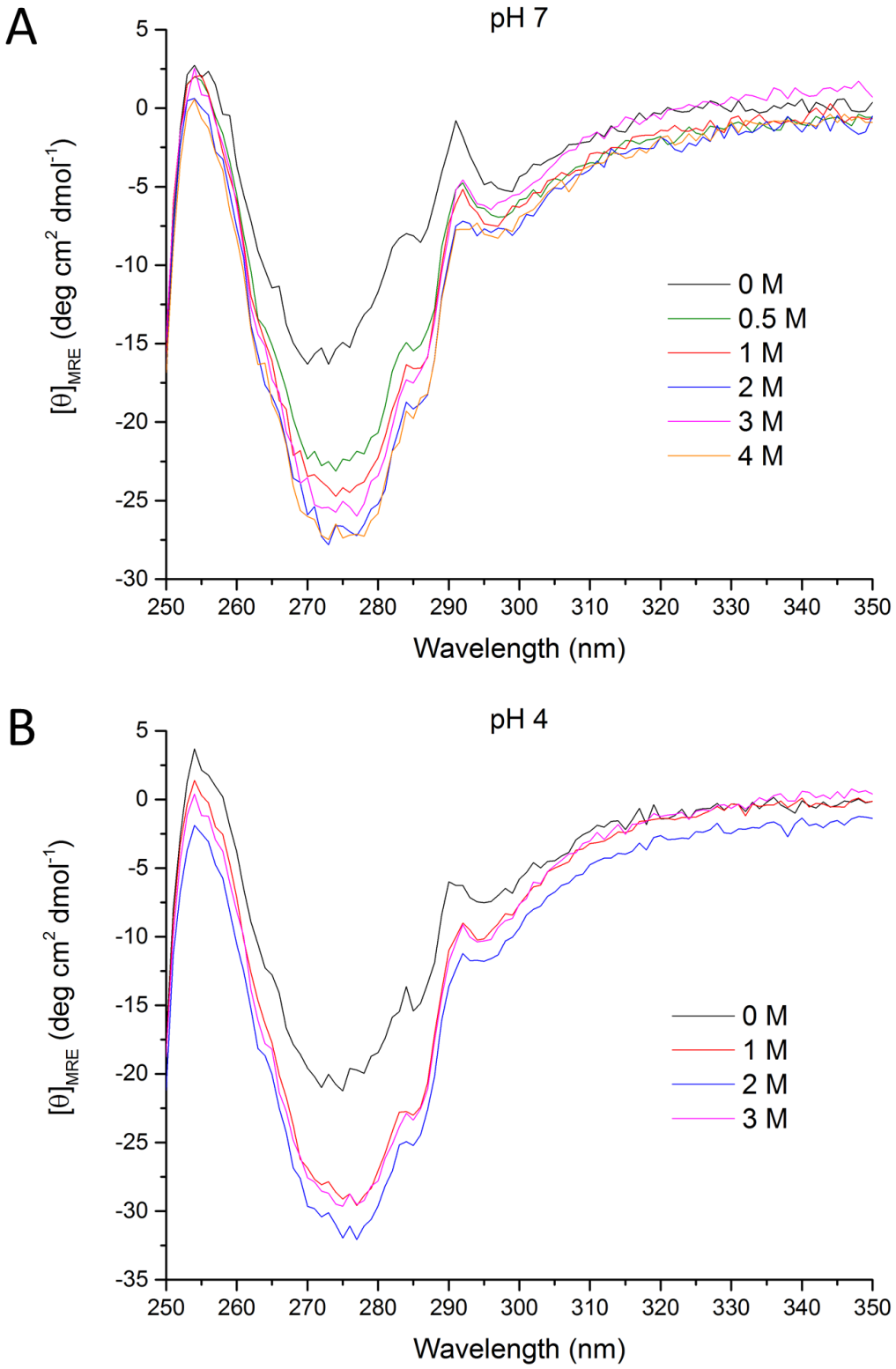


Figure 4.10 Near-UV CD spectra of G-CSF C3 at different urea concentrations. Spectra of G-CSF C3 in 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide pH 7 buffer (I = 80 mM) were acquired in the presence of 0 (black), 0.5 (green), 1 (red), 2 (blue), 3 (pink) and 4 M (orange) urea using 45-57 µM G-CSF C3 (A), while spectra of G-CSF C3 in 20 mM sodium phosphate, 20 mM sodium acetate pH 4 buffer (I = 43 mM) were acquired in the presence of the same urea concentrations, excluding 0.5 and 4 M urea, using 36-52 µM G-CSF C3 (B).

4.2.10 Assessing the structural integrity of G-CSF C3 between 25 and 37 ^{o}C

Comparison of the secondary and tertiary structures of G-CSF C3 in the presence and absence of urea was carried out at 25 ^{o}C, revealing a change in the environment of one or more of the three Tyr side chains. Additionally, equilibrium denaturation analyses were used to calculate the thermodynamic stability of G-CSF C3 at 25 ^{o}C, as well as relating the fractions of native G-CSF C3 at 25 ^{o}C to the fractions of soluble G-CSF C3 calculated from aggregation assays carried out at 37 ^{o}C. Relating the conformational change observed at 25 ^{o}C to the increased aggregation observed for G-CSF C3 at pH 7 in the presence of 4 M urea and at pH 4 in the presence of 3 M urea at 37 ^{o}C relies on there being no conformational change of G-CSF C3 between 25 and 37 ^{o}C.

To probe for conformational changes over this temperature range, ^{1}H-^{15}N HSQC spectra of G-CSF C3 in pH 7 buffer with or without 4 M urea were acquired at temperatures of 25-37 ^{o}C with 3 ^{o}C increments, using protein concentrations of 65 and 63 µM in 0 and 4 M urea, respectively (Figure 4.11A and Figure 4.12A). ^{1}H-^{15}N HSQC spectra of G-CSF C3 in pH 4 buffer with or without 3 M urea were acquired over the same temperatures, using protein concentrations of 85 and 78 µM in 0 and 3 M urea, respectively (Figure 4.13A and Figure 4.14A). This temperature titration of ^{1}H-^{15}N HSQC spectra also allowed assignments of backbone amide protons to be transferred from 25 to 37 ^{o}C for aggregation assays, monitored by changes in HSQC peak intensity (see Chapter 5).

Chemical shifts of backbone amide proton and nitrogen nuclei were used to calculate CSPs (see Section 2.9.6) for G-CSF C3 at 25 ^{o}C relative to 37 ^{o}C (Figure 4.11B-Figure 4.14B). The CSPs observed are small in amplitude and are due to the movement of peaks in the proton dimension of ^{1}H-^{15}N HSQC spectra upfield towards the proton resonance frequency of water (Figure 4.11A-Figure 4.14A), which is calculated to be 4.75 and 4.63 ppm at 25 and 37 ^{o}C, respectively, at pH 5.5. Although the proton resonance frequency of water will vary with pH, this variation is only 0.02 ppm per pH unit, smaller than the variation caused by temperature[319]. CSPs were not calculated for Ile96 from ^{1}H-^{15}N HSQC spectra of G-CSF C3 at pH 4 or 7 due to overlap of the HSQC peak for this residue with peaks for the side chains of Gln, Arg and Lys residues (Figure 3.17A), which made the peak for Ile96 difficult to track over different temperatures.

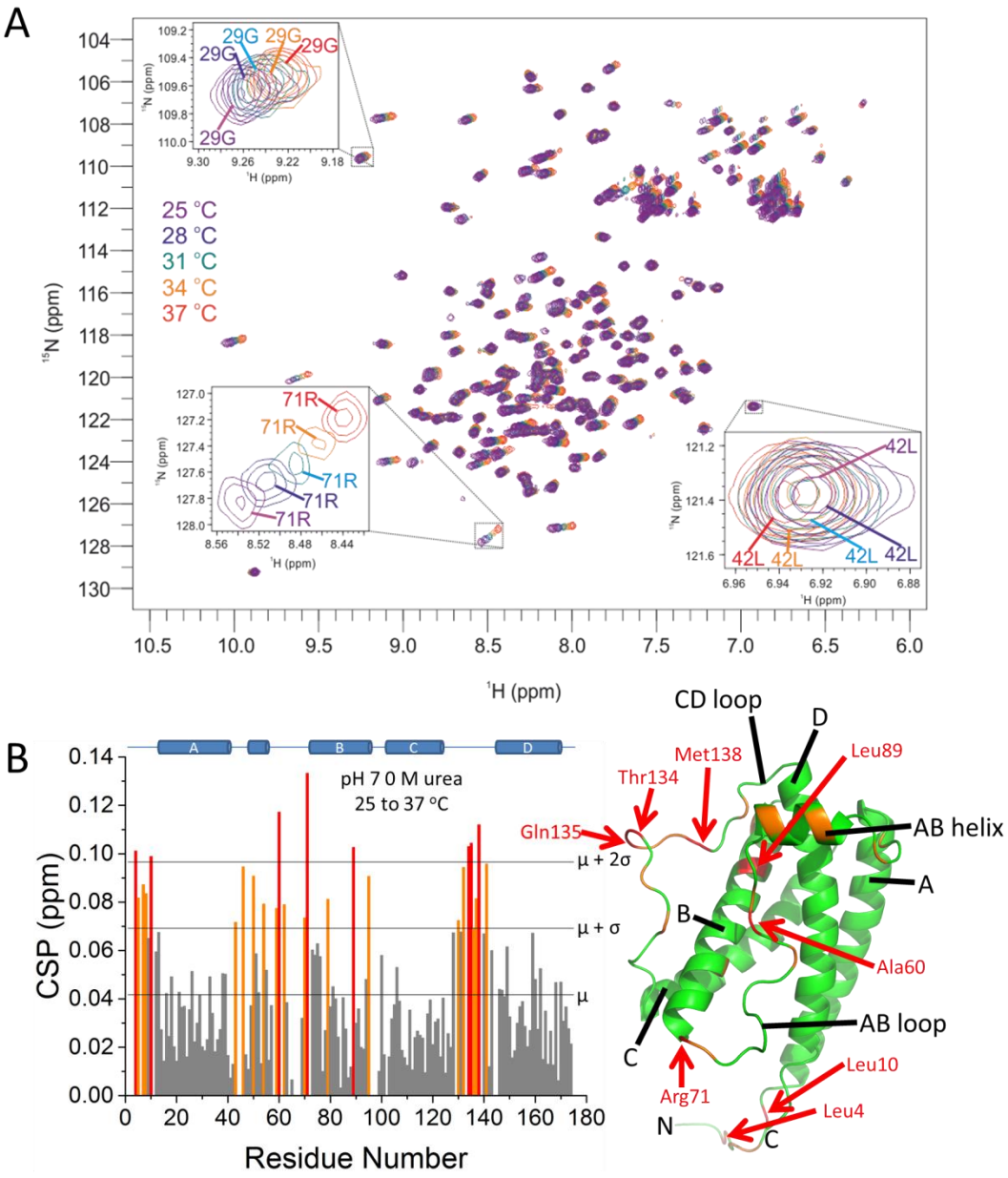


Figure 4.11 Thermal titration of G-CSF C3 at pH 7 without urea from 25 to 37 °C monitored by ^{1}H-^{15}N HSQC spectra. Spectra of 65 µM G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 0.018 % (*w/v*) sodium azide, 10 % (*v/v*) D_2O pH 7 buffer (I = 72 mM) were acquired at 25 (purple), 28 (blue), 31 (teal), 34 (orange) and 37 °C (red) (A). CSPs were calculated using backbone amide proton and nitrogen chemical shifts of G-CSF C3 at 25 and 37 °C (B). Horizontal lines show the mean, mean + σ and mean + 2σ CSP values. Orange and red bars correspond to residues exhibiting CSPs greater than the mean + σ and mean + 2σ, respectively. A secondary structure schematic is shown above for reference. Right: Residues with significant CSPs are shown mapped onto the structure of G-CSF wt (PDB 1GNC[283]) using the same colour scheme, constructed using PyMOL[289]. Red residues are also labelled. Spectra were acquired using a spectrometer operating at a proton resonance frequency of 750 MHz equipped with a cryogenic probe.

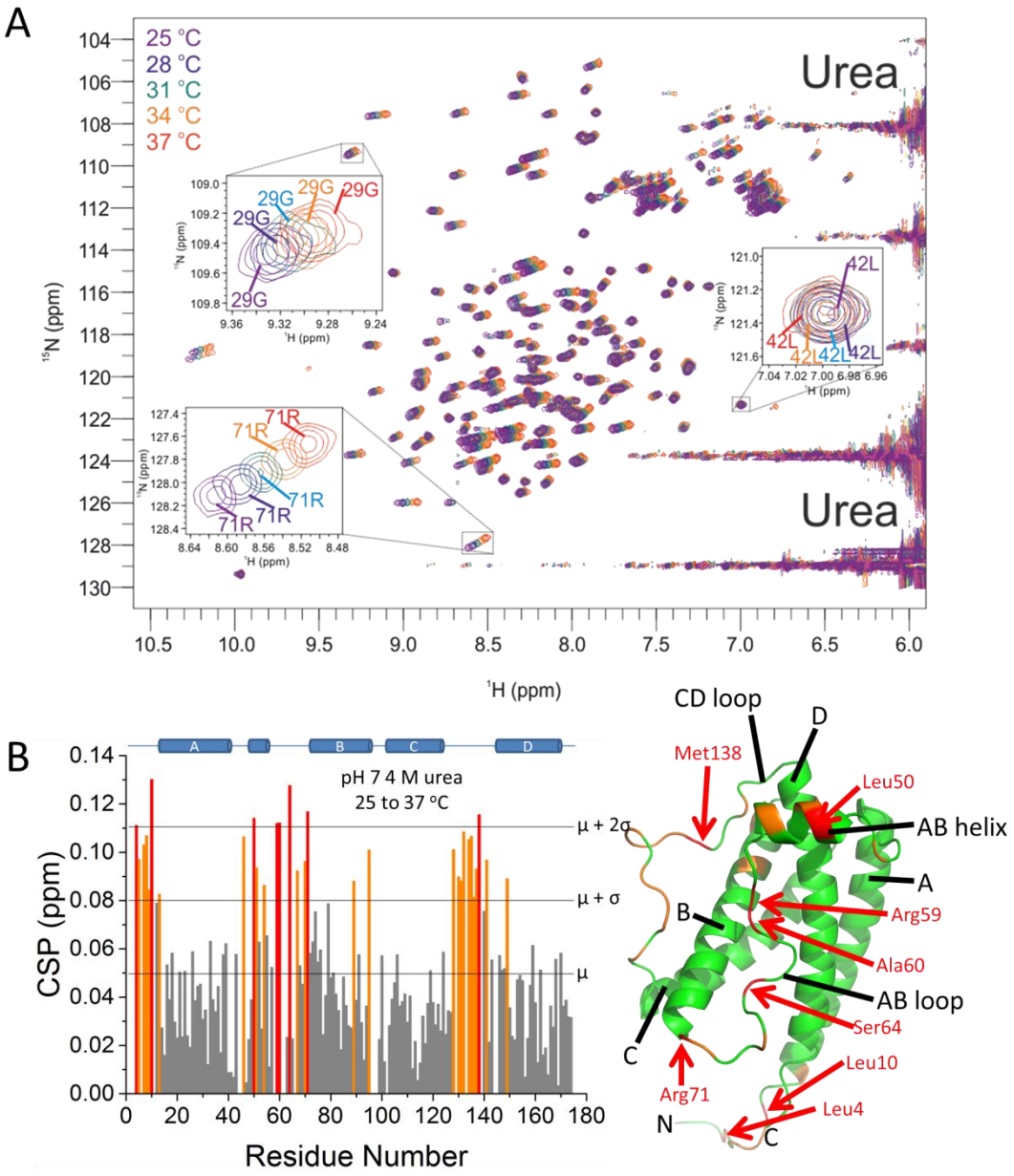


Figure 4.12 Thermal titration of G-CSF C3 at pH 7 with 4 M urea from 25 to 37 °C monitored by ^{1}H-^{15}N HSQC spectra. Spectra of 63 µM G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 0.018 % (*w/v*) sodium azide, 10 % (*v/v*) D_2O pH 7 buffer (I = 72 mM) were acquired in the presence of 4 M urea at 25 (purple), 28 (blue), 31 (teal), 34 (orange) and 37 °C (red). Extra peaks on the right hand side of the spectrum due to urea are labelled (A). CSPs were calculated using backbone amide proton and nitrogen chemical shifts of G-CSF C3 at 25 and 37 °C (B). Horizontal lines show the mean, mean + σ and mean + 2σ CSP values. Orange and red bars correspond to residues exhibiting CSPs greater than the mean + σ and mean + 2σ, respectively. A secondary structure schematic is shown above for reference. Right: Residues with significant CSPs are shown mapped onto the structure of G-CSF wt (PDB 1GNC[283]) using the same colour scheme, constructed using PyMOL[289]. Red residues are also labelled. Spectra were acquired using a spectrometer operating at a proton resonance frequency of 750 MHz equipped with a cryogenic probe.

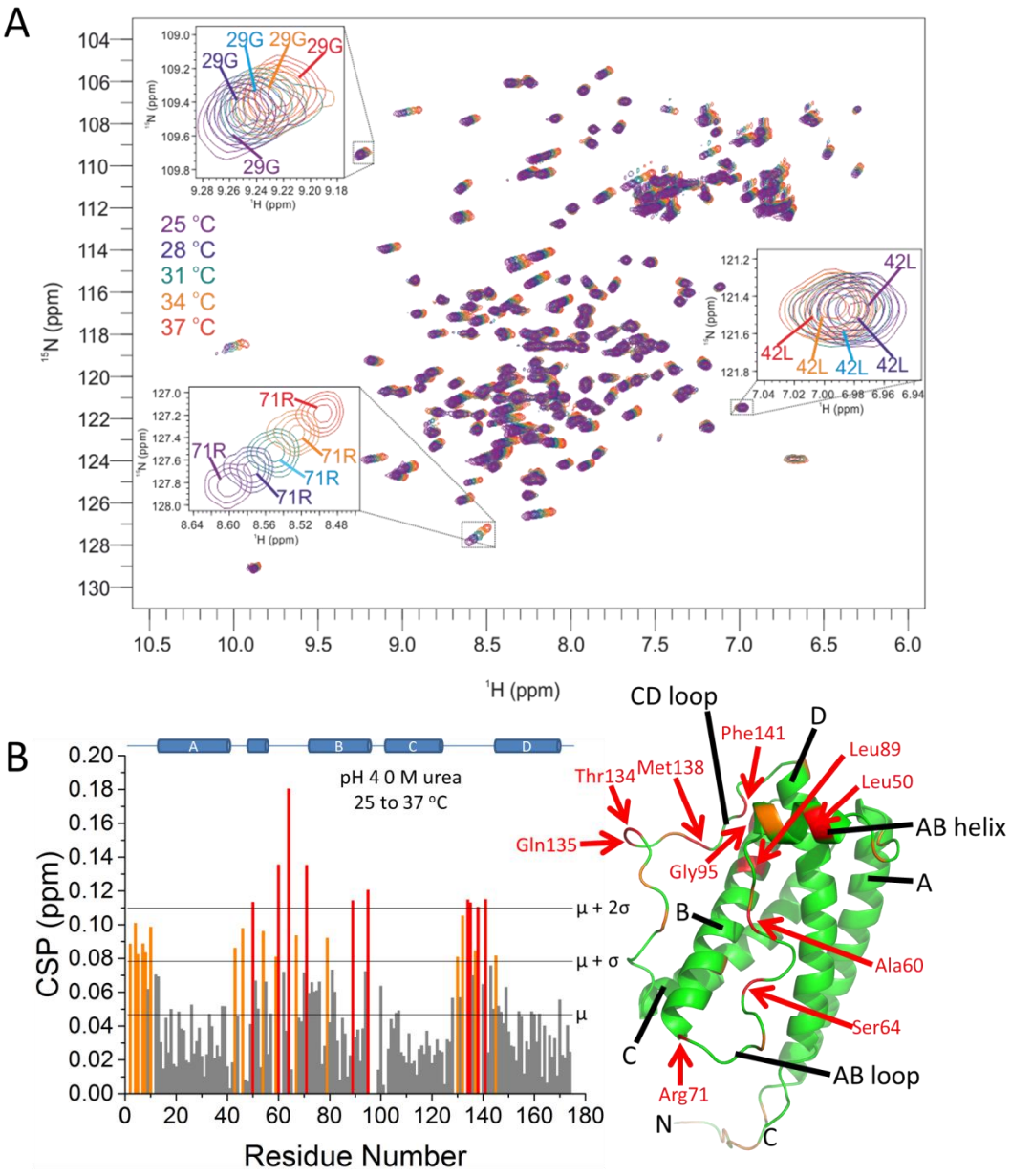


Figure 4.13 Thermal titration of G-CSF C3 at pH 4 without urea from 25 to 37 °C monitored by ^{1}H-^{15}N HSQC spectra. Spectra of 85 μM G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 10 % (*v/v*) D_2O pH 4 buffer (I = 39 mM) were acquired at 25 (purple), 28 (blue), 31 (teal), 34 (orange) and 37 °C (red) (A). CSPs were calculated using backbone amide proton and nitrogen chemical shifts of G-CSF C3 at 25 and 37 °C (B). Horizontal lines show the mean, mean + σ and mean + 2σ CSP values. Orange and red bars correspond to residues exhibiting CSPs greater than the mean + σ and mean + 2σ, respectively. A secondary structure schematic is shown above for reference. Right: Residues with significant CSPs are shown mapped onto the structure of G-CSF wt (PDB 1GNC[283]) using the same colour scheme, constructed using PyMOL[289]. Red residues are also labelled. Spectra were acquired using a spectrometer operating at a proton resonance frequency of 750 MHz equipped with a cryogenic probe.

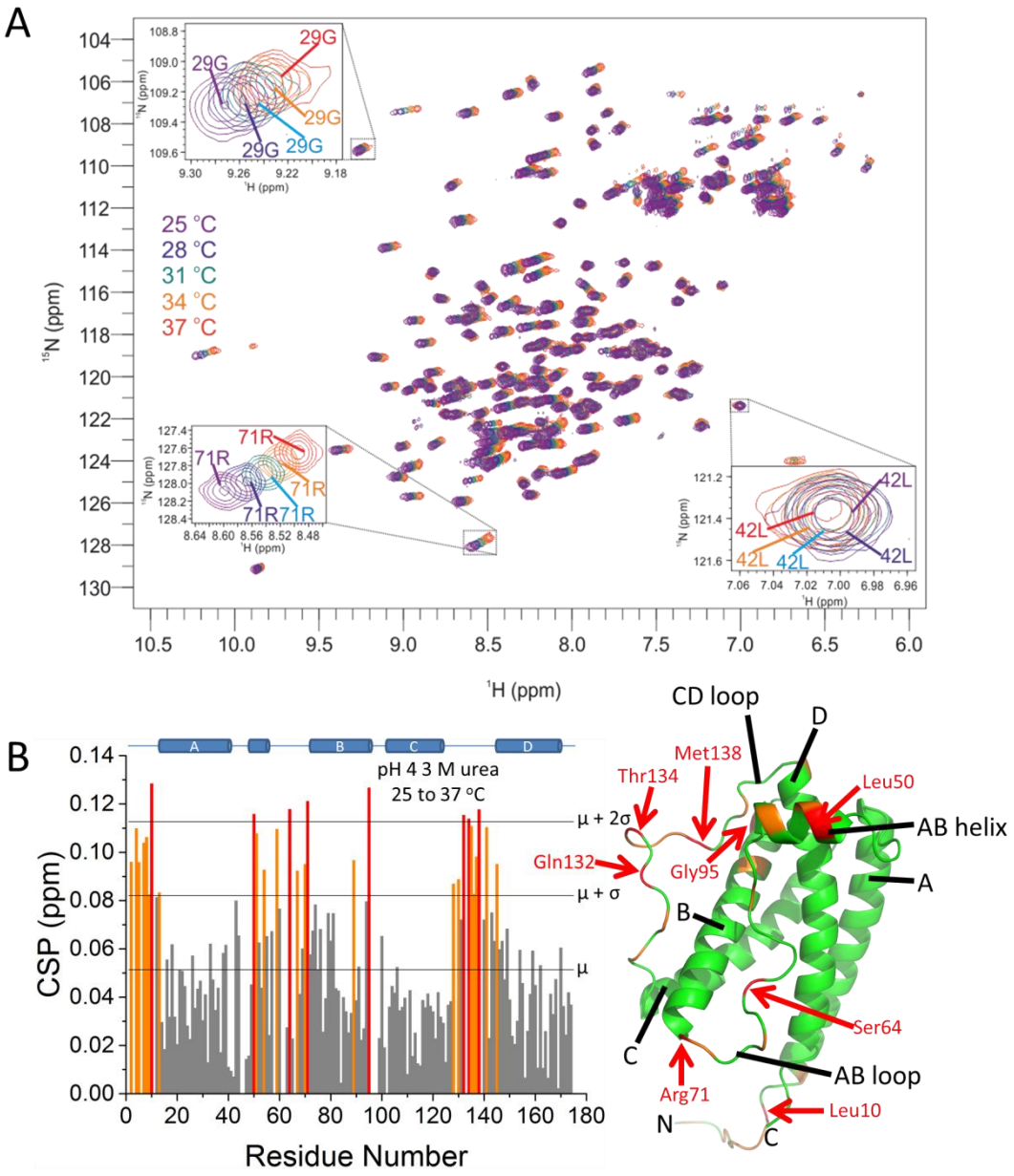


Figure 4.14 Thermal titration of G-CSF C3 at pH 4 with 3 M urea from 25 to 37 ^{o}C monitored by ^{1}H-^{15}N HSQC spectra. Spectra of 78 μM G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 10 % (*v/v*) D_2O pH 4 buffer (I = 39 mM) were acquired in the presence of 3 M urea at 25 (purple), 28 (blue), 31 (teal), 34 (orange) and 37 ^{o}C (red) (A). CSPs were calculated using backbone amide proton and nitrogen chemical shifts of G-CSF C3 at 25 and 37 ^{o}C (B). Horizontal lines show the mean, mean + σ and mean + 2σ CSP values. Orange and red bars correspond to residues exhibiting CSPs greater than the mean + σ and mean + 2σ, respectively. A secondary structure schematic is shown above for reference. Right: Residues with significant CSPs are shown mapped onto the structure of G-CSF wt (PDB 1GNC[283]) using the same colour scheme, constructed using PyMOL[289]. Red residues are also labelled. Spectra were acquired using a spectrometer operating at a proton resonance frequency of 750 MHz equipped with a cryogenic probe.

The migration of peaks towards the water resonance frequency with increasing temperature is observable in the ^{1}H-^{15}N HSQC spectra of G-CSF C3 at pH 4 and 7 both in the presence and absence of urea (Figure 4.11A-Figure 4.14A) and is due to an increase in the hydrogen exchange rate of amide protons with water protons in the solvent[335]. The change in hydrogen exchange rate also affects the amide nitrogen chemical shift, causing movement of HSQC peaks in the nitrogen dimension as well as proton. This accounts for the largest CSPs mainly being observed for amide protons of residues in unstructured regions of G-CSF C3, which are more often involved in hydrogen bonds with water molecules than other residues. As well as depending on buffer conditions such as pH and temperature[335], the rate of hydrogen exchange also depends on the amino acid sequence[336], which accounts for differences in CSPs observed for residues that are close together in the amino acid sequence.

CSPs could not be calculated from ^{1}H-^{15}N HSQC spectra of G-CSF C3 at pH 7 for Phe145 in the absence or presence of 4 M urea (Figure 4.11B and Figure 4.12B, respectively), as the peak for this residue is broadened beyond detection at 37 ^{o}C. In the absence of urea, CSPs could not be calculated for Ser67, Gln68 and Leu72 for G-CSF C3 at pH 7 (Figure 4.11B) because these residues are also broadened beyond detection at 37 ^{o}C. Broadening of these residues could suggest changes in their dynamics at higher temperature at pH 7 that are affecting their longitudinal and transverse relaxation rates (see Section 2.9.8). CSPs also could not be calculated for residue Glu46 of G-CSF C3 at pH 4 in the presence of 3 M urea (Figure 4.14B), as this residue is too broad for detection at 25 ^{o}C but is observable at 37 ^{o}C, again implying a possible change in the dynamics of this residue at higher temperature under these conditions.

Overall, thermal titration of G-CSF C3 from 25 to 37 ^{o}C monitored using ^{1}H-^{15}N HSQC spectra suggests no conformational changes as a result of this temperature increase, based on calculation of CSPs. This supports the notion of a change in the environment of one or more Tyr side chains observed in sub-denaturing urea concentrations at 25 ^{o}C as being potentially linked to G-CSF C3 aggregation at 37 ^{o}C, which is promoted in sub-denaturing urea concentrations.

4.2.11 Investigating exposed hydrophobic surfaces of G-CSF C3 by 8-anilino-1-napthalenesulphonic acid (ANS) fluorescence

The fluorescence emission properties of ANS depend on the polarity of the environment surrounding the dye. In water, the maximum emission intensity of ANS fluorescence is observed at 505 nm[394]. However, upon increased hydrophobicity of the environment around the dye e.g. due to binding to a hydrophobic protein surface, the fluorescence quantum yield increases significantly. For instance, the quantum yield of ANS increases from 0.004 to 0.75 upon binding to BSA in water[395]. The increase in fluorescence emission intensity is accompanied by a blue-shift in the wavelength at which the maximum emission intensity is observed to around 480 nm[396].

The increased fluorescence emission intensity and blue-shift in the wavelength of maximum emission upon increased hydrophobicity of the environment around ANS has led to its use in identifying partially folded states of proteins, including not only the population of equilibrium and kinetic intermediates[355] and molten globule states[394] in protein folding, but also amorphous aggregates[394] and oligomeric species that represent early-stage intermediates in amyloid fibril formation[79]. ANS has also been used to assess the surface hydrophobicity of native proteins and is generally used to probe for exposed hydrophobic surfaces[394].

Although commonly used to probe for exposed hydrophobic surfaces, the dominant mechanism of ANS binding to protein surfaces is thought to be electrostatic, specifically through ion pairing between the negatively charged sulphonate group and positively charged side chains of Arg, Lys and His at acidic pH[397-399]. There are also most likely complementary interactions such as van der Waals or hydrophobic interactions that help stabilise binding[398,399]. Binding of ANS to Arg and to a lesser extent Lys side chains has been shown to cause an increase in fluorescence emission intensity and a blue-shift in the wavelength of maximum emission, although these changes are lower in magnitude than those brought about by binding to hydrophobic surfaces and have been demonstrated using at least ten thousand-fold molar excesses of single Arg or Lys amino acids to ANS, or the use of poly-Arg or poly-Lys peptides[398]. Generally, ANS binding to protein surfaces through electrostatic interactions will not cause the characteristic increase in fluorescence

intensity and blue-shift in the wavelength of maximum emission unless also accompanied by a change in the hydrophobicity of the environment of the dye[397,398].

ANS fluorescence of G-CSF C3 in the presence and absence of urea was used to probe for changes in the amount of exposed hydrophobic surface area that may provide interfaces for protein aggregation. 250 µM ANS was incubated with 2.5 µM G-CSF C3 in pH 7 buffer, or 5 µM G-CSF C3 in pH 4 buffer, in the presence of 0-7 M urea. Urea concentrations of 0-1 M were used at 0.2 M increments, as well as 1-7 M at 1 M increments. 1 ml samples were set up in 1.5 ml Eppendorf tubes and incubated quiescently at 25 ^{o}C overnight and fluorescence emission spectra acquired at 25 ^{o}C the next day (Figure 4.15).

ANS fluorescence emission spectra in pH 7 buffer in 0-7 M urea with or without G-CSF C3 appear similar (Figure 4.15A and C, respectively), suggesting no change in the polarity of the environment around ANS as a result of adding G-CSF C3. Although there are slight differences in the emission intensities of ANS with or without G-CSF C3 at a wavelength of 475 nm in 0.2-1 M urea (Figure 4.15E), these differences are smaller than the characteristic changes in fluorescence emission associated with a change in the hydrophobicity of the environment of ANS[395,396], suggesting no interaction of ANS with exposed hydrophobic surfaces of G-CSF C3.

ANS fluorescence emission spectra in pH 4 buffer with G-CSF C3 show higher intensities and are blue-shifted compared to spectra acquired without G-CSF C3 (Figure 4.15B and D, respectively), implying increased hydrophobicity of the environment of ANS in the presence of G-CSF C3. Intriguingly, the intensity of emission of ANS with G-CSF C3 at 475 nm decreases dramatically as the urea concentration increases from 0-1 M and then appears to remain low up to urea concentrations of 7 M (Figure 4.15F), even though ANS has been demonstrated to bind to proteins at higher urea concentrations[399]. This implies reduced hydrophobicity of the environment around ANS in the presence of G-CSF C3 at higher urea concentrations and perhaps contrasts with the increased aggregation of G-CSF C3 observed at higher urea concentrations, which is assumed to be due to increased exposure of hydrophobic surface area.

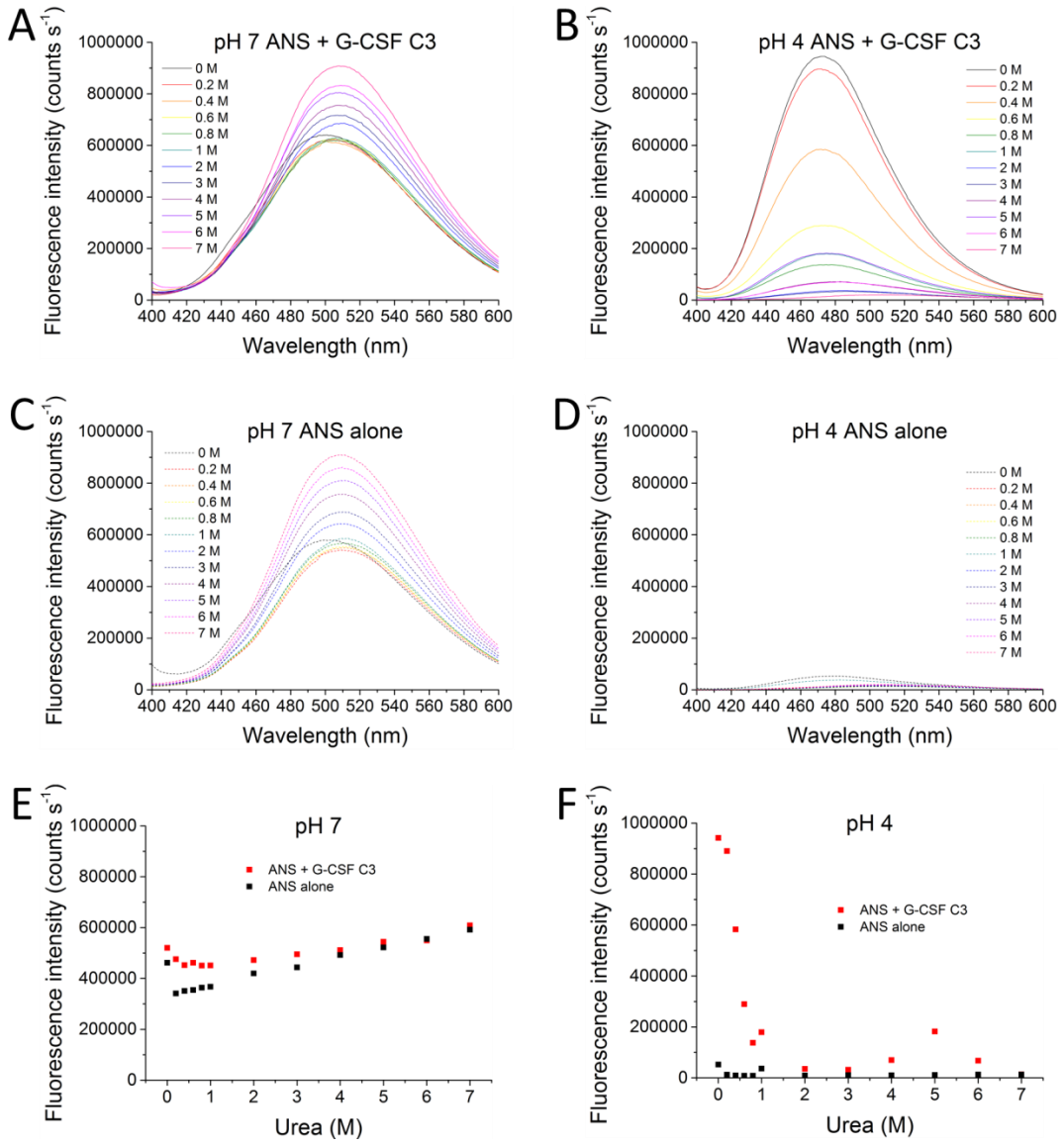


Figure 4.15 ANS fluorescence emission spectra at different urea concentrations with or without G-CSF C3. Spectra of 250 µM ANS in 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide pH 7 buffer (I = 80 mM) were acquired with (solid lines in A) or without (dashed lines in C) 2.5 µM G-CSF C3 in the presence of 0 (black), 0.2 (red), 0.4 (orange), 0.6 (yellow), 0.8 (green), 1 (teal), 2 (blue), 3 (navy blue), 4 (purple), 5 (violet), 6 (magenta) and 7 M (pink) urea. Spectra of 250 µM ANS in 20 mM sodium phosphate, 20 mM sodium acetate pH 4 buffer (I = 43 mM) were acquired with (solid lines in B) or without (dashed lines in D) 5 µM G-CSF C3 in the presence of the same urea concentrations. The fluorescence intensity at a wavelength of 475 nm in different urea concentrations is shown for ANS samples with or without G-CSF C3 (red and black, respectively) at pH 7 (E) and 4 (F).

Incubation of G-CSF C3 with ANS in pH 4 buffer in the absence of urea resulted in protein precipitation. In the presence of urea concentrations above 1 M insoluble aggregates could not be observed, suggesting that formation of G-CSF C3 aggregates with ANS incorporated may account for the increased hydrophobicity in the environment around the dye.

4.3 Discussion

Despite the reduced aggregation propensity predicted for G-CSF C3 compared to the wild-type protein (Figure 4.1), the aggregation behaviours of these two proteins appear to be similar. Aggregation of G-CSF C3 at pH 4 requires higher protein concentrations to cause comparable rates of aggregation to those observed at pH 7 during quiescent incubation at 37 ^{o}C (Figure 4.2), suggesting that G-CSF C3 is less aggregation-prone at pH 4. This matches the reported behaviour for wild-type G-CSF in that it is less aggregation-prone at acidic pH[185,276,303] e.g. in its formulation buffer as Filgrastim[276], but aggregates readily at physiological pH and temperature[142,185,214,307,308,391]. Additionally, the proportions of aggregated G-CSF C3 at pH 7 at 1.7 mg/ml after certain time periods of incubation at 37 ^{o}C (Figure 4.2) were similar to proportions of aggregated protein present after incubation of 1[142] and 1.5[185,308] mg/ml G-CSF wt under the same conditions for similar time periods (see Section 4.2.2).

The proportion of aggregated G-CSF C3 at 5.5 mg/ml after 68 h quiescent incubation at pH 7 and 37 ^{o}C was 49 % (Figure 4.2), which is very similar to reported values around 50 % of aggregated wild-type G-CSF at 5 mg/ml incubated for 48 h under the same conditions[142,214]. This suggests that although the rate of aggregation of 5.5 mg/ml G-CSF C3 may be slower than 5 mg/ml wild-type G-CSF at pH 7, ultimately these proteins will aggregate to the same extent. A similar effect has previously been observed as a result of substitution of the unpaired Cys18 residue, which reduced the rate of aggregation but ultimately aggregated to the same extent as wild-type G-CSF[142]. Therefore, the slower rate of G-CSF C3 aggregation relative to wild-type G-CSF may be the result of the C18G substitution.

Aggregation of G-CSF C3 at pH 4 and 7 appears to be accelerated by addition of urea concentrations that are too low to cause global unfolding (Figure 4.3A and B, respectively). The effect of adding 5, 6 and 7 M urea on the rate of G-CSF C3 aggregation at pH 7 (Figure 4.3A) is similar to the effect of 2, 2.5 and 3.5 M GdnHCl on the rate of aggregation of wild-type G-CSF at pH 7[142], in the sense that the fastest rates of G-CSF wt and C3 aggregation are observed in the presence of 2.5 M GdnHCl and 6 M urea, respectively. At higher denaturant concentrations the rates of G-CSF wt and C3 aggregation are reduced again, presumably due to solubilisation of the unfolded state.

Interestingly, addition of sub-denaturing urea concentrations also appears to increase the rate of G-CSF C3 aggregation at pH 4 (Figure 4.3B), but to a lesser extent than observed at pH 7 (Figure 4.3A). The denaturant dependence of G-CSF wt aggregation at pH 4 has not been reported, so it is unknown whether the wild-type protein would also exhibit increased aggregation at acidic pH in the presence of denaturant.

Surprisingly, the morphologies of G-CSF C3 aggregates formed at pH 4 and 7 appear very different (Figure 4.5). G-CSF C3 aggregates formed at pH 7 appear to be largely amorphous in the absence (Figure 4.5A and B) or presence (Figure 4.5C and D) of 4 M urea, while aggregates at pH 4 appear to more closely resemble protofibrils i.e. late-stage intermediates in amyloid fibril formation[79] (Figure 4.5E-H), particularly when formed in the presence of 3 M urea (Figure 4.5G and H). Consequently, it is unclear whether the lower rate of aggregation of G-CSF C3 observed at pH 4 (Figure 4.3B) relative to at pH 7 (Figure 4.3A) in the presence of sub-denaturing urea concentrations is entirely due to electrostatic repulsion between G-CSF C3 monomers at acidic pH, or is affected by the different morphology of aggregates formed at acidic and neutral pH (Figure 4.5).

Aggregation assays of G-CSF C3 at pH 7 in the presence of different urea concentrations appear to result in different end-points of aggregation, based on different plateaus of A_{280} values over time during aggregation assays (Figure 4.3A). Conversely, aggregation assays of G-CSF wt at pH 7 in the presence of different GdnHCl concentrations appear to progress towards the same end-point[142]. The ionic strength of the buffer used does not appear to affect the rate of aggregation of G-CSF wt at neutral pH in the absence of denaturant, based on comparison of wild-type G-CSF aggregation assays in different studies using similar protein concentrations but buffers of different ionic strengths[142,185,308]. However, the overall charge of GdnHCl that is absent from urea could be affecting the size of the aggregates formed.

EM images show that aggregation of G-CSF C3 at pH 7 in the presence of 4 M urea results in mixtures of amorphous and protofibrillar aggregates (Figure 4.5D), while G-CSF C3 aggregates formed in the absence of urea appear to be entirely amorphous (Figure 4.5A). This could mean that the aggregation assays for G-CSF C3 carried out in different urea concentrations result in different proportions of aggregates of different sizes being formed, which in turn lead to different end-points

of aggregation assays because larger aggregates will pellet more easily (Figure 4.3A). Conversely, aggregation assays for G-CSF wt carried out in different GdnHCl concentrations could result in aggregates of the same size or simply larger aggregates, owing to a possible effect of electrostatic interactions on the size of aggregates formed, meaning these aggregation assays lead to the same end-point. Unfortunately, the size of G-CSF wt aggregates formed in the presence of denaturant has not been reported, meaning this hypothesis cannot be confirmed.

Equilibrium denaturation analyses of G-CSF C3 at pH 4 and 7 at 25 ^{o}C are shown in Figure 4.4A and B, respectively overlaid with the fractions of soluble G-CSF C3 remaining after incubation of G-CSF C3 at pH 4 and 7 at 37 ^{o}C in the presence of 0-7 M urea, which were calculated from the aggregation assays shown in Figure 4.3. Overlaying of the fractions of folded and soluble G-CSF C3 revealed urea concentrations at which G-CSF C3 appears to be native but insoluble aggregates are formed (4 M urea at pH 7 and 3 M urea at pH 4, Figure 4.4A and B, respectively). Moreover, DLS analysis of aggregation assays revealed that soluble G-CSF C3 aggregates form more readily at pH 7 in the presence of 4 M urea and at pH 4 in the presence of 3 M urea, even before incubation of samples at 37 ^{o}C (Figure 4.6 and Figure 4.7, respectively). Therefore, many subsequent experiments were performed in urea concentrations of 0-4 M at pH 7 and 0-3 M at pH 4.

Equilibrium denaturation analyses of G-CSF C3 at pH 4 and 7 at 25 ^{o}C were also used to calculate ΔG^{o}_{UN} and M_{UN} values (Table 4.1), representing the free energy of unfolding and the dependence of this free energy on denaturant concentration, respectively[312]. ΔG^{o}_{UN} values for G-CSF C3 at pH 4 and 7 were calculated to be 37.0 and 40.0 kJ mol^{-1}, respectively from fluorescence emission measurements at 348 and 320 nm, respectively, which are similar to ΔG_{UN} values reported for wild-type G-CSF of 37.7 kJ mol^{-1} at pH 4 at 20 ^{o}C[304] and 40.2 kJ mol^{-1} at pH 7 at 25 ^{o}C[142], suggesting G-CSF C3 has a similar thermodynamic stability to the wild-type protein at pH 4 and 7.

ΔG^{o}_{UN} and M_{UN} values could only be calculated for G-CSF C3 at pH 7 using fluorescence emission intensities and ellipticities measured in 2-9 M urea due to the significant increases in signal observed in urea concentrations up to 2 M (Figure 4.4A). Far-UV CD spectra of G-CSF C3 at pH 7 in 0-4 M urea did not reveal any significant change in ellipticity in the presence of urea (Figure 4.8A), suggesting that the increased ellipticity observed during equilibrium denaturation analysis of G-

CSF C3 at pH 7 in 0-2 M urea may be exaggerated by linear extrapolation of the pre-transition baseline of this equilibrium curve (Figure 4.4A). Conversely, fluorescence emission spectra of G-CSF C3 at pH 7 in 0-4 M urea showed an increase in intensity across all wavelengths measured in urea concentrations of 0-2 M and to a lesser extent 2-4 M (Figure 4.9A), corroborating the increase in fluorescence emission intensity observed during equilibrium denaturation analysis of G-CSF C3 at pH 7 in 0-2 M urea (Figure 4.4A).

Equilibrium denaturation analysis of G-CSF C3 at pH 4 revealed significant increases in the fluorescence emission intensity at 301 nm in urea concentrations up to 1 M (Figure 4.4B), which prevented use of these intensities in calculation of ΔG^{o}_{UN} and M_{UN} values. However, fluorescence emission intensities measured at 348 nm and ellipticities did not show any significant change in urea concentrations that are too low to cause global denaturation (Figure 4.4B). Fluorescence emission spectra of G-CSF C3 at pH 4 in 0-3 M urea showed an increase in intensity around 300 nm but no significant change around 345 nm. This would suggest a change in the environment of one or more of the three Tyr side chains in G-CSF C3 at pH 4 upon addition of urea, although Trp fluorescence quenching is observed for G-CSF C3 at pH 4 (Figure 3.11C), making it difficult to determine if a change in the environment of Trp119 was also occurring as a result of adding urea.

Near-UV CD spectra were used to further probe the changes in tertiary structure observed by fluorescence emission spectra in sub-denaturing urea concentrations. Near-UV CD spectra of G-CSF C3 at both pH 7 and 4 showed increased negativity in the region around 275 nm upon addition of urea (Figure 4.10A and B, respectively), indicative of a change in the environment of Tyr side chains. Conversely, far-UV CD spectra of G-CSF C3 at pH 7 and 4 in different urea concentrations did not reveal any significant change in ellipticity in the presence of urea (Figure 4.8A and B, respectively), suggesting no detectable changes in the secondary structure of G-CSF C3 upon addition of urea. A change in the environment of one or more of the three Tyr side chains of G-CSF C3 upon addition of urea concentrations that are too low to cause global denaturation is most likely also responsible for the increases in fluorescence emission intensity observed in spectra of G-CSF C3.

The biophysical experiments that revealed a change in the environment of Tyr side chains of G-CSF C3 in sub-denaturing urea concentrations were carried out at 25

^{o}C, whereas aggregation assays showing increased aggregation of G-CSF C3 in the presence of sub-denaturing urea concentrations were carried out at 37 ^{o}C. Thermal titrations of G-CSF C3 at pH 7 in the absence or presence of 4 M urea (Figure 4.11A and Figure 4.12A respectively) and at pH 4 in the absence or presence of 3 M urea (Figure 4.13A and Figure 4.14A respectively) monitored by ^{1}H-^{15}N HSQC spectra were carried out to ensure there are no significant conformational differences between G-CSF C3 at 25 and 37 ^{o}C. Calculation of CSPs revealed only small changes in proton and nitrogen chemical shifts (Figure 4.11B-Figure 4.14B), attributable to the increased hydrogen exchange rate of amide protons with solvent protons at higher temperatures[335]. This supports the theory that a change in the environment of Tyr side chains of G-CSF C3 detected at 25 ^{o}C is linked to aggregation at 37 ^{o}C in sub-denaturing urea concentrations.

In the study by Raso *et al.*[142], wild-type G-CSF showed accelerated aggregation in the presence of 1.5 M GdnHCl at pH 7 and has been reported to exhibit increased fluorescence emission under these conditions, which has been attributed to a conformational change of the AB loop altering the environment of the Trp59 side chain. Trp59 was identified as the source of increased wild-type G-CSF fluorescence by sequentially substituting each of the two Trp residues using W59Q and W119F mutants[142], the former of which should exhibit the same fluorescence emission spectra as G-CSF C3 based on the number and location of aromatic residues in these proteins. However, the G-CSF W59Q mutant (G-CSF C3 analogue) was not reported to exhibit increased emission intensity in the presence of sub-denaturing GdnHCl concentrations[142]. This would suggest that the W59R substitution found in G-CSF C3 is causing the increased fluorescence emission intensity observed in sub-denaturing urea concentrations and may be linked to a conformational change of the AB loop.

Fluorescence emission spectra of folded and unfolded G-CSF wt showed quenching of Trp fluorescence at neutral and to a far greater extent acidic pH (Figure 3.11B and A, respectively). The structure of G-CSF wt suggests that quenching of Trp fluorescence at neutral pH is most likely due to the proximity of the side chains of Trp59 and Tyr86 (Figure 3.12A). Therefore, the increased fluorescence emission of Trp59 in wild-type G-CSF observed by Raso *et al*.[142] upon addition of 1.5 M GdnHCl at pH 7 is most likely due to an increase in distance between the side chains of Trp59 and Tyr86, which could lessen the quenching effect of Tyr86 on Trp59 fluorescence emission.

This could mean that the increased fluorescence emission intensity of G-CSF C3 observed in low urea concentrations is due to reduced quenching of Tyr86 by Arg59, implying a change in the environment of the Tyr86 side chain in low urea concentrations that may be linked to aggregation. However, changes in the environments of the side chains of Tyr40 and Tyr166 may not have been detected by Raso *et al*.[142] as this study was primarily focussed on changes in the environment of Trp residues of wild-type G-CSF under conditions that promote aggregation. The conformational change undergone by G-CSF C3 in aggregation-prone conditions is probed at amino acid resolution by NMR in Chapter 5 in order to identify the Tyr residue or residues that may be linked to promoting G-CSF C3 aggregation.

ANS fluorescence emission in the presence of G-CSF C3 was investigated over a range of urea concentrations to probe for exposure of hydrophobic surfaces that could act as aggregation interfaces. The fluorescence emission spectra of ANS with or without G-CSF C3 at pH 7 were similar in 0-7 M urea (Figure 4.15A and C, respectively), suggesting no interaction of ANS with hydrophobic surfaces.

The fluorescence emission spectra of ANS with G-CSF C3 at pH 4 in the absence of urea were significantly more intense than spectra of ANS alone at pH 4 (Figure 4.15B and D, respectively) and the wavelength at which maximum emission intensity was observed was blue-shifted in the presence of G-CSF C3. This suggests a change in the hydrophobicity of the environment around ANS in the presence of G-CSF C3 at pH 4, which was not observed in the presence of urea concentrations of 1 M and above. However, protein precipitation was observed when G-CSF C3 was incubated with ANS at pH 4 in the absence of urea but not in the presence of urea concentrations above 1 M urea, suggesting that the change in hydrophobicity of ANS may be due to incorporation of ANS into aggregates.

There is no change in ANS fluorescence at pH 7 with or without G-CSF C3, which coupled with the structural similarity of G-CSF C3 conformations at pH 4 and 7, shown by CSPs calculated from ^{1}H-^{15}N HSQC spectra (Figure 3.19), suggests that ANS binding at pH 4 is driven by electrostatic interactions between the negatively charged sulphonate group and positively charged side chains, which is the dominant mechanism of ANS binding to protein surfaces[397-399]. This binding event may then promote aggregation by neutralising a positive charge on the surface of G-CSF C3 and adding a hydrophobic naphthalene group, whereby incorporation of ANS into aggregates causes the changes in ANS fluorescence emission spectra that are

characteristic of a change in the polarity of the environment of ANS. ANS has also previously been shown to accelerate aggregation of a therapeutic protein, although aggregation was attributed to an increase in hydrophobic interactions between ANS and the protein at aggregation-prone temperatures, compared to predominantly electrostatic interactions at lower temperatures[187].

Reduced binding of ANS to G-CSF C3 at pH 4 upon addition of urea could be due to a conformational change that disturbs an exposed positively charged surface that binds ANS, which in turn may help elucidate conformational changes in urea that promote G-CSF C3 aggregation. In addition, the hypothesised positively charged surface that may be neutralised upon ANS binding could be responsible for reducing the aggregation propensity of G-CSF C3 at acidic pH in the absence of ANS. For these reasons ANS binding by G-CSF C3 merited further investigation by NMR in Chapter 5.

In Chapter 5, the conformational changes of G-CSF C3 that may promote aggregation are characterised at amino acid resolution by NMR. In addition, the dynamics of G-CSF C3 are investigated under conditions that favour and disfavour aggregation through measurement of transverse relaxation rates. Hydrogen exchange NMR is also used to probe for changes in solvent exposure of residues in the presence and absence of urea concentrations that promote aggregation, along with aggregation assays during which ^{1}H-^{15}N HSQC peak intensities are monitored over time in an attempt to identify aggregation interfaces. ANS binding is also further investigated by NMR to probe for conformational changes that disturb ANS binding and could be linked to aggregation. In summary, the work in this chapter showed that G-CSF C3 exhibits similar aggregation behaviour to the wild-type protein and aggregation may be linked to a change in the environment of one or more of the three Tyr side chains found in G-CSF C3.

Chapter 5 Elucidating the conformational changes that promote G-CSF C3 aggregation

5.1 Introduction

The work presented in Chapter 4 showed that aggregation of G-CSF C3 at 37 ^{o}C is promoted in urea concentrations that are too low to cause global denaturation. In these sub-denaturing urea concentrations a change in the environment of Tyr40, Tyr86 and/or Tyr166 has been suggested by near-UV CD and fluorescence emission spectra (Figure 4.10 and Figure 4.9, respectively). Previously, a study of the aggregation of wild-type G-CSF at 37 ^{o}C had revealed a change in the fluorescence properties of Trp59 as being potentially linked to increased aggregation in sub-denaturing GdnHCl concentrations at pH 7 and 25 ^{o}C[142].

The proximity of the side-chains of Trp59 and Tyr86 (Figure 3.12A) suggests that the increased fluorescence emission intensity of Trp59 is most likely due to reduced quenching by Tyr86, resulting from the conformational rearrangement of the AB loop proposed by Raso *et al*.[142], which would increase the distance between these side-chains. A conformation of wild-type G-CSF with native α-helical structure, but non-native Trp59 fluorescence, has also been identified during a study of wild-type G-CSF folding at pH 4 and 25 ^{o}C and proposed as a late stage kinetic folding intermediate[304]. This suggests that a conformational change in the AB loop of G-CSF C3 could be responsible for the changes in near-UV CD and fluorescence emission spectra observed in the presence of low urea concentrations but, more importantly, could be linked to aggregation. However, previous studies of G-CSF wt folding and aggregation were largely focussed on studying the changes in the environments of Trp59 and Trp119 side chains, and hence changes in the environment of Tyr40, Tyr86 and/or Tyr166 may not have been detected.

In this chapter, conformational changes of G-CSF C3 in sub-denaturing urea concentrations are characterised at amino acid resolution by monitoring urea titrations by ^{1}H-^{15}N HSQC spectra and calculation of CSPs. The dynamics of residues of G-CSF C3 are also investigated in urea concentrations that favour and disfavour aggregation by transverse relaxation rates (R_2) of backbone amide nitrogen nuclei. In addition, hydrogen-deuterium exchange is used to the measure

dynamics of G-CSF C3 residues over a longer timescale (minutes and hours) and to probe for regions of increased solvent exposure in aggregation-prone conditions. Identification of aggregation interfaces by monitoring peak intensities in ^{1}H-^{15}N HSQC spectra during G-CSF C3 aggregation assays in the presence and absence of urea is also described.

ANS was used in Chapter 4 in an attempt to reveal exposure of hydrophobic surfaces in aggregation-prone conditions, which proved not to be the case. However, differences in the fluorescence emission spectra of ANS in the presence and absence of G-CSF C3 at pH 4 were identified, but upon addition of urea concentrations of 1 M or higher this effect was eradicated (Figure 4.15F). These changes in the fluorescence properties of ANS are hypothesised to be caused indirectly by electrostatic interactions between the negatively charged sulphonate group and a positively charged surface on G-CSF C3 at pH 4 that may be involved in reducing aggregation of G-CSF C3 at acidic pH in the absence of ANS (Figure 4.2). Upon addition of ANS insoluble aggregates are observable by eye, suggesting that this positively charged surface is neutralised and ANS is incorporated into the resulting aggregates, which causes the changes observed in its fluorescence emission spectra. Consequently, ANS binding to G-CSF C3 at pH 4 and 7 has been further investigated by acquisition of ^{1}H-^{15}N HSQC spectra to investigate how ANS binds to the protein.

In Chapter 5 10 % (*v/v*) D_2O was added to all samples before performing NMR experiments. Therefore, all experiments in this chapter were carried out in pH 4 or 7 buffers in 18 mM sodium phosphate, 18 mM sodium acetate, 10 % (*v/v*) D_2O pH 4 buffer or 18 mM sodium phosphate, 18 mM sodium acetate, 0.018 % (*w/v*) sodium azide, 10 % (*v/v*) D_2O pH 7 buffer unless otherwise stated. The different pH values plus urea and protein concentrations used for measuring CSPs and R_2 rates are outlined in Table 5.1.

Experiment	pH	Urea concentration / M	G-CSF C3 concentration / μM
CSP	7	0	458
		0.5	81
		4	208
	4	0	473
		0.5	103
		3	173
R_2	7	0	130
		0.5	135
		4	134
	4	0	162
		0.5	151
		3	163

Table 5.1 Conditions used for the measurement of CSPs and backbone ^{15}N R_2 rates. CSPs and R_2 rates were measured at pH 4 and 7 in the presence of the urea concentrations shown using the G-CSF C3 concentrations shown. However, CSPs resulting from the addition of 3 or 4 M urea at pH 4 and 7, respectively were measured largely by urea titration of ^{1}H-^{15}N HSQC spectra using 103 and 81 μM ^{15}N G-CSF C3 at pH 4 and 7, respectively, but also by acquisition of HNCA, CBCA(CO)NH and ^{1}H-^{15}N HSQC spectra using 173 and 208 μM ^{13}C ^{15}N G-CSF C3 at pH 4 in 3 M urea and at pH 7 in 4 M urea, respectively.

5.2 Results

5.2.1 Characterising conformational changes of G-CSF C3 under aggregation-prone conditions using CSPs

Conformational changes of G-CSF C3 at pH 4 and 7 resulting from the addition of sub-denaturing urea concentrations were determined at amino acid resolution by calculation of CSPs (see Section 2.9.6) relative to G-CSF C3 in the absence of urea. In order to calculate CSPs for G-CSF C3 at pH 7 in the presence of 4 M urea and at pH 4 in the presence of 3 M urea (Figure 5.2 and Figure 5.4, respectively), chemical

shift assignments for backbone amide proton and nitrogen nuclei of G-CSF C3 in pH 4 and 7 buffers in the absence of urea were transferred to ^{1}H-^{15}N HSQC spectra of G-CSF C3 acquired in the presence of these urea concentrations. This was carried out largely by acquisition of ^{1}H-^{15}N HSQC spectra using 81 µM G-CSF C3 in pH 7 buffer in the presence of 0-4 M urea with 0.5 M increments, as well as 103 µM G-CSF C3 in pH 4 buffer in the presence of 0-3 M urea with 0.5 M increments (see Section 2.9.3). In addition, assignments were aided by acquisition of HNCA, CBCA(CO)NH and ^{1}H-^{15}N HSQC spectra using 208 µM G-CSF C3 in pH 7 buffer with 4 M urea and 176 µM G-CSF C3 in pH 4 buffer with 3 M urea (see Section 2.9.1).

Near-UV CD and fluorescence emission spectra of G-CSF C3 at pH 4 and 7 in the presence of sub-denaturing urea concentrations (Figure 4.10 and Figure 4.9, respectively) showed that the change in the environment of one or more Tyr side chains occurs at low urea concentrations i.e. 1 M urea at pH 4 and 0.5 M urea at pH 7, although at higher urea concentrations the environment of Tyr side chains appears to be perturbed to a greater extent. These urea concentrations have not been observed to cause increased aggregation of G-CSF C3 (Figure 4.3). Therefore, CSPs were also calculated for G-CSF C3 in the presence of 0.5 M urea at pH 4 and 7 (Figure 5.1 and Figure 5.3, respectively).

In the absence of urea, the backbone amide proton and nitrogen atoms of 160 residues of G-CSF C3 were assigned at pH 4 (see Section 3.2.7.2). The experiments above enabled assignment of the backbone amide proton and nitrogen atoms of 159 residues of G-CSF C3 at pH 4 in the presence of 3 M urea, with Ile96 being the only residue that could not be assigned. For G-CSF C3 at pH 7 in the absence of urea, 157 residues were assigned (see Section 3.2.7.2), while in the presence of 4 M urea 155 residues were assigned, but Leu62 and Ile96 could not be assigned. For G-CSF C3 at pH 4 in the presence of 0.5 M urea, 158 residues were assigned, with Leu62 and Ile96 not being assigned. At pH 7 in the presence of 0.5 M urea, 151 residues were assigned but Leu62, Ser64, Gln68, Leu72, Ile96 and Phe145 could not be assigned. Consequently, any possible conformational change associated with these residues in the urea concentrations outlined could not be assessed because CSPs could not be calculated.

Generally, addition of urea results in the movement of peaks in the proton dimension of ^{1}H-^{15}N HSQC spectra downfield away from the proton resonance frequency of

water, which is calculated to be 4.75 ppm at 25 ^{o}C at pH 5.5. Although the proton resonance frequency of water will vary with pH, this variation is only 0.02 ppm per pH unit[319].

Addition of urea results in reduction of the intrinsic exchange rate of amide protons with water protons in the solvent[335]. The change in hydrogen exchange rate also affects the amide nitrogen chemical shift, causing movement of HSQC peaks in the nitrogen dimension as well as proton. Urea reduces the base catalysed rate of hydrogen exchange by reducing the concentration of hydroxyl anions in the solution, which occurs as a result of urea repressing the ionisation of water to protons and hydroxyl anions i.e. K_w[335,400]. Hydroxyl anions are predominantly involved in catalysis of hydrogen exchange reactions above pH 3 in the absence of urea, meaning urea reduces the intrinsic hydrogen exchange rate at these pH values[335].

The general effect of adding urea of causing peaks to be shifted downfield in the proton and nitrogen dimensions of the ^{1}H-^{15}N HSQC spectra of G-CSF C3 is greater at pH 7 in the presence of 4 M urea (Figure 5.2A) than at pH 4 in the presence of 3 M urea (Figure 5.4A), which is evident from the changes in the peak position of the Trp119 indole side chain in Figure 5.2A and Figure 5.4A, respectively. This is because the intrinsic rate of base catalysed hydrogen exchange increases at higher pH, such that the effect of urea on reducing hydrogen exchange is more significant at higher pH. The effect of urea on the intrinsic hydrogen exchange rate is also concentration dependent, i.e. the change in the peak position of the Trp119 indole side chain of G-CSF C3 at pH 7 is greater in the presence of 4 M urea (Figure 5.2A) than in the presence of 0.5 M urea (Figure 5.1A)[335].

Any conformational changes of G-CSF C3 that occur in sub-denaturing urea concentrations should give rise to greater CSPs than the subtle changes that arise from changes in the intrinsic rate of hydrogen exchange, particularly as a result of the change in the environment of one or more Tyr side chains of G-CSF C3 as aromatic ring current effects cause large proton chemical shift changes[332]. However, any conformational changes of G-CSF C3 in sub-denaturing urea concentrations may be more easily observed at pH 4 compared with pH 7 due to the lower intrinsic rate of hydrogen exchange[335].

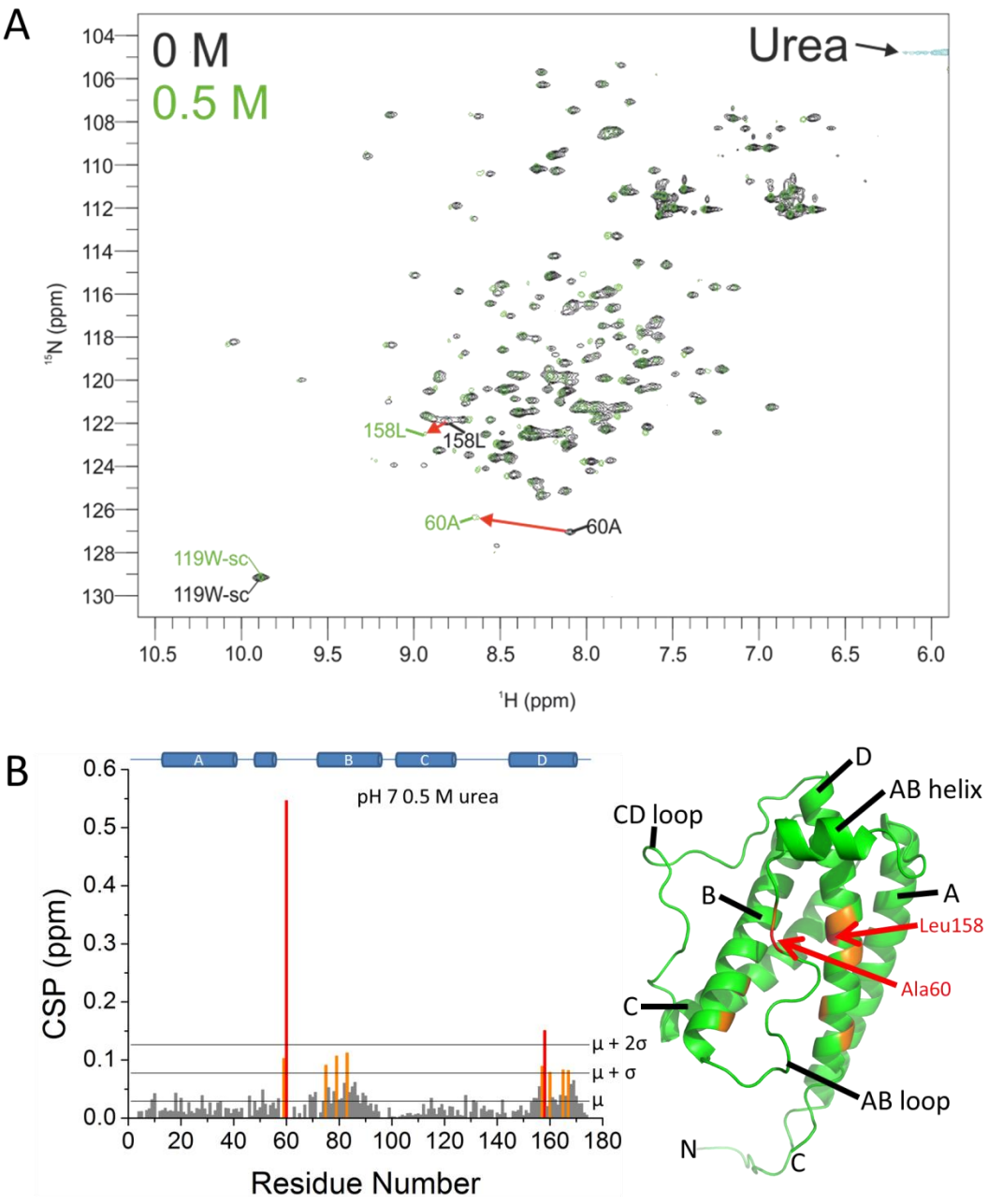


Figure 5.1 CSPs for G-CSF C3 at pH 7 resulting from addition of 0.5 M urea. The ^{1}H-^{15}N HSQC spectra of 458 μM G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 0.018 % (*w/v*) sodium azide, 10 % (*v/v*) D_2O pH 7 buffer (I = 72 mM) in the absence of urea (black) and 81 μM G-CSF C3 in the presence of 0.5 M urea (green) are shown in A, which were used for calculation of the CSPs shown in B. Extra peaks on the right hand side of the spectrum due to urea are labelled, along with peaks corresponding to the side chain (sc) of Trp119. CSPs were calculated using backbone amide proton and nitrogen chemical shifts of G-CSF C3 in the presence and absence of 0.5 M urea (B). Red arrows in A show the changes in peak position of residues that exhibit significantly large CSPs upon addition of 0.5 M urea, which are coloured red in B. Horizontal lines show the mean, mean + σ and mean + 2σ CSP values. Orange and red bars correspond to residues exhibiting CSPs greater than the mean + σ and mean + 2σ, respectively. A secondary structure schematic is shown above for reference. Right: Residues with significant CSPs are shown mapped onto the structure of G-CSF wt (PDB 1GNC[283]) using the same colour scheme, constructed using PyMOL[289]. Red residues are also labelled. Spectra were acquired at 25 ^{o}C using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

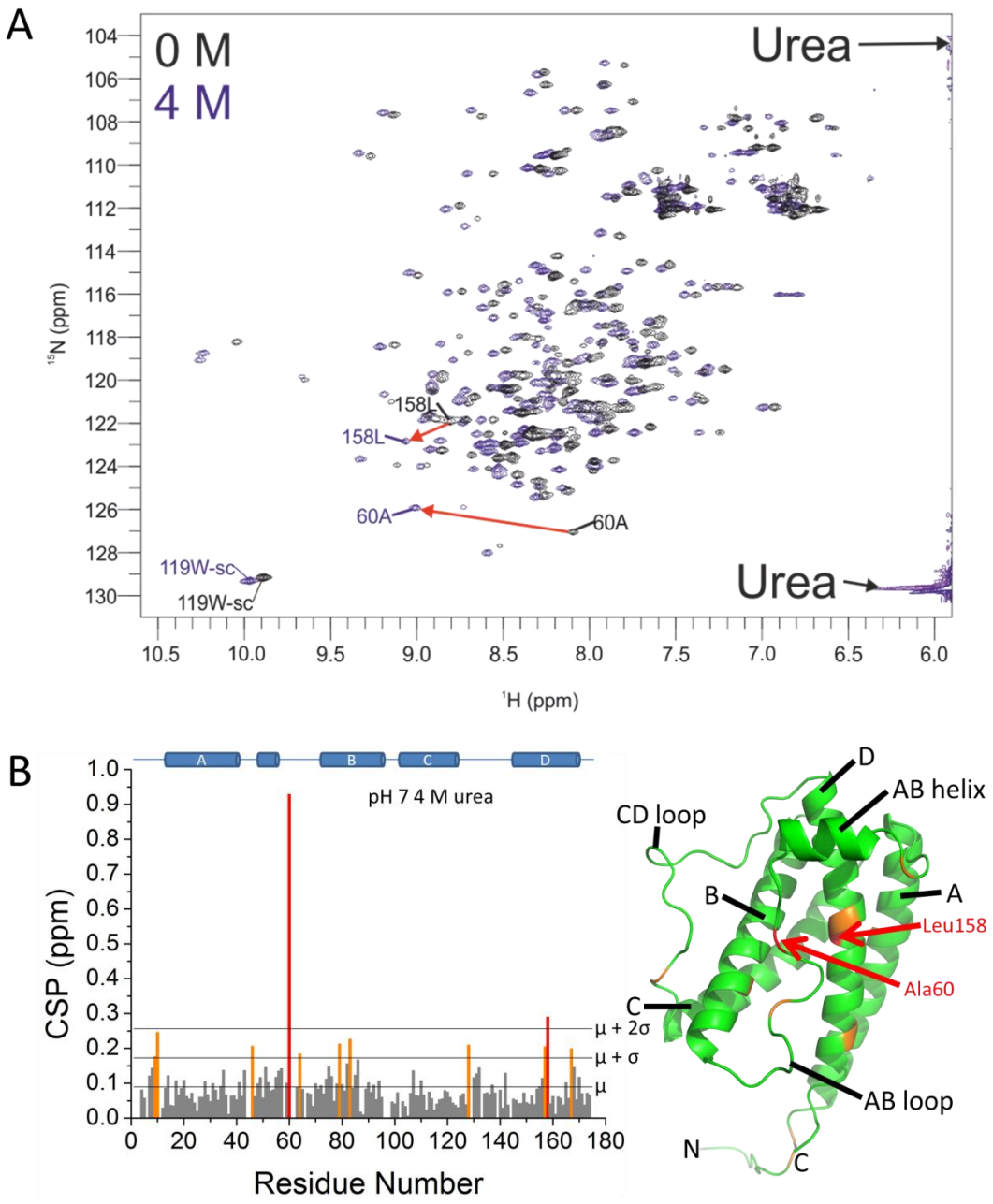


Figure 5.2 CSPs for G-CSF C3 at pH 7 resulting from addition of 4 M urea. The ^{1}H-^{15}N HSQC spectra of 458 μM G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 0.018 % (*w/v*) sodium azide, 10 % (*v/v*) D_2O pH 7 buffer (I = 72 mM) in the absence of urea (black) and 208 μM G-CSF C3 in the presence of 4 M urea (purple) are shown in A, which were used for calculation of the CSPs shown in B. Extra peaks on the right hand side of the spectrum due to urea are labelled, along with peaks corresponding to the side chain (sc) of Trp119. CSPs were calculated using backbone amide proton and nitrogen chemical shifts of G-CSF C3 in the presence and absence of 4 M urea (B). Red arrows in A show the changes in peak position of residues that exhibit significantly large CSPs upon addition of 4 M urea, which are coloured red in B. Horizontal lines show the mean, mean + σ and mean + 2σ CSP values. Orange and red bars correspond to residues exhibiting CSPs greater than the mean + σ and mean + 2σ, respectively. A secondary structure schematic is shown above for reference. Right: Residues with significant CSPs are shown mapped onto the structure of G-CSF wt (PDB 1GNC[283]) using the same colour scheme, constructed using PyMOL[289]. Red residues are also labelled. Spectra were acquired at 25 °C using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

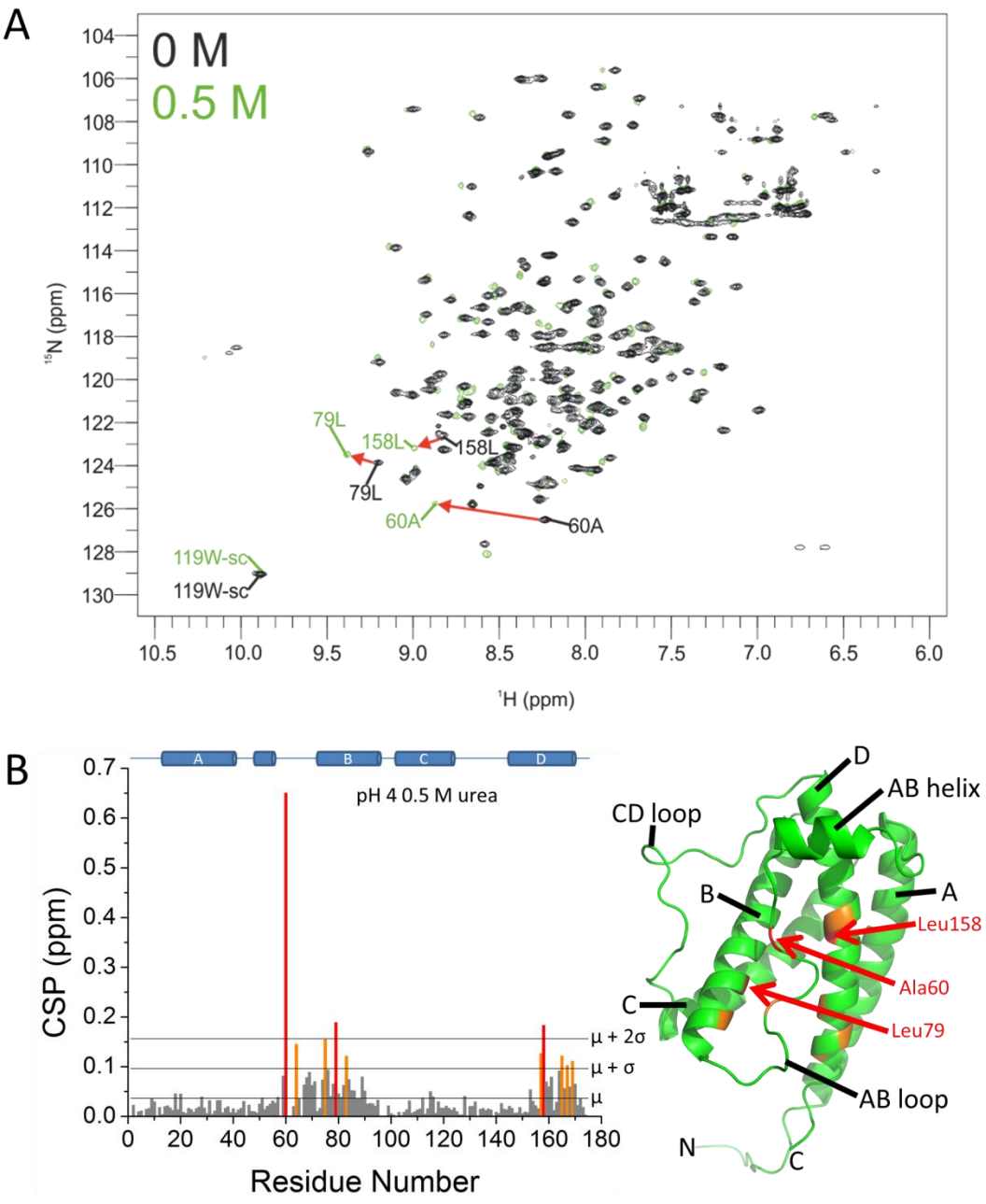


Figure 5.3 CSPs for G-CSF C3 at pH 4 resulting from addition of 0.5 M urea. The ^{1}H-^{15}N HSQC spectra of 473 μM G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 10 % (*v/v*) D_2O pH 4 buffer (I = 39 mM) in the absence of urea (black) and 103 μM G-CSF C3 in the presence of 0.5 M urea (green) are shown in A, which were used for calculation of the CSPs shown in B. Peaks corresponding to the side chain (sc) of Trp119 are labelled. CSPs were calculated using backbone amide proton and nitrogen chemical shifts of G-CSF C3 in the presence and absence of 0.5 M urea (B). Red arrows in A show the changes in peak position of residues that exhibit significantly large CSPs upon addition of 0.5 M urea, which are coloured in red in B. Horizontal lines show the mean, mean + σ and mean + 2σ CSP values. Orange and red bars correspond to residues exhibiting CSPs greater than the mean + σ and mean + 2σ, respectively. A secondary structure schematic is shown above for reference. Right: Residues with significant CSPs are shown mapped onto the structure of G-CSF wt (PDB 1GNC[283]) using the same colour scheme, constructed using PyMOL[289]. Red residues are also labelled. Spectra were acquired at 25 °C using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

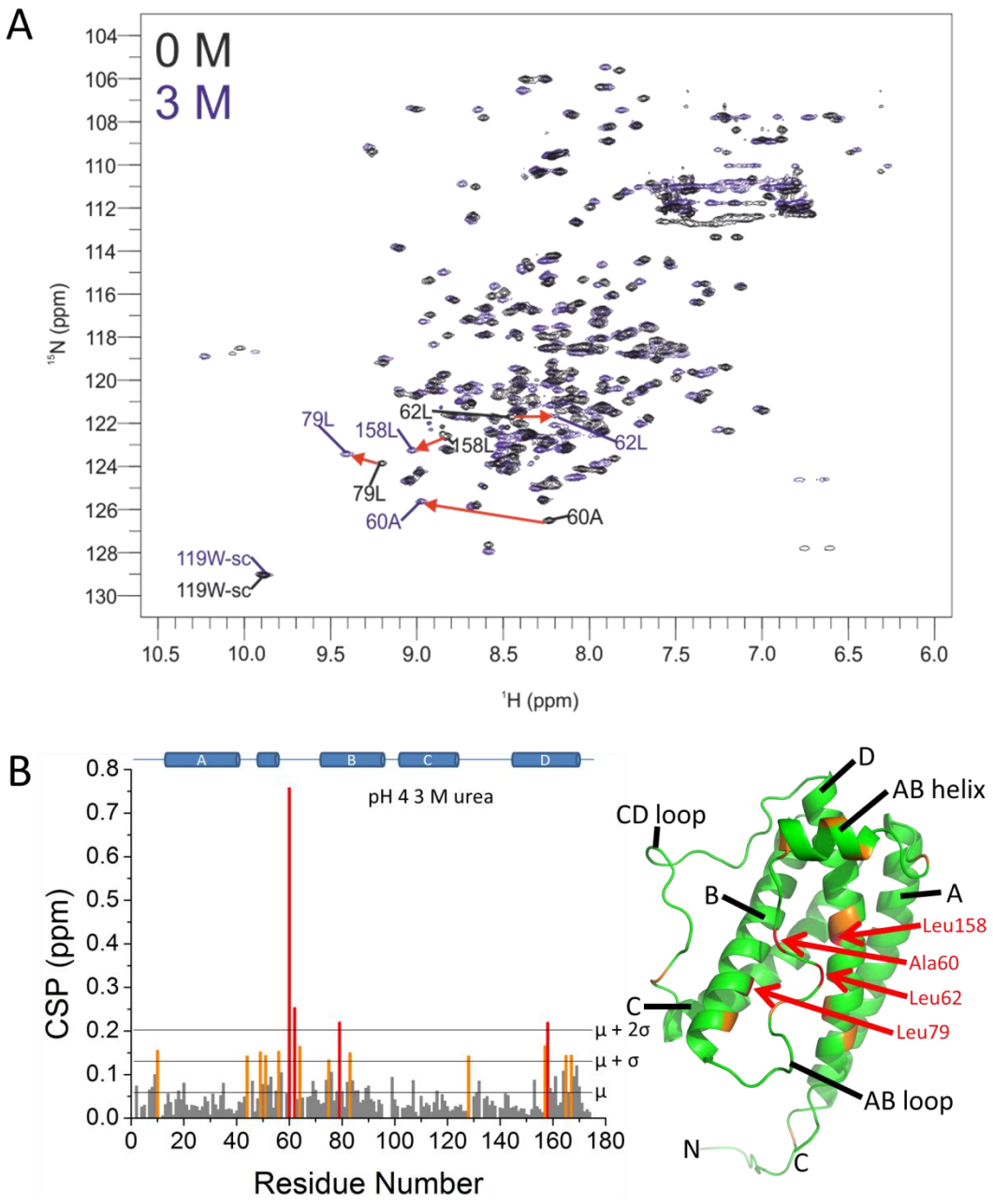


Figure 5.4 CSPs for G-CSF C3 at pH 4 resulting from addition of 3 M urea. The ^{1}H-^{15}N HSQC spectra of 473 μM G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 10 % (*v/v*) D_2O pH 4 buffer (I = 39 mM) in the absence of urea (black) and 173 μM G-CSF C3 in the presence of 3 M urea (purple) are shown in A, which were used for calculation of the CSPs shown in B. Peaks corresponding to the side chain (sc) of Trp119 are labelled. CSPs were calculated using backbone amide proton and nitrogen chemical shifts of G-CSF C3 in the presence and absence of 3 M urea (B). Red arrows in A show the changes in peak position of residues that exhibit significantly large CSPs upon addition of 3 M urea, which are coloured in red in B. Horizontal lines show the mean, mean + σ and mean + 2σ CSP values. Orange and red bars correspond to residues exhibiting CSPs greater than the mean + σ and mean + 2σ, respectively. A secondary structure schematic is shown above for reference. Right: Residues with significant CSPs are shown mapped onto the structure of G-CSF wt (PDB 1GNC[283]) using the same colour scheme, constructed using PyMOL[289]. Red residues are also labelled. Spectra were acquired at 25 °C using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

CSPs for residues Ala60 and Leu158 are among the largest observed for G-CSF C3 at both pH 4 and 7 in the presence of urea (coloured red in Figure 5.1B-Figure 5.4B). Additionally, Leu79 exhibits a large CSP at pH 4 in the presence of 0.5 and 3 M urea (Figure 5.3B and Figure 5.4B, respectively), while Leu62 also exhibits a large CSP at pH 4 in the presence of 3 M urea (Figure 5.4B). Leu62 of G-CSF C3 could not be assigned at pH 4 in the presence of 0.5 M urea, making it unclear if this residue experiences a significant change in its chemical environment upon addition of 0.5 M as well as 3 M urea. Leu79 also exhibits a significant CSP at pH 7 in the presence of 0.5 and 4 M urea (coloured orange in Figure 5.1B and Figure 5.2B, respectively). Residues Leu83, His157 and Arg167 also exhibit significant CSPs at pH 4 and 7 in the presence of urea.

Overall, the vast majority of significant and large CSPs observed for G-CSF C3 at pH 4 and 7 upon addition of urea are observed for residues in the AB loop or residues in helices B and D that are proximal spatially to regions of the latter half of the AB loop (Figure 5.1B-Figure 5.4B), suggesting that a conformational change in this region results from addition of urea at sub-denaturing concentrations. Surprisingly, similar residues of G-CSF C3 at pH 4 and 7 exhibit significantly large CSPs in the presence of 0.5 M (Figure 5.3B and Figure 5.1B, respectively) and in urea concentrations that promote aggregation (Figure 5.4B and Figure 5.3B, respectively). This suggests that the conformational change in the AB loop occurs at lower urea concentrations than those required to promote aggregation of G-CSF C3, which agrees with the results of near-UV CD (Figure 4.10) and fluorescence emission spectra (Figure 4.9) which also suggest a conformational change in urea concentrations lower than those required to promote G-CSF C3 aggregation.

The program CMView[401] was used to determine whether residues exhibiting large CSPs are clustered around a particular Tyr side chain in the structure of G-CSF wt (PDB 1GNC[283]), modelled in PyMOL[289]. CMView[401] predicts a contact between two residues if there are atoms in these residues that are within 8 Å[401]. All of the residues exhibiting large CSPs (coloured red in Figure 5.1B-Figure 5.4B) were predicted to have contacts with Tyr86 (Figure 5.5), while none of these residues were predicted to have contacts with Tyr40 or Tyr166. This strongly suggests that a change in the environment of Tyr86 is occurring in low urea concentrations, which is responsible for the changes in near-UV CD (Figure 4.10) and fluorescence emission (Figure 4.9) spectra of G-CSF C3 in sub-denaturing urea concentrations and could be linked to aggregation. Moreover, many of the significant and large CSPs are

located on the latter half of the AB loop, supporting the notion that a conformational change in the AB loop is responsible for the change in the environment of the side chain of Tyr86.

Interestingly, a conformational change in the AB loop has been suggested previously to be linked to aggregation of wild-type G-CSF in low GdnHCl concentrations at pH 7, reported by increased fluorescence emission intensity of Trp59[142]. This suggests that G-CSF C3 aggregation could be linked to the same conformational change that has been hypothesised to promote aggregation of wild-type G-CSF.

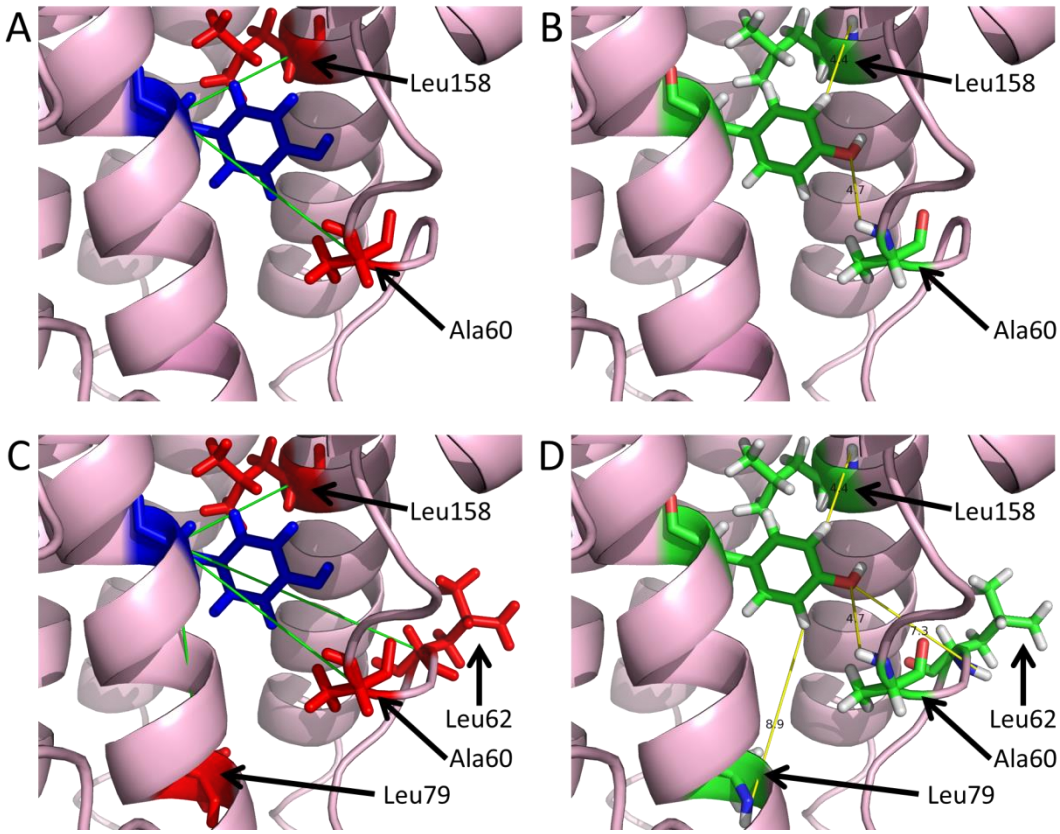


Figure 5.5 Proximity of the side chain of Tyr86 in G-CSF C3 to residues that exhibit large CSPs in the presence of urea concentrations that promote aggregation. Contacts predicted between Tyr86 (blue) and residues that exhibit large CSPs (red) by the program CMView[401] are represented by green lines drawn between the C_α atoms of residues, although contacts are not necessarily predicted between these particular atoms. Predicted contacts are shown for residues of G-CSF C3 that exhibit large CSPs at pH 7 in the presence of 4 M urea i.e. Ala60 and Leu158 (A) and at pH 4 in the presence of 3 M urea i.e. Ala60, Leu62, Leu79 and Leu158 (C). Distances between the amide protons of residues exhibiting large CSPs at pH 7 in the presence of 4 M urea (B) and at pH 4 in the presence of 3 M urea (D) and the nearest atom of the side chain of Tyr86 are shown in Å connected by yellow dotted lines. This figure was constructed using the structure of G-CSF wt (PDB 1GNC[283]) in PyMOL[289].

5.2.2 Probing for changes in the dynamics of G-CSF C3 under aggregation-prone conditions using amide nitrogen transverse relaxation rates (R_2)

The effect of sub-denaturing urea concentrations on the dynamics of G-CSF C3 was determined next by measurement of amide nitrogen R_2 rates. Transverse relaxation refers to decoherence of magnetisation in the xy plane[324]. The low gyromagnetic ratio of ^{15}N nuclei coupled to the high gyromagnetic ratio of protons[319] means that relaxation of backbone ^{15}N nuclei is predominantly due to dipole-dipole interactions between amide proton and nitrogen nuclei[319,340]. Therefore, measurement of backbone ^{15}N nuclei longitudinal and transverse relaxation rates (denoted R_1 and R_2, respectively) can be used to investigate local polypeptide fluctuations on the ps-ns timescale. However, R_2 rates are also affected by the rotational correlation time (τ_c) and chemical exchange resulting from larger amplitude protein motions on the µs-ms timescale e.g. conformational exchange[341]. A more comprehensive description of the theory behind transverse relaxation is given in Section 2.9.8.

For measurement of R_2 rates in pH 7 buffer 130, 135 and 134 µM G-CSF C3 were used in the presence of 0, 0.5 and 4 M urea, respectively. For R_2 rate measurements in pH 4 buffer 162, 151 and 163 µM G-CSF C3 were used in the presence of 0, 0.5 and 3 M urea, respectively. R_2 rates were measured in the presence of 0.5 M urea, as well as urea concentrations which promote aggregation, in order to probe for changes in the dynamics of G-CSF C3 which only occur in higher urea concentrations and so are more likely to relate to aggregation.

R_2 rates are measured using a variation of the HSQC pulse sequence (Figure 2.1), which contains a CPMG pulse train[342,343] between the first INEPT block and the t_1 evolution period in the indirect dimension. By varying the number of repetitions of the CPMG block time delays of different lengths are created and ^{1}H-^{15}N HSQC spectra with different amide proton peak intensities are acquired, which are used to calculate R_2 rates for each amide nitrogen[341]. Time delays of 16.1, 32.2, 48.4, 64.5, 80.6, 96.7, 112.8 and 129.0 ms were used for measurement of G-CSF C3 R_2 rates. Amide proton peak intensities at each of these relaxation delays were fitted to a single exponential decay function, with errors being calculated from the difference of duplicate peak intensities from the mean intensity acquired using relaxation delays of 32.2 and 96.7 ms. Monte Carlo error analysis was subsequently performed to calculate the error in the rates obtained from fitting of peak intensities (see

Section 2.9.8). Plots of peak intensity at each relaxation delay outlined and fitting of these peak intensities to a single exponential decay function are shown in Appendices 7.9 and 7.10 respectively for G-CSF C3 at pH 4 and 7 in the absence of urea. Some examples of these plots are shown in Figure 5.8A-E for G-CSF C3 at pH 7 in the absence or presence of 4 M urea.

After measurement of backbone ^{15}N R_2 rates for G-CSF C3 at pH 7 in the presence of 0, 0.5 and 4 M urea (Figure 5.6A and Figure 5.7A, respectively) and at pH 4 in the presence of 0, 0.5 and 3 M urea (Figure 5.9A and Figure 5.10A, respectively) ratios of R_2 rates measured without urea to R_2 rates measured with urea were calculated (Figure 5.6B-Figure 5.10B). Residues with R_2 ratios significantly greater than the mean + σ ratio are coloured red in Figure 5.6B-Figure 5.10B, while residues with R_2 ratios significantly lower than the mean – σ ratio are coloured blue. Residues are only highlighted red if their R_2 ratio minus the error in this calculated ratio is greater than the mean + σ ratio and are only highlighted blue if their R_2 ratio plus the error in this calculated ratio is lower than the mean – σ ratio.

Generally, R_2 rates are lower for residues of G-CSF C3 in unstructured regions, such as the N- and C-termini and the AB and CD loops (Figure 5.6A-Figure 5.10A), suggesting that these regions of G-CSF C3 are more dynamic on the ps-ns timescale. Conversely, residues in helical sections of G-CSF C3 have higher R_2 rates, suggesting that these regions are more rigid. Through measurement of NMR relaxation rates, the N- and C- termini of wild-type G-CSF as well as the CD loop have been shown previously to be more dynamic than the rest of G-CSF on the sub-nanosecond timescale[283,288].

Despite the fact the AB loop of G-CSF C3 is longer than the CD loop (residues 42-71 in the AB loop compared with residues 125-144 in the CD loop), the R_2 rates of residues in the AB loop are higher than those in the CD loop (Figure 5.6A-Figure 5.10A), suggesting that the CD loop is more dynamic than the AB loop. This has been observed previously for the AB and CD loops of G-CSF wt and is most likely due to the short helical section in the AB loop plus the two disulphide bonds linking residues Cys37-Cys43 and Cys65-Cys75, which restrain the AB loop by linking residues at the N- and C-termini of this loop to helices A and B, respectively (Figure 1.11)[288].

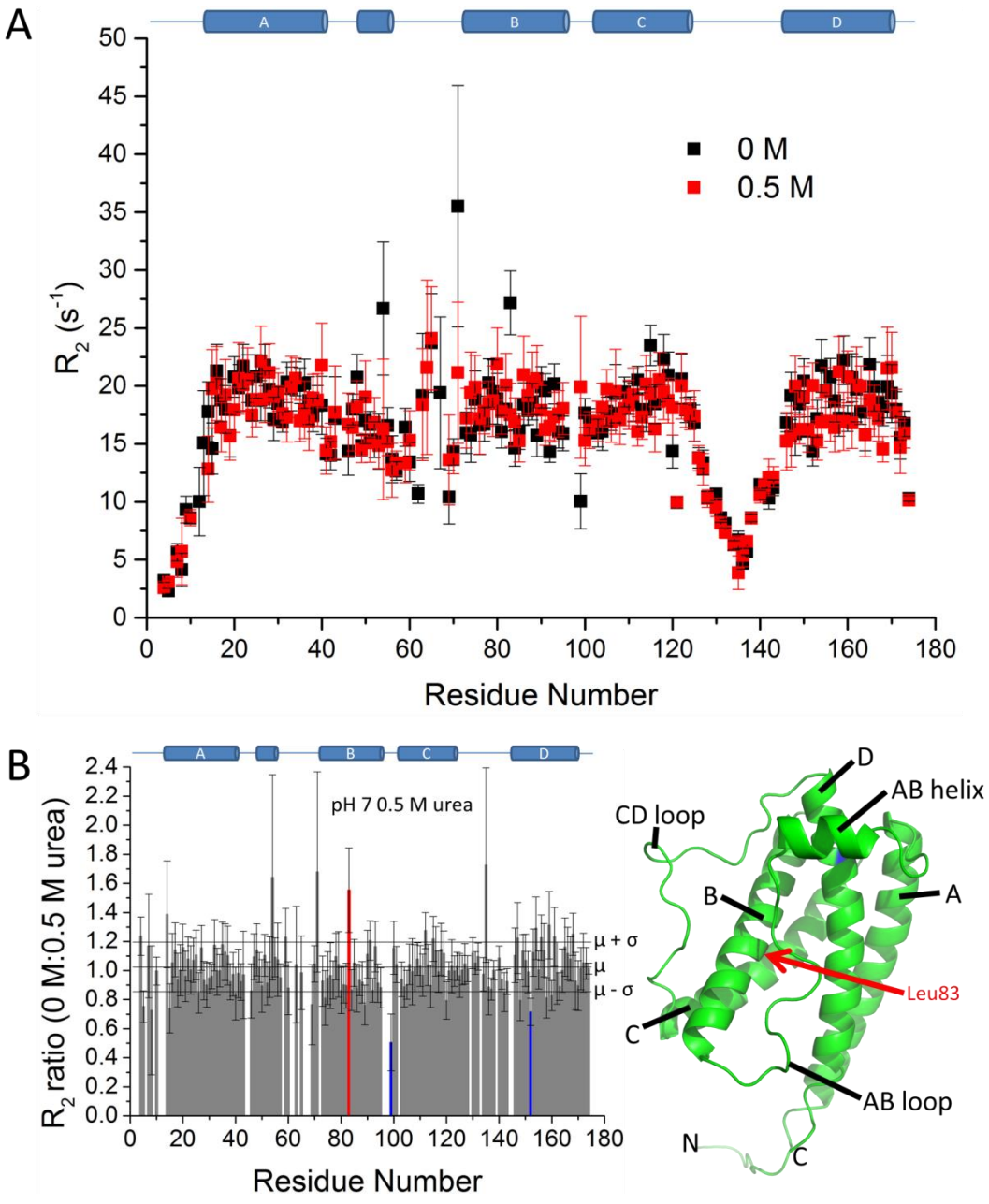


Figure 5.6 Backbone ^{15}N R_2 rates for G-CSF C3 at pH 7 in the absence or presence of 0.5 M urea. R_2 rates measured using 130 μM G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 0.018 % (*w/v*) sodium azide, 10 % (*v/v*) D_2O pH 7 buffer (I = 72 mM) in the absence of urea (black) and 135 μM G-CSF C3 in the presence of 0.5 M urea (red) are shown in A with a secondary structure schematic above for reference. The ratios of R_2 rates measured in 0 M urea to rates measured in 0.5 M urea are shown in B with a secondary structure schematic above for reference. Horizontal lines show the mean, mean + σ and mean - σ R_2 ratios. The red bar corresponds to a residue exhibiting an R_2 ratio significantly greater than the mean + σ value and blue bars correspond to residues exhibiting R_2 ratios significantly lower than the mean - σ value. Right: Residues with significant R_2 ratios are shown mapped onto the structure of G-CSF wt (PDB 1GNC[283]) using the same colour scheme, constructed using PyMOL[289]. The red residue is also labelled. R_2 rates were measured at 25 ^{o}C using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

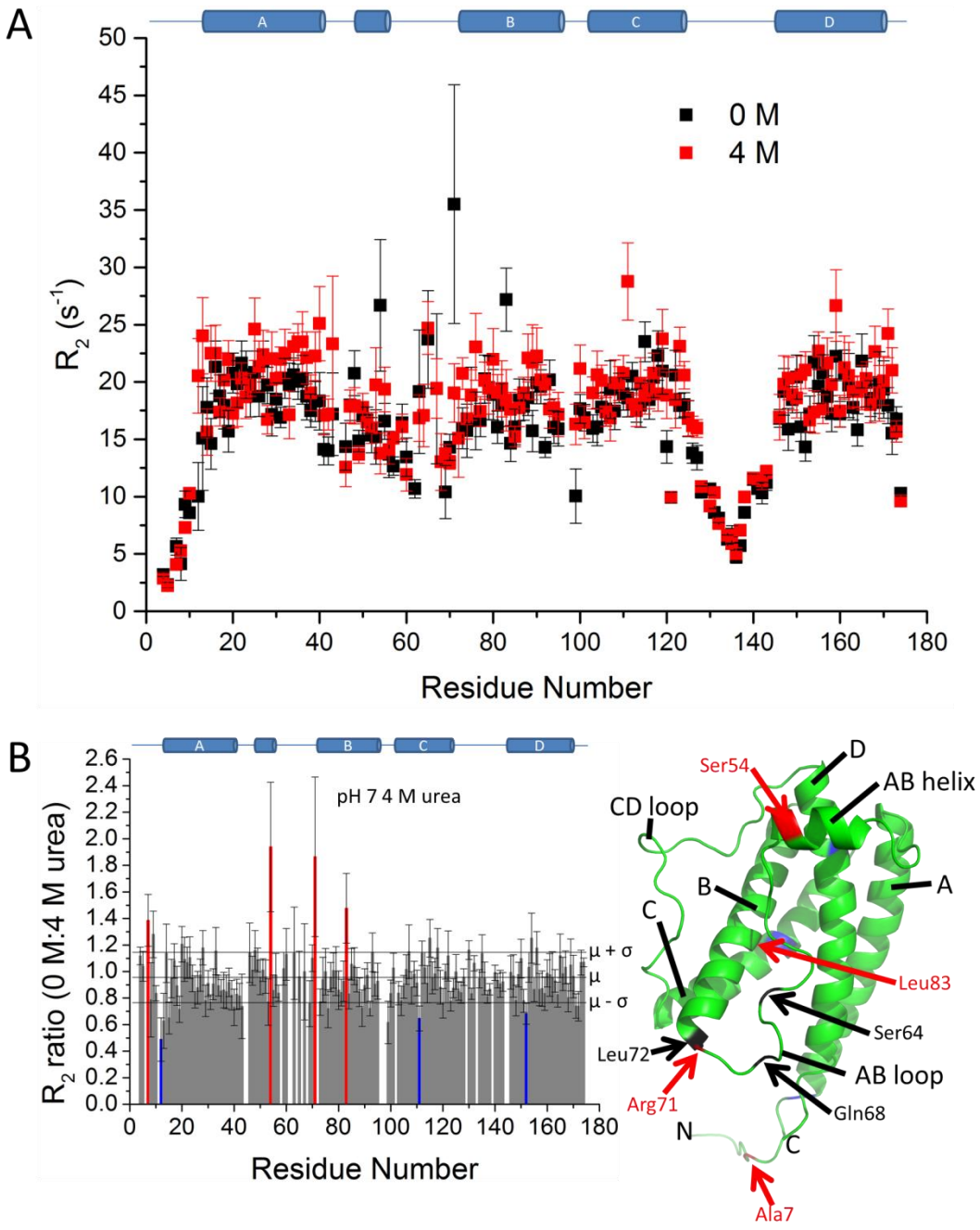


Figure 5.7 Backbone ^{15}N R_2 rates for G-CSF C3 at pH 7 in the absence or presence of 4 M urea. R_2 rates measured using 130 μM G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 0.018 % (*w/v*) sodium azide, 10 % (*v/v*) D_2O pH 7 buffer (I = 72 mM) in the absence of urea (black) and 134 μM G-CSF C3 in the presence of 4 M urea (red) are shown in A with a secondary structure schematic above for reference. The ratios of R_2 rates measured in 0 M urea to rates measured in 4 M urea are shown in B with a secondary structure schematic above for reference. Horizontal lines show the mean, mean + σ and mean - σ R_2 ratios. Red bars correspond to residues exhibiting R_2 ratios significantly greater than the mean + σ value and blue bars correspond to residues exhibiting R_2 ratios significantly lower than the mean - σ value. Right: Residues with significant R_2 ratios are shown mapped onto the structure of G-CSF wt (PDB 1GNC[283]) using the same colour scheme, constructed using PyMOL[289]. Red residues are also labelled. R_2 rates for residues coloured in black and labelled can only be measured reliably in the presence of 4 M urea. R_2 rates were measured at 25 °C using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

Ratios of R_2 rates measured at pH 7 in the absence of urea to rates measured in the presence of 0.5 M urea suggest that only one residue becomes significantly more dynamic in the presence of urea i.e. Leu83, which is located in helix B (Figure 5.6B). Conversely, ratios of R_2 rates at pH 7 calculated in the presence of 4 M urea suggest that three residues within the AB loop become significantly more dynamic, these being Ser54, Arg71 and Leu83. Ser54 and Arg71 are located on the helical section of the AB loop and in the region between this helical section and helix B, respectively (Figure 5.7B).

There are also some residues for which R_2 rates can only be measured reliably in the presence of 4 M urea as the peaks in ^{1}H-^{15}N HSQC spectra for these residues in the absence of urea are too broad (Figure 5.8). This suggests that upon addition of 4 M urea these residues become more dynamic, giving rise to narrower peaks in ^{1}H-^{15}N HSQC spectra and enabling measurement of accurate R_2 rates, in accordance with the relationship between the width of peaks at half height in NMR spectra ($\nu_{1/2}$) and the R_2 rate (Equation 5.1)[324]. These residues are Ser64, Gln68 and Leu72 and are found in the latter half of the AB loop (coloured black in Figure 5.7B), along with Ser54 and Arg71 (coloured red in Figure 5.7B), suggesting that increased dynamics of the residues in this region are observed upon addition of 4 M urea, but not in the presence of 0.5 M urea. Therefore, increased mobility of the latter half of the AB loop region of G-CSF C3 at pH 7 resulting from addition of 4 M urea could be linked to increased aggregation observed under the same conditions.

$$\nu_{1/2} = \frac{R_2}{\pi} \qquad \textbf{Equation 5.1}$$

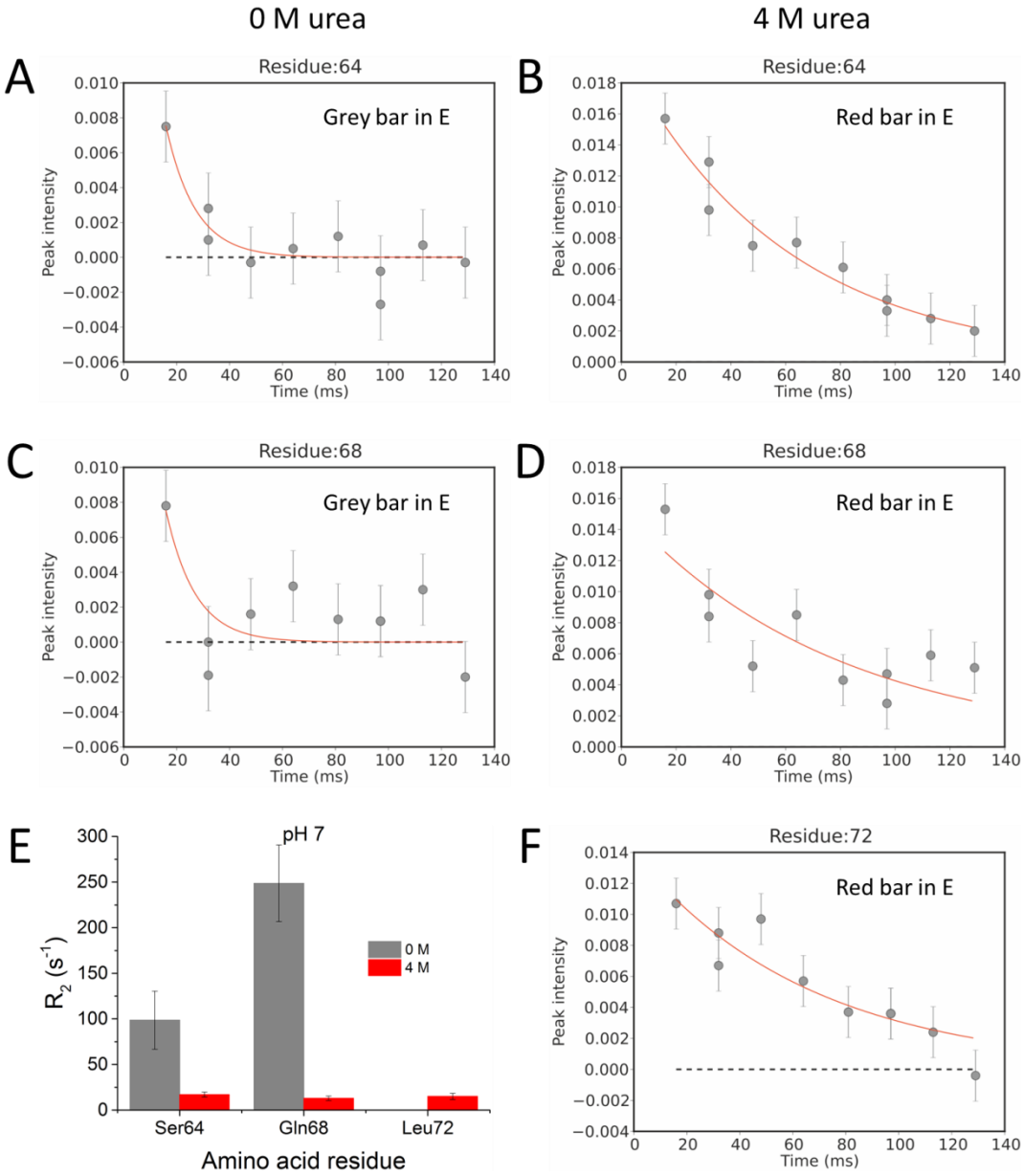


Figure 5.8 Backbone ^{15}N R_2 rates for residues of G-CSF C3 at pH 7 that exhibit broad ^{1}H-^{15}N HSQC peaks in the absence of urea. Plots of peak intensity over time used for calculation of R_2 rates for G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 0.018 % (*w/v*) sodium azide, 10 % (*v/v*) D_2O pH 7 buffer (I = 72 mM) in the absence or presence of 4 M urea are shown for residues Ser64 (A and B, respectively) and Gln68 (C and D, respectively). In the absence of urea, the peak for Leu72 is too broad to detect in the ^{1}H-^{15}N HSQC spectrum, such that an R_2 rate could not be measured and so a plot of peak intensity over time is only shown for this residue in the presence of 4 M urea (F). Peak intensity is shown in arbitrary units and a horizontal dotted line is drawn at zero peak intensity. R_2 rates were measured using 130 µM G-CSF C3 at pH 7 in the absence of urea and 134 µM G-CSF C3 in the presence of 4 M urea. R_2 rates calculated from these plots are shown in E for residues of G-CSF C3 at pH 7 measured in the absence or presence of 4 M urea (grey and red, respectively).

Ratios of R_2 rates measured at pH 4 in the absence of urea to rates measured in the presence of 0.5 M urea suggest that two residues within the latter half of the AB loop become significantly more dynamic, these being Ala69 and Leu70, as well as Leu72 which is the N-terminal residue of helix B (Figure 5.9B). Interestingly, Tyr86 on helix B also appears to be significantly more dynamic in the presence of 0.5 M urea, although this effect is not observed in 3 M urea (Figure 5.10B).

In the presence of 3 M urea, residues Gly56, Arg59, Leu70 and Arg71 within the latter half of the AB loop of G-CSF C3 at pH 4 are significantly more dynamic. This could suggest that in the presence of 0.5 M urea the last few residues of the C-terminal end of the AB loop become more dynamic, whereas upon addition of 3 M urea the entire latter half of the AB loop, encompassing the C-terminus of the short helical section to the C-terminal end of the loop becomes more dynamic. The larger amplitude of the protein motion suggested in 3 M compared to 0.5 M urea may account for the increased aggregation of G-CSF C3 observed at pH 4 in the presence of 3 M but not 0.5 M urea.

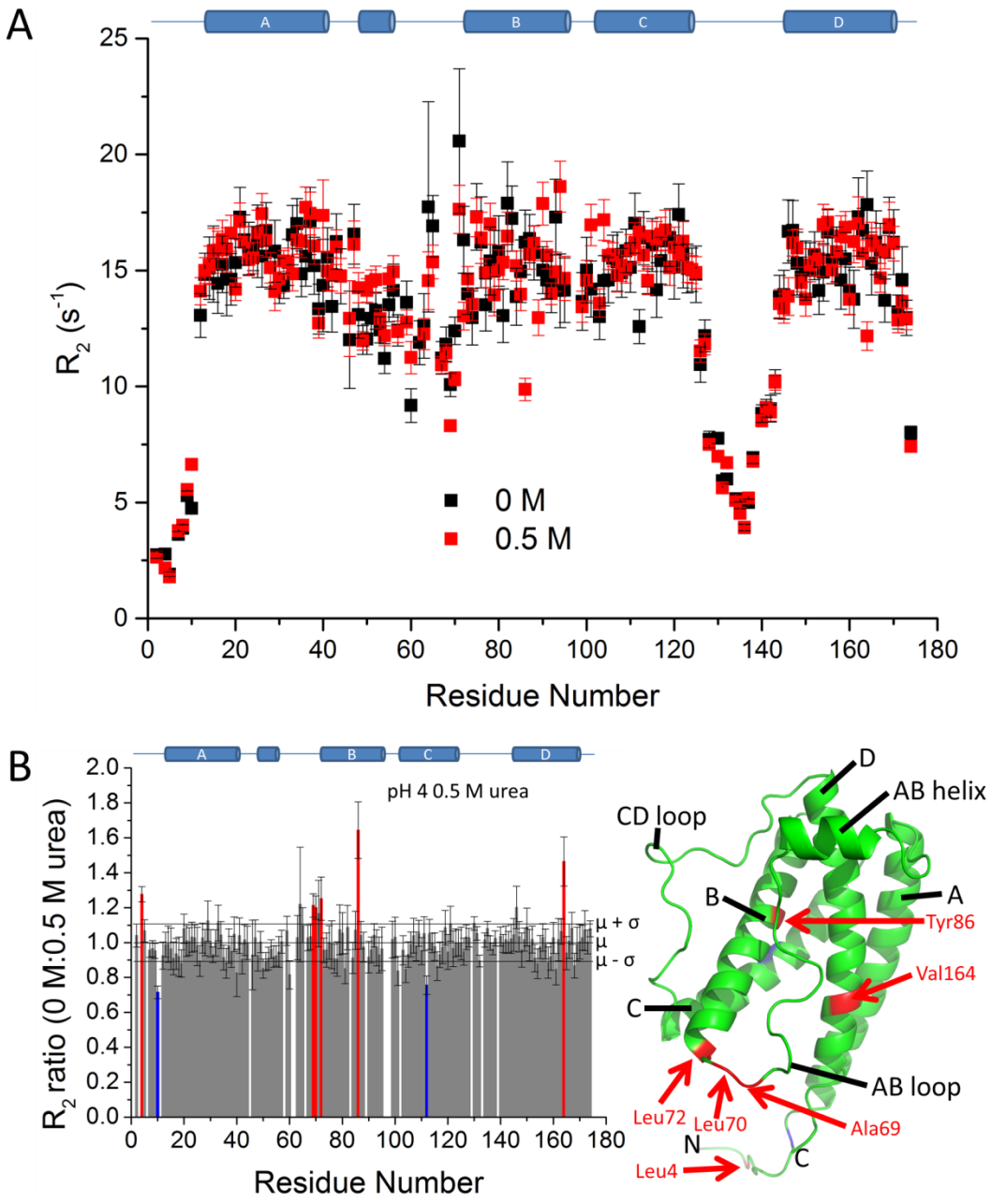


Figure 5.9 Backbone ^{15}N R_2 rates for G-CSF C3 at pH 4 in the absence or presence of 0.5 M urea. R_2 rates measured using 162 μM G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 10 % (*v/v*) D_2O pH 4 buffer (I = 39 mM) in the absence of urea (black) and 151 μM G-CSF C3 in the presence of 0.5 M urea (red) are shown in A with a secondary structure schematic above for reference. The ratios of R_2 rates measured in 0 M urea to rates measured in 0.5 M urea are shown in B with a secondary structure schematic above for reference. Horizontal lines show the mean, mean + σ and mean – σ R_2 ratios. Red bars correspond to residues exhibiting R_2 ratios significantly greater than the mean + σ value and blue bars correspond to residues exhibiting R_2 ratios significantly lower than the mean – σ value. Right: Residues with significant R_2 ratios are shown mapped onto the structure of G-CSF wt (PDB 1GNC[283]) using the same colour scheme, constructed using PyMOL[289]. Red residues are also labelled. R_2 rates were measured at 25 ^{o}C using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

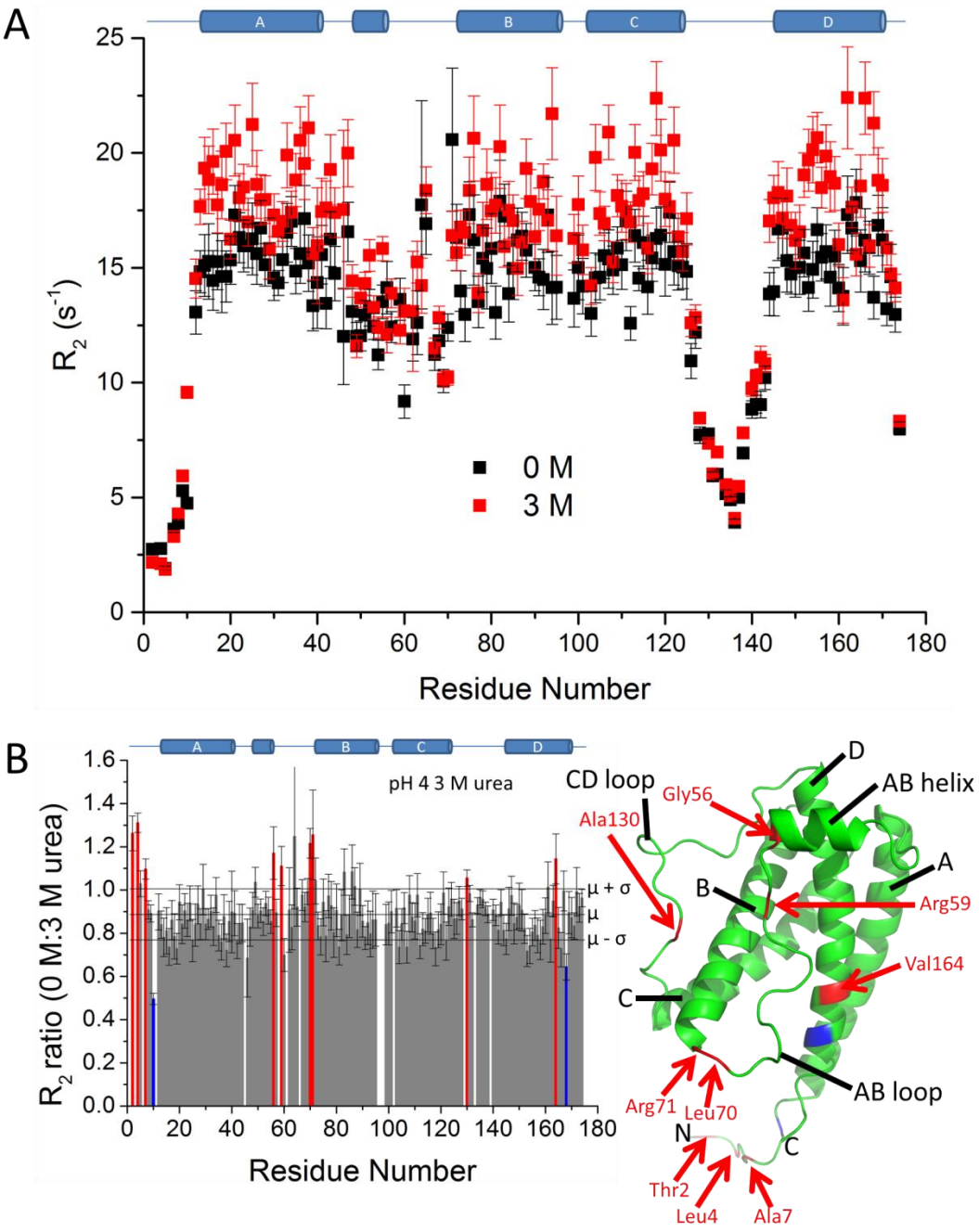


Figure 5.10 Backbone ^{15}N R_2 rates for G-CSF C3 at pH 4 in the absence or presence of 3 M urea. R_2 rates measured using 162 μM G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 10 % (*v/v*) D_2O pH 4 buffer (I = 39 mM) in the absence of urea (black) and 163 μM G-CSF C3 in the presence of 3 M urea (red) are shown in A with a secondary structure schematic above for reference. The ratios of R_2 rates measured in 0 M urea to rates measured in 3 M urea are shown in B with a secondary structure schematic above for reference. Horizontal lines show the mean, mean + σ and mean - σ R_2 ratios. Red bars correspond to residues exhibiting R_2 ratios significantly greater than the mean + σ value and blue bars correspond to residues exhibiting R_2 ratios significantly lower than the mean – σ value. Right: Residues with significant R_2 ratios are shown mapped onto the structure of G-CSF wt (PDB 1GNC[283]) using the same colour scheme, constructed using PyMOL[289]. Red residues are also labelled. R_2 rates were measured at 25 °C using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

5.2.3 Probing for changes in the solvent exposure of G-CSF C3 under aggregation-prone conditions using hydrogen/deuterium exchange

Native state hydrogen/deuterium exchange of backbone amide protons on G-CSF C3 was used to measure dynamics over longer timescales than accessible by R_2 rates, as well as probe for changes in solvent exposure in the presence and absence of urea concentrations that promote aggregation. Native state hydrogen exchange monitored by NMR has been used to identify partially unfolded forms (PUFs) of proteins that are sparsely populated[402,403] and often undetectable in kinetic and thermodynamic studies[404-408]. Native state hydrogen exchange has also been used to identify PUFs involved in aggregation[89] as well as more general increases in dynamics and solvent exposure that promote aggregation[92,409,410]. A more comprehensive description of the theory behind hydrogen/deuterium exchange is given in Section 2.9.7.

Considering that the aggregation of wild-type G-CSF at pH 7 has been proposed to occur from a native-like species[142,185,303,307,308], which has also been hypothesised to correspond to a late stage folding intermediate observed during kinetic folding studies of G-CSF wt at pH 4[142,304], native state hydrogen exchange is an appropriate technique to probe for subtle conformational changes in the native state of G-CSF C3 that give rise to aggregation.

Native state hydrogen exchange monitored by mass spectrometry has been carried out for G-CSF wt previously at pH 7 at temperatures which favour and disfavour aggregation i.e. 37 and 4 ^{o}C, respectively, revealing a general increase in solvent exposure of helices at 37 ^{o}C and increased protection of helices at 4 ^{o}C[307]. However, increased intrinsic hydrogen exchange rates[335], coupled with the increased thermal motion of proteins at higher temperatures, make it unclear to what extent the general increase in solvent exposure of helices at higher temperature relates to the minimal conformational change needed to promote G-CSF aggregation.

Native state hydrogen exchange monitored by mass spectrometry[309] and infrared (IR) spectroscopy[185,308] have also been carried out for G-CSF wt in the presence of sucrose and benzyl alcohol, which have been shown to discourage[185,307,308] and promote[185] wild-type G-CSF aggregation, respectively. These experiments showed a general increase in solvent exposure of helices in the presence of benzyl

alcohol[185,309] and increased protection from solvent in the presence of sucrose[185,307-309]. Sucrose is a commonly used excipient in protein drug formulations[178] which causes compaction of the native conformation and destabilises partially unfolded states of proteins through preferential exclusion of water from the protein surface[198]. Conversely, benzyl alcohol is an antimicrobial preservative[99] which has been shown to cause aggregation and interact with partially unfolded species of proteins[186-188]. Therefore, it is perhaps unsurprising that increased protection of helices from hydrogen exchange is observed for G-CSF in the presence of sucrose[185,307-309], which causes compaction of the native conformation, while increased solvent exposure is observed in the presence of benzyl alcohol[185,309], most likely due to partial unfolding[186,187]. Consequently, these global increases and decreases in solvent exposure of the native conformation may mask the minimal conformational fluctuation required for G-CSF aggregation.

For initiation of hydrogen/deuterium exchange reactions, concentrated stocks of G-CSF C3 in pH 7 and 4 buffers were diluted ten-fold with deuterated pD 7.8 and pD 4.8 buffers, respectively, to a final concentration of 52 µM G-CSF C3. pD 7.8 buffer with or without 4.44 M urea was used for initiating hydrogen exchange in the presence and absence of a final concentration of 4 M urea, while pD 4.8 buffer with or without 3.33 M urea was used for hydrogen exchange reactions in the presence and absence of 3 M urea. pD 4.8 and 7.8 buffers exhibited pH meter readings of 4.4 and 7.4, respectively[338].

After dilution, hydrogen exchange of amide protons was monitored by recording ^{1}H-^{15}N band-selective optimised-flip-angle short-transient (SOFAST) HMQC spectra[339] every 5 mins over a 15 h period for pD 7.8 reactions and every 20 mins over a 24 h period for pD 4.8 reactions. Spectra with different acquisition times were taken to account for the increased intrinsic rate of hydrogen exchange at neutral pH[335]. ^{1}H-^{15}N SOFAST HMQC spectra reduce the acquisition time from hours to minutes through band-selective excitation of amide protons, which results in faster longitudinal relaxation due to dipolar interactions with non-perturbed protons[339]. The longitudinal relaxation rate (R_1) determines the time before nuclei can be excited again by a radiofrequency pulse[324], meaning more scans can be carried out over a shorter time period by acquiring ^{1}H-^{15}N SOFAST HMQC spectra[339] compared with ^{1}H-^{15}N HSQC spectra.

Changes in the amide proton peak intensities of residues over time were fitted to a single exponential decay function with a y-axis offset. However, many peaks in the ^{1}H-^{15}N SOFAST HMQC spectrum of G-CSF C3 after dilution into deuterated pD 7.8 buffer were not detectable at the beginning of acquisition. This is the result of exchange of amide protons with deuterium occurring in the dead-time between initiation of the hydrogen exchange reaction and beginning acquisition, which was 7-8 mins.

Of the residues for which rate constants of hydrogen exchange could be reliably measured for G-CSF C3 amide protons at pD 7.8, protection factors (PFs) were calculated (see Section 2.9.7). Plots and fits used for calculation of the PFs shown in Figure 5.11A are shown in Appendix 7.12 for residues of G-CSF C3 in the absence of urea. Protection factors are the ratio of the intrinsic hydrogen exchange rates, calculated using CIntX[337], to the observed rates. The ratios of ln(PF) in 0 M urea to ln(PF) in 4 M urea were calculated to probe for regions of G-CSF C3 with significantly altered dynamics or solvent exposure in the presence of 4 M urea.

Overall, ln(PF) for residues of G-CSF C3 at pD 7.8 are lower in the presence of 4 M urea (Figure 5.11A). However, there are no residues that appear to be significantly less protected than other residues in the presence of 4 M urea (Figure 5.11B), suggesting a general increase in solvent exposure of these residues in helices of G-CSF C3. This would suggest that the effect of 4 M urea on G-CSF C3 at pD 7.8 is to increase the general exposure of helical residues to solvent, which is comparable to the reported effect of benzyl alcohol on the solvent exposure of wild-type G-CSF[185,309]. Consequently, no specific regions of G-CSF C3 could be detected to undergo partial unfolding or exhibit increased solvent exposure at pD 7.8 upon addition of 4 M urea.

Additionally, ln(PF) calculated for residues in the absence of urea are close to the ln(PF) representing global unfolding (shown as a green line in Figure 5.11A), suggesting that solvent exchange of the residues detected at pD 7.8 is reporting on global unfolding rather than partial unfolding events that are potentially linked to aggregation.

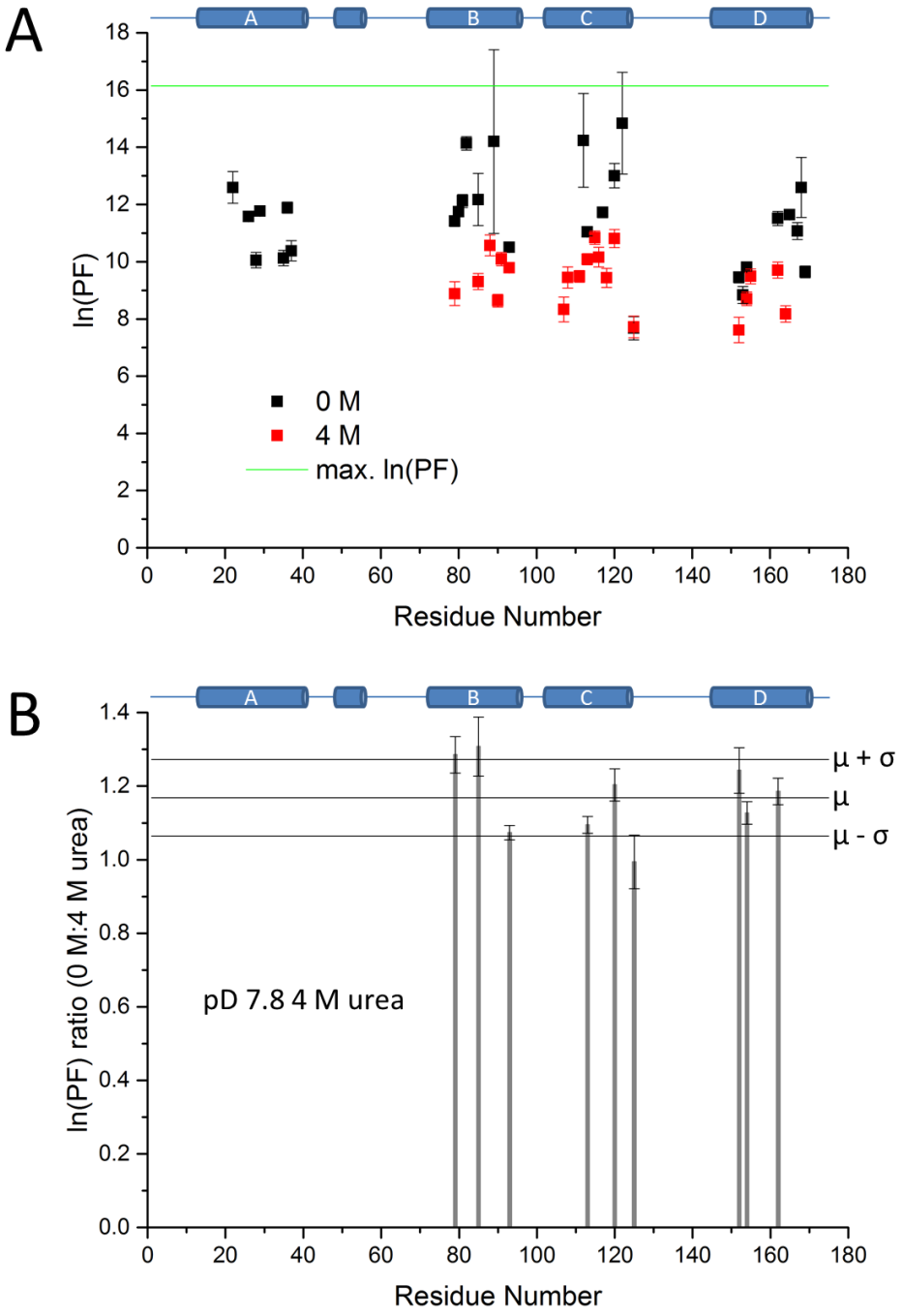


Figure 5.11 PFs for G-CSF C3 calculated from hydrogen/deuterium exchange rates at pD 7.8 in the absence or presence of 4 M urea. The ln(PF)s shown in A are calculated as the ratio of intrinsic rates (calculated using CIntX[337]) to observed rates for 52 μM G-CSF C3 in 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide, 90 % (*v/v*) D_2O pD 7.8 buffer (I = 89 mM) in the absence or presence of 4 M urea (black and red, respectively) and a secondary structure schematic is shown above for reference. The theoretical maximum ln(PF) was calculated using a ΔG^{o}_{UN} value of 40 kJ mol^{-1}, measured by equilibrium denaturation analysis of G-CSF C3 at pH 7 (Table 4.1) and is shown as a horizontal green line in A. The ratios of ln(PF) calculated in 0 M urea to ln(PF) calculated in 4 M urea are shown in B with a secondary structure schematic above for reference. Horizontal lines show the mean, mean + σ and mean - σ ln(PF) ratios. Hydrogen/deuterium exchange rates were measured at 25 ^{o}C using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

Hydrogen exchange rates could be calculated for more residues of G-CSF C3 at pD 4.8 than at pD 7.8, owing to the slower exchange rate at acidic pH[335]. However, for residues in largely or entirely unstructured regions of G-CSF C3 i.e. the AB loop plus the N- and C-termini and CD loop, respectively, the majority of amide proton peak intensities were undetectable at the beginning of acquisition, having exchanged in the dead-time of the experiment.

The only residues in these unstructured regions that could be detected in ^{1}H-^{15}N SOFAST HMQC spectra had already undergone exchange of amide protons with deuterium, producing plots of largely unchanged peak intensity over time that could not be fitted to a single exponential decay function with a y-axis offset (Figure 5.12B). Peaks for these residues were visible even after exchange of 90 % of amide protons with deuterium because the peak intensities for the remaining 10 % of unexchanged amide protons were great enough to be observed.

Conversely, many amide protons for residues of G-CSF C3 did not undergo a significant amount of exchange with deuterium at pD 4.8 during the 24 h acquisition period, again resulting in plots of largely unchanged peak intensity over time that could not be fitted to a single exponential decay function with a y-axis offset (Figure 5.12A). Residues that had undergone negligible exchange of amide protons with deuterium were distinguishable from residues that were detectable even after exchange of 90 % of amide protons by their higher relative peak intensities (Figure 5.12A and B, respectively).

PFs were only calculated for residues for which reliable hydrogen exchange rates could be calculated. The reliability of exchange rates was determined from the amplitude of the error relative to the calculated rate as well as visual inspection of the plots of peak intensity over time used for fitting to a single exponential decay function with a y-axis offset. Plots and fits used for calculation of the PFs shown in Figure 5.13A can be found in Appendix 7.11 for residues of G-CSF C3 in the absence of urea, plus two examples of these plots are shown in Figure 5.12C and D.

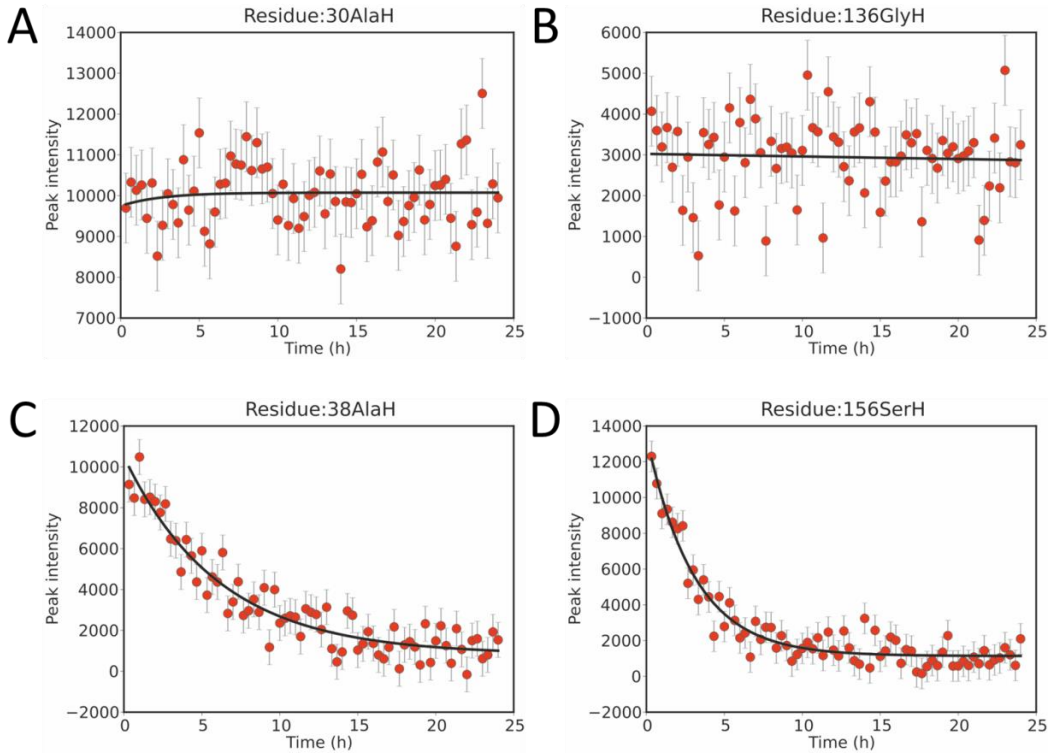


Figure 5.12 Plots of peak intensity over time for residues of G-CSF C3 undergoing hydrogen-deuterium exchange at pD 4.8 in the absence of urea. Examples of hydrogen exchange curves for 52 µM G-CSF C3 are shown for a residue that undergoes negligible hydrogen exchange over 24 h (A), a residue that is 90 % exchanged during the dead-time of the experiment but can still be observed from the peak intensity of the remaining 10 % of amide protons (B), and residues that undergo hydrogen exchange on a timescale that allows calculation of accurate exchange rates and protection factors (C and D). Peak intensity is shown in arbitrary units at 20 minute intervals, as well as fitting of these peak intensities to a single exponential decay function with a y-axis offset. Error bars represent the experimental noise calculated in CCPN Analysis[321]. Hydrogen/deuterium exchange rates were measured at 25 ^{o}C using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

Overall, a significant difference between the ln(PF) at pD 4.8 with or without 3 M urea could only be detected for one residue of G-CSF C3 relative to other residues measured (Figure 5.13B). Considering the number of other residues of G-CSF C3 for which accurate PFs could not be calculated, it is uncertain whether the increased protection of Leu100 from solvent in the presence of 3 M urea is significant. Therefore, no specific regions of G-CSF C3 could be detected to undergo partial unfolding or exhibit increased solvent exposure at pD 4.8 upon addition of 3 M urea, similarly to the results of hydrogen exchange monitored for G-CSF C3 at pD 7.8. However, the ln(PF) calculated for residues at pD 4.8 are significantly lower than the ln(PF) representing global unfolding (shown as a green line in Figure 5.13A), suggesting solvent exchange of the residues detected at pD 4.8 is reporting on local unfolding, in contrast to the solvent exchange observed for residues at pD 7.8.

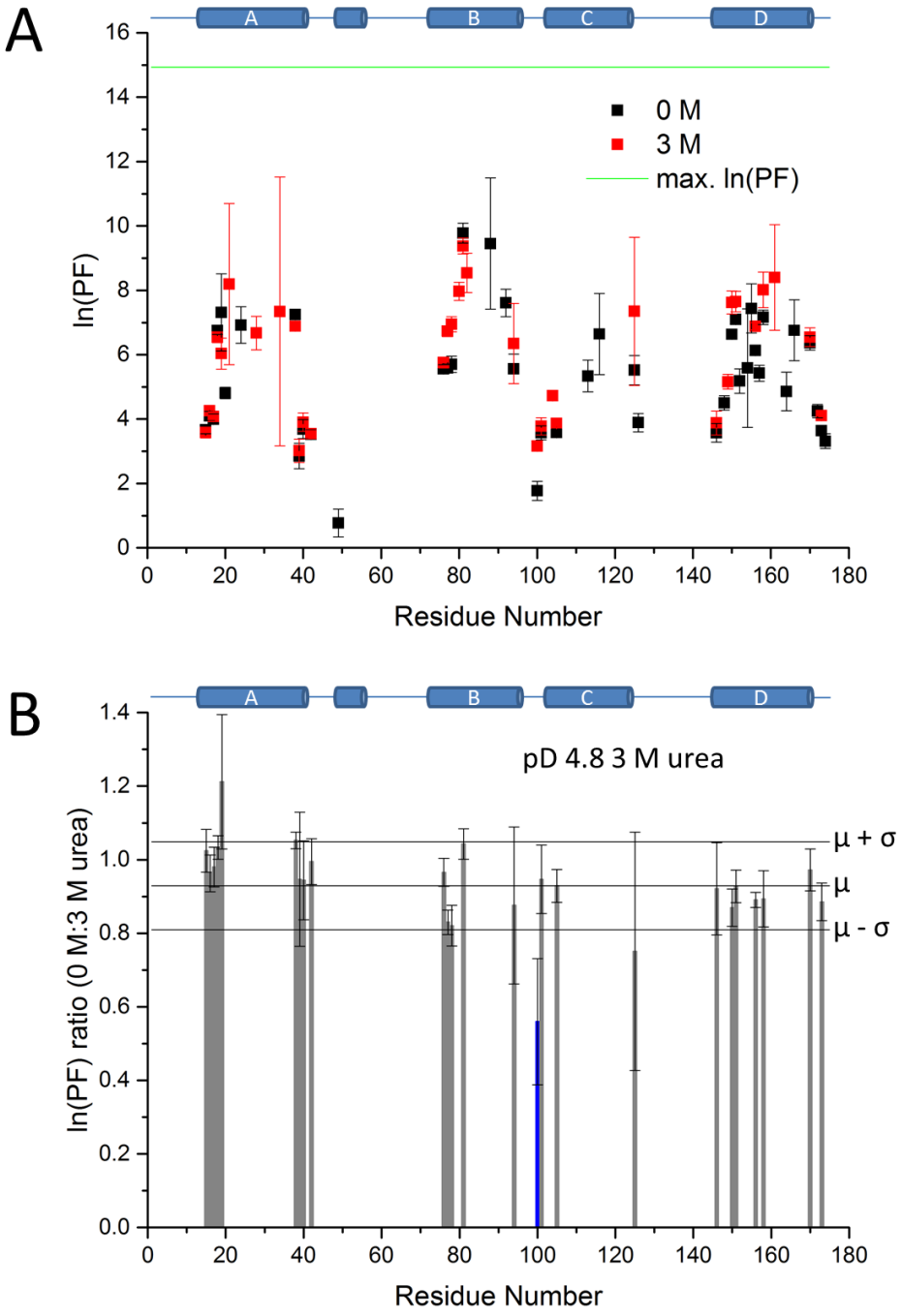


Figure 5.13 PFs for G-CSF C3 calculated from hydrogen/deuterium exchange rates at pD 4.8 in the absence or presence of 3 M urea. The ln(PF)s shown in A are calculated as the ratio of intrinsic rates (calculated using CIntX[337]) to observed rates for 52 μM G-CSF C3 in 20 mM sodium phosphate, 20 mM sodium acetate, 90 % (*v/v*) D_2O pD 4.8 buffer (I = 46 mM) in the absence or presence of 3 M urea (black and red, respectively) and a secondary structure schematic is shown above for reference. The theoretical maximum ln(PF) was calculated using a ΔG^{o}_{UN} value of 37 kJ mol^{-1}, measured by equilibrium denaturation analysis of G-CSF C3 at pH 4 (Table 4.1) and is shown as a horizontal green line in A. The ratios of ln(PF) calculated in 0 M urea to ln(PF) calculated in 3 M urea are shown in B with a secondary structure schematic above for reference. Horizontal lines show the mean, mean + σ and mean - σ ln(PF) ratios. The blue bar corresponds to a residue exhibiting an ln(PF) ratio significantly lower than the mean – σ value. Hydrogen/deuterium exchange rates were measured at 25 ^{o}C using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

5.2.4 Searching for aggregation interfaces of G-CSF C3 using ^{1}H-^{15}N HSQC spectra

In an attempt to determine the protein-protein interaction interfaces involved in G-CSF C3 aggregation, aggregation assays were set up and monomer depletion was monitored by the changes in peak intensities of ^{1}H-^{15}N HSQC spectra versus time. The peak intensities of all G-CSF C3 residues should decrease over time during aggregation due to the formation of soluble aggregates. These larger species will exhibit slower molecular tumbling according to Stokes' Law (Equation 5.2). In Equation 5.2, τ_c is the rotational correlation time i.e. the time taken for the particle to rotate by one radian, η is the viscosity of the sample, R_h is the hydrodynamic radius of the particle, k_B is Boltzmann's constant and T is temperature[319].

$$\tau_c = \frac{4\pi\eta R_h^3}{3k_B T} \qquad \textbf{Equation 5.2}$$

As soluble aggregates are formed and the τ_c of these species increases, the R_2 rate of relaxation will also increase (Figure 2.4). Consequently, the peaks in ^{1}H-^{15}N HSQC spectra will decrease in intensity exponentially due to the relationship between R_2 rate and peak width (Equation 5.1)[324]. Eventually, peaks will be broadened beyond detection, which is referred to as lifetime broadening, causing the change in peak intensity over time to reach a plateau.

The R_2 rate is also contributed to by chemical exchange[341], which will arise as a result of the amide proton of a particular residue being in different chemical environments. Therefore, residues at the aggregation interface will be broadened by chemical exchange as well as lifetime broadening and as a result of slower molecular tumbling, meaning that the peak intensities of these residues should decay faster than residues not involved with the aggregation interface. Calculating the decay rate of peak intensity over the course of an aggregation assay should thus provide an indication of residues at the aggregation interface.

Assignments of backbone amide nitrogen and proton nuclei were transferred from 25 to 37 ^{o}C by thermal titration of G-CSF C3 over these temperatures monitored through ^{1}H-^{15}N HSQC spectra. For aggregation assays monitored by ^{1}H-^{15}N HSQC

spectra, 300 µl samples of G-CSF C3 were set up in a Shigemi tube, using 66 and 67 µM G-CSF C3 in pH 7 buffer in the presence and absence of 4 M urea, respectively and 83 and 85 µM G-CSF C3 in pH 4 buffer in the presence and absence of 3 M urea, respectively (see Section 2.9.4). These samples were incubated at 37 ^{o}C for 24 h and ^{1}H-^{15}N HSQC spectra recorded every 30 mins.

The resulting decays in amide proton peak intensities during aggregation of G-CSF C3 at pH 7 in the absence and presence of 4 M urea were fitted to a single exponential decay function with a y-axis offset to calculate decay rates. Examples of the plots and fits used for calculation of decay rates for residues of G-CSF C3 in the absence or presence of 4 M urea are shown in Figure 5.14A and B, respectively. All of the plots and fits used for calculation of the decay rates shown in Figure 5.14C can be found in Appendix 7.13 for residues of G-CSF C3 in the absence of urea.

The resulting decays in amide proton peak intensities during aggregation of G-CSF C3 at pH 4 in the absence and presence of 3 M urea could not be fitted to a single exponential decay function with a y-axis offset so these data are not shown. However, these experiments are mentioned because insoluble aggregates from all G-CSF C3 aggregation assay samples monitored by ^{1}H-^{15}N HSQC spectra were snap frozen and later analysed by EM in order to characterise their morphology (see Sections 2.8 and 4.2.5).

Surprisingly, the rates of decay for amide proton peaks during aggregation of G-CSF C3 at pH 7 are greater in the absence of urea than in the presence of 4 M urea (Figure 5.14C and D, respectively), even though G-CSF C3 aggregates more readily in the presence of 4 M urea (Figure 4.3A and Figure 4.6). This is most likely due to the size of the soluble aggregates formed in the absence and presence of 4 M urea (Figure 4.6C and D, respectively).

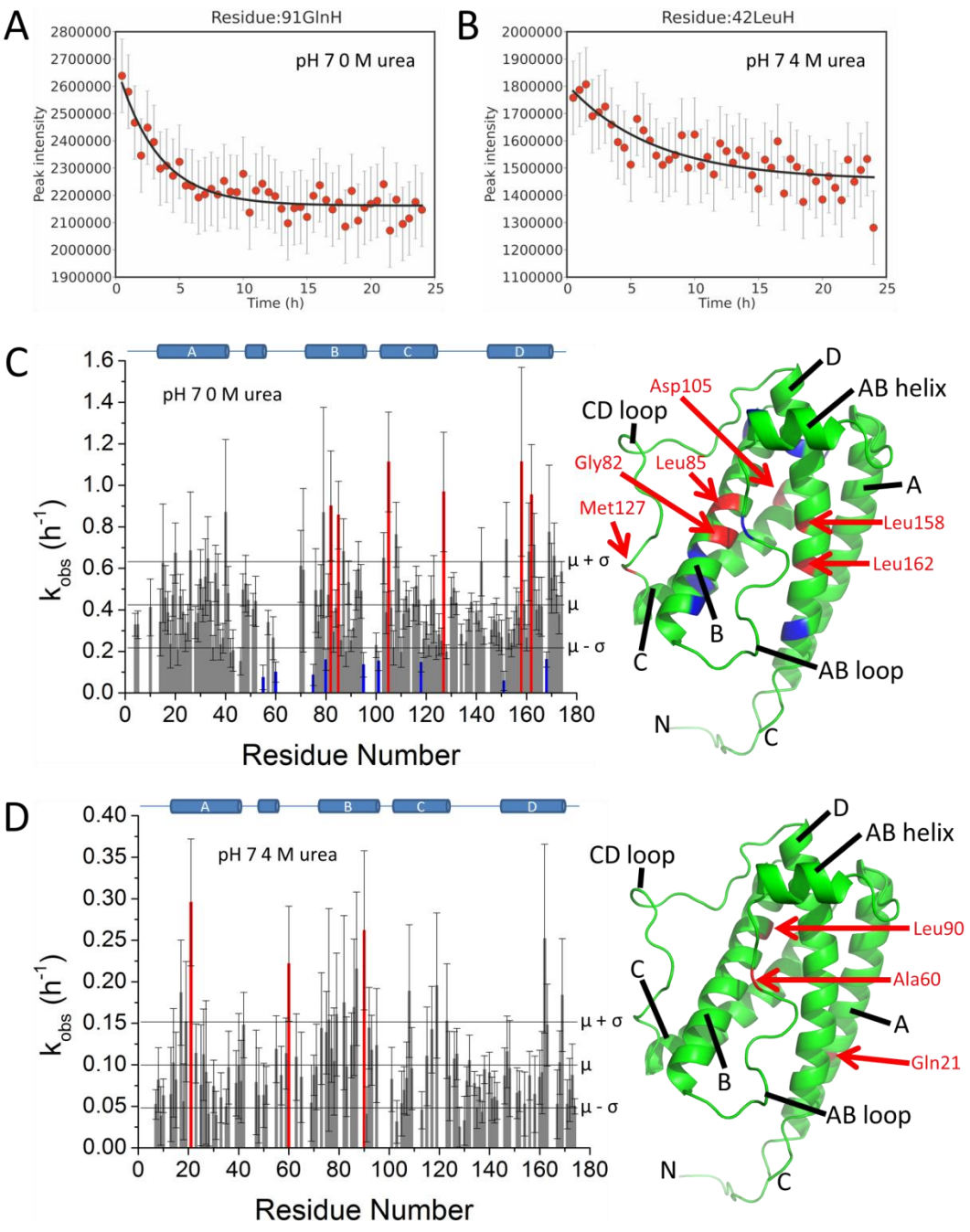


Figure 5.14 Decay rates in amide proton peak intensities during aggregation assays of G-CSF C3 at pH 7 in the absence or presence of 4 M urea. Observed decay rates (k_{obs}) were measured using 66 µM G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 0.018 % (*w/v*) sodium azide, 10 % (*v/v*) D_2O pH 7 buffer (I = 72 mM) in 0 M urea and 67 µM G-CSF C3 in 4 M urea incubated at 37 °C. Examples of the plots and fits used for calculating k_{obs} in the absence or presence of 4 M urea are shown in A and B, respectively. k_{obs} rates measured in the absence or presence of 4 M urea are shown plotted against residue number in C and D, respectively. Horizontal lines show the mean, mean + σ and mean – σ k_{obs} rates and a secondary structure schematic is shown above for reference. Red bars correspond to residues exhibiting k_{obs} rates significantly greater than the mean + σ value and blue bars correspond to residues exhibiting k_{obs} rates significantly lower than the mean – σ value. Right: Residues with significant k_{obs} rates are shown mapped onto the structure of G-CSF wt (PDB 1GNC[283]) using the same colour scheme, constructed using PyMOL[289]. Red residues are also labelled. k_{obs} rates were measured at 37 °C using a spectrometer operating at a proton resonance frequency of 750 MHz equipped with a cryogenic probe.

In the absence of urea, G-CSF C3 at pH 7 is largely monomeric before incubation at 37 ^{o}C, as shown by analytical SEC (Figure 3.9) and DLS (Figure 4.6A, C and E). After incubation of G-CSF C3 at 37 ^{o}C quiescently at pH 7 in the absence of urea, DLS showed that there was increased deviation of the autocorrelation function from the fit to a single exponential decay function at early time periods (Figure 4.6A), which corresponded to an increase in the width of the main peak in regularisation analysis (Figure 4.6C and E). This is indicative of the formation of species with larger R_h values i.e. soluble aggregates.

In the presence of 4 M urea G-CSF C3 was observed by DLS to form larger species than those formed in the absence of urea even before incubation at 37 ^{o}C (Figure 4.6B, D and F), with this incubation period simply shifting the distribution of species to even larger R_h values (Figure 4.6D and F). Therefore, the transition from soluble aggregates formed in the presence of 4 M urea to larger soluble aggregates formed during incubation at 37 ^{o}C may not be detectable in ^{1}H-^{15}N HSQC spectra, while the transition from largely monomeric G-CSF C3 to oilgomeric species at pH 7 without urea can be detected, resulting in faster rates of decay for amide proton peak intensities.

Residues of G-CSF C3 exhibiting significantly high decay rates during aggregation at pH 7 in the absence and presence of 4 M urea are coloured red in Figure 5.14C and D, respectively. The three residues that exhibit significantly high decay rates in the presence of 4 M urea are not clustered in a particular region (Figure 5.14D). In addition, the decay rates encountered in the presence of 4 M urea are lower generally than the rates observed in the absence of urea, suggesting that they may not be reporting on interactions at a possible aggregation interface.

Of the six residues of G-CSF C3 that exhibit significantly high decay rates during aggregation at pH 7 in the absence of urea, four are located around the AB loop that has already been shown to exhibit increased dynamics in aggregation-prone conditions (see Section 5.2.2), these residues being Gly82 and Leu85 in helix B and Leu158 and Leu162 in helix D. Interestingly, these residues in helix B are close to Tyr86, the side chain of which has been shown to undergo an alteration in its chemical environment under aggregation-prone conditions at both pH 4 and 7 (Figure 5.5). Moreover, Leu158 is one of the residues that exhibits a large CSP in sub-denaturing urea concentrations (Figure 5.1B-Figure 5.4B), again as a result of a change in the environment of the side chain of Tyr86.

This could suggest that the region around Tyr86 is acting as an aggregation interface, which would also agree with the changes in the near-UV CD and fluorescence emission spectra observed for G-CSF C3 in sub-denaturing urea concentrations (Figure 4.10 and Figure 4.9, respectively). Near-UV CD spectra showed increased negativity around 275 nm upon addition of sub-denaturing urea concentrations, suggesting that the side chain of Tyr86 was becoming less mobile (Figure 4.10). Likewise, fluorescence emission spectra showed increased emission intensity upon adding these same urea concentrations, suggested Tyr86 was becoming less solvent exposed (Figure 4.9).

5.2.5 Mapping ANS binding to G-CSF C3 using ^{1}H-^{15}N HSQC spectra

In Chapter 4 the fluorescence emission spectra of ANS in the presence and absence of G-CSF C3 were used to probe for exposed hydrophobic surfaces that could be linked to aggregation. The fluorescence emission spectra of 250 µM ANS at pH 7 were largely unchanged in the presence of 2.5 µM G-CSF C3 in 0-7 M urea, suggesting no change in the hydrophobicity of the surface detected by ANS (Figure 4.15A and C). Conversely, the fluorescence emission spectra of 250 µM ANS at pH 4 showed a significant increase in emission intensity and a blue shift in the emission maxima in the presence of 5 µM G-CSF C3, suggesting increased hydrophobicity of the surface detected by dye binding[395,396] in the presence of G-CSF C3 (Figure 4.15B and D). Intriguingly, the fluorescence emission spectra of ANS at pH 4 in the presence of urea concentrations of 1 M and above were similar with and without G-CSF C3, suggesting that this change in hydrophobicity was not occurring at these urea concentrations.

The structural similarity of G-CSF C3 at pH 4 and 7 suggested by calculation of CSPs (Figure 3.19), coupled to the overall positive charge on G-CSF C3 at pH 4 that is not apparent at pH 7 (based on a pI of 6.67 predicted from the amino acid sequence[117]), led to the hypothesis that the binding of ANS to G-CSF C3 at pH 4 was driven by electrostatic rather than hydrophobic interactions. Previous studies have shown that ion pairing between the sulphonate group of ANS and positively charged side chains of Lys, Arg and His at acidic pH is the dominant mechanism of binding to protein surfaces[397-399]. However, such electrostatic interactions with a protein surface do not usually cause the characteristic increase in emission intensity

and blue shift in the emission maximum of ANS that are associated with an increased hydrophobicity of its environment[398] (see Section 4.2.11).

The changes in fluorescence emission spectra of ANS resulting from incubation with G-CSF C3 at pH 4 were attributed to aggregation because protein precipitation was evident in this sample and ANS has previously been shown to cause protein aggregation[187]. Moreover, no insoluble aggregate could be observed in samples of G-CSF C3 and ANS at pH 4 in the presence of urea concentrations of 1 M or greater, in-line with the similar fluorescence emission spectra of ANS observed with or without G-CSF C3 under these conditions. Therefore, it was postulated that ANS binding to G-CSF C3 at pH 4 may block a positively charged site that is involved in inhibiting aggregation at acidic pH and replacing this site with a hydrophobic naphthalene group that promotes aggregation.

To elucidate the ANS binding behaviour of G-CSF C3, ^{1}H-^{15}N HSQC spectra were acquired using 85 µM G-CSF C3 at pH 4 and 65 µM G-CSF C3 at pH 7 in the presence of 700 µM ANS at 25 ^{o}C (see Section 2.9.5). These spectra were compared with spectra acquired using the same concentrations of G-CSF C3 in the absence of ANS at 25 ^{o}C and CSPs calculated (see Section 2.9.6).

Due to the lower protein concentrations required for measuring fluorescence emission spectra compared to ^{1}H-^{15}N HSQC spectra (at least 2 and 50 µM, respectively), the same molar ratio of G-CSF C3 to ANS could not be used during acquisition of ^{1}H-^{15}N HSQC spectra that was used previously for acquiring fluorescence emission spectra. Therefore, the molar excess of ANS to G-CSF C3 used for acquiring ^{1}H-^{15}N HSQC spectra was roughly nine- and ten-fold at pH 4 and 7, respectively, compared with the fifty- and one hundred-fold molar excesses of ANS to G-CSF C3 at pH 4 and 7, respectively, that were used for acquiring fluorescence emission spectra.

Upon mixing of G-CSF C3 and ANS at pH 4 aggregation could be observed, which meant that ^{1}H-^{15}N HSQC spectra were acquired immediately after brief manual mixing of ANS and G-CSF C3. Conversely, samples of ANS and G-CSF C3 used for acquisition of fluorescence emission spectra were incubated at 25 ^{o}C overnight before acquisition as aggregation did not significantly affect the emission intensity of spectra.

At pH 7, Arg71 of G-CSF C3 exhibits the largest CSP as a result of adding ANS, suggesting that ANS binds to this residue (Figure 5.15B). Additionally, residues neighbouring Arg71 in the amino acid sequence and proximal to this residue in the tertiary structure of G-CSF C3 exhibit large CSPs, coloured in red in Figure 5.15B, which also suggests that ANS binds in this region. Arg71 also exhibits a large CSP upon addition of ANS at pH 4 (Figure 5.16B), suggesting that ANS also binds Arg71 at pH 4.

The peaks for Arg71 of G-CSF C3 at pH 4 and 7 are far less intense in the presence of ANS, but aggregation of G-CSF C3 at pH 4 upon addition of ANS makes it unclear if the reduced intensity of this peak relates to ANS binding or a decrease in the concentration of soluble G-CSF C3 (Figure 5.16A). However, aggregation of G-CSF C3 was not observed at pH 7 after adding ANS, yet this peak is considerably weaker in the presence of ANS, while other peaks appear to exhibit comparable intensities in the presence and absence of ANS, suggesting that broadening of the peak for Arg71 occurs as a result of ANS binding (Figure 5.15A).

At pH 4, His157 of G-CSF C3 exhibits a large CSP as a result of adding ANS, suggesting that ANS also binds to this residue (Figure 5.16B). Ala60 and Ser67 also exhibit large CSPs, which is most likely the result of ANS binding to His157. Interestingly, His157 does not exhibit a significant CSP as a result of addition of ANS at pH 7 (Figure 5.15B), suggesting that ANS binds to this residue at pH 4 but not at pH 7. This is consistent with both the pKa of the His side-chain (6.0[117]) and the dominant mechanism of ANS binding being ion pairing between the sulphonate group of ANS and positively charged side chains[397-399].

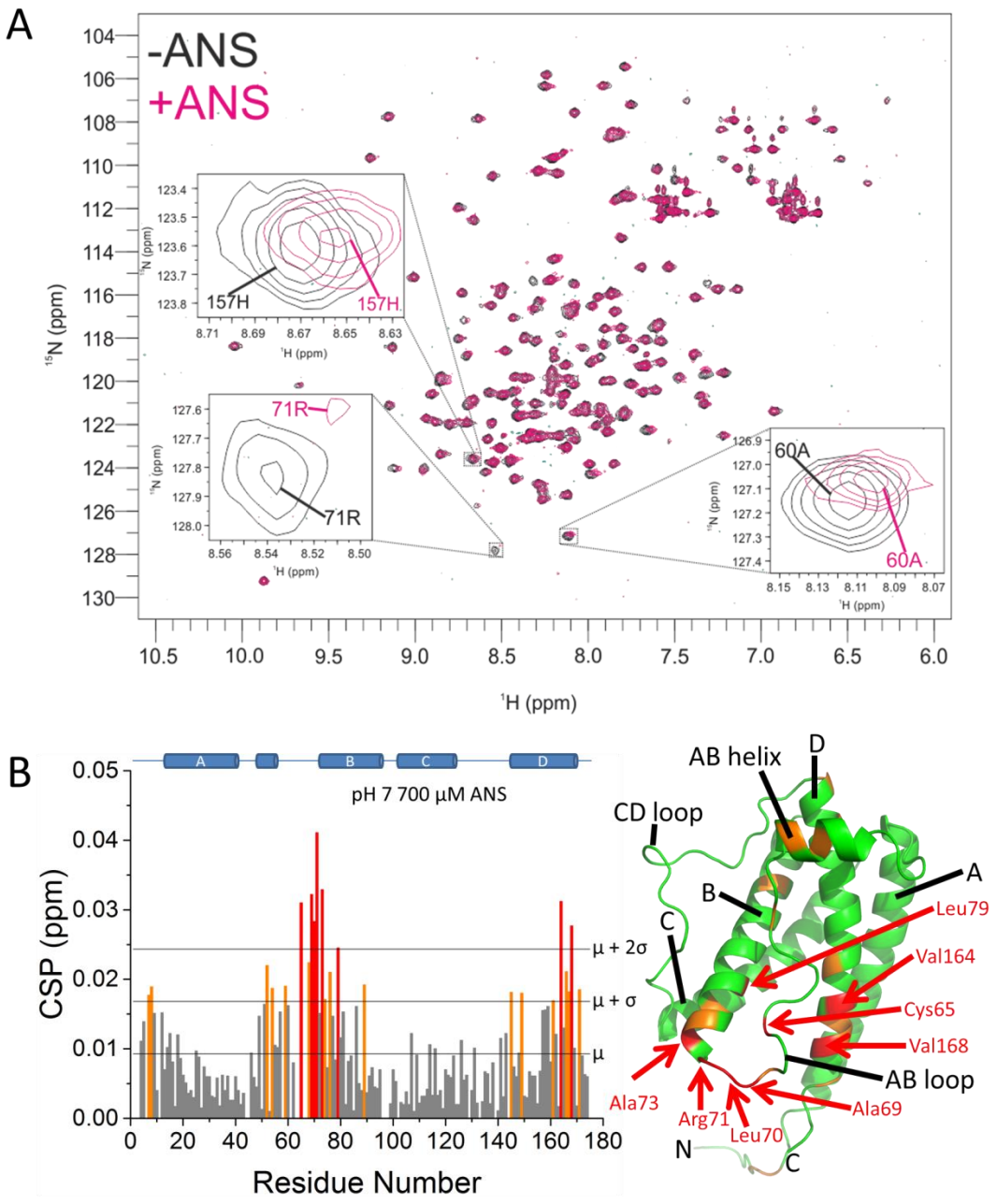


Figure 5.15 ANS binding to G-CSF C3 at pH 7. The ^{1}H-^{15}N HSQC spectra of 65 µM G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 0.018 % (*w/v*) sodium azide, 10 % (*v/v*) D_2O pH 7 buffer (I = 72 mM) in the absence or presence of 700 µM ANS (black and pink, respectively) are shown in A, which were used for calculation of the CSPs shown in B. CSPs were calculated using backbone amide proton and nitrogen chemical shifts of G-CSF C3 in the presence and absence of 700 µM ANS (B). Residues Ala60, Arg71 and His157 are shown magnified in A as these residues give rise to relatively large CSPs upon addition of 700 µM ANS at pH 4 (Figure 5.16B). Of these residues, only Arg71 exhibits a significantly large CSP at pH 7, which is coloured in red in B. Horizontal lines show the mean, mean + σ and mean + 2σ CSP values. Orange and red bars correspond to residues exhibiting CSPs greater than the mean + σ and mean + 2σ, respectively. A secondary structure schematic is shown above for reference. Residues with significant CSPs are shown mapped onto the structure of G-CSF wt (PDB 1GNC[283]) using the same colour scheme, constructed using PyMOL[289]. Red residues are also labelled. Spectra were acquired at 25 ^{o}C using a spectrometer operating at a proton resonance frequency of 750 MHz equipped with a cryogenic probe.

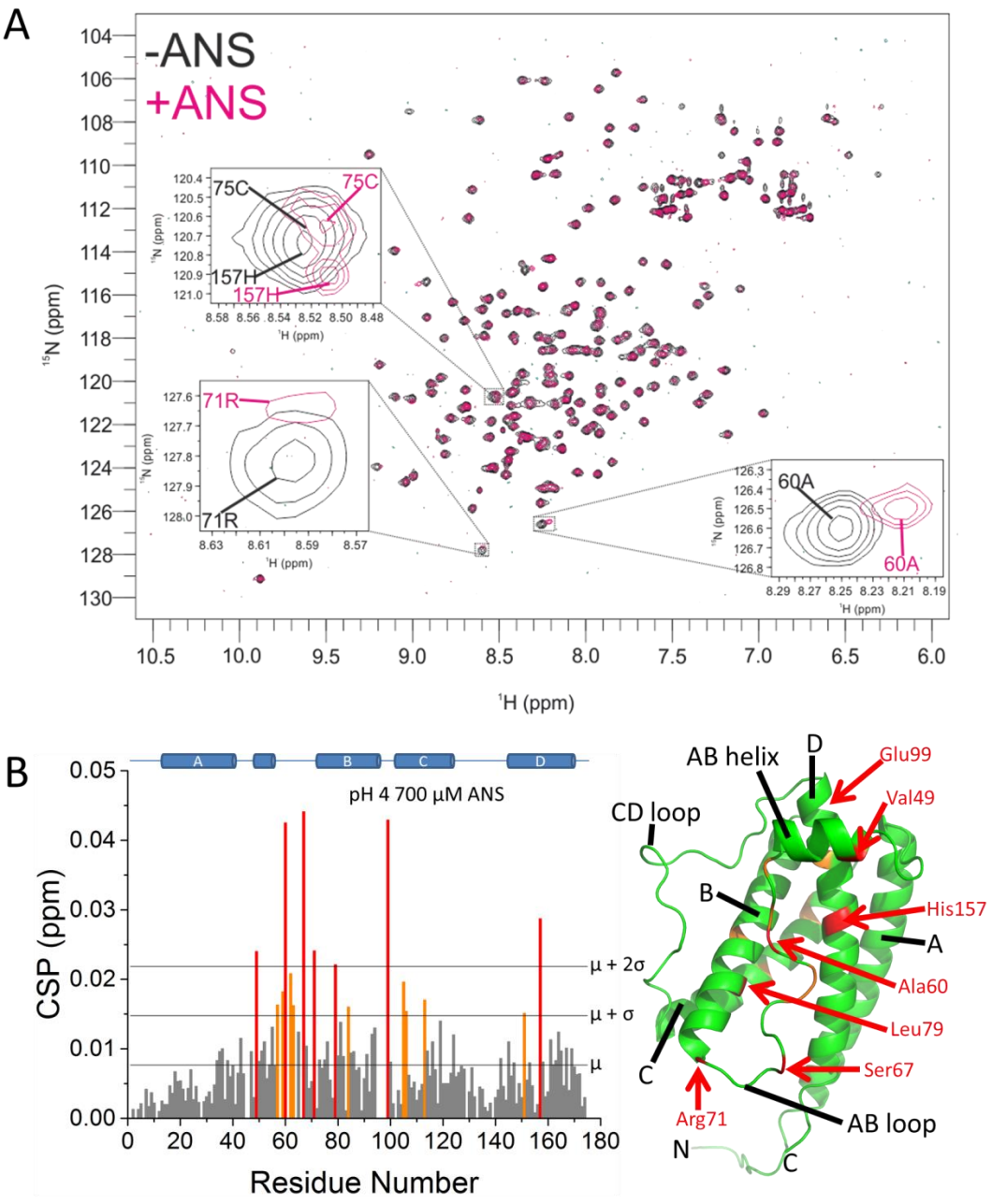


Figure 5.16 ANS binding to G-CSF C3 at pH 4. The ^{1}H-^{15}N HSQC spectra of 85 µM G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 10 % (*v/v*) D_2O pH 4 buffer (I = 39 mM) in the absence or presence of 700 µM ANS (black and pink, respectively) are shown in A, which were used for calculation of the CSPs shown in B. CSPs were calculated using backbone amide proton and nitrogen chemical shifts of G-CSF C3 in the presence and absence of 700 µM ANS (B). Peaks for residues Ala60, Arg71 and His157 are shown magnified in A as these residues give rise to relatively large CSPs and so are coloured in red in B. The peak for Cys75 overlaps with that of His157 in the absence of ANS and so is also shown magnified in A, but does not exhibit a significant CSP in B. Horizontal lines show the mean, mean + σ and mean + 2σ CSP values. Orange and red bars correspond to residues exhibiting CSPs greater than the mean + σ and mean + 2σ, respectively. A secondary structure schematic is shown above for reference. Residues with significant CSPs are shown mapped onto the structure of G-CSF wt (PDB 1GNC[283]) using the same colour scheme, constructed using PyMOL[289]. Red residues are also labelled. Spectra were acquired at 25 °C using a spectrometer operating at a proton resonance frequency of 750 MHz equipped with a cryogenic probe.

The observation of G-CSF C3 aggregation upon mixing with ANS at pH 4 but not at pH 7, coupled to the binding of ANS to residue His157 of G-CSF C3 at pH 4 but not at pH 7 possibly reveals a role for this residue in the reduced aggregation propensity of G-CSF C3 at pH 4 (Figure 4.2). Reduced aggregation of wild-type G-CSF has also been reported at acidic pH[185,276,303], which has been attributed to increased electrostatic repulsion between monomeric G-CSF, owing to the overall positive charge on G-CSF wt at acidic pH[303]. The proposed role of His157 in inhibiting G-CSF C3 aggregation at acidic pH is further supported by the suggested roles of the nearby residues Leu158 and Leu162 in the aggregation interface, which were identified during monitoring of G-CSF C3 aggregation at pH 7 in the absence of urea using ^{1}H-^{15}N HSQC spectra (Figure 5.14A).

The binding of ANS to G-CSF C3 at pH 4 in the presence of urea was not investigated using ^{1}H-^{15}N HSQC spectra. However, fluorescence emission spectra of ANS with and without G-CSF C3 at pH 4 in the presence of urea concentrations of 1 M or higher are similar (Figure 4.15B, D and F). In addition, aggregation could not be observed during incubation of G-CSF C3 with ANS at these urea concentrations, which suggests that ANS is not able to bind His157 in urea concentrations higher than 1 M. This could mean that the increased dynamics of the AB loop of G-CSF C3 in sub-denaturing urea concentrations is hindering ANS binding to His157, either through steric hindrance of the side chain of this residue or through disruption of the proximal arrangement of the positively charged side chains from His53, Arg59 and His157 (Figure 3.21).

The ability of ANS to bind G-CSF C3 at pH 4 has been shown to decrease gradually as the urea concentration is raised from 0-1 M in 0.2 M increments (Figure 4.15F). This may be accounted for by a gradual increase in the dynamics of the AB loop, as R_2 measurements would suggest that part of the latter half of the AB loop becomes more dynamic in the presence of only 0.5 M urea at pH 4 (Figure 5.9B), which may partially inhibit ANS binding. However, in the presence of urea concentrations that promote aggregation i.e. 3 M urea the entire latter half of this loop is more dynamic (Figure 5.10B), which completely prevents ANS binding.

5.3 Discussion

Characterisation of the conformational change of G-CSF C3 that occurs upon addition of urea concentrations that are too low to cause global unfolding at amino acid resolution was achieved by calculation of CSPs, revealing a change in the environment of residues of the AB loop and residues on helices B and D that are proximal to this loop. This conformational change occurred in the presence of 0.5 M urea (Figure 5.1B and Figure 5.3B) as well as urea concentrations that promote aggregation (Figure 5.2B and Figure 5.4B). Residues exhibiting the largest CSPs e.g. Ala60 and Leu158 have been shown to have contacts with Tyr86 (Figure 5.5), which shows that a change in the environment of the side chain of Tyr86 is responsible for changes in the near-UV CD and fluorescence emission spectra of G-CSF C3 observed upon addition of sub-denaturing urea concentrations, as outlined in Chapter 4 (Figure 4.10 and Figure 4.9, respectively).

The change in the environment of Tyr86 is observed in the presence of urea concentrations that have been shown to promote G-CSF C3 aggregation, as well as 0.5 M urea, based on analysis of ^{1}H-^{15}N HSQC (Figure 5.1-Figure 5.4), near-UV CD (Figure 4.10) and fluorescence emission spectra (Figure 4.9). Such low urea concentrations do not promote aggregation (Figure 4.3), which raises the possibility that a change in the environment of Tyr86 may be coincidental with increased aggregation observed in the presence of urea. However, the observation that 0.5 M urea does not promote G-CSF C3 aggregation is based on formation of insoluble aggregates, whereas the effect of such low urea concentrations on soluble G-CSF C3 aggregate formation has not been tested. Considering that G-CSF C3 has been shown to undergo soluble aggregate formation by DLS at pH 4 and 7 in the absence of urea (Figure 4.7A, C and F and Figure 4.6A, C and F, respectively), it is possible that urea concentrations as low as 0.5 M may promote soluble G-CSF C3 aggregate formation. Alternatively, the concentration of the aggregation-prone species in 0.5 M urea may be too low to initiate aggregation as protein aggregation is concentration-dependent[41].

Measurement of R_2 rates for backbone amide nitrogen nuclei of G-CSF C3 allowed changes in the local dynamics of the polypeptide backbone to be elucidated in the presence of 0.5 M urea (Figure 5.6 and Figure 5.9), as well as in higher urea concentrations at which G-CSF C3 exhibited a higher propensity to aggregate

(Figure 5.7, Figure 5.8 and Figure 5.10). At pH 7, residue Leu83 in helix B was shown to have a significantly reduced R_2 rate in both 0.5 and 4 M urea (Figure 5.6B and Figure 5.7B, respectively), suggesting that this residue becomes more dynamic in sub-denaturing urea concentrations. However, in the presence of 4 M urea residues Ser54 and Arg71 in the latter half of the AB loop were also shown to be more dynamic (Figure 5.7B). Additionally, accurate R_2 rates could only be measured for residues Ser64, Gln68 and Leu72 in the presence of 4 M urea (Figure 5.8F), suggesting that these residues in the latter half of the AB loop are also more dynamic in the presence of 4 M urea. However, the low signal-to-noise ratio of the peaks for these residues in the absence of urea, resulting from their high R_2 rates (Equation 5.1)[324] makes it difficult to measure reliably the extent to which their R_2 rate had increased.

At pH 4, residues Ala69 and Leu70 of the AB loop as well as Leu72 of helix B, located at the point of connection between the AB loop and helix B were shown to become significantly more dynamic in the presence of 0.5 M urea (Figure 5.9B). In the presence of 3 M urea, residues Gly56 and Arg59 in the latter half of the AB loop exhibited significantly increased dynamics, along with residues Leu70 and Arg71 (Figure 5.10B). Overall, measurement of R_2 rates for G-CSF C3 in different urea concentrations would suggest that in the presence of 0.5 M urea there is an increase in the dynamics of residues at the C-terminal end of the loop AB as well as the N-terminal end of helix B, whereas in the presence of urea concentrations that promote aggregation residues throughout the latter half of the AB loop become more dynamic, perhaps implying that a greater proportion of the AB loop of G-CSF C3 becomes more dynamic in order to cause aggregation.

Monitoring of G-CSF C3 aggregation at pH 7 in the absence of urea using ^{1}H-^{15}N HSQC spectra showed that residues Gly82, Leu85, Leu158 and Leu162 were among the six residues that exhibited significantly high rates of decay (Figure 5.14A). During soluble aggregate formation, the peaks of all amide protons of G-CSF C3 will be reduced in intensity due to the slower molecular tumbling exhibited by larger species (Equation 5.2)[319]. The faster rates of decay for the outlined residues results from the additional contribution of chemical exchange to the R_2 rate[341], which is caused by these residues experiencing two different chemical environments i.e. in monomeric G-CSF C3 and at the aggregation interface of oligomeric species.

The proximity of residues Gly82 and Leu85 to Tyr86 suggests that the change in the chemical environment of the Tyr86 side chain that occurs in sub-denaturing urea concentrations may result from this region of helix B being involved in the aggregation interface. This would support the results of near-UV CD spectra acquired in the presence of sub-denaturing urea concentrations, which show an increased negativity of the absorption minimum around 275 nm (Figure 4.10), suggesting Tyr86 is becoming less mobile upon adding urea. Moreover, fluorescence emission spectra acquired in the presence of sub-denaturing urea concentrations show increased emission intensity (Figure 4.9), implying reduced solvent exposure of the Tyr86 side chain upon adding urea. Additionally, Leu158 exhibits one of the largest CSPs in the presence of low urea concentrations (Figure 5.1B-Figure 5.4B), resulting from its proximity to Tyr86 (Figure 5.5), which supports the notion that Leu158 and possibly the nearby residue Leu162 are also part of the aggregation interface.

Mapping the binding of ANS to G-CSF C3 using ^{1}H-^{15}N HSQC spectra also indicates that the region around Leu158 may be involved in an aggregation interface. ANS was shown to bind to Arg71 of G-CSF C3 at both pH 4 and 7 based on calculation of CSPs and broadening of this peak (Figure 5.16 and Figure 5.15, respectively). However, the fluorescence emission spectra of ANS at pH 7 with and without G-CSF C3 are similar, regardless of the concentration of urea present (Figure 4.15A and C), suggesting that binding of ANS to Arg71 is not accompanied by a change in the polarity of the environment around the dye at pH 7. This agrees with previous studies which have demonstrated that the dominant mechanism of ANS binding to protein surfaces is through ion pairing interactions between the sulphonate group of ANS and positively charged side chains[397-399].

Conversely, ANS at pH 4 without urea shows a dramatic increase in fluorescence emission intensity and a blue shift in the wavelength at which maximum emission is observed in the presence of G-CSF C3 compared with the absence of G-CSF C3 (Figure 4.15B and D). This is indicative of increased hydrophobicity of the environment around ANS[395,396] and is attributed to insoluble aggregate formation, which was observed during incubation of G-CSF C3 and ANS at pH 4 before analysis by both ^{1}H-^{15}N HSQC and fluorescence emission spectra. ANS was shown to bind to His157 of G-CSF C3 at pH 4 but not at pH 7 based on calculation of CSPs (Figure 5.16B and Figure 5.15B, respectively), suggesting that the protonated side chain of His157 may be involved in inhibition of aggregation at pH 4 compared with

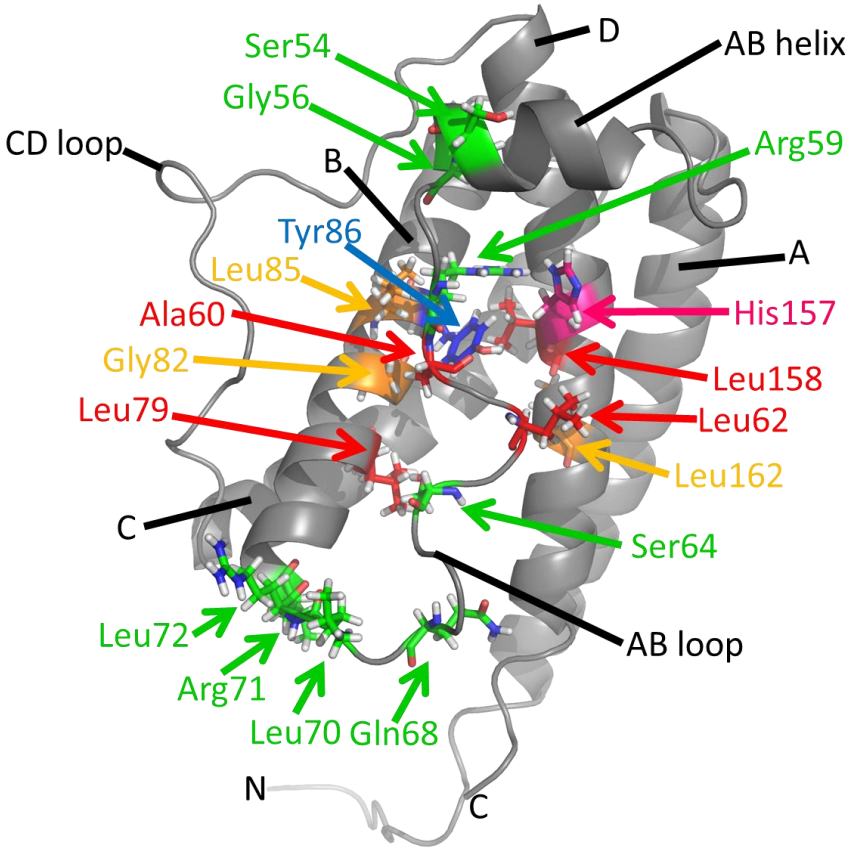
pH 7 (Figure 4.2) and more generally that the region of G-CSF C3 around His157 may be involved in an aggregation interface.

In the presence of urea concentrations of 0-1 M at 0.2 M increments a gradual decrease in the fluorescence emission intensity of ANS is observed with G-CSF C3 (Figure 4.15F), which coincides with no formation of insoluble aggregates (observable by eye) at pH 4 with ANS in the presence of urea concentrations of 1 M or greater. This would also suggest that ANS is not able to bind His157 as effectively in the presence of urea concentrations that have been shown to promote G-CSF C3 aggregation in the absence of ANS, most likely as a result of the increased dynamics of the latter half of the AB loop. Moreover, ANS is able to bind to G-CSF C3 at pH 4 in the presence of urea concentrations below 1 M, albeit to a lesser extent than in the absence of urea. This supports the theory that there is an increase in the dynamics of the AB loop in the presence of urea concentrations that promote aggregation compared to lower urea concentrations that do not appear to affect the amount of insoluble aggregate formation.

The mechanism by which this increased mobility of the AB loop would hinder ANS binding to His157 is uncertain, but could relate to steric hindrance of the His157 side chain by the loop. Alternatively, movement of the AB loop may disrupt the triad of positively charged side chains formed by His53, Arg59 and His157 (Figure 3.21), which could be involved in facilitating binding of ANS to His157. Contrary to this argument, the CSPs exhibited by His53 are not significantly large at pH 4 or 7 (Figure 5.16B and Figure 5.15B, respectively), indicating that this residue is not directly involved ANS binding, although Arg59 does exhibit a significant CSP at both pH 4 and 7, suggesting a possible role of Arg59 in ANS interaction along with His157.

Overall, the work in Chapter 5 has shown that there is a change in the environment of Tyr86 and increased dynamics in the latter half of the AB loop that is proximal spatially to this residue in aggregation-prone conditions. Aggregation assays monitored by NMR have also indicated that residues close to Tyr86 may be part of an aggregation interface, which is supported by the role of His157 in ANS binding and presumably inhibiting aggregation at pH 4. The experiments and key findings from Chapter 5 are summarised in Figure 5.17 and the theorised involvement of Tyr86 in G-CSF C3 aggregation is further explored in Section 5.3.1 using the aggregation prediction program CamSol[63].

A



B

Experiment	Residues identified	Observation
CSPs	Ala60, Leu62, Leu79, Leu158	Change in the environment of Tyr86 under aggregation-prone conditions
R_2 relaxation	Ser54, Gly56, Arg59, Ser64, Gln68, Leu70, Arg71, Leu72	Increased dynamics in latter half of AB loop under aggregation-prone conditions
NMR aggregation assay	Gly82, Leu85, Leu158, Leu162	Involved in aggregation interface at pH 7 in the absence of urea
ANS binding	Arg71, His157	ANS binds to His157 at pH 4 and causes aggregation but not pH 7 ANS does not bind or cause aggregation at pH 4 in >1 M urea

Figure 5.17 Illustrative summary of the key features of G-CSF C3 aggregation. Important residues identified from the experiments listed in B are shown mapped onto the structure of G-CSF wt (PDB 1GNC[283]) in A using the same colour scheme, constructed using PyMOL[289]. Residues exhibiting large CSPs in aggregation-prone conditions, resulting from a change in their chemical environment relative to Tyr86, at pH 4 and/or 7 are coloured/labelled in red in A and listed in B. Residues in the latter half of the AB loop that exhibit increased dynamics (decreased R_2 rates) in urea concentrations that cause aggregation are coloured/labelled in green in A and listed in B. Residues involved in protein-protein interactions at the aggregation interface are coloured/labelled in orange in A and listed in B. Residues involved in ANS binding are coloured/labelled in pink in A and listed in B. Residues Arg71 and Leu158 were identified in more than one experiment but are only shown in one colour.

5.3.1 Tyr86 is predicted to be intrinsically aggregation-prone

CamSol[63] was used to compare the intrinsic solubility of the amino acid sequence of G-CSF C3 with the structurally-corrected solubility scores (Figure 5.18A and B, respectively), which project these intrinsic scores onto the surface of a protein structure in order to correct for the solvent accessible surface area (SASA) of residues (see Section 4.2.1)[63]. Additionally, structurally corrected scores were calculated for a hypothetical G-CSF C3 structure that lacked the AB loop in order to assess the degree of protection from solvent provided by this loop to residues that may be intrinsically insoluble (Figure 5.18C and D).

Tyr86 is predicted to significantly reduce the intrinsic solubility of G-CSF C3 at pH 4 and 7, according to the score of this residue being below -1 (Figure 5.18A). Residues Gly82 and Leu85 also exhibit scores of predicted intrinsic solubility below -1 at pH 7, with Leu85 also scoring very close to -1 at pH 4, suggesting that these residues are also predicted to promote general aggregation of G-CSF C3 at pH 7 and to a lesser extent at pH 4. Whilst His157 and Leu158 are not predicted to significantly affect the intrinsic solubility of G-CSF C3 at pH 4 or 7, Leu162 shows a score more negative than -1 at pH 4 and very close to -1 at pH 7, suggesting that this residue is predicted to promote general aggregation.

Interestingly, the scores of all of the highlighted residues are very close or equal to zero when corrected for the structure of G-CSF wt (Figure 5.18B), suggesting that these residues do not significantly contribute to the solubility of the structure of G-CSF C3. The changes observed in the solubility scores of these residues when projected onto the structure of G-CSF C3 relates to their low SASA.

Carrying out CamSol[63] predictions of the solubility of the structure of G-CSF C3 without the AB loop results in increased negativity of scores for residues Gly82, Leu85 and Tyr86 at both pH 4 and 7 (Figure 5.18C and D, respectively). The increases in negativity observed are of similar magnitude for these residues, suggesting that the AB loop shields them from solvent exposure. These predictions help strengthen the theory that increased dynamics of the AB loop promotes G-CSF C3 aggregation by exposing Tyr86 and surrounding residues such as Gly82 and Leu85, resulting in this region of helix B being involved in the aggregation interface.

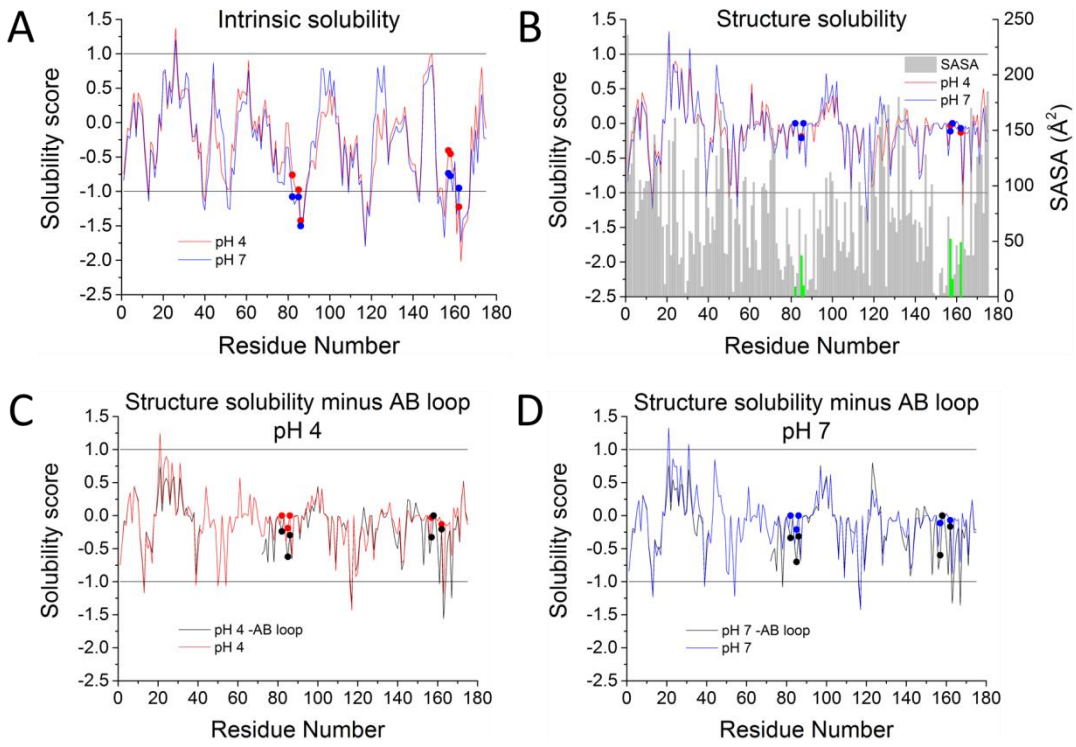


Figure 5.18 Predictions of the intrinsic and structural solubility of G-CSF C3. CamSol[63] solubility scores above 1 indicate residues predicted to increase solubility, while scores below -1 indicate residues predicted to decrease solubility and so promote aggregation[63] of G-CSF C3 at pH 4 and 7 (red and blue, respectively). Intrinsic solubility scores predicted from the amino acid sequence of G-CSF C3 are shown in A. Solubility scores corrected for structure were calculated using the G-CSF wt structure (PDB 1GNC[283]), with residues being mutated in PyMOL[289] for calculation of G-CSF C3 solubility scores (B-D). The solvent accessible surface area (SASA) of residues is plotted in B as grey bars on the right-hand y-axis with values for residues Gly82, Leu85, Tyr86, His157, Leu158 and Leu162 shown in green. SASAs were obtained from the DSSP[411] file associated with the structure of G-CSF wt (PDB 1GNC[283]), acquired using the MRS server[412]. For calculation of solubility scores corrected for structure without the AB loop at pH 4 and 7 (black and pink, respectively), residues 42-71 were deleted from the G-CSF wt structure (PDB 1GNC[283]) in PyMOL[289] (C and D). Scores predicted for residues Gly82, Leu85, Tyr86, His157, Leu158 and Leu162 with or without the AB loop are shown as filled data points.

Solubility scores for residues Leu158 and Leu162 remain close to zero even after deletion of the AB loop, suggesting possible protection of these residues from solvent by a different region of G-CSF C3. However, during monitoring of G-CSF C3 aggregation at pH 7 in the absence of urea by ^{1}H-^{15}N HSQC spectra peaks for the amide protons of these residues exhibit significantly faster rates of decay than other residues, along with Gly82 and Leu85 (Figure 5.14A). This could indicate a role of residues around Tyr86 on helix B in the early interactions that drive aggregation, followed by additional interactions involving Leu158 and Leu162 and possibly surrounding residues on helix D that help stabilise aggregates.

Residues Glu163 and Val164 exhibit the most negative intrinsic solubility scores at pH 4 and 7 (Figure 5.18A). These residues are also predicted to promote generic aggregation in the context of a tertiary structure of G-CSF C3 with the AB loop removed at pH 4 (Figure 5.18C), while Glu163 is predicted to promote aggregation in this same context at both pH 4 and 7. Furthermore, these residues exhibit increasingly negative scores in the structure of G-CSF C3 as a result of removal of the AB loop (Figure 5.18C and D, respectively), implying a degree of protection of these residues from solvent by this loop that may be disrupted in urea concentrations that promote aggregation.

Surprisingly, the solubility score for His157 becomes more negative as a result of deletion of the AB loop, implying a degree of protection of His157 from solvent exposure by this loop. This is in contrast to decreased ANS binding by G-CSF C3 at pH 4 in the presence of urea concentrations that significantly increase AB loop dynamics, which would imply an increase in the protection of His157 from solvent as a result of a conformational change of the AB loop. Therefore, the reason for the decreased binding of ANS to His157 of G-CSF C3 at pH 4 in the presence of urea most likely does not relate to solvent exposure, but may be the result of disrupting the proximal arrangement of positively charged side chains from His53, Arg59 and His157 (Figure 3.21), although mapping of ANS binding using CSPs does not indicate any direct interaction between His53 and ANS at pH 4 or 7 (Figure 5.16B and Figure 5.15B, respectively).

In Chapter 6, the key findings in this study relating to G-CSF C3 folding, dynamics and aggregation are summarised and discussed in relation to existing knowledge of wild-type G-CSF. Future perspectives on this work are also envisaged for further studying the link between G-CSF C3 folding, dynamics and aggregation, as well as relating these characteristics back to wild-type G-CSF and attempting to clarify the effects of the substitutions found in the variant C3 on recombinant expression level. Overall, the work in this chapter has identified a change in the environment of the side chain of Tyr86 and residues in and around the AB loop of G-CSF C3 in sub-denaturing urea concentrations. The conformational change of the AB loop is due to increased dynamics, which exposes an intrinsically aggregation-prone segment of G-CSF C3 encompassing Tyr86 on helix B that is predicted to drive aggregation.

Chapter 6 Conclusions and future perspectives

The work in Chapter 5 has identified the conformational fluctuation in the native state of G-CSF C3 that promotes aggregation in denaturant concentrations that are too low to cause global unfolding, this being increased dynamics in the latter half of the AB loop that changes the environment of the proximal side chain of Tyr86. This subtle conformational change is hypothesised to promote aggregation of G-CSF C3 by increasing the exposure of the region around Tyr86 on helix B to solvent, which is predicted by CamSol[63] to be intrinsically aggregation-prone (Figure 5.18A). Additionally, Aggrescan[62], Zyggregator[64] and TANGO[65] also predict this region of the amino acid sequence of G-CSF C3 to be amongst the most significantly aggregation-prone (Figure 4.1A, D and E, respectively). Moreover, the predicted aggregation propensity of this region in the context of the structure of wild-type G-CSF is increased upon removal of the AB loop (Figure 5.18C and D), suggesting a degree of protection provided by this loop.

The proposed role of the region around Tyr86 in G-CSF C3 aggregation is supported by monitoring of G-CSF C3 aggregation at pH 7 in the absence of urea using ^{1}H-^{15}N HSQC spectra, which revealed significantly higher rates of decay for peak intensities of residues Gly82 and Leu85 of helix B, as well as nearby Leu158 Leu162 of helix D (Figure 5.14A). Plus, near-UV CD spectra of G-CSF C3 in sub-denaturing urea concentrations indicate reduced mobility of the side chain of Tyr86 (Figure 4.10), rather than the increased mobility that would be expected from increased solvent exposure. Furthermore, fluorescence emission spectra under the same conditions point towards reduced solvent exposure of Tyr86 (Figure 4.9).

The work in Chapter 4 showed that G-CSF C3 exhibits similar aggregation behaviour to that reported for the wild-type protein in terms of the dependence of aggregation on pH, protein concentration (see Section 4.2.2) and denaturant concentrations that are too low to cause global denaturation (see Section 4.2.3). In addition, analysis of the thermodynamic stability of G-CSF C3 at pH 4 and 7 in urea resulted in calculation of similar ΔG^{o}_{UN} values to those reported for the wild-type protein in GdnHCl at the same pH values (see Section 4.2.4).

Overall, the similarities of the folding and aggregation properties of these proteins, as well as their structural similarity, suggest strongly that the subtle conformational

change elucidated for G-CSF C3 aggregation also applies to the wild-type protein. Indeed, aggregation of G-CSF wt in low GdnHCl concentrations at pH 7 has been proposed previously to be linked to a conformational rearrangement of the AB loop, revealed by increased fluorescence emission intensity of Trp59[142], the fluorescence of which is quenched by Tyr86 at neutral pH in the absence of denaturant (see Section 3.2.6). Although calculation of CSPs from ^{1}H-^{15}N HSQC spectra of G-CSF wt and C3 in pH 4 buffer could not be used to rule out a subtle change in conformation around the W59R substitution of G-CSF, no differences in conformation beyond the local environments of the four substitutions could be observed (Figure 3.19).

Despite the fact that the role of increased dynamics of the AB loop of G-CSF C3 in promoting aggregation has been established, the timescale for motions of this loop that affect aggregation is uncertain. R_2 rates reflect the local fluctuations of polypeptides (ps-ns motions) but are also affected by chemical exchange and τ_c, which can result from larger amplitude motions such as conformational exchange (µs-ms motions). Conversely, R_1 rates and heteronuclear nOe relaxation rates are determined by local polypeptide fluctuations only[341]. Therefore, measurement of these additional relaxation rates under the same conditions as those used for measuring R_2 rates would allow greater definition of the timescale of these motions of the AB loop.

A possible avenue for producing an aggregation-resistant version of G-CSF would be to reduce the mobility of this loop, for instance by introducing an additional disulphide linkage between the latter half of the loop and helix B to further restrict its motion. However, this could greatly reduce the yield of native G-CSF, either during refolding of recombinantly expressed G-CSF from insoluble inclusion bodies or during disulphide formation in the periplasm of *E. coli*, as disulphide bonds must also be formed between Cys37-Cys43 and Cys65-Cys75 at the N- and C-terminal ends of the AB loop, respectively in order for G-CSF to attain its native conformation. Additionally, residues in the helical section of the AB loop are involved in the minor site of interaction of G-CSF with its receptor in the ligand-receptor 2:2 complex (see Section 1.6.2), which is required for cellular signal transduction. Consequently, attempting to fix the conformation of the AB loop of G-CSF may hinder its functional activity.

Considering no residues of helix B appear to be involved in receptor binding (see Section 1.6.2), there is potential for mutation of Tyr86 to a charged residue such as Arg or Glu to try and reduce the aggregation propensity of G-CSF C3. The functional activities of G-CSF wt and C3 are equivalent, demonstrated by respective EC_{50} values of 6 and 2.6 µM measured in cell proliferation assays[150]. G-CSF wt and C3 exhibit these equivalent functional activities even though G-CSF C3 contains the substitutions W59R and F84L, which are in close proximity to Tyr86, further supporting the notion that this local region of G-CSF is not involved in function.

Although not directly involved in receptor binding, Tyr86 may be involved in the hydrophobic core of G-CSF, stabilising the native conformation that is required for receptor recognition. Tyr86 is buried within the structure of G-CSF as shown by its low solvent accessibility (Figure 5.18B), plus the residues around Tyr86 in the sequence are largely hydrophobic (GLLLYQGLL) suggesting that this residue may be involved in hydrophobic interactions in the core of G-CSF and this presumably makes this region more aggregation-prone.

In order to assess whether the region encompassing Tyr86 is involved in stabilising the native conformation, the single-residue local frustration (srlf) index of G-CSF C3 was calculated (Figure 6.1) using the Frustratometer[413]. This prediction program is based on the principle of minimal frustration, which states that the energy of a protein decreases as it assumes conformations that are more like the ground (native) state. The reduction in energy is of a greater magnitude than can be accounted for by chance and so allows an unfolded polypeptide to traverse the energy landscape to the native state, overcoming kinetic traps and the reduced entropy of the unfolded polypeptide[19]. However, within the ensemble of native conformations there may still be frustration in the energy landscape, arising from a particular arrangement of amino acids that is not optimal for the stability of the native conformation but may be essential for function. In this way, proteins are able to harness the roughness of their energy landscape, facilitating dynamic motions between different functional conformations[31,32].

The Frustratometer[413] quantifies the degree of local frustration in a protein structure by assessing the contacts made by a particular residue and then how energetically favourable these contacts are relative to the contacts that would be made by every other residue occupying this position. Minimally frustrated residues exhibit scores greater than 0.78 and are predicted to be involved in the folding core of a protein,

having formed contacts that are among the most energetically favoured. On the other hand, maximally frustrated residues exhibit scores more negative than -1, having formed contacts that are energetically unfavourable and so are predicted to be in regions of a protein required for function[413]. As a result of the predicted link between locally frustrated regions of proteins and their function, the structure of G-CSF C3 both free in solution and in complex with the G-CSF receptor were submitted for local frustration calculation.

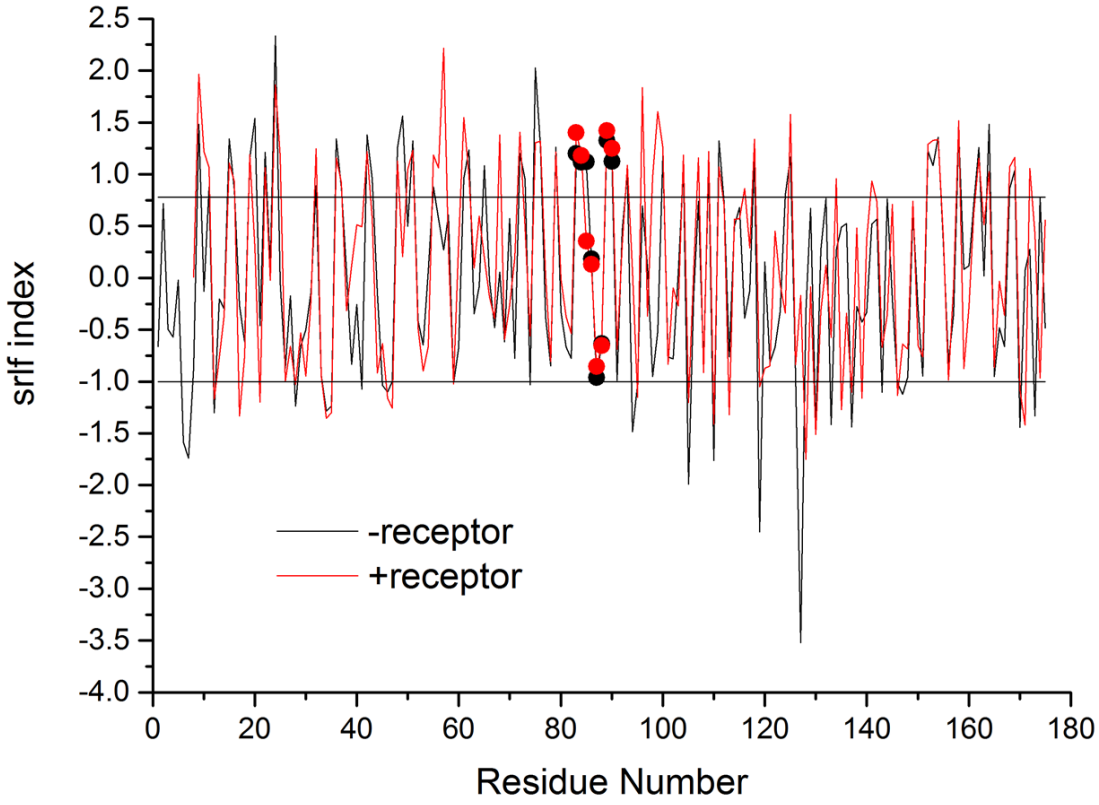


Figure 6.1 Local frustration scores for residues in the G-CSF C3 structure. Single-residue local frustration (srlf) scores were predicted using the Frustratometer[413] and the G-CSF wt structure (PDB 1GNC[283]), with residues being mutated in PyMOL[289] for calculation of G-CSF C3 srlf scores (black). For calculation of srlf scores of G-CSF C3 bound to the G-CSF receptor (red) the crystal structure of G-CSF wt in complex with this receptor was used (PDB 2D9Q[287]) after deletion of the receptor in PyMOL[289]. Scores for residues 83-90 are shown as filled data points.

Although the predicted frustration score for Tyr86 appears neutral, the surrounding residues 83-85 and 89-90 produce scores above 0.78, suggesting that these residues are minimally frustrated (Figure 6.1). This means that the contacts made by these residues are predicted to be among the most energetically favourable that can be made by residues in these positions and may be important in stabilising the native conformation of G-CSF. Consequently, mutation of Tyr86 to a charged residue could affect the structure and thermodynamic stability of the native

conformation of G-CSF, potentially affecting its receptor binding activity or even counteracting the reduced propensity for aggregation hypothesised.

The work in Chapter 4 showed that the substitutions found in G-CSF C3 do not appear to alter aggregation behaviour significantly in relation to the wild-type protein, but they dramatically increase the recombinant expression level of G-CSF, regardless of whether the protein is being expressed as soluble native protein in the periplasm of *E. coli*[150] or as insoluble aggregates in the cytoplasm (Figure 3.2). This suggests that a change in another factor relating to the recombinant expression of these proteins (e.g. translation levels) is caused by one or more of these substitutions, which is discussed in more detail in Section 3.3.

The substitutions found in G-CSF C3 were identified from screening of a randomly generated library of G-CSF DNA constructs[150]. As such, it is unknown if all four of these substitutions are required for the observed improvement in recombinant expression. Recombinant expression trials of G-CSF constructs containing only single substitutions would help identify the mutation and subsequently the region of G-CSF responsible for the observed increase in heterologous expression level, which is presumed to result from factors other than protein aggregation (see Section 3.3). In addition, this would allow removal of unnecessary mutations from G-CSF variants, providing a system more closely related to the wild-type protein for further investigation of the link between G-CSF folding, dynamics, aggregation and recombinant expression.

Aggregation of therapeutic proteins from the native state is not only observed for G-CSF. IFN-α-2a is a four helical cytokine used in cancer and viral disease treatment[131] that aggregates in the presence of benzyl alcohol, attributed to partial unfolding of the native state[188]. IFN-γ is another four helical cytokine that exhibits antitumor and anti-infective activities[267] and has been shown to aggregate in low denaturant concentrations[410,414] via partial unfolding of the native state[410].

Protein aggregation from native-like species is also linked to a range of diseases. For instance, protein aggregation from partial unfolding of the native state has been shown for γS-crystallin[89,90], which is involved in cataract formation[91]. TTR has also been shown to aggregate from a native-like state[92,93], causing a range of amyloidoses[94,415]. In addition, a variant of the protein DJ-1 has been linked to a rare form of inherited Parkinson's disease and has been shown to result from transient

exposure of hydrophobic surfaces in the otherwise native conformation of this protein[409].

Although the factors surrounding the immunogenicity of protein drugs are wide ranging[155], aggregates are considered to be one of the main causes[156,157]. Little is known about the mechanism of the immunological response to recombinantly produced human therapeutic proteins[157], despite the fact that such immunogenic responses are potentially life-threatening[156,164,416], which coupled to the difficulties in detecting and characterising the wide range of aggregate sizes that occur[162,165,166] and uncertainty in the aggregate species responsible for eliciting immunogenic responses[156], means that every effort must be made to prevent aggregation in protein drug formulations. Currently, pharmaceutical protein preparation is based generally on trial-and-error exploration of a wide range of buffer conditions[131,210]. However, with a better understanding of the mechanisms of protein drug aggregation and the conformational substates involved, more rational approaches can be developed to avoid population of these states.

Chapter 7 Appendices

7.1 DNA sequences of G-CSF constructs

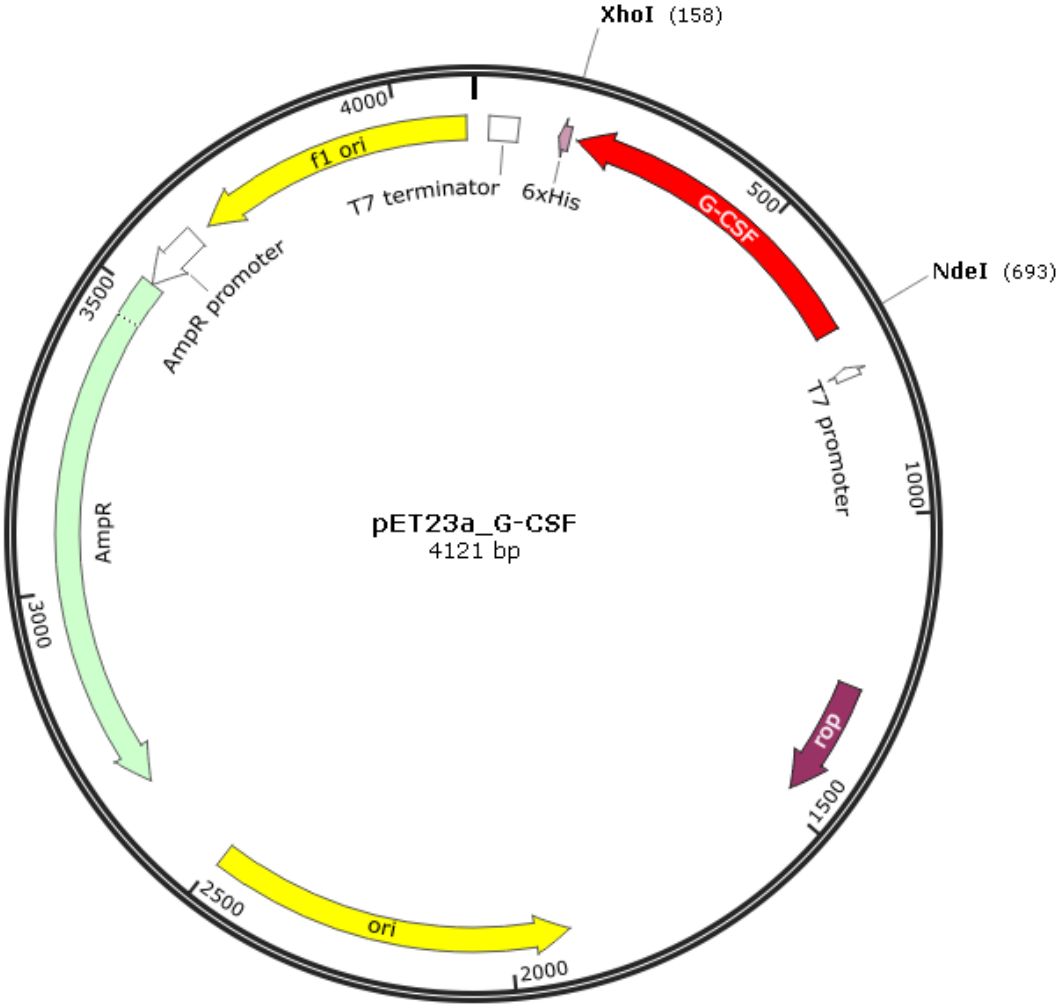
```
1   ATGACTCCTC TCGGTCCTGC ATCTAGCCTT CCGCAAAGCT TCCTGCTGAA
51  GTGCCTTGAA CAGGTCCGCA AAATCCAGGG TGATGGCGCT GCCCTGCAGG
101 AGAAACTGTG TGCGACCTAC AAACTGTGCC ATCCGGAAGA ACTGGTGCTG
151 TTGGGCCACT CTCTGGGCAT TCCGTGGGCA CCCCTGAGTA GCTGTCCATC
201 CCAAGCCCTG CAATTAGCCG GCTGCTTATC GCAGCTGCAC TCAGGACTCT
251 TCCTGTATCA GGGCCTGTTA CAGGCGTTAG AGGGGATTTC CCCGGAACTG
301 GGTCCCACAC TCGATACGCT TCAGTTGGAT GTTGCGGACT TTGCAACCAC
351 CATCTGGCAG CAGATGGAGG AACTGGGGAT GGCACCAGCG TTGCAACCGA
401 CGCAAGGTGC GATGCCGGCC TTTGCGAGTG CCTTTCAGCG TCGTGCTGGA
451 GGTGTACTCG TGGCTTCACA TCTGCAGTCG TTCCTGGAAG TGAGCTATCG
501 CGTTCTGCGC CATTTGGCGC AACCGTAATA A
```

Appendix 7.1 DNA sequence of G-CSF wt. Start and stop codons are underlined. Codons encoding different amino acids in G-CSF wt and C3 are highlighted in red. All other nucleotide variations between G-CSF wt and C3 DNA sequences are silent and not highlighted. The nucleotide at the beginning of each line is numbered on the left.

```
1   ATGACCCCGT TAGGCCCTGC CTCCTCTCTT CCGCAATCCT TCCTGCTGAA
51  AGGCCTGGAA CAGGTGCGCA AAATCCAGGG TGATGGTGCA GCACTGCAAG
101 AGAAGCTGTG TGCGACCTAC AAACTGTGCC ATCCGGAAGA GCTGGTGTTG
151 CTCGGCCATA GCCTGGGTAT TCCGCGTGCA CCCCTCAGCA GCTGTCCGTC
201 ACAGGCGTTA CGTCTGGCCG GTTGCCTTAG TCAGCTGCAC TCTGGACTGT
251 TGCTGTATCA GGGGTTACTC CAAGCCCTGG AAGGCATTTC GCCGGAACTG
301 GGTCCAACTC TGGATACGCT GCAGTTGGAC GTTGCGGATT TTGCCACGAC
351 CATCTGGCAG CAGATGGAGG AACTTGGCAT GGCTCCAGCG CTGCAACCCA
401 CACAAGGGGC AATGCCTGCG TTTGCTTCAG CGTTTCAACG GCGTGCGGGA
451 GGCGTACTGG TTGCCAGTCA CTTGCAGAGC TTCCTCGAAG TCTCGTATCG
501 CGTGTTACGC CATCTTGCTC AGCCGTAATA A
```

Appendix 7.2 DNA sequence of G-CSF C3. Start and stop codons are underlined. Codons encoding different amino acids in G-CSF wt and C3 are highlighted in red. All other nucleotide variations between G-CSF wt and C3 DNA sequences are silent and not highlighted. The nucleotide at the beginning of each line is numbered on the left.

7.2 Plasmid map of G-CSF constructs



Appendix 7.3 Plasmid map of G-CSF in pET23a.

7.3 Amino acid sequences of G-CSF proteins

```
1   MTPLGPASSL PQSFLLKCLE QVRKIQGDGA ALQEKLCATY KLCHPEELVL
51  LGHSLGIPWA PLSSCPSQAL QLAGCLSQLH SGLFLYQGLL QALEGISPEL
101 GPTLDTLQLD VADFATTIWQ QMEELGMAPA LQPTQGAMPA FASAFQRRAG
151 GVLVASHLQS FLEVSYRVLR HLAQP
```

Appendix 7.4 Amino acid sequence of G-CSF wt. Mutated residues in G-CSF wt and C3 amino acid sequences are highlighted in red. Cys residues linked by disulphide bonds are highlighted in the same colour (blue or green). The amino acid at the beginning of each line is numbered on the left.

```
1   MTPLGPASSL PQSFLLKGLE QVRKIQGDGA ALQEKLCATY KLCHPEELVL
51  LGHSLGIPRA PLSSCPSQAL RLAGCLSQLH SGLLLYQGLL QALEGISPEL
101 GPTLDTLQLD VADFATTIWQ QMEELGMAPA LQPTQGAMPA FASAFQRRAG
151 GVLVASHLQS FLEVSYRVLR HLAQP
```

Appendix 7.5 Amino acid sequence of G-CSF C3. Mutated residues in G-CSF wt and C3 amino acid sequences are highlighted in red. Cys residues linked by disulphide bonds are highlighted in the same colour (blue or green). The amino acid at the beginning of each line is numbered on the left.

7.4 Backbone assignments of G-CSF proteins

7.4.1 Backbone assignments of G-CSF wt at pH 4

Residue Number	Amino acid	^{1}H / ppm	^{15}N / ppm
2	Thr	8.68	121.12
4	Leu	8.33	122.56
5	Gly	8.22	109.69
7	Ala	8.45	123.42
8	Ser	8.19	114.26
9	Ser	8.04	116.37
10	Leu	8.4	122.09
12	Gln	9.1	125.53
13	Ser	8.84	112.46
14	Phe	7.06	122.95
15	Leu	7.71	121.48
16	Leu	8.47	117.22
17	Lys	7.53	120.85
18	Cys	8.12	116.72
19	Leu	8.14	118.98
20	Glu	7.63	118.43
21	Gln	8.03	121.23
22	Val	8.36	119.82

23	Arg	7.57	117.84
24	Lys	7.96	122.17
25	Ile	8.25	119.59
26	Gln	8.84	119.99
27	Gly	8.01	107.72
28	Asp	8.43	124.18
29	Gly	9.29	109.54
30	Ala	8.23	125.73
31	Ala	7.91	121.18
32	Leu	7.59	118.69
33	Gln	8.46	118
34	Glu	8.69	120.4
35	Lys	8.07	120.52
36	Leu	8.59	122.09
37	Cys	7.96	118.2
38	Ala	8.81	123.3
39	Thr	8.17	110.37
40	Tyr	7.12	115.7
41	Lys	7.68	114.49
42	Leu	6.98	121.48
43	Cys	8.03	118.63
44	His	8.53	117.12
46	Glu	10.11	118.66
47	Glu	7.65	117.43
48	Leu	7.7	118.46
49	Val	7.16	119.42
50	Leu	8.52	121.96
51	Leu	7.97	122.31
52	Gly	8.31	105.52
53	His	8.04	118.63
54	Ser	8.5	116.9
55	Leu	8.09	119.71
56	Gly	7.6	107.02
57	Ile	7.41	119.62
59	Trp	7.82	120.19
60	Ala	7.25	126.21
62	Leu	8.55	120.78
63	Ser	7.52	116.35
64	Ser	9.74	118.51
65	Cys	7.95	118.2
67	Ser	8.31	113.29
68	Gln	7.87	118.14
69	Ala	8.35	123.55
70	Leu	7.77	120.69
71	Gln	8.75	125.67
72	Leu	8.66	126.25
73	Ala	8.86	120.22
74	Gly	7.88	105.95
75	Cys	8.24	120.51
76	Leu	8.78	117.72
77	Ser	8.18	112.15
78	Gln	8.06	122.86

79	Leu	8.99	124.53
80	His	8.97	116.6
81	Ser	9.15	115
82	Gly	8.62	111.3
83	Leu	8.12	121.4
84	Phe	8.15	119.44
85	Leu	7.97	123.04
86	Tyr	8.12	116.86
87	Gln	8.87	120.18
88	Gly	7.71	107.38
89	Leu	8.44	123.72
90	Leu	8.44	115.69
91	Gln	7.97	119.67
92	Ala	7.88	121.75
93	Leu	7.24	115.92
94	Glu	7.92	115.39
95	Gly	8.98	107.38
96	Ile	7.59	110.76
99	Glu	8.95	115.4
100	Leu	7.77	121.3
101	Gly	7.91	108.34
103	Thr	7.56	114.41
104	Leu	8.23	124.9
105	Asp	9.06	120.46
106	Thr	7.9	117.02
107	Leu	7.2	120.51
108	Gln	8.71	116.97
109	Leu	8.2	121.32
110	Asp	8.43	119.78
111	Val	8.61	122.73
112	Ala	8.37	122.92
113	Asp	8.67	117.97
114	Phe	7.89	121.82
115	Ala	8.87	121.49
116	Thr	8.73	116.46
117	Thr	8.44	121.41
118	Ile	8.24	120.42
119	Trp	8.77	122
120	Gln	8.95	116.05
121	Gln	7.58	119.5
122	Met	8.25	118.38
123	Glu	8.23	118.62
124	Glu	7.86	120.66
125	Leu	7.39	116.27
126	Gly	7.72	108.22
127	Met	7.97	116.26
128	Ala	7.71	123.52
130	Ala	8.24	122.42
131	Leu	7.94	119.9
132	Gln	8.36	122.77
134	Thr	8.19	114.94
135	Gln	8.67	124.34

136	Gly	8.32	111.29
137	Ala	8.11	122.85
138	Met	8.35	121.83
140	Ala	7.91	122.31
141	Phe	8.4	120.57
142	Ala	8.2	124.93
143	Ser	7.85	111.65
144	Ala	9.03	124.64
145	Phe	8.72	117.16
146	Gln	7.89	116.06
147	Arg	7.97	117.47
148	Arg	7.85	118.12
149	Ala	8.29	120.12
150	Gly	8.28	106.44
151	Gly	7.92	109.01
152	Val	7.36	121.06
153	Leu	8.12	120.01
154	Val	9.41	119.95
155	Ala	8.25	119.77
156	Ser	8.09	112.39
157	His	8.15	120.97
158	Leu	8.7	122.55
159	Gln	8.2	117.76
160	Ser	7.66	115.68
161	Phe	8.39	124.7
162	Leu	9.08	120.29
163	Glu	8.13	121.42
164	Val	7.48	119.48
165	Ser	8.13	115.96
166	Tyr	8.79	121.8
167	Arg	7.52	117.24
168	Val	8.22	120.27
169	Leu	8.37	119.62
170	Arg	8.56	117.8
171	His	7.72	116.83
172	Leu	7.78	118.09
173	Ala	7.67	120.15
174	Gln	7.76	118.79

Appendix 7.6 Backbone assignments of G-CSF wt at pH 4. Assignments of backbone amide proton and nitrogen nuclei were carried out using previous assignments[283,288]. The ^{1}H-^{15}N HSQC spectrum used for assignments was acquired using 111 µM ^{15}N G-CSF wt in 18 mM sodium phosphate, 18 mM sodium acetate, 10 % (*v/v*) D_2O pH 4 (I = 39 mM) buffer at 25 ^{o}C, using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

7.4.2 Backbone assignments of G-CSF C3 at pH 4

Residue Number	Amino acid	^{1}H / ppm	^{15}N / ppm	^{13}C / ppm	$^{13}C_{\alpha}$ / ppm	$^{13}C_{\beta}$ / ppm
1	Met	-	-	172.23	55.12	32.65
2	Thr	8.61	121.2	174.03	60.15	-
3	Pro	-	-	176.66	63.2	32.11
4	Leu	8.27	122.64	177.63	55.25	42.61
5	Gly	8.14	109.76	-	44.77	-
6	Pro	-	-	177.04	63.37	32.11
7	Ala	8.37	123.68	177.69	52.67	19.16
8	Ser	8.13	114.35	174.02	58.56	64.13
9	Ser	7.99	116.56	173.63	57.97	-
10	Leu	8.28	122.58	174.21	52.35	-
11	Pro	-	-	177.99	62.79	32.05
12	Gln	8.92	124.43	177.95	59.12	28.42
13	Ser	8.6	112.5	177.01	60.93	61.51
14	Phe	7.12	122.49	176.55	60.21	39.28
15	Leu	7.43	120.02	178.23	57.93	41.9
16	Leu	8.26	116.92	180.26	58.02	41.34
17	Lys	7.54	120.29	178.92	58.95	32.11
18	Gly	8.54	107.96	174.8	47.18	-
19	Leu	8.12	121.15	179.32	58.66	41.52
20	Glu	7.4	118.66	179.41	59.04	28.41
21	Gln	8.37	121.16	178.9	59.36	27.97
22	Val	8.34	120.1	177.58	67.45	31.38
23	Arg	7.53	118.3	179.83	58.89	29.98
24	Lys	7.89	122.11	179.08	59.29	32.59
25	Ile	8.15	119.42	178.8	65.14	37.9
26	Gln	8.77	119.88	180.34	59.52	28.86
27	Gly	8.03	107.83	177.01	47.37	-
28	Asp	8.4	124.36	179.38	56.01	39.38
29	Gly	9.19	109.54	175.15	46.9	-
30	Ala	8.2	125.7	180.44	54.91	17.67
31	Ala	7.86	121.45	180.45	55.09	17.76
32	Leu	7.5	118.79	177.87	58.39	41.59
33	Gln	8.35	117.99	179.23	60.16	29.26
34	Glu	8.63	120.45	179.04	59.41	28.05
35	Lys	8.04	120.58	180.34	59.36	32.89
36	Leu	8.48	121.9	179	59.13	41.66
37	Cys	7.87	118.19	176.96	57.2	37.99
38	Ala	8.75	123.4	179.47	55.34	18.88
39	Thr	8.1	110.46	175.3	65.62	69.94
40	Tyr	7.05	115.82	174.71	57.04	39.68
41	Lys	7.61	114.53	176.22	57.66	28.68
42	Leu	6.92	121.55	175.4	53.55	42.2
43	Cys	7.96	118.58	173.62	57.11	47.8
44	His	8.39	116.96	173.55	52.46	-
45	Pro	-	-	177.95	65.53	31.86
46	Glu	9.96	118.65	177.98	59.21	27.64
47	Glu	7.59	117.62	177.18	57.17	29.62

48	Leu	7.62	118.64	177	54.26	42.46
49	Val	7.14	119.54	178.06	65.62	32.04
50	Leu	8.34	122.13	179.59	57.83	40.98
51	Leu	7.93	122.27	179.21	57.18	41.32
52	Gly	8.3	106.16	175.68	47.97	-
53	His	7.94	118.65	176.82	57.54	28.25
54	Ser	8.32	116.73	175.71	60.8	63.07
55	Leu	8.12	120.15	177.27	55.54	42.98
56	Gly	7.61	107.05	174.78	46.69	-
57	Ile	7.74	120.12	174.3	58.96	-
58	Pro	-	-	175	62.72	32.54
59	Arg	8.2	118.07	175.15	54.52	31.84
60	Ala	8.16	126.65	175	49.51	-
61	Pro	-	-	176.92	63.12	32.8
62	Leu	8.38	121.85	179.32	54.38	43.18
63	Ser	7.26	115.65	176.08	60.3	-
64	Ser	10	118.9	173.69	59.38	63.04
65	Cys	7.87	118.73	171.4	54.01	-
66	Pro	-	-	176.84	64.42	31.36
67	Ser	8.28	114.89	175.18	61.18	62.11
68	Gln	8.07	117.92	175.05	56.03	29.11
69	Ala	7.98	124.25	175.17	52.11	17.95
70	Leu	7.59	121.91	177.29	56.72	42.24
71	Arg	8.52	127.78	176.22	54.62	29.8
72	Leu	8.58	125.96	177.61	59.82	42.02
73	Ala	8.83	120.19	180.71	56.22	17.86
74	Gly	7.76	105.76	175.78	47.18	-
75	Cys	8.45	120.71	176.32	58.23	38.99
76	Leu	8.75	118.06	178.83	58.75	40.5
77	Ser	7.92	112.09	177.57	62.38	62.75
78	Gln	8.09	122.55	180.94	59.75	28.53
79	Leu	9.13	124	178.65	58.63	42.16
80	His	8.85	117.11	176.11	62.44	28.83
81	Ser	9.03	114.02	177.62	62.25	62.51
82	Gly	8.59	111.17	174.53	47.74	-
83	Leu	8	121.89	178.54	58.43	41.88
84	Leu	7.46	118.53	180.46	58	40.81
85	Leu	7.79	123.78	179.19	58.3	40.65
86	Tyr	8.03	116.87	177.87	64.4	37.55
87	Gln	8.83	120.61	178.59	60.56	29.1
88	Gly	7.87	106.53	177.11	47.47	-
89	Leu	8.45	124.1	178.1	58.02	43.19
90	Leu	8.42	116.08	180.36	57.83	41.8
91	Gln	8.03	119.79	179.03	59.18	28.3
92	Ala	7.79	121.67	178.06	54.13	17.96
93	Leu	7.24	116.11	176.8	55.79	41.14
94	Glu	7.84	115.59	176.74	56.57	26.62
95	Gly	8.93	107.57	173.85	46.4	-
96	Ile	7.54	111.27	-	-	-
98	Pro	-	-	179.08	65.82	31.67
99	Glu	8.86	115.49	177.55	58.81	28.51
100	Leu	7.71	121.11	177.09	54.51	42.73

101	Gly	7.81	108.36	-	49.45	-
102	Pro	-	-	179.3	65.9	32.01
103	Thr	7.47	114.68	177.05	66.71	67.09
104	Leu	8.15	124.87	178.5	58.04	41.22
105	Asp	8.94	120.86	178.33	57.15	38.93
106	Thr	7.87	117.4	174.93	66.94	-
107	Leu	7.25	120.72	177.88	57.99	42.37
108	Gln	8.53	116.78	178.37	60.52	29.51
109	Leu	8.02	121.5	179.28	58.25	41.69
110	Asp	8.32	119.72	180.02	57.13	38.92
111	Val	8.49	122.6	177.04	67.58	-
112	Ala	8.18	122.92	180.7	55.49	18.34
113	Asp	8.53	118.01	178.66	56.9	39.76
114	Phe	7.86	121.41	177.34	57.35	39.79
115	Ala	8.73	121.76	180.2	56.35	17.34
116	Thr	8.49	116.24	176.08	66.93	67.43
117	Thr	8.26	121.14	177.47	67.01	-
118	Ile	8.26	120.96	177.16	65.2	32.69
119	Trp	8.61	121.86	177.48	60.63	29.77
120	Gln	8.71	116.43	178.87	58.98	28.22
121	Gln	7.54	119.18	178.52	57.83	29.1
122	Met	8.19	118.64	178.45	60.15	32.98
123	Glu	8.13	118.6	180.28	58.82	28.13
124	Glu	7.85	120.63	177.78	58.75	28.15
125	Leu	7.3	116.52	177.05	54.61	42.42
126	Gly	7.65	108.32	175.18	46.54	-
127	Met	7.85	116.85	174.34	55.24	33.75
128	Ala	7.86	124.34	174.74	50.31	-
129	Pro	-	-	175.98	62.52	31.96
130	Ala	8.19	123.09	177.75	52.67	19.25
131	Leu	8.04	120.94	176.75	54.8	42.78
132	Gln	8.36	123.27	173.84	53.45	-
133	Pro	-	-	176.73	62.92	32.13
134	Thr	8.17	115.63	174.41	61.94	69.94
135	Gln	8.46	124	176.07	55.8	29.87
136	Gly	8.22	110.5	172.78	44.78	-
137	Ala	8.06	123.21	177.66	52.18	19.23
138	Met	8.33	121.65	174.16	52.27	-
139	Pro	-	-	174.25	62.27	31.48
140	Ala	7.85	122.48	176.71	50.91	19.52
141	Phe	8.34	120.49	175.51	56.53	39.37
142	Ala	8.19	125.03	177.32	53.99	20.08
143	Ser	7.76	111.62	174.32	56.08	66.8
144	Ala	-	-	179.9	55.94	17.64
144	Ala	8.97	124.76	-	-	-
145	Phe	8.63	117.27	176.97	61.87	38.88
146	Gln	7.79	116.16	178.44	59.83	28.54
147	Arg	7.88	117.55	177.35	59.81	30.58
148	Arg	7.75	118.14	176.92	59.63	30.36
149	Ala	8.14	120.15	179.94	54.71	18.54
150	Gly	8.18	106.15	175.16	47.28	-
151	Gly	7.82	109.03	173.96	48.14	-

152	Val	7.29	121.05	178.89	66.52	31.55
153	Leu	7.85	120.35	178.94	58.3	43.17
154	Val	9.12	119.33	177.49	67.79	31.67
155	Ala	8.14	119.8	179.25	55.97	17.45
156	Ser	8	112.83	178.09	61.64	62.38
157	His	8.46	120.88	177.65	57.96	-
158	Leu	8.76	122.74	177.97	57.82	41.28
159	Gln	8.15	118.15	179.65	59.54	27.77
160	Ser	7.69	115.83	176.48	62.25	62.45
161	Phe	8.12	125.01	178.7	61.38	39.53
162	Leu	9.03	120.75	178.64	57.03	41.45
163	Glu	7.98	121.06	179.48	59.63	28.54
164	Val	7.33	119.77	178.11	66.69	31.11
165	Ser	7.97	115.48	174.84	63.41	-
166	Tyr	8.65	121.36	176.13	62.14	38.48
167	Arg	7.57	117.74	179.73	59.92	30.01
168	Val	8.37	120.6	178.03	66.37	32.27
169	Leu	8.02	118.64	178.88	58.5	40.95
170	Arg	8.49	117.46	178.75	58.74	29.24
171	His	7.85	117	176.57	58.1	28.12
172	Leu	7.7	117.93	177.39	56.46	42.4
173	Ala	7.62	120.56	177.2	52.88	19.12
174	Gln	7.66	118.91	172.2	53.54	-

Appendix 7.7 Backbone assignments of G-CSF C3 at pH 4. Assignments of backbone amide proton and nitrogen nuclei plus alpha, beta and carbonyl carbon atoms were carried out using an ^{1}H-^{15}N HSQC spectrum in combination with HNCA, HN(CO)CA, CBCA(CO)NH, HNCO and HN(CA)CO triple-resonance spectra. Spectra were acquired using 473 µM ^{13}C ^{15}N G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 10 % (*v/v*) D_2O pH 4 buffer (I = 39 mM) at 25 ^{o}C, using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

7.4.3 Backbone assignments of G-CSF C3 at pH 7

Residue Number	Amino acid	^{1}H / ppm	^{15}N / ppm	^{13}C / ppm	$^{13}C_{\alpha}$ / ppm	$^{13}C_{\beta}$ / ppm
3	Pro	-	-	176.71	63.24	32.1
4	Leu	8.26	122.34	177.62	55.24	42.7
5	Gly	8.13	109.61	-	44.77	-
6	Pro	-	-	177.04	63.38	32.08
7	Ala	8.38	123.57	177.66	52.69	19.34
8	Ser	8.11	114.26	174.02	58.57	64.19
9	Ser	8.01	116.48	173.65	57.99	64.37
10	Leu	8.26	122.63	174.28	52.41	-
11	Pro	-	-	177.89	62.8	-
12	Gln	8.87	123.98	177.75	59.05	29.01
13	Ser	8.58	112.53	176.91	60.91	61.42
14	Phe	7.17	122.45	176.44	60.3	-
15	Leu	7.42	119.74	178.22	57.87	-
16	Leu	8.23	116.61	180.24	57.97	-
17	Lys	7.55	120.35	178.92	58.98	-
18	Gly	8.56	107.77	174.83	47.18	-
19	Leu	8.08	121.07	179.25	58.72	-
20	Glu	7.34	118.64	179.8	59.18	-
21	Gln	8.34	120.9	178.73	59.4	-
22	Val	8.36	119.75	177.55	67.48	-
23	Arg	7.51	118.02	179.76	58.79	-
24	Lys	7.88	121.96	179.16	59.48	-
25	Ile	8.15	119.45	178.74	65.09	-
26	Gln	8.78	119.75	180.48	59.58	-
27	Gly	8.01	107.49	176.94	47.36	-
28	Asp	8.35	124.43	179.41	56.21	39.67
29	Gly	9.2	109.61	175.07	46.86	-
30	Ala	8.19	125.37	180.44	54.9	17.67
31	Ala	7.86	121.39	180.49	55	17.69
32	Leu	7.52	119.09	177.83	58.4	-
33	Gln	8.3	118.01	179.22	60.14	-
34	Glu	8.59	120.79	179.2	59.78	29.26
35	Lys	8.02	120.11	180.5	59.44	-
36	Leu	8.49	121.86	179.11	59.19	-
37	Cys	7.85	118.23	177.06	57.12	37.91
38	Ala	8.79	123.28	179.57	55.33	18.9
39	Thr	8.1	110.33	175.32	65.7	69.34
40	Tyr	7.08	115.74	174.52	57.05	-
41	Lys	7.63	114.56	176.03	57.61	28.75
42	Leu	6.86	121.27	175.38	53.53	-
43	Cys	8.11	118.87	-	57.61	-
45	Pro	-	-	178.19	65.59	-
46	Glu	9.97	118.26	178.5	59.63	-
47	Glu	7.52	117.22	177.59	57.55	-
48	Leu	7.6	117.81	177.17	54.48	-
49	Val	7.14	119.53	178.24	65.97	32.01
50	Leu	8.25	120.96	179.62	57.91	42.54

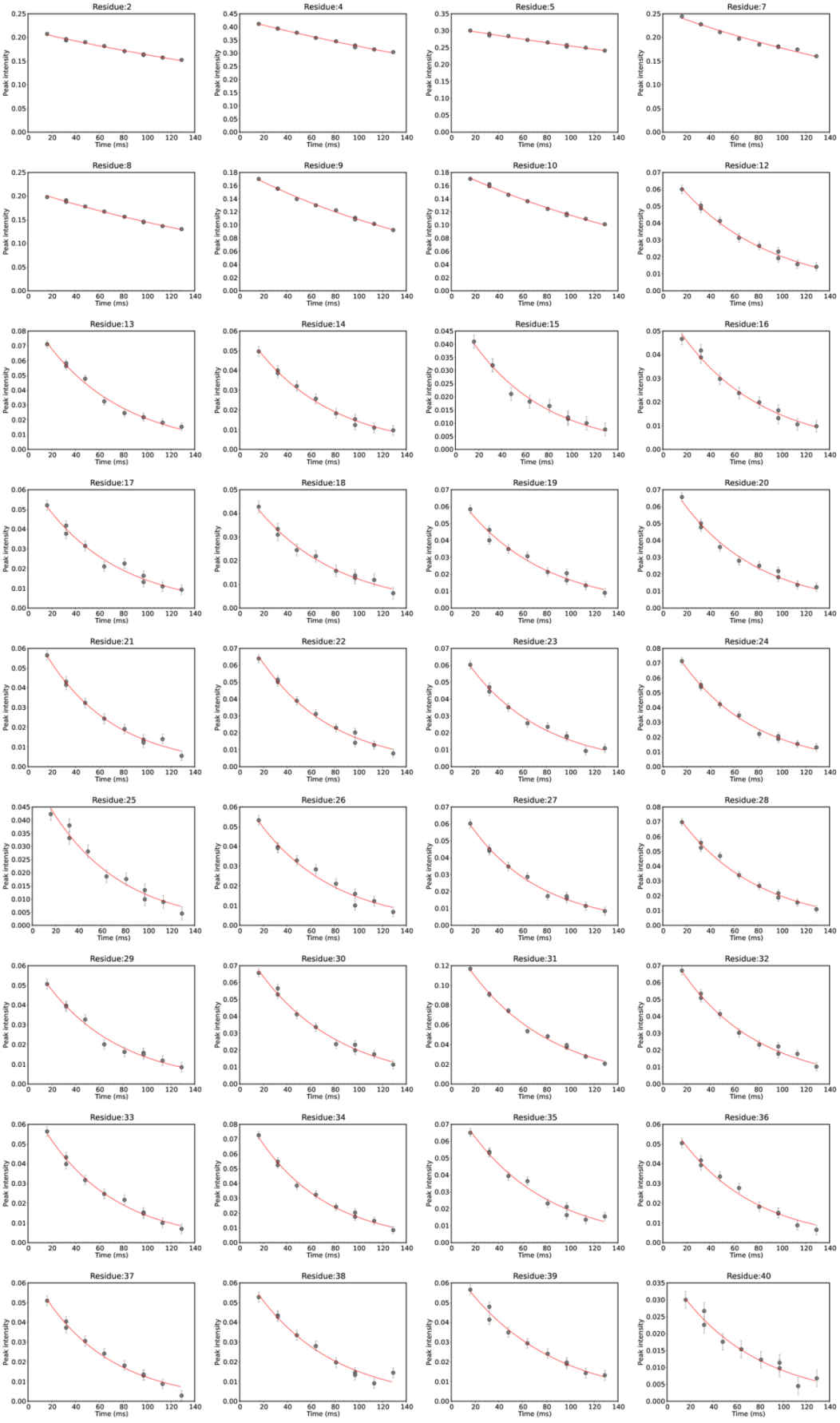
51	Leu	7.75	121.34	179.02	57.12	41.71
52	Gly	8.18	106.32	175.76	48.02	-
53	His	7.82	119.94	177.98	58.39	30.9
54	Ser	8.2	116.18	176.01	60.78	-
55	Leu	8.13	120.01	177.34	55.62	-
56	Gly	7.68	107.1	174.85	46.58	-
57	Ile	7.69	120.07	173.98	59.23	38.22
58	Pro	-	-	-	62.49	-
59	Arg	8.01	116.67	175.79	54.04	32.01
60	Ala	8.02	127.08	175.19	49.72	19.5
61	Pro	-	-	-	63	-
62	Leu	8.35	121.54	179.3	54.38	-
63	Ser	7.26	115.38	-	60.35	-
64	Ser	10	118.87	173.66	59.5	-
65	Cys	7.87	118.87	-	54.13	-
66	Pro	-	-	177.08	64.66	-
67	Ser	8.26	115.36	175.29	61.22	62.27
68	Gln	8.1	117.78	175.09	56.4	28.89
69	Ala	7.91	124.24	175.22	52.01	18.02
70	Leu	7.58	122.2	177.31	56.72	-
71	Arg	8.45	127.71	-	54.65	-
72	Leu	8.57	125.74	177.45	59.97	-
73	Ala	8.79	119.97	180.6	56.06	17.95
74	Gly	7.73	105.4	175.65	47.3	-
75	Cys	8.5	120.45	176.2	58.32	-
76	Leu	8.68	118	178.52	58.82	-
77	Ser	7.78	111.6	177.47	62.27	62.73
78	Gln	8.01	122.69	180.66	59.63	-
79	Leu	9.05	123.96	178.43	58.56	-
80	His	8.63	118.76	176.6	63.24	-
81	Ser	8.68	111.93	178.05	61.94	62.62
82	Gly	8.49	110.45	174.23	47.59	-
83	Leu	8.04	122.28	178.81	58.49	-
84	Leu	7.67	118.46	180.67	58.02	-
85	Leu	7.79	123.76	179.09	58.35	-
86	Tyr	8.05	116.87	177.9	64.13	-
87	Gln	8.84	120.55	178.49	60.51	29.05
88	Gly	7.85	106.28	177.16	47.45	-
89	Leu	8.52	124.12	178.09	58.05	-
90	Leu	8.45	115.98	180.63	57.88	-
91	Gln	8.04	119.8	179.12	59.32	28.18
92	Ala	7.83	121.75	177.78	54.07	17.93
93	Leu	7.19	115.72	176.7	55.87	-
94	Glu	7.87	116.09	176.78	57.21	27.76
95	Gly	9.07	107.71	173.64	46.59	38.22
96	Ile	7.53	110.3	-	-	-
98	Pro	-	-	179	65.9	-
99	Glu	8.93	115.16	177.69	58.99	-
100	Leu	7.73	121.29	177.14	54.54	42.64
101	Gly	7.81	108.46	-	49.59	-
102	Pro	-	-	179.35	65.79	-
103	Thr	7.47	114.69	176.89	66.76	67.33

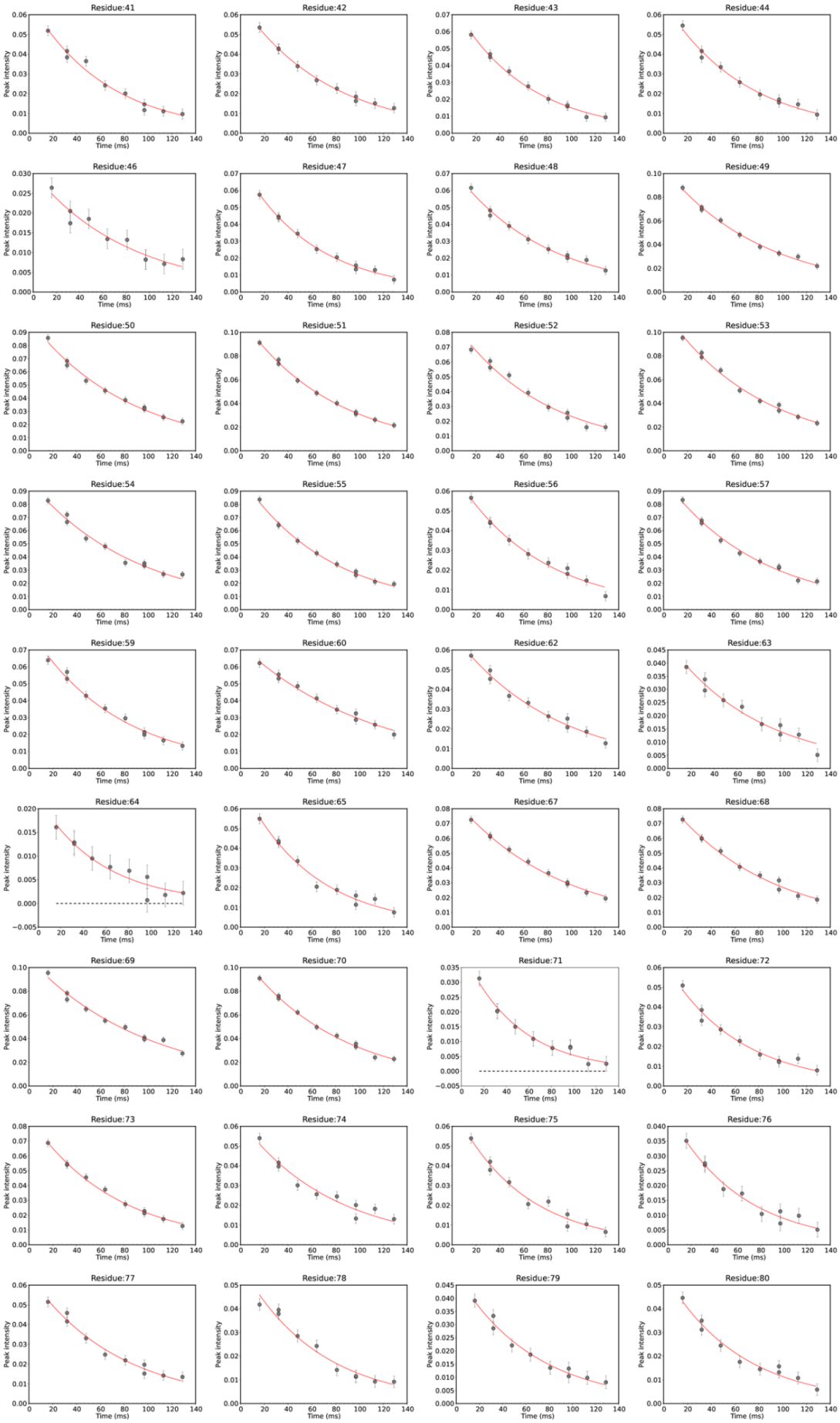
104	Leu	8.05	125.16	178.74	58.04	-
105	Asp	8.86	121.66	178.77	57.91	40.59
106	Thr	7.78	116.7	175.05	66.87	-
107	Leu	7.31	120.71	177.92	58.01	-
108	Gln	8.49	116.46	178.39	60.48	29.08
109	Leu	7.98	121.34	179.37	58.19	41.54
110	Asp	8.31	119.82	179.76	57.24	-
111	Val	8.51	122.54	176.99	67.55	-
112	Ala	8.1	122.51	180.64	55.41	18.39
113	Asp	8.42	118.62	179	57.43	41.23
114	Phe	7.89	121.17	177.34	57.4	-
115	Ala	8.65	121.87	179.86	56.25	17.24
116	Thr	8.42	115.58	176.14	67.02	67.85
117	Thr	8.15	120.47	177.49	66.94	67.95
118	Ile	8.24	121.46	176.83	65.26	42.8
119	Trp	8.82	121.83	177.77	61.5	-
120	Gln	8.67	115.91	178.83	59.08	-
121	Gln	7.62	119.21	178.48	57.78	-
122	Met	8.13	118.55	178.42	60.43	-
123	Glu	8.14	119.82	180.69	59.62	29.57
124	Glu	7.94	121.33	178.28	59.54	29.63
125	Leu	7.31	116.08	177.09	54.66	42.34
126	Gly	7.76	108.48	175.33	46.62	-
127	Met	7.91	116.48	174.38	55.26	33.96
128	Ala	7.9	123.8	174.94	50.78	17.26
129	Pro	-	-	175.73	62.49	-
130	Ala	8.22	122.59	177.53	52.51	19.67
131	Leu	8.04	120.52	176.85	54.64	43.12
132	Gln	8.42	123	173.99	53.59	29.02
133	Pro	-	-	176.68	63.03	32.07
134	Thr	8.13	115.19	174.46	61.77	69.98
135	Gln	8.46	123.65	176.2	55.86	29.91
136	Gly	8.22	110.21	172.8	44.78	-
137	Ala	8.04	123.02	177.66	52.19	19.37
138	Met	8.31	121.54	174.17	52.26	33.7
139	Pro	-	-	174.11	62.29	31.4
140	Ala	7.81	122.29	176.63	50.74	19.83
141	Phe	8.4	120.47	175.52	56.7	39.7
142	Ala	8.22	124.73	177.17	53.72	19.98
143	Ser	7.7	111.26	174.35	56.12	66.84
145	Phe	8.67	117.2	179.72	61.9	-
146	Gln	7.81	115.91	178.4	59.81	-
147	Arg	7.83	117.51	177.27	59.82	-
148	Arg	7.74	118.02	176.88	59.69	-
149	Ala	8.12	119.94	179.73	54.7	18.31
150	Gly	8.19	105.74	175.02	47.27	-
151	Gly	7.82	108.67	173.78	48.07	-
152	Val	7.27	120.98	178.89	66.51	-
153	Leu	7.83	119.94	178.32	58.23	-
154	Val	9.06	118.4	177.45	67.44	-
155	Ala	8.05	119.22	179.21	56	17.34
156	Ser	7.75	113.34	177.92	61.63	62.07

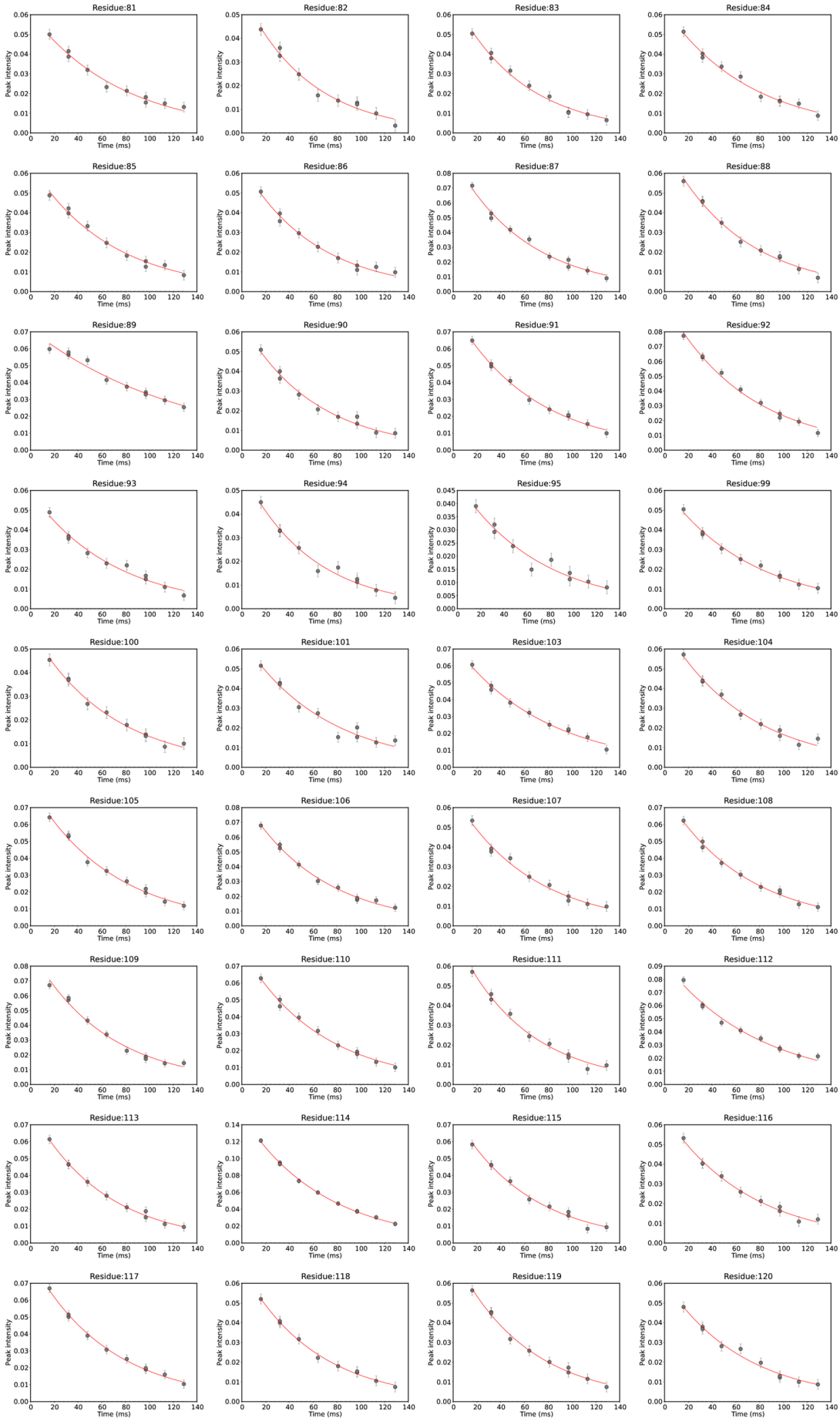
157	His	8.61	123.52	178.95	58.53	-
158	Leu	8.74	121.93	178.28	57.92	-
159	Gln	8.22	118.12	176.8	59.68	-
160	Ser	7.81	115.91	176.61	62.13	62.44
161	Phe	8.21	125	178.65	61.27	-
162	Leu	9.08	121.04	178.53	57.11	41.4
163	Glu	7.96	121.18	179.61	59.8	-
164	Val	7.27	119.62	178.04	66.66	-
165	Ser	8.02	115.45	174.81	63.31	-
166	Tyr	8.63	121.12	176.13	62.12	-
167	Arg	7.54	117.48	179.59	59.86	-
168	Val	8.35	120.42	177.57	66.46	-
169	Leu	7.97	118.43	178.79	58.49	-
170	Arg	8.37	117.04	178.86	58.8	-
171	His	7.92	119.04	177.71	59.1	30.21
172	Leu	7.74	117.47	177.5	56.4	42.77
173	Ala	7.62	120.47	177.12	52.69	19.35
174	Gln	7.6	119.25	172.51	53.85	29.11

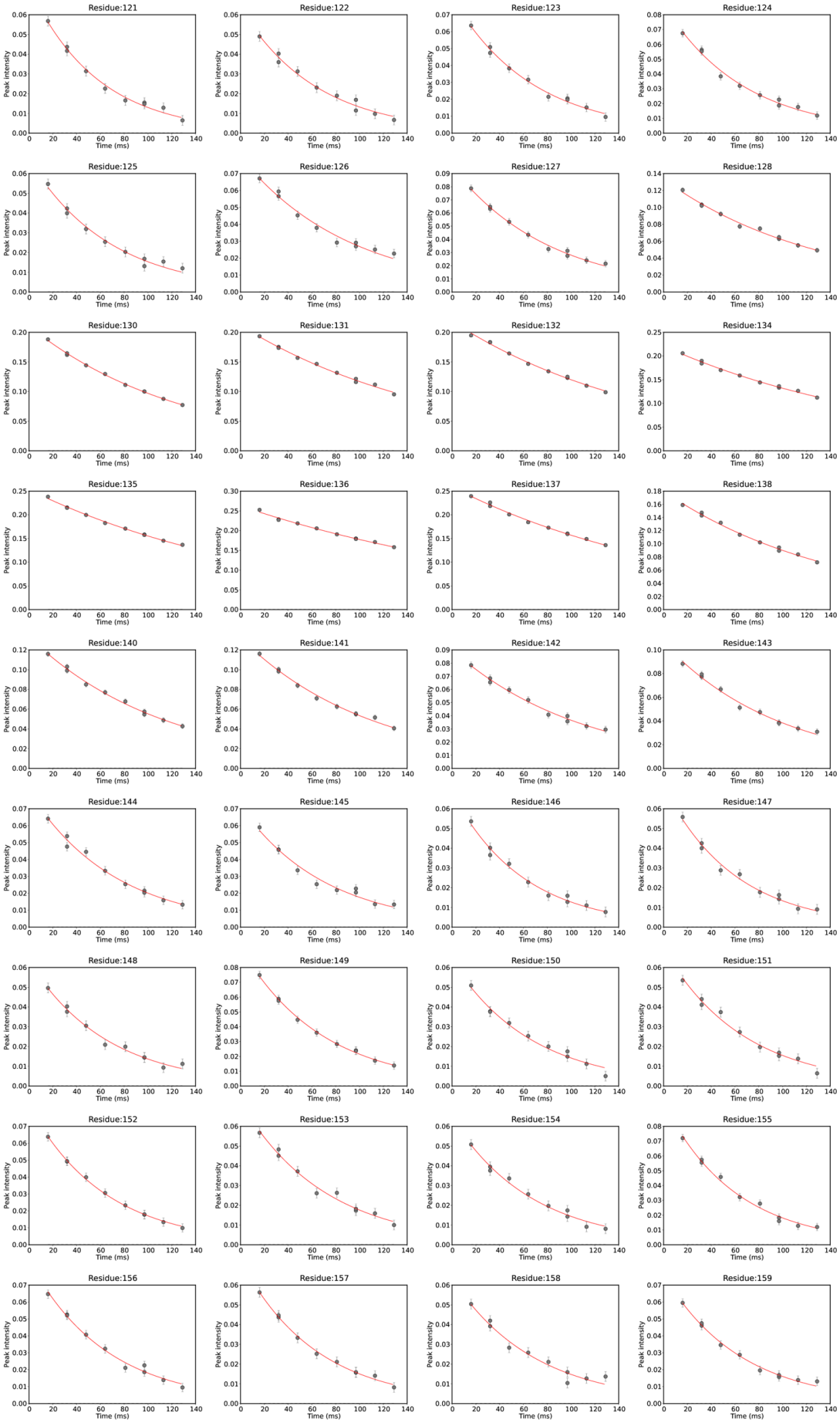
Appendix 7.8 Backbone assignments of G-CSF C3 at pH 7. Assignments of backbone amide proton and nitrogen nuclei plus alpha, beta and carbonyl carbon atoms were carried out using an ^{1}H-^{15}N HSQC spectrum in combination with HNCA, HN(CO)CA, CBCA(CO)NH, HNCO and HN(CA)CO triple-resonance spectra. Spectra were acquired using 458 µM ^{13}C ^{15}N G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 0.018 % (*w/v*) sodium azide, 10 % (*v/v*) D_2O pH 7 buffer (I = 72 mM) at 25 ^{o}C, using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

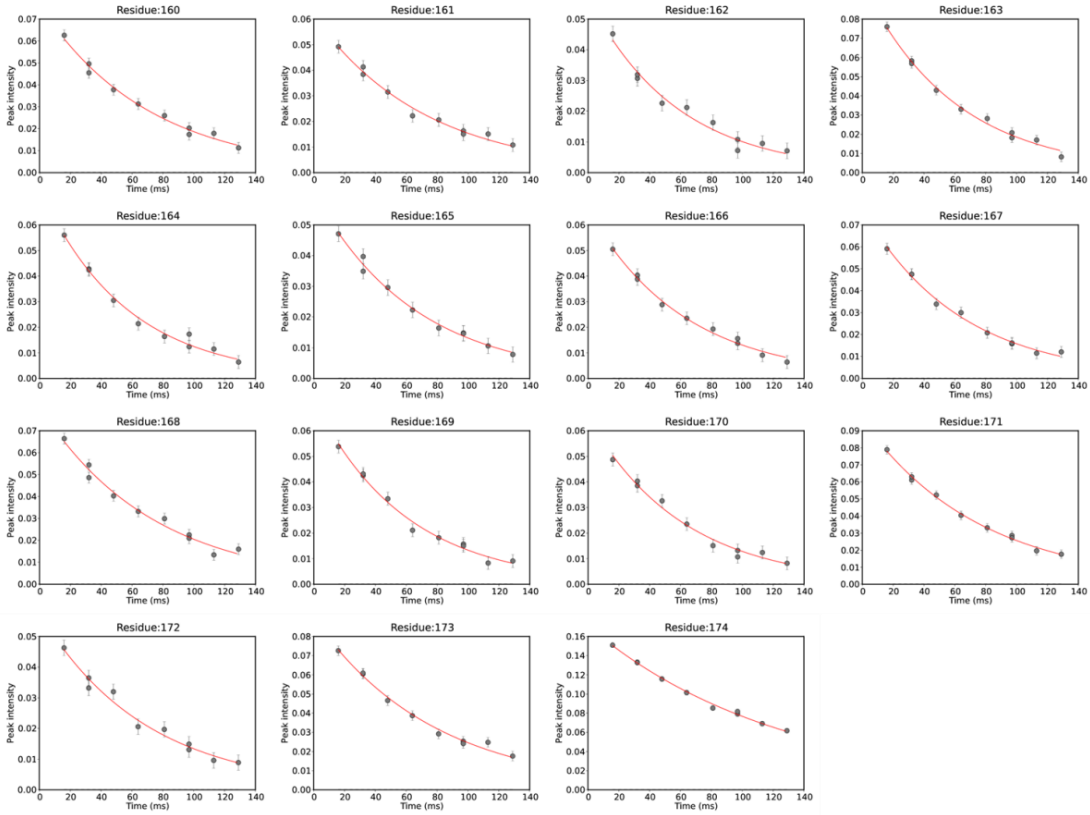
7.5 R_2 curves for G-CSF C3 at pH 4




Residue:41
Residue:42
Residue:43
Residue:44
Residue:46
Residue:47
Residue:48
Residue:49
Residue:50
Residue:51
Residue:52
Residue:53
Residue:54
Residue:55
Residue:56
Residue:57
Residue:59
Residue:60
Residue:62
Residue:63
Residue:64
Residue:65
Residue:67
Residue:68
Residue:69
Residue:70
Residue:71
Residue:72
Residue:73
Residue:74
Residue:75
Residue:76
Residue:77
Residue:78
Residue:79
Residue:80
Peak intensity
Time (ms)

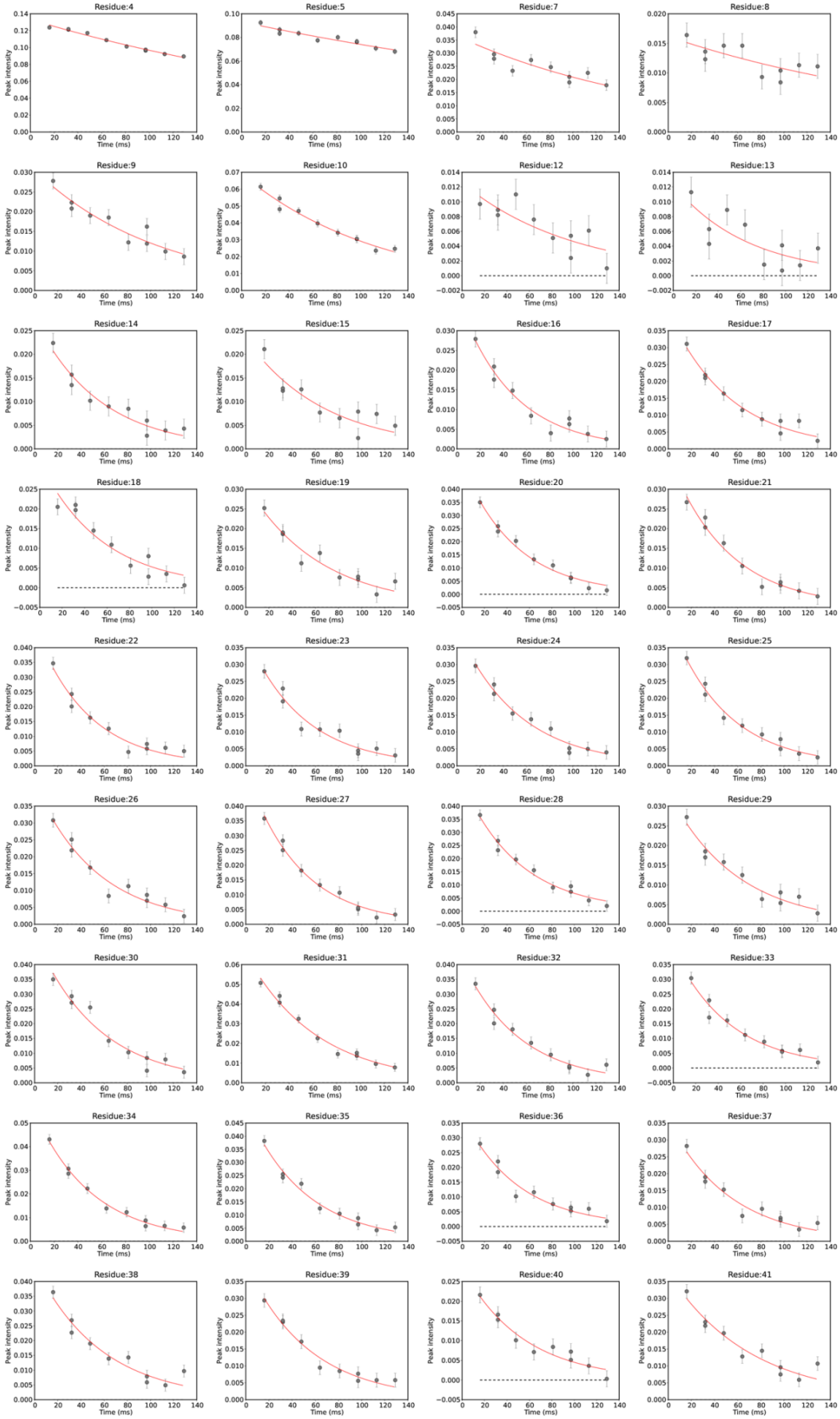

Residue:81
Residue:82
Residue:83
Residue:84
Residue:85
Residue:86
Residue:87
Residue:88
Residue:89
Residue:90
Residue:91
Residue:92
Residue:93
Residue:94
Residue:95
Residue:99
Residue:100
Residue:101
Residue:103
Residue:104
Residue:105
Residue:106
Residue:107
Residue:108
Residue:109
Residue:110
Residue:111
Residue:112
Residue:113
Residue:114
Residue:115
Residue:116
Residue:117
Residue:118
Residue:119
Residue:120
Peak intensity
Time (ms)

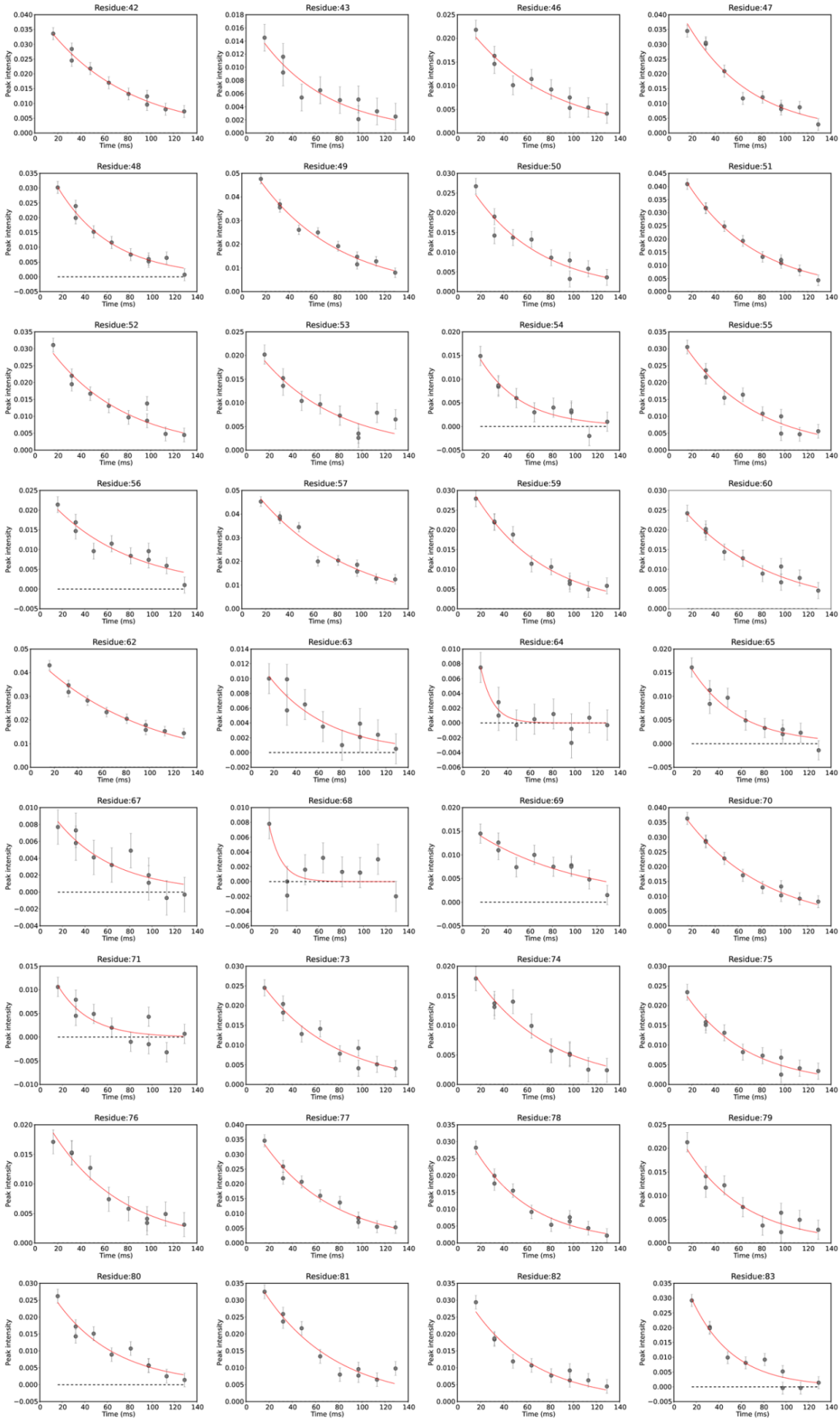

Residue:121
Residue:122
Residue:123
Residue:124
Residue:125
Residue:126
Residue:127
Residue:128
Residue:130
Residue:131
Residue:132
Residue:134
Residue:135
Residue:136
Residue:137
Residue:138
Residue:140
Residue:141
Residue:142
Residue:143
Residue:144
Residue:145
Residue:146
Residue:147
Residue:148
Residue:149
Residue:150
Residue:151
Residue:152
Residue:153
Residue:154
Residue:155
Residue:156
Residue:157
Residue:158
Residue:159
Peak intensity
Time (ms)

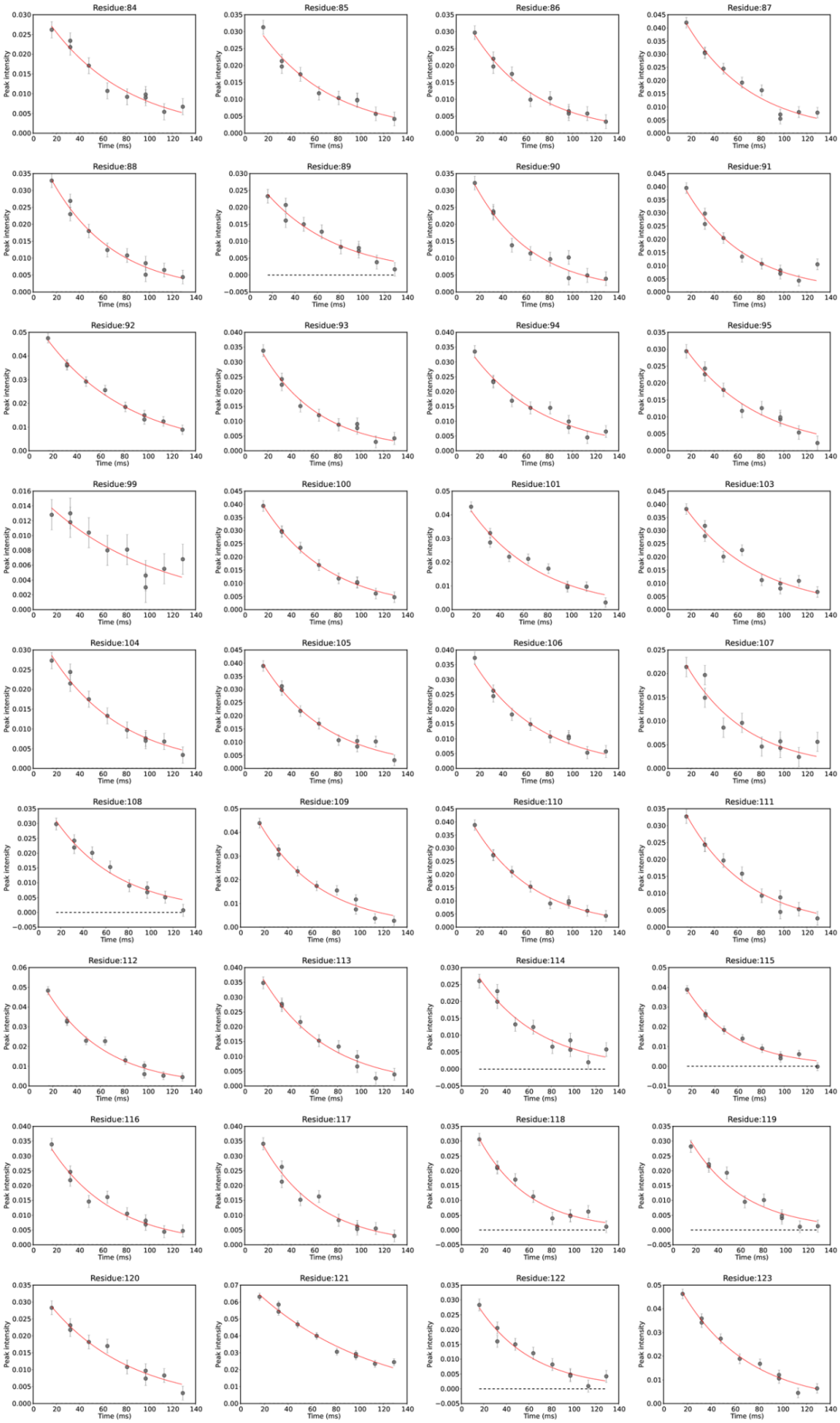


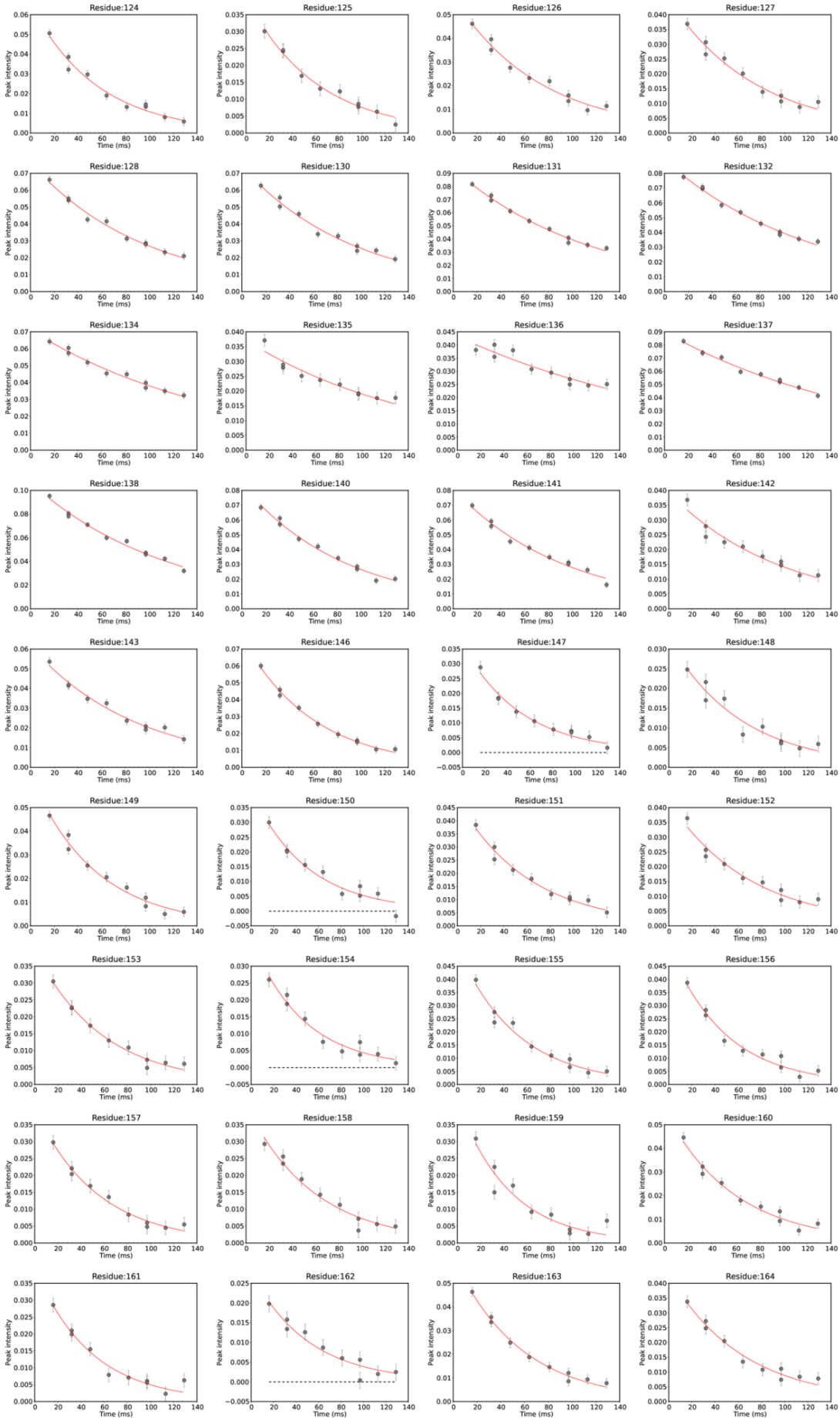
Appendix 7.9 Plots of peak intensity over time used for calculation of backbone ^{15}N R_2 rates for G-CSF C3 at pH 4. Peak intensity is shown in arbitrary units after relaxation delays of 16.1, 32.2, 48.4, 64.5, 80.6, 96.7, 112.8 and 129.0 ms, as well as fitting of these peak intensities to a single exponential decay function. Error bars represent the difference of duplicate measurements from the mean peak intensity at relaxation delays of 32.2 and 96.7 ms. A horizontal dotted line is drawn at zero peak intensity. R_2 rates were measured using 162 µM ^{15}N G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 10 % (*v/v*) D_2O pH 4 buffer (I = 39 mM) at 25 ^{o}C, using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

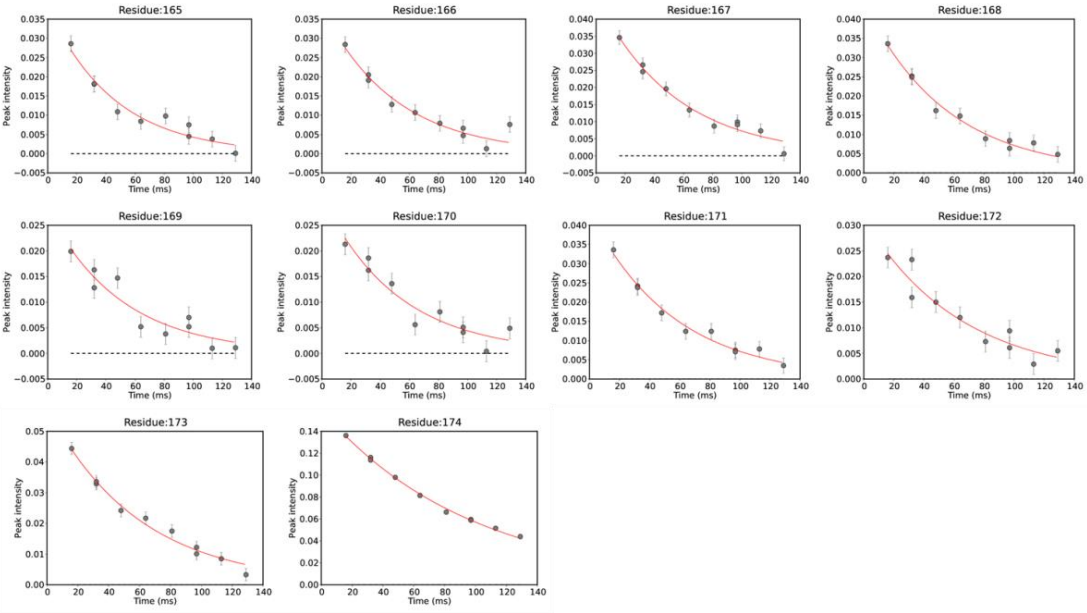
7.6 R_2 curves for G-CSF C3 at pH 7




Residue:42
Residue:43
Residue:46
Residue:47
Residue:48
Residue:49
Residue:50
Residue:51
Residue:52
Residue:53
Residue:54
Residue:55
Residue:56
Residue:57
Residue:59
Residue:60
Residue:62
Residue:63
Residue:64
Residue:65
Residue:67
Residue:68
Residue:69
Residue:70
Residue:71
Residue:73
Residue:74
Residue:75
Residue:76
Residue:77
Residue:78
Residue:79
Residue:80
Residue:81
Residue:82
Residue:83
Peak intensity
Time (ms)

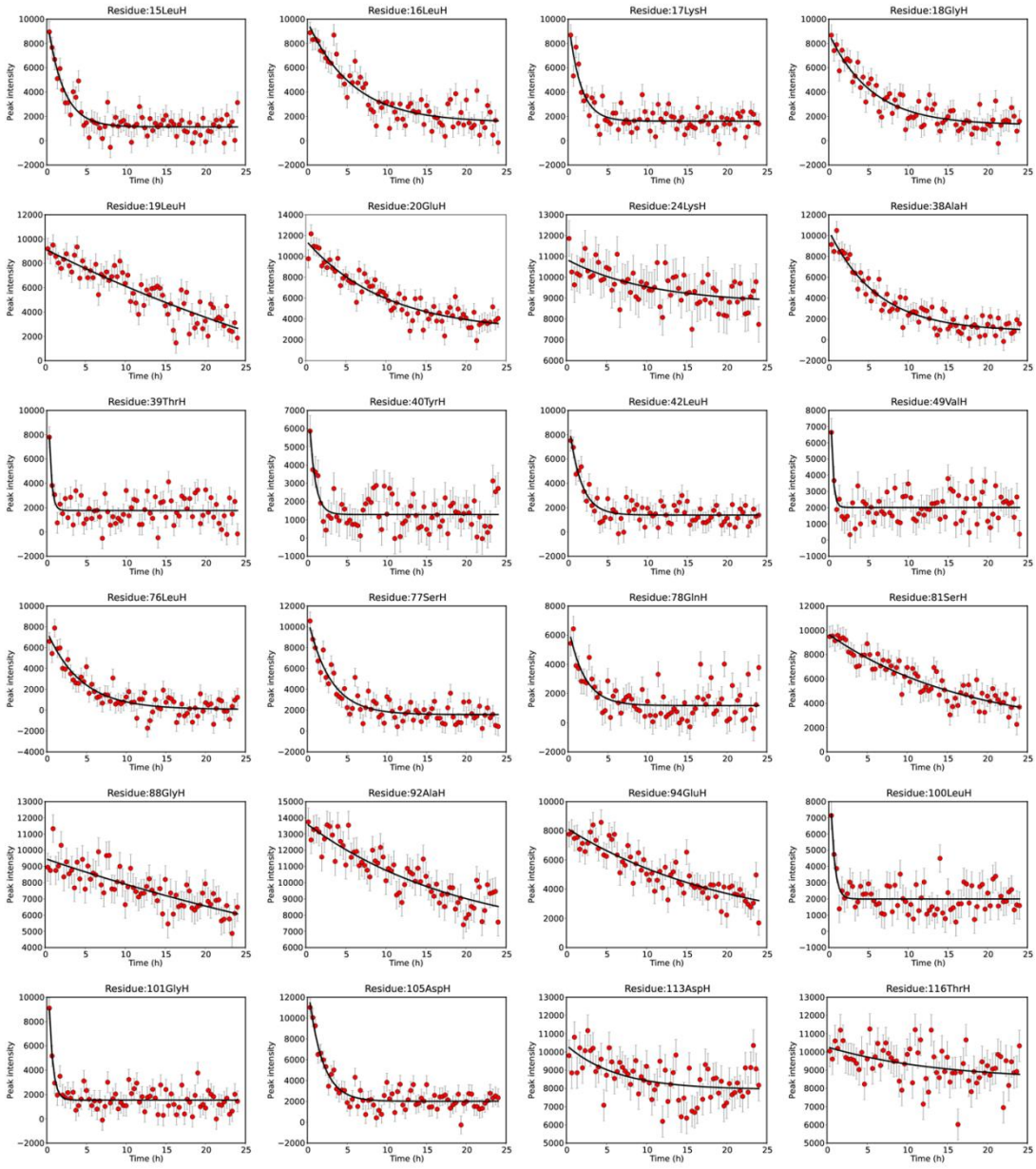

Residue:84
Residue:85
Residue:86
Residue:87
Residue:88
Residue:89
Residue:90
Residue:91
Residue:92
Residue:93
Residue:94
Residue:95
Residue:99
Residue:100
Residue:101
Residue:103
Residue:104
Residue:105
Residue:106
Residue:107
Residue:108
Residue:109
Residue:110
Residue:111
Residue:112
Residue:113
Residue:114
Residue:115
Residue:116
Residue:117
Residue:118
Residue:119
Residue:120
Residue:121
Residue:122
Residue:123
Peak intensity
Time (ms)

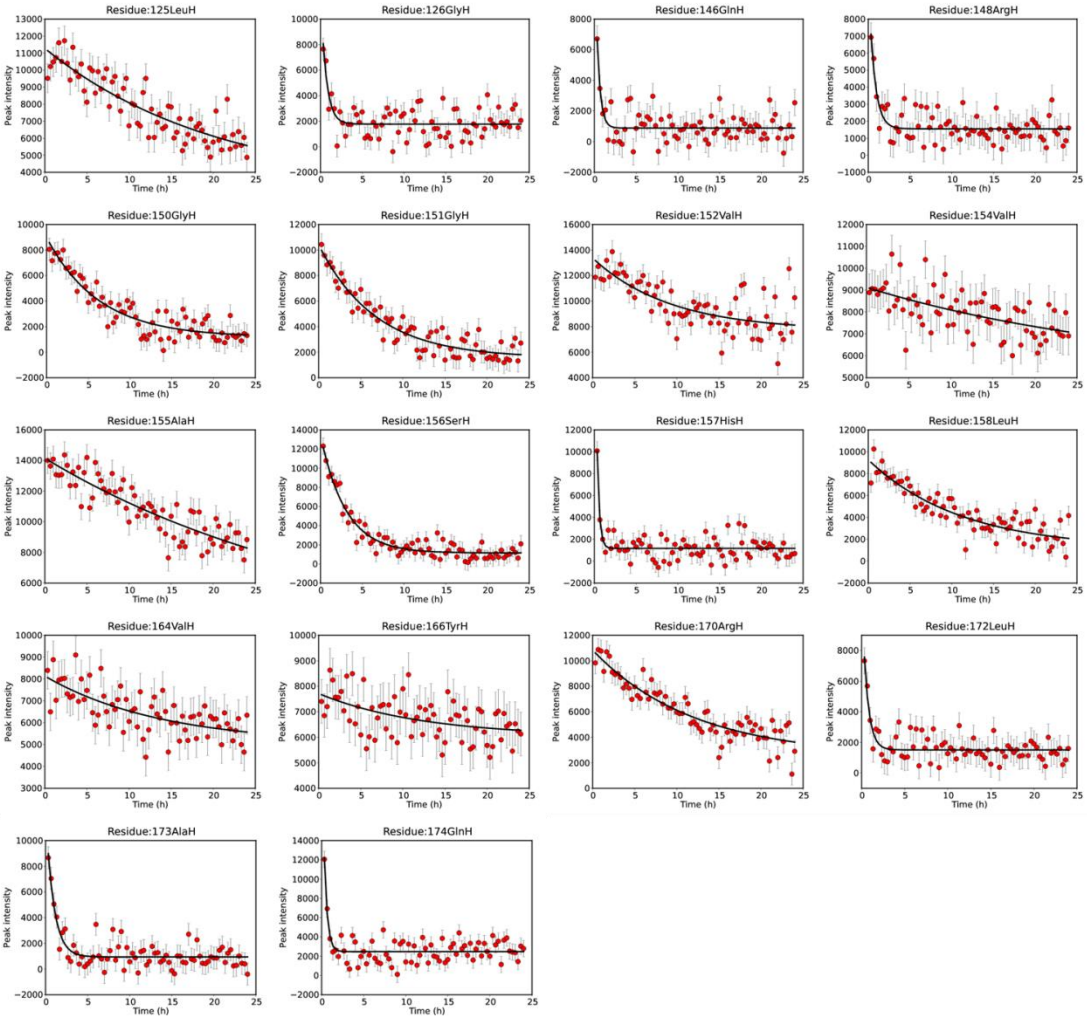

Residue:124
Residue:125
Residue:126
Residue:127
Residue:128
Residue:130
Residue:131
Residue:132
Residue:134
Residue:135
Residue:136
Residue:137
Residue:138
Residue:140
Residue:141
Residue:142
Residue:143
Residue:146
Residue:147
Residue:148
Residue:149
Residue:150
Residue:151
Residue:152
Residue:153
Residue:154
Residue:155
Residue:156
Residue:157
Residue:158
Residue:159
Residue:160
Residue:161
Residue:162
Residue:163
Residue:164
Peak intensity
Time (ms)



Appendix 7.10 Plots of peak intensity over time used for calculation of backbone ^{15}N R_2 rates for G-CSF C3 at pH 7. Peak intensity is shown in arbitrary units after relaxation delays of 16.1, 32.2, 48.4, 64.5, 80.6, 96.7, 112.8 and 129.0 ms, as well as fitting of these peak intensities to a single exponential decay function. Error bars represent the difference of duplicate measurements from the mean peak intensity at relaxation delays of 32.2 and 96.7 ms. A horizontal dotted line is drawn at zero peak intensity. R_2 rates were measured using 130 µM ^{15}N G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 0.018 % (*w/v*) sodium azide, 10 % (*v/v*) D_2O pH 7 buffer (I = 72 mM) at 25 ^{o}C, using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

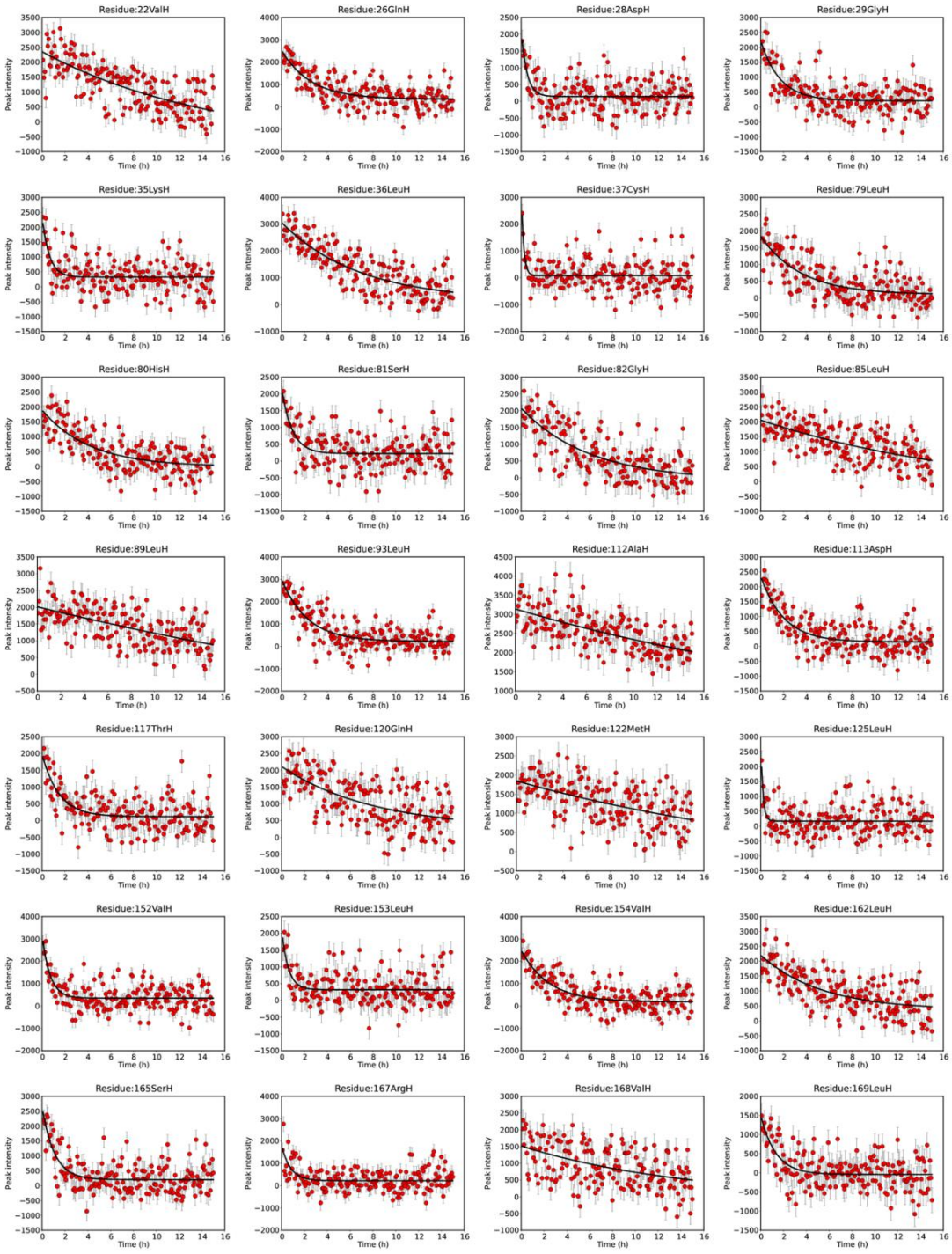
7.7 Hydrogen/deuterium exchange curves for G-CSF C3 at pD 4.8





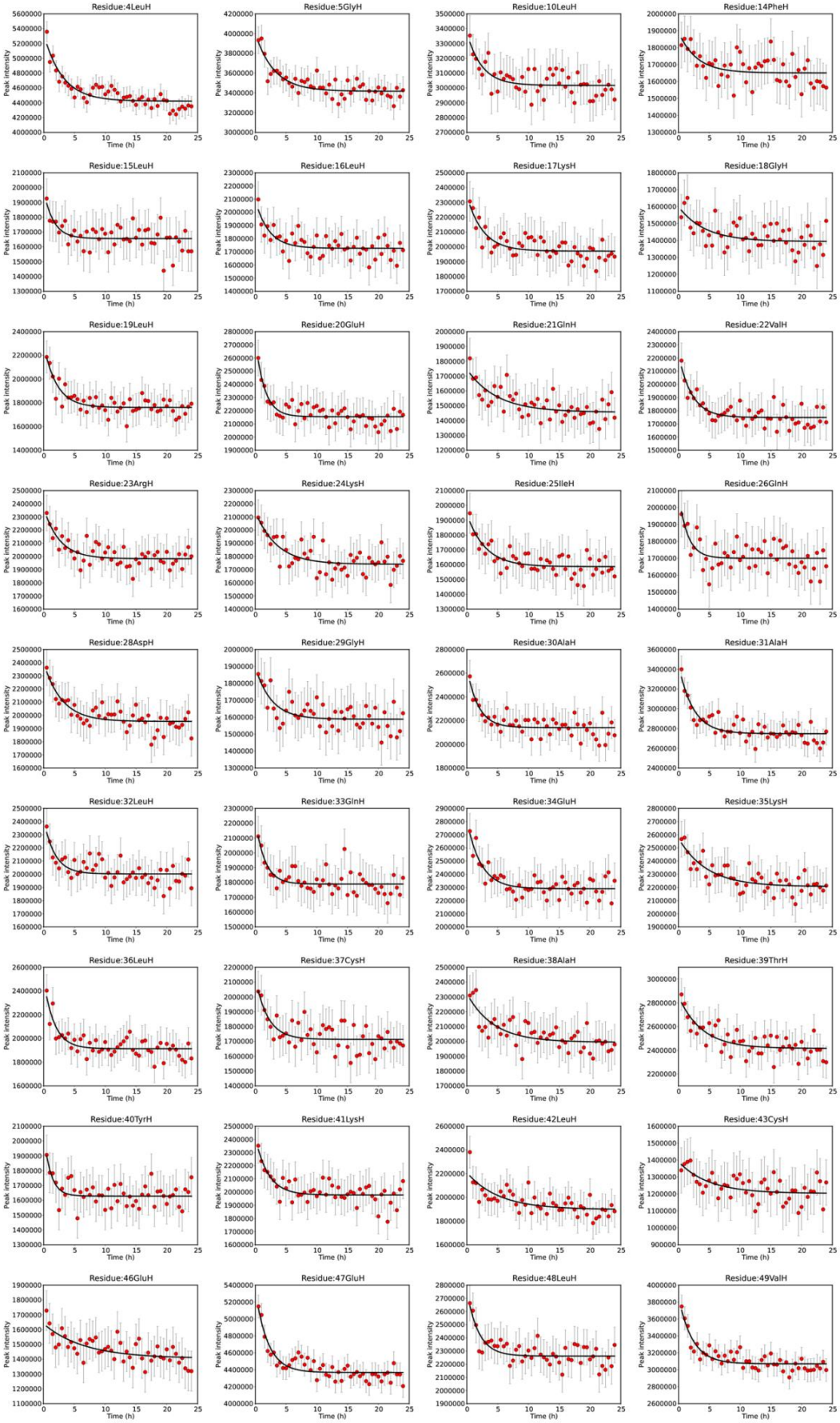
Appendix 7.11 Plots of peak intensity over time used for calculation of amide proton hydrogen exchange rates for G-CSF C3 at pD 4.8. Peak intensity is shown in arbitrary units at 20 minute intervals, as well as fitting of these peak intensities to a single exponential decay function with a y-axis offset. Error bars represent the experimental noise calculated in CCPN Analysis[321]. Hydrogen exchange rates were measured using 52 µM ^{15}N G-CSF C3 in 20 mM sodium phosphate, 20 mM sodium acetate, 90 % (*v/v*) D_2O pD 4.8 buffer (I = 46 mM) at 25 ^{o}C, using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

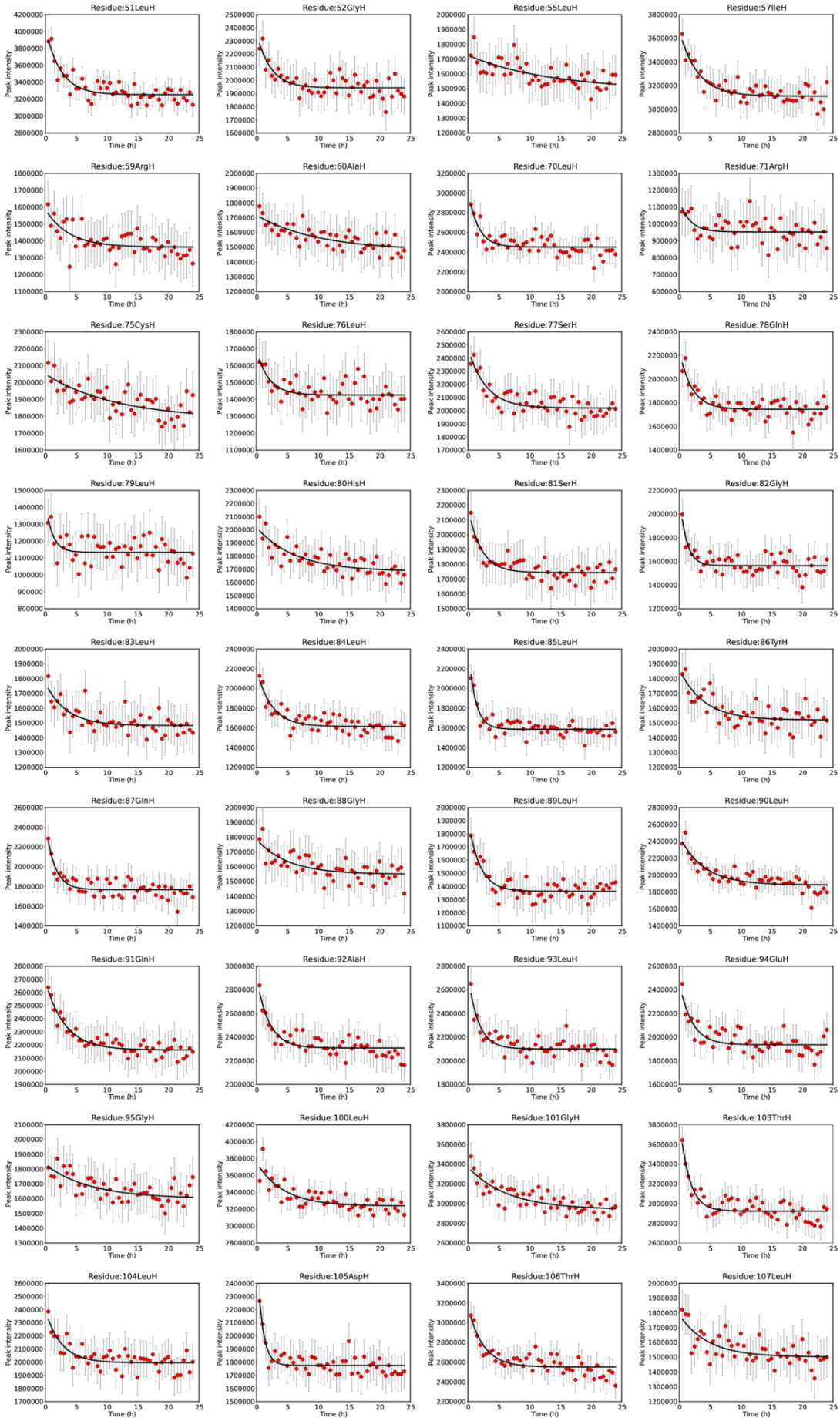
7.8 Hydrogen/deuterium exchange curves for G-CSF C3 at pD 7.8

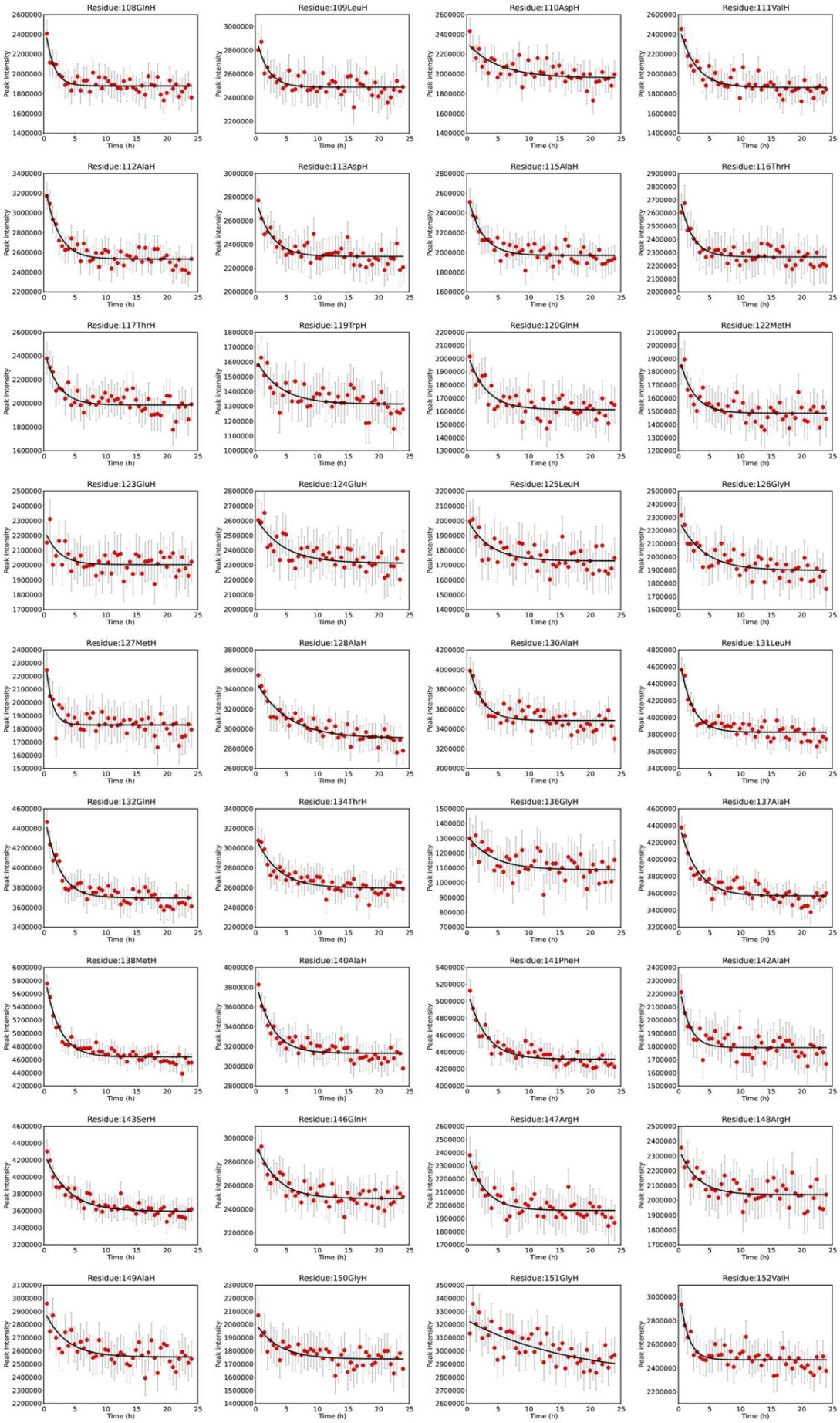


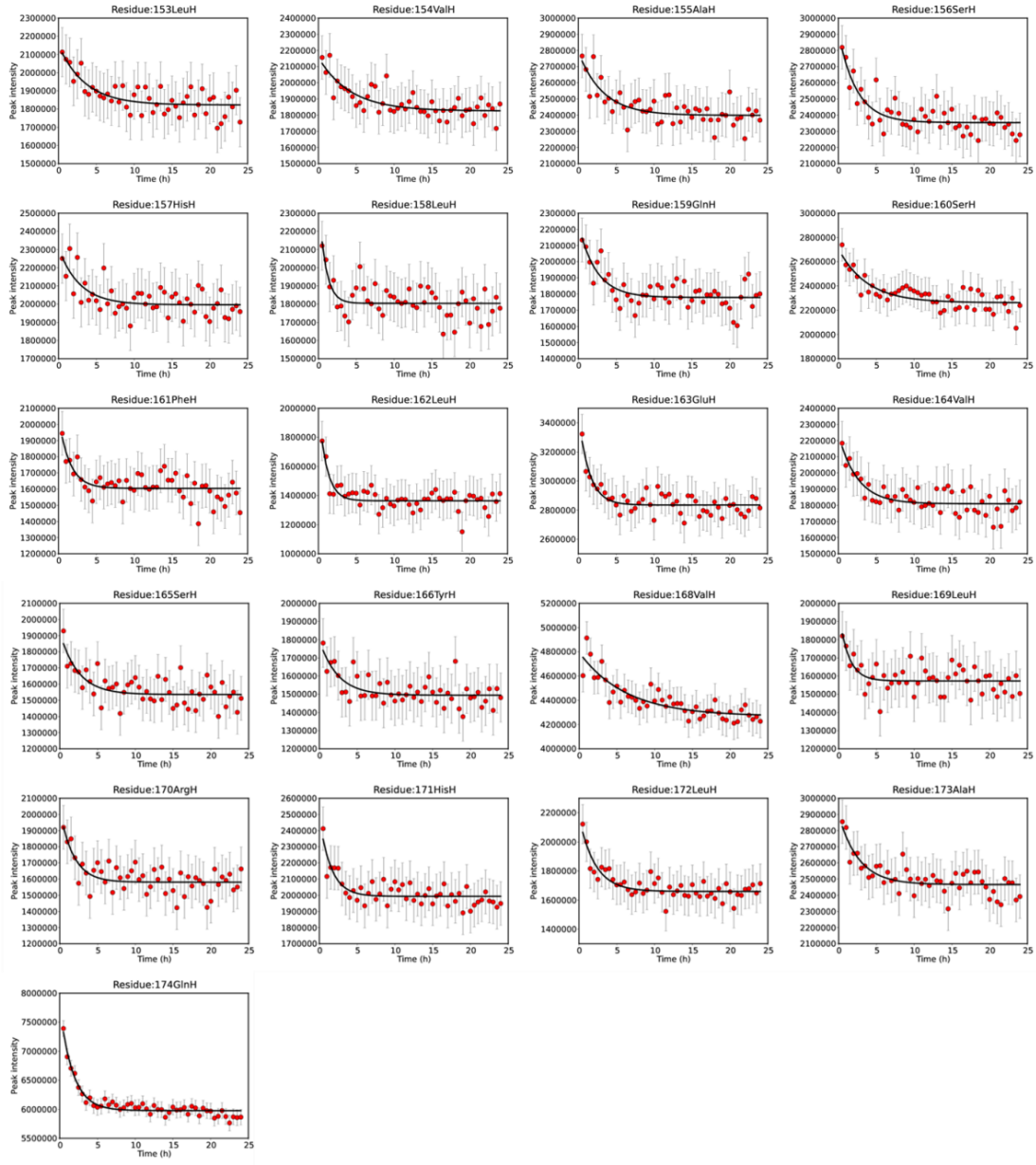
Appendix 7.12 Plots of peak intensity over time used for calculation of amide proton hydrogen exchange rates for G-CSF C3 at pD 7.8. Peak intensity is shown in arbitrary units at 5 minute intervals, as well as fitting of these peak intensities to a single exponential decay function with a y-axis offset. Error bars represent the experimental noise calculated in CCPN Analysis[321]. Hydrogen exchange rates were measured using 52 µM ^{15}N G-CSF C3 in 20 mM sodium phosphate, 20 mM sodium acetate, 0.02 % (*w/v*) sodium azide, 90 % (*v/v*) D_2O pD 7.8 buffer (I = 89 mM) at 25 ^{o}C, using a spectrometer operating at a proton resonance frequency of 600 MHz equipped with a room temperature probe.

7.9 Aggregation assay curves for G-CSF C3 at pH 7




Residue:51LeuH
Residue:52GlyH
Residue:55LeuH
Residue:57IleH
Residue:59ArgH
Residue:60AlaH
Residue:70LeuH
Residue:71ArgH
Residue:75CysH
Residue:76LeuH
Residue:77SerH
Residue:78GlnH
Residue:79LeuH
Residue:80HisH
Residue:81SerH
Residue:82GlyH
Residue:83LeuH
Residue:84LeuH
Residue:85LeuH
Residue:86TyrH
Residue:87GlnH
Residue:88GlyH
Residue:89LeuH
Residue:90LeuH
Residue:91GlnH
Residue:92AlaH
Residue:93LeuH
Residue:94GluH
Residue:95GlyH
Residue:100LeuH
Residue:101GlyH
Residue:103ThrH
Residue:104LeuH
Residue:105AspH
Residue:106ThrH
Residue:107LeuH
Peak intensity
Time (h)


Residue:108GlnH
Residue:109LeuH
Residue:110AspH
Residue:111ValH
Residue:112AlaH
Residue:113AspH
Residue:115AlaH
Residue:116ThrH
Residue:117ThrH
Residue:119TrpH
Residue:120GlnH
Residue:122MetH
Residue:123GluH
Residue:124GluH
Residue:125LeuH
Residue:126GlyH
Residue:127MetH
Residue:128AlaH
Residue:130AlaH
Residue:131LeuH
Residue:132GlnH
Residue:134ThrH
Residue:136GlyH
Residue:137AlaH
Residue:138MetH
Residue:140AlaH
Residue:141PheH
Residue:142AlaH
Residue:143SerH
Residue:146GlnH
Residue:147ArgH
Residue:148ArgH
Residue:149AlaH
Residue:150GlyH
Residue:151GlyH
Residue:152ValH
Peak intensity
Time (h)



Appendix 7.13 Plots of peak intensity over time used for calculation of amide proton decay rates during G-CSF C3 aggregation at pH 7. Peak intensity is shown in arbitrary units at 30 minute intervals, as well as fitting of these peak intensities to a single exponential decay function with a y-axis offset. Error bars represent the experimental noise calculated in CCPN Analysis[321]. Decay rates were measured using 66 µM ^{15}N G-CSF C3 in 18 mM sodium phosphate, 18 mM sodium acetate, 0.018 % (*w/v*) sodium azide, 10 % (*v/v*) D_2O pH 7 buffer (I = 72 mM) at 37 ^{o}C, using a spectrometer operating at a proton resonance frequency of 750 MHz equipped with a cryogenic probe.

References

1. Anfinsen, C. B. *et al.* The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc Natl Acad Sci U S A* 1961, **47**, p.1309-1314.
2. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* 1973, **181**, p.223-230.
3. Levinthal, C. Are there pathways for protein folding? *J Chim Phys* 1968, **65**, p.44-45.
4. Wetlaufer, D. B. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci U S A* 1973, **70**, p.697-701.
5. Kim, P. S. and Baldwin, R. L. Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. *Annu Rev Biochem* 1982, **51**, p.459-489.
6. Baldwin, R. L. The nature of protein folding pathways: The classical versus the new view. *J Biomol NMR* 1995, **5**, p.103-109.
7. Karplus, M. and Weaver, D. L. Diffusion-collision model for protein folding. *Biopolymers* 1979, **18**, p.1421-1437.
8. Tanford, C. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *J Am Chem Soc* 1962, **84**, p.4240-4247.
9. Baldwin, R. L. How does protein folding get started? *Trends Biochem Sci* 1989, **14**, p.291-294.
10. Daggett, V. and Fersht, A. R. Is there a unifying mechanism for protein folding? *Trends Biochem Sci* 2003, **28**, p.18-25.
11. Fersht, A. R. Nucleation mechanisms in protein folding. *Curr Opin Struct Biol* 1997, **7**, p.3-9.
12. Fersht, A. R. Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc Natl Acad Sci U S A* 1995, **92**, p.10869-10873.
13. Otzen, D. E. *et al.* Structure of the transition state for the folding/unfolding of the barley chymotrypsin inhibitor 2 and its implications for mechanisms of protein folding. *Proc Natl Acad Sci U S A* 1994, **91**, p.10422-10425.
14. Itzhaki, L. S., Otzen, D. E. and Fersht, A. R. The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering

methods: evidence for a nucleation-condensation mechanism for protein folding. *J Mol Biol* 1995, **254**, p.260-288.

15. Morris, E. R. and Searle, M. S. Overview of protein folding mechanisms: experimental and theoretical approaches to probing energy landscapes. *Curr Protoc Protein Sci* 2012, **Chapter 28**, p.Unit 28.22.21-22.
16. Dill, K. A. and Chan, H. S. From Levinthal to pathways to funnels. *Nat Struct Biol* 1997, **4**, p.10-19.
17. Onuchic, J. N., Luthey-Schulten, Z. and Wolynes, P. G. Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem* 1997, **48**, p.545-600.
18. Bartlett, A. I. and Radford, S. E. An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms. *Nat Struct Mol Biol* 2009, **16**, p.582-588.
19. Bryngelson, J. D. and Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci U S A* 1987, **84**, p.7524-7528.
20. Cheung, M. S., Garcia, A. E. and Onuchic, J. N. Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc Natl Acad Sci U S A* 2002, **99**, p.685-690.
21. Onuchic, J. N. and Wolynes, P. G. Theory of protein folding. *Curr Opin Struct Biol* 2004, **14**, p.70-75.
22. Clementi, C., Nymeyer, H. and Onuchic, J. N. Topological and energetic factors: what determines the structural details of the transition state ensemble and “en-route” intermediates for protein folding? an investigation for small globular proteins. *J Mol Biol* 2000, **298**, p.937-953.
23. Koga, N. and Takada, S. Roles of native topology and chain-length scaling in protein folding: A simulation study with a Gō-like model. *J Mol Biol* 2001, **313**, p.171-180.
24. Chavez, L. L., Onuchic, J. N. and Clementi, C. Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. *J Am Chem Soc* 2004, **126**, p.8426-8432.
25. Ferreiro, D. U. *et al.* The energy landscape of modular repeat proteins: topology determines folding mechanism in the ankyrin family. *J Mol Biol* 2005, **354**, p.679-692.
26. Plaxco, K. W., Simons, K. T. and Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* 1998, **277**, p.985-994.

27. Zarrine-Afsar, A., Larson, S. M. and Davidson, A. R. The family feud: do proteins with similar structures fold via the same pathway? *Curr Opin Struct Biol* 2005, **15**, p.42-49.
28. Shoichet, B. K. *et al.* A relationship between protein stability and protein function. *Proc Natl Acad Sci U S A* 1995, **92**, p.452-456.
29. Dill, K. A. Dominant forces in protein folding. *Biochemistry* 1990, **29**, p.7133-7155.
30. DePristo, M. A., Weinreich, D. M. and Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet* 2005, **6**, p.678-687.
31. Frauenfelder, H., Sligar, S. and Wolynes, P. The energy landscapes and motions of proteins. *Science* 1991, **254**, p.1598-1603.
32. Hegler, J. A., Weinkam, P. and Wolynes, P. G. The spectrum of biomolecular states and motions. *HFSP Journal* 2008, **2**, p.307-313.
33. Jackson, S. E. How do small single-domain proteins fold? *Fold Des* 1998, **3**, p.R81-91.
34. Brockwell, D. J. and Radford, S. E. Intermediates: ubiquitous species on folding energy landscapes? *Curr Opin Struct Biol* 2007, **17**, p.30-37.
35. Bai, Y. Hidden intermediates and levinthal paradox in the folding of small proteins. *Biochem Biophys Res Commun* 2003, **305**, p.785-788.
36. Balchin, D., Hayer-Hartl, M. and Hartl, F. U. *In vivo* aspects of protein folding and quality control. *Science* 2016, **353**.
37. Dobson, C. M. Protein folding and misfolding. *Nature* 2003, **426**, p.884-890.
38. Tsong, T. Y., Baldwin, R. L. and Elson, E. L. The sequential unfolding of ribonuclease A: Detection of a fast initial phase in the kinetics of unfolding. *Proc Natl Acad Sci U S A* 1971, **68**, p.2712-2715.
39. Ikai, A. and Tanford, C. Kinetic evidence for incorrectly folded intermediate states in the refolding of denatured proteins. *Nature* 1971, **230**, p.100-102.
40. Jahn, T. R. and Radford, S. E. The Yin and Yang of protein folding. *FEBS Journal* 2005, **272**, p.5962-5970.
41. Chiti, F. and Dobson, C. M. Protein misfolding, functional amyloid, and human disease. *Annu Rev Biochem* 2006, **75**, p.333-366.
42. Ellis, R. J. and Minton, A. P. Protein aggregation in crowded environments. *Biol Chem* 2006, **387**, p.485-497.
43. Gruebele, M. Downhill protein folding: evolution meets physics. *C R Biol* 2005, **328**, p.701-712.

44. Krishna, M. M. G. *et al.* Hydrogen exchange methods to study protein folding. *Methods* 2004, **34**, p.51-64.
45. Mittermaier, A. and Kay, L. E. New tools provide new insights in NMR studies of protein dynamics. *Science* 2006, **312**, p.224-228.
46. Hartl, F. U. Molecular chaperones in cellular protein folding. *Nature* 1996, **381**, p.571-580.
47. Kaiser, C. M. *et al.* The ribosome modulates nascent protein folding. *Science* 2011, **334**, p.1723-1727.
48. Evans, M. S., Sander, I. M. and Clark, P. L. Cotranslational folding promotes β-helix formation and avoids aggregation *in vivo*. *J Mol Biol* 2008, **383**, p.683-692.
49. Schiene, C. and Fischer, G. Enzymes that catalyse the restructuring of proteins. *Curr Opin Struct Biol* 2000, **10**, p.40-45.
50. Oh, E. *et al.* Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor *in vivo*. *Cell* 2011, **147**, p.1295-1308.
51. Preissler, S. and Deuerling, E. Ribosome-associated chaperones as key players in proteostasis. *Trends Biochem Sci* 2012, **37**, p.274-283.
52. Clerico, E. M. *et al.* How Hsp70 molecular machines interact with their substrates to mediate diverse physiological functions. *J Mol Biol* 2015, **427**, p.1575-1588.
53. Hayer-Hartl, M., Bracher, A. and Hartl, F. U. The GroEL-GroES chaperonin machine: a nano-cage for protein folding. *Trends Biochem Sci* 2016, **41**, p.62-76.
54. Lopez, T., Dalton, K. and Frydman, J. The mechanism and function of group II chaperonins. *J Mol Biol* 2015, **427**, p.2919-2930.
55. Ciechanover, A. and Kwon, Y. T. Degradation of misfolded proteins in neurodegenerative diseases: therapeutic targets and strategies. *Exp Mol Med* 2015, **47**, e147.
56. Rampelt, H. *et al.* Metazoan Hsp70 machines use Hsp110 to power protein disaggregation. *The EMBO Journal* 2012, **31**, p.4221-4235.
57. Nillegoda, N. B. *et al.* Crucial HSP70 co-chaperone complex unlocks metazoan protein disaggregation. *Nature* 2015, **524**, p.247-251.
58. Mogk, A., Kummer, E. and Bukau, B. Cooperation of Hsp70 and Hsp100 chaperone machines in protein disaggregation. *Front Mol Biosci* 2015, **2**, Article 22.

59. Jahn, T. R. and Radford, S. E. Folding versus aggregation: Polypeptide conformations on competing pathways. *Arch Biochem Biophys* 2008, **469**, p.100-117.
60. Dobson, C. M. Principles of protein folding, misfolding and aggregation. *Semin Cell Dev Biol* 2004, **15**, p.3-16.
61. Uversky, V. N. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* 2002, **11**, p.739-756.
62. de Groot, N. S. *et al.* Prediction of "hot spots" of aggregation in disease-linked polypeptides. *BMC Struct Biol* 2005, **5**, p.1-15.
63. Sormanni, P., Aprile, F. A. and Vendruscolo, M. The CamSol method of rational design of protein mutants with enhanced solubility. *J Mol Biol* 2015, **427**, p.478-490.
64. Tartaglia, G. G. *et al.* Prediction of aggregation-prone regions in structured proteins. *J Mol Biol* 2008, **380**, p.425-436.
65. Fernandez-Escamilla, A.-M. *et al.* Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotech* 2004, **22**, p.1302-1306.
66. Belli, M., Ramazzotti, M. and Chiti, F. Prediction of amyloid aggregation in vivo. *EMBO reports* 2011, **12**, p.657-663.
67. Sunde, M. and Blake, C. The structure of amyloid fibrils by electron microscopy and X-ray diffraction. *Adv Protein Chem* 1997, **50**, p.123-159.
68. Fink, A. L. Protein aggregation: folding aggregates, inclusion bodies and amyloid. *Fold Des* 1998, **3**, p.R9-R23.
69. West, M. W. *et al. De novo* amyloid proteins from designed combinatorial libraries. *Proc Natl Acad Sci U S A* 1999, **96**, p.11211-11216.
70. Steward, A., Adhya, S. and Clarke, J. Sequence conservation in Ig-like domains: the role of highly conserved proline residues in the fibronectin type III superfamily. *J Mol Biol* 2002, **318**, p.935-940.
71. Parrini, C. *et al.* Glycine residues appear to be evolutionarily conserved for their ability to inhibit aggregation. *Structure* 2005, **13**, p.1143-1151.
72. Rousseau, F., Serrano, L. and Schymkowitz, J. W. How evolutionary pressure against protein aggregation shaped chaperone specificity. *J Mol Biol* 2006, **355**, p.1037-1047.
73. Monsellier, E. and Chiti, F. Prevention of amyloid-like aggregation as a driving force of protein evolution. *EMBO reports* 2007, **8**, p.737-742.
74. Dobson, C. M. The structural basis of protein folding and its links with human disease. *Philos Trans R Soc Lond B Biol Sci* 2001, **356**, p.133-145.

75. Fändrich, M. and Dobson, C. M. The behaviour of polyamino acids reveals an inverse side chain effect in amyloid structure formation. *The EMBO Journal* 2002, **21**, p.5682-5690.
76. Thomas, P. J., Qu, B.-H. and Pedersen, P. L. Defective protein folding as a basis of human disease. *Trends Biochem Sci* 1995, **20**, p.456-459.
77. Bullock, A. N. and Fersht, A. R. Rescuing the function of mutant p53. *Nat Rev Cancer* 2001, **1**, p.68-76.
78. Pepys, M. B. Amyloidosis. *Annu Rev Med* 2006, **57**, p.223-241.
79. Fändrich, M. Oligomeric intermediates in amyloid formation: structure determination and mechanisms of toxicity. *J Mol Biol* 2012, **421**, p.427-440.
80. Breydo, L. and Uversky, V. N. Structural, morphological, and functional diversity of amyloid oligomers. *FEBS Lett* 2015, **589**, p.2640-2648.
81. Mannini, B. *et al.* Toxicity of protein oligomers is rationalized by a function combining size and surface hydrophobicity. *ACS Chem Biol* 2014, **9**, p.2309-2317.
82. Bucciantini, M. *et al.* Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* 2002, **416**, p.507-511.
83. Knowles, T. P. J. *et al.* An analytical solution to the kinetics of breakable filament assembly. *Science* 2009, **326**, p.1533-1537.
84. Ben-Zvi, A., Miller, E. A. and Morimoto, R. I. Collapse of proteostasis represents an early molecular event in *Caenorhabditis elegans* aging. *Proc Natl Acad Sci U S A* 2009, **106**, p.14914-14919.
85. Gejyo, F. *et al.* Serum levels of β2-microglobulin as a new form of amyloid protein in patients undergoing long-term hemodialysis. *N Engl J Med* 1986, **314**, 585.
86. McParland, V. J. *et al.* Structural properties of an amyloid precursor of β2-microglobulin. *Nat Struct Mol Biol* 2002, **9**, p.326-331.
87. Platt, G. W. *et al.* Dynamics in the unfolded state of β2-microglobulin studied by NMR. *J Mol Biol* 2005, **346**, p.279-294.
88. Jahn, T. R. *et al.* Amyloid formation under physiological conditions proceeds via a native-like folding intermediate. *Nat Struct Mol Biol* 2006, **13**, p.195-201.
89. Lee, S. *et al.* A single destabilizing mutation (F9S) promotes concerted unfolding of an entire globular domain in γS-crystallin. *J Mol Biol* 2010, **399**, p.320-330.
90. Mahler, B. *et al.* Characterization of a transient unfolding intermediate in a core mutant of γS-crystallin. *J Mol Biol* 2011, **405**, p.840-850.

91. Sinha, D. *et al.* A temperature-sensitive mutation of *Crygs* in the murine *Opj* cataract. *J Biol Chem* 2001, **276**, p.9308-9315.

92. Liu, K., Kelly, J. W. and Wemmer, D. E. Native state hydrogen exchange study of suppressor and pathogenic variants of transthyretin. *J Mol Biol* 2002, **320**, p.821-832.

93. Lim, K. H. *et al.* Localized structural fluctuations promote amyloidogenic conformations in transthyretin. *J Mol Biol* 2013, **425**, p.977-988.

94. Benson, M. D. and Uemichi, T. Transthyretin amyloidosis. *Amyloid* 1996, **3**, p.44-56.

95. Hörnberg, A. *et al.* A comparative analysis of 23 structures of the amyloidogenic protein transthyretin. *J Mol Biol* 2000, **302**, p.649-669.

96. Rader, R. A. (Re)defining biopharmaceutical. *Nat Biotech* 2008, **26**, p.743-751.

97. Walsh, G. Biopharmaceutical benchmarks 2014. *Nat Biotech* 2014, **32**, p.992-1000.

98. Wadhwa, M. and Thorpe, R. Haematopoietic growth factors and their therapeutic use. *Thromb Haemost* 2008, **99**, p.863-873.

99. Meyer, B. K. *et al.* Antimicrobial preservative use in parenteral products: Past and present. *J Pharm Sci* 2007, **96**, p.3155-3167.

100. Walsh, G. Biopharmaceuticals: recent approvals and likely directions. *Trends Biotechnol* 2005, **23**, p.553-558.

101. Leader, B., Baca, Q. J. and Golan, D. E. Protein therapeutics: a summary and pharmacological classification. *Nat Rev Drug Discov* 2008, **7**, p.21-39.

102. Aggarwal, S. What's fueling the biotech engine - 2012 to 2013. *Nat Biotech* 2014, **32**, p.32-39.

103. Goeddel, D. V. *et al.* Expression in *Escherichia coli* of chemically synthesized genes for human insulin. *Proc Natl Acad Sci U S A* 1979, **76**, p.106-110.

104. Walsh, G. Therapeutic insulins and their large-scale manufacture. *Appl Microbiol Biotechnol* 2005, **67**, p.151-159.

105. Fuchs, H. J. *et al.* Effect of aerosolized recombinant human DNase on exacerbations of respiratory symptoms and on pulmonary function in patients with cystic fibrosis. *N Engl J Med* 1994, **331**, p.637-642.

106. Vogel, C. L. *et al.* Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer. *J Clin Oncol* 2002, **20**, p.719-726.

107. Olsen, E. *et al.* Pivotal phase III trial of two dose levels of denileukin diftitox for the treatment of cutaneous T-cell lymphoma. *J Clin Oncol* 2001, **19**, p.376-388.

108. Szmuness, W. *et al.* Hepatitis B vaccine: demonstration of efficacy in a controlled clinical trial in a high-risk population in the United States. *N Engl J Med* 1980, **303**, p.833-841.

109. MacKenzie, I. Z. *et al.* Efficacy and safety of a new, chromatographically purified rhesus (D) immunoglobulin. *Eur J Obstet Gynecol Reprod Biol* 2004, **117**, p.154-161.

110. Sliwkowski, M. X. and Mellman, I. Antibody therapeutics in cancer. *Science* 2013, **341**, p.1192-1198.

111. Hodi, F. S. *et al.* Improved survival with ipilimumab in patients with metastatic melanoma. *N Engl J Med* 2010, **363**, p.711-723.

112. Campos-Neto, A. *et al.* Evaluation of DPPD, a single recombinant *Mycobacterium tuberculosis* protein as an alternative antigen for the Mantoux test. *Tuberculosis* 2001, **81**, p.353-358.

113. Urnovitz, H. B., Sturge, J. C. and Gottfried, T. D. Increased sensitivity of HIV-1 antibody detection. *Nat Med* 1997, **3**, p.1258-1258.

114. Walsh, G. Biopharmaceutical benchmarks. *Nat Biotechnol* 2000, **18**, p.831-833.

115. McKeage, K. and Goa, K. L. Insulin Glargine. *Drugs* 2001, **61**, p.1599-1624.

116. Jones, P. T. *et al.* Replacing the complementarity-determining regions in a human antibody with those from a mouse. *Nature* 1985, **321**, p.522-525.

117. Berg, J., Tymoczko, J. and Stryer, L. Biochemistry. W. H. Freeman, New York, 6th Edition 2002.

118. McCafferty, J. *et al.* Phage antibodies: filamentous phage displaying antibody variable domains. *Nature* 1990, **348**, p.552-554.

119. Hoogenboom, H. R. *et al.* Multi-subunit proteins on the surface of filamentous phage: methodologies for displaying antibody (Fab) heavy and light chains. *Nucleic Acids Res* 1991, **19**, p.4133-4137.

120. Lee, C. C., Perchiacca, J. M. and Tessier, P. M. Toward aggregation-resistant antibodies by design. *Trends Biotechnol* 2013, **31**, p.612-620.

121. Cho, H.-S. *et al.* Structure of the extracellular region of HER2 alone and in complex with the Herceptin Fab. *Nature* 2003, **421**, p.756-760.

122. Clackson, T. *et al.* Making antibody fragments using phage display libraries. *Nature* 1991, **352**, p.624-628.

123. Roodveldt, C., Aharoni, A. and Tawfik, D. S. Directed evolution of proteins for heterologous expression and stability. *Curr Opin Struct Biol* 2005, **15**, p.50-56.

124. Magliery, T. J., Lavinder, J. J. and Sullivan, B. J. Protein stability by number: high-throughput and statistical approaches to one of protein science's most difficult problems. *Curr Opin Chem Biol* 2011, **15**, p.443-451.

125. Magliery, T. J. and Regan, L. Combinatorial approaches to protein stability and structure. *Eur J Biochem* 2004, **271**, p.1595-1608.

126. Shusta, E. V. *et al.* Yeast polypeptide fusion surface display levels predict thermal stability and soluble secretion efficiency. *J Mol Biol* 1999, **292**, p.949-956.

127. Park, S. *et al.* Limitations of yeast surface display in engineering proteins of high thermostability. *Protein Eng Des Sel* 2006, **19**, p.211-217.

128. Waldo, G. S. *et al.* Rapid protein-folding assay using green fluorescent protein. *Nat Biotech* 1999, **17**, p.691-695.

129. Cabantous, S., Terwilliger, T. C. and Waldo, G. S. Protein tagging and detection with engineered self-assembling fragments of green fluorescent protein. *Nat Biotech* 2005, **23**, p.102-107.

130. Åslund, F. and Beckwith, J. The thioredoxin superfamily: redundancy, specificity, and gray-area genomics. *J Bacteriol* 1999, **181**, p.1375-1379.

131. Lipiäinen, T. *et al.* Formulation and stability of cytokine therapeutics. *J Pharm Sci* 2015, **104**, p.307-326.

132. Saunders, J. C. *et al.* An *in vivo* platform for identifying inhibitors of protein aggregation. *Nat Chem Biol* 2016, **12**, p.94-101.

133. Foit, L. *et al.* Optimizing protein stability *in vivo*. *Mol Cell* 2009, **36**, p.861-871.

134. Espargaró, A. *et al.* The *in vivo* and *in vitro* aggregation properties of globular proteins correlate with their conformational stability: the SH3 case. *J Mol Biol* 2008, **378**, p.1116-1131.

135. Rodnina, M. V. The ribosome in action: Tuning of translational efficiency and protein folding. *Protein Sci* 2016.

136. Tartaglia, G. G. *et al.* A relationship between mRNA expression levels and protein solubility in *E. coli*. *J Mol Biol* 2009, **388**, p.381-389.

137. Hanes, J. and Plückthun, A. *In vitro* selection and evolution of functional proteins by using ribosome display. *Proc Natl Acad Sci U S A* 1997, **94**, p.4937-4942.

138. Roberts, R. W. and Szostak, J. W. RNA-peptide fusions for the *in vitro* selection of peptides and proteins. *Proc Natl Acad Sci U S A* 1997, **94**, p.12297-12302.

139. Kotz, J. D., Bond, C. J. and Cochran, A. G. Phage-display as a tool for quantifying protein stability determinants. *Eur J Biochem* 2004, **271**, p.1623-1629.

140. Jespers, L. *et al.* Aggregation-resistant domain antibodies selected on phage by heat denaturation. *Nat Biotech* 2004, **22**, p.1161-1165.

141. Dudgeon, K., Rouet, R. and Christ, D. Rapid prediction of expression and refolding yields using phage display. *Protein Eng Des Sel* 2013, **26**, p.671-674.

142. Raso, S. W. *et al.* Aggregation of granulocyte-colony stimulating factor *in vitro* involves a conformationally altered monomeric state. *Protein Sci* 2005, **14**, p.2246-2257.

143. Jung, S., Honegger, A. and Plückthun, A. Selection for improved protein stability by phage display. *J Mol Biol* 1999, **294**, p.163-180.

144. Martin, A., Sieber, V. and Schmid, F. X. *In vitro* selection of highly stabilized protein variants with optimized surface. *J Mol Biol* 2001, **309**, p.717-726.

145. Bai, Y. and Feng, H. Selection of stably folded proteins by phage-display with proteolysis. *Eur J Biochem* 2004, **271**, p.1609-1614.

146. Galan, A. *et al.* Library-based display technologies: where do we stand? *Mol Biosyst* 2016.

147. Thom, G. *et al.* Probing a protein-protein interaction by *in vitro* evolution. *Proc Natl Acad Sci U S A* 2006, **103**, p.7619-7624.

148. Jermutus, L. *et al.* Tailoring *in vitro* evolution for protein affinity or stability. *Proc Natl Acad Sci U S A* 2001, **98**, p.75-80.

149. Matsuura, T. and Plückthun, A. Selection based on the folding properties of proteins with ribosome display. *FEBS Lett* 2003, **539**, p.24-28.

150. Buchanan, A. *et al.* Improved drug-like properties of therapeutic proteins by directed evolution. *Protein Eng Des Sel* 2012, **25**, p.631-638.

151. Monro, R. and Vazquez, D. Ribosome-catalysed peptidyl transfer: effects of some inhibitors of protein synthesis. *J Mol Biol* 1967, **28**, p.161-165.

152. Monro, R. E. and Marcker, K. A. Ribosome-catalysed reaction of puromycin with a formylmethionine-containing oligonucleotide. *J Mol Biol* 1967, **25**, p.347-350.

153. Millward, S. W. *et al.* Design of cyclic peptides that bind protein surfaces with antibody-like affinity. *ACS Chem Biol* 2007, **2**, p.625-634.

154. Howell, S. M. *et al.* Serum stable natural peptides designed by mRNA display. *Sci Rep* 2014, **4**, 6008.
155. Singh, S. K. Impact of product-related factors on immunogenicity of biotherapeutics. *J Pharm Sci* 2011, **100**, p.354-387.
156. Rosenberg, A. S. Effects of protein aggregates: An immunologic perspective. *The AAPS Journal* 2006, **8**, p.E501-E507.
157. Sauerborn, M. *et al.* Immunological mechanism underlying the immune response to recombinant human protein therapeutics. *Trends Pharmacol Sci* 2010, **31**, p.53-59.
158. Fathallah, A. M. *et al.* The effect of small oligomeric protein aggregates on the immunogenicity of intravenous and subcutaneous administered antibodies. *J Pharm Sci* 2015, **104**, p.3691-3702.
159. Parkin, J. and Cohen, B. An overview of the immune system. *Lancet* 2001, **357**, p.1777-1789.
160. Xiang, S. D. *et al.* Pathogen recognition and development of particulate vaccines: Does size matter? *Methods* 2006, **40**, p.1-9.
161. Ponce, R. *et al.* Immunogenicity of biologically-derived therapeutics: Assessment and interpretation of nonclinical safety studies. *Regul Toxicol Pharmacol* 2009, **54**, p.164-182.
162. Berkowitz, S. A. *et al.* Analytical tools for characterizing biopharmaceuticals and the implications for biosimilars. *Nat Rev Drug Discov* 2012, **11**, p.527-540.
163. Schellekens, H. and Jiskoot, W. Erythropoietin-associated PRCA: still an unsolved mystery. *J Immunotoxicol* 2006, **3**, p.123-130.
164. Jiskoot, W. *et al.* Protein instability and immunogenicity: roadblocks to clinical application of injectable protein delivery systems for sustained release. *J Pharm Sci* 2012, **101**, p.946-954.
165. Hamrang, Z., Rattray, N. J. W. and Pluen, A. Proteins behaving badly: emerging technologies in profiling biopharmaceutical aggregation. *Trends Biotechnol* 2013, **31**, p.448-458.
166. Philo, J. S. Is any measurement method optimal for all aggregate sizes and types? *The AAPS Journal* 2006, **8**, p.E564-E571.
167. Zölls, S. *et al.* Particles in therapeutic protein formulations, Part 1: overview of analytical methods. *J Pharm Sci* 2012, **101**, p.914-935.
168. Carpenter, J. F. *et al.* Potential inaccurate quantitation and sizing of protein aggregates by size exclusion chromatography: Essential need to use

orthogonal methods to assure the quality of therapeutic protein products. *J Pharm Sci* 2010, **99**, p.2200-2208.

169. Mahler, H.-C. *et al.* Protein aggregation: Pathways, induction factors and analysis. *J Pharm Sci* 2009, **98**, p.2909-2934.
170. Filipe, V., Hawe, A. and Jiskoot, W. Critical evaluation of Nanoparticle Tracking Analysis (NTA) by NanoSight for the measurement of nanoparticles and protein aggregates. *Pharm Res* 2010, **27**, p.796-810.
171. Hawe, A. *et al.* Taylor Dispersion Analysis compared to Dynamic Light Scattering for the size analysis of therapeutic peptides and proteins and their aggregates. *Pharm Res* 2011, **28**, p.2302-2310.
172. Patel, A. R., Lau, D. and Liu, J. Quantification and characterization of micrometer and submicrometer subvisible particles in protein therapeutics by use of a suspended microchannel resonator. *Anal Chem* 2012, **84**, p.6833-6840.
173. Coulter, B. The sizing of non-spherical, submicron particles using Polarization Intensity Differential Scattering (PIDS). Application Information A-1995A), Beckman Coulter 1995.
174. Carpenter, J. F. *et al.* Overlooking subvisible particles in therapeutic protein products: Gaps that may compromise product quality. *J Pharm Sci* 2009, **98**, p.1201-1205.
175. Wang, W., Nema, S. and Teagarden, D. Protein aggregation—Pathways and influencing factors. *Int J Pharm* 2010, **390**, p.89-99.
176. Baneyx, F. Recombinant protein expression in *Escherichia coli*. *Curr Opin Biotechnol* 1999, **10**, p.411-421.
177. Burgess, R. R. Refolding solubilized inclusion body proteins. *Methods Enzymol* 2009, **463**, p.259-282.
178. Wang, W. Protein aggregation and its inhibition in biopharmaceutics. *Int J Pharm* 2005, **289**, p.1-30.
179. Brorson, K. *et al.* Bracketed generic inactivation of rodent retroviruses by low pH treatment for monoclonal antibodies and recombinant proteins. *Biotechnol Bioeng* 2003, **82**, p.321-329.
180. Speed, M. A., King, J. and Wang, D. I. Polymerization mechanism of polypeptide chain aggregation. *Biotechnol Bioeng* 1997, **54**, p.333-343.
181. Schein, C. H. Solubility as a function of protein structure and solvent components. *Biotechnology (N Y)* 1990, **8**, p.308-317.
182. Manning, M. C. *et al.* Stability of protein pharmaceuticals: an update. *Pharm Res* 2010, **27**, p.544-575.

183. Kueltzo, L. A. *et al.* Effects of solution conditions, processing parameters, and container materials on aggregation of a monoclonal antibody during freeze-thawing. *J Pharm Sci* 2008, **97**, p.1801-1812.

184. Bee, J. S. *et al.* Response of a concentrated monoclonal antibody formulation to high shear. *Biotechnol Bioeng* 2009, **103**, p.936-943.

185. Thirumangalathu, R. *et al.* Effects of pH, temperature, and sucrose on benzyl alcohol-induced aggregation of recombinant human granulocyte colony stimulating factor. *J Pharm Sci* 2006, **95**, p.1480-1497.

186. Zhang, Y. *et al.* Mechanism for benzyl alcohol-induced aggregation of recombinant human interleukin-1 receptor antagonist in aqueous solution. *J Pharm Sci* 2004, **93**, p.3076-3089.

187. Roy, S. *et al.* Temperature dependence of benzyl alcohol- and 8-anilinonaphthalene-1-sulfonate-induced aggregation of recombinant human interleukin-1 receptor antagonist. *Biochemistry* 2006, **45**, p.3898-3911.

188. Bis, R. L. *et al.* Role of benzyl alcohol in the unfolding and aggregation of interferon α-2a. *J Pharm Sci* 2015, **104**, p.407-415.

189. Redecke, L. *et al.* UV-light-induced conversion and aggregation of prion proteins. *Free Radic Biol Med* 2009, **46**, p.1353-1361.

190. Davies, M. J. Singlet oxygen-mediated damage to proteins and its consequences. *Biochem Biophys Res Commun* 2003, **305**, p.761-770.

191. Thirumangalathu, R. *et al.* Silicone oil- and agitation-induced aggregation of a monoclonal antibody in aqueous solution. *J Pharm Sci* 2009, **98**, p.3167-3181.

192. Cabra, V. *et al.* The effect of sulfhydryl groups and disulphide linkage in the thermal aggregation of Z19 α-zein. *Biochim Biophys Acta* 2008, **1784**, p.1028-1036.

193. Wilmarth, P. *et al.* Age-related changes in human crystallins determined from comparative analysis of post-translational modifications in young and aged lens: does deamidation contribute to crystallin insolubility? *J Proteome Res* 2006, **5**, p.2554-2566.

194. Harms, M. J. *et al.* Laser light-scattering evidence for an altered association of βB1-crystallin deamidated in the connecting peptide. *Protein Sci* 2004, **13**, p.678-686.

195. Bouma, B. *et al.* Glycation induces formation of amyloid cross-β structure in albumin. *J Biol Chem* 2003, **278**, p.41810-41819.

196. Liu, D. *et al.* Structure and stability changes of human IgG1 Fc as a consequence of methionine oxidation. *Biochemistry* 2008, **47**, p.5088-5100.

197. Wang, W. Advanced protein formulations. *Protein Sci* 2015, **24**, p.1031-1039.

198. Timasheff, S. N. The control of protein stability and association by weak interactions with water: how do solvents affect these processes? *Annu Rev Biophys Biomol Struct* 1993, **22**, p.67-97.

199. Qoronfleh, M. W., Hesterberg, L. K. and Seefeldt, M. B. Confronting high-throughput protein refolding using high pressure and solution screens. *Protein Expr Purif* 2007, **55**, p.209-224.

200. Tsumoto, K. *et al.* Role of arginine in protein refolding, solubilization, and purification. *Biotechnol Prog* 2004, **20**, p.1301-1308.

201. Arakawa, T. and Tsumoto, K. The effects of arginine on refolding of aggregated proteins: not facilitate refolding, but suppress aggregation. *Biochem Biophys Res Commun* 2003, **304**, p.148-152.

202. Chen, J. *et al.* Cooperative effects of urea and l-arginine on protein refolding. *Protein Expr Purif* 2009, **66**, p.82-90.

203. Zhao, D. *et al.* Interaction of arginine with protein during refolding process probed by amide H/D exchange mass spectrometry and isothermal titration calorimetry. *Biochim Biophys Acta* 2015, **1854**, p.39-45.

204. Reddy, K. *et al.* L-Arginine increases the solubility of unfolded species of hen egg white lysozyme. *Protein Sci* 2005, **14**, p.929-935.

205. Arakawa, T. *et al.* Suppression of protein interactions by arginine: a proposed mechanism of the arginine effects. *Biophys Chem* 2007, **127**, p.1-8.

206. Randolph, T. W. and Jones, L. S. Surfactant-protein interactions. *Pharm Biotechnol* 2002, **13**, p.159-175.

207. Hoffmann, C. *et al.* Insights into protein-polysorbate interactions analysed by means of isothermal titration and differential scanning calorimetry. *Eur Biophys J* 2009, **38**, p.557-568.

208. Joshi, O. and McGuire, J. Adsorption behavior of lysozyme and Tween 80 at hydrophilic and hydrophobic silica-water interfaces. *Appl Biochem Biotechnol* 2009, **152**, p.235-248.

209. Panyukov, Y. V. *et al.* Surfactant-induced amorphous aggregation of tobacco mosaic virus coat protein: a physical methods approach. *Macromol Biosci* 2008, **8**, p.199-209.

210. Schiffter, H. Pharmaceutical proteins—Structure, stability, and formulation. Comprehensive biotechnology. Academic Press, Burlington, Vermont 2011, p.521-541.

211. Harris, J. M. and Chess, R. B. Effect of pegylation on pharmaceuticals. *Nat Rev Drug Discov* 2003, **2**, p.214-221.

212. Rodriguez-Martinez, J. A. *et al.* Stabilization of α-chymotrypsin upon PEGylation correlates with reduced structural dynamics. *Biotechnol Bioeng* 2008, **101**, p.1142-1149.

213. Li, Y. *et al.* Characterization of the self-association of human interferon-α2b, albinterferon-α2b, and pegasys. *J Pharm Sci* 2012, **101**, p.68-80.

214. Rajan, R. S. *et al.* Modulation of protein aggregation by polyethylene glycol conjugation: GCSF as a case study. *Protein Sci* 2006, **15**, p.1063-1075.

215. da Silva Freitas, D. and Abrahão-Neto, J. Biochemical and biophysical characterization of lysozyme modified by PEGylation. *Int J Pharm* 2010, **392**, p.111-117.

216. Harris, J. M., Martin, N. E. and Modi, M. Pegylation. *Clin Pharmacokinet* 2001, **40**, p.539-551.

217. Gaberc-Porekar, V. *et al.* Obstacles and pitfalls in the PEGylation of therapeutic proteins. *Curr Opin Drug Discov Devel* 2008, **11**, 242.

218. Solá, R. J. and Griebenow, K. Effects of glycosylation on the stability of protein pharmaceuticals. *J Pharm Sci* 2009, **98**, p.1223-1245.

219. Solá, R. J., Al-Azzam, W. and Griebenow, K. Engineering of protein thermodynamic, kinetic, and colloidal stability: Chemical glycosylation with monofunctionally activated glycans. *Biotechnol Bioeng* 2006, **94**, p.1072-1079.

220. Solá, R., Rodriguez-Martinez, J. and Griebenow, K. Modulation of protein biophysical properties by chemical glycosylation: biochemical insights and biomedical implications. *Cell Mol Life Sci* 2007, **64**, p.2133-2152.

221. Solá, R. J. and Griebenow, K. Chemical glycosylation: new insights on the interrelation between protein structural mobility, thermodynamic stability, and catalysis. *FEBS Lett* 2006, **580**, p.1685-1690.

222. Chennamsetty, N. *et al.* Design of therapeutic proteins with enhanced stability. *Proc Natl Acad Sci U S A* 2009, **106**, p.11937-11942.

223. Perchiacca, J. M. *et al.* Aggregation-resistant domain antibodies engineered with charged mutations near the edges of the complementarity-determining regions. *Protein Eng Des Sel* 2012, **25**, p.591-602.

224. Dudgeon, K. *et al.* General strategy for the generation of human antibody variable domains with increased aggregation resistance. *Proc Natl Acad Sci U S A* 2012, **109**, p.10879-10884.

225. Wu, S.-J. *et al.* Structure-based engineering of a monoclonal antibody for improved solubility. *Protein Eng Des Sel* 2010, **23**, p.643-651.

226. Lauer, T. M. *et al.* Developability Index: A rapid *in silico* tool for the screening of antibody aggregation propensity. *J Pharm Sci* 2012, **101**, p.102-115.

227. Chennamsetty, N. *et al.* Aggregation-prone motifs in human immunoglobulin G. *J Mol Biol* 2009, **391**, p.404-413.

228. Buck, P. M. *et al.* Computational methods to predict therapeutic protein aggregation. Therapeutic Proteins: Methods and Protocols. Humana Press, New Jersey 2012, p.425-451.

229. Buck, P. M., Kumar, S. and Singh, S. K. Insights into the potential aggregation liabilities of the b12 Fab fragment via elevated temperature molecular dynamics. *Protein Eng Des Sel* 2013, **26**, p.195-205.

230. Presnell, S. R. and Cohen, F. E. Topological distribution of four-alpha-helix bundles. *Proc Natl Acad Sci U S A* 1989, **86**, p.6592-6596.

231. Sprang, S. R. and Fernando Bazan, J. Cytokine structural taxonomy and mechanisms of receptor engagement. *Curr Opin Struct Biol* 1993, **3**, p.815-827.

232. Rozwarski, D. A. *et al.* Structural comparisons among the short-chain helical cytokines. *Structure* 1994, **2**, p.159-173.

233. Hill, E. E., Morea, V. and Chothia, C. Sequence conservation in families whose members have little or no sequence similarity: the four-helical cytokines and cytochromes. *J Mol Biol* 2002, **322**, p.205-233.

234. Jiang, X. *et al.* Structure of the active core of human stem cell factor and analysis of binding to its receptor kit. *The EMBO journal* 2000, **19**, p.3192-3203.

235. Pandit, J. *et al.* Three-dimensional structure of dimeric human recombinant macrophage colony-stimulating factor. *Science* 1992, **258**, p.1358-1362.

236. Milburn, M. V. *et al.* A novel dimer configuration revealed by the crystal structure at 2.4 Å resolution of human interleukin-5. *Nature* 1993, **363**, p.172-176.

237. Savvides, S. N., Boone, T. and Karplus, P. A. Flt3 ligand structure and unexpected commonalities of helical bundles and cystine knots. *Nat Struct Mol Biol* 2000, **7**, p.486-491.

238. Radhakrishnan, R. *et al.* Zinc mediated dimer of human interferon-α2b revealed by X-ray crystallography. *Structure* 1996, **4**, p.1453-1463.

239. Karpusas, M. *et al.* The crystal structure of human interferon β at 2.2-Å resolution. *Proc Natl Acad Sci U S A* 1997, **94**, p.11813-11818.

240. Ealick, S. *et al.* Three-dimensional structure of recombinant human interferon-gamma. *Science* 1991, **252**, p.698-702.

241. Zdanov, A. *et al.* Crystal structure of interleukin-10 reveals the functional dimer with an unexpected topological similarity to interferon γ. *Structure* 1995, **3**, p.591-601.

242. Diederichs, K., Boone, T. and Karplus, P. Novel fold and putative receptor binding site of granulocyte-macrophage colony-stimulating factor. *Science* 1991, **254**, p.1779-1782.

243. Clark, S. C. and Kamen, R. The human hematopoietic colony-stimulating factors. *Science* 1987, **236**, p.1229-1237.

244. Brandhuber, B. *et al.* Three-dimensional structure of interleukin-2. *Science* 1987, **238**, p.1707-1709.

245. Wlodaver, A., Pavlovsky, A. and Gustchina, A. Crystal structure of human recombinant interleukin-4 at 2.25 Å resolution. *FEBS Lett* 1992, **309**, p.59-64.

246. Yokota, T. *et al.* Molecular biology of interleukin 4 and interleukin 5 genes and biology of their products that stimulate B cells, T cells and hemopoietic cells. *Immunol Rev* 1988, **102**, p.137-187.

247. Fichelson, S. The FLT3/FLK2 ligand: structure, functions and prospects. *Eur Cytokine Netw* 1998, **9**, p.7-22.

248. Syed, R. S. *et al.* Efficiency of signalling through cytokine receptors depends critically on receptor orientation. *Nature* 1998, **395**, p.511-516.

249. Cheetham, J. C. *et al.* NMR structure of human erythropoietin and a comparison with its receptor bound conformation. *Nat Struct Mol Biol* 1998, **5**, p.861-866.

250. Graber, S. E. and Krantz, S. Erythropoietin and the control of red cell production. *Annu Rev Med* 1978, **29**, p.51-66.

251. Hill, C. P., Osslund, T. D. and Eisenberg, D. The structure of granulocyte-colony-stimulating factor and its relationship to other growth factors. *Proc Natl Acad Sci U S A* 1993, **90**, p.5167-5171.

252. Roberts, A. W. G-CSF: A key regulator of neutrophil production, but that's not all! *Growth Factors* 2005, **23**, p.33-41.

253. Chantalat, L. *et al.* The crystal structure of wild-type growth hormone at 2.5 Å resolution. *Protein Pept Lett* 1995, **2**, p.333-340.

254. Chawla, R. K., Parks, J. S. and Rudman, D. Structural variants of human growth hormone: biochemical, genetic, and clinical aspects. *Annu Rev Med* 1983, **34**, p.519-545.

255. Somers, W., Stahl, M. and Seehra, J. S. 1.9 Å crystal structure of interleukin 6: implications for a novel mode of receptor dimerization and signaling. *The EMBO journal* 1997, **16**, p.989-997.

256. Putoczki, T. L., Dobson, R. C. and Griffin, M. D. The structure of human interleukin-11 reveals receptor-binding site features and structural differences from interleukin-6. *Acta Crystallogr D Biol Crystallogr* 2014, **70**, p.2277-2285.

257. Du, X. and Williams, D. Interleukin-11: a multifunctional growth factor derived from the hematopoietic microenvironment. *Blood* 1994, **83**, p.2023-2030.

258. Robinson, R. *et al.* The crystal structure and biological function of leukemia inhibitory factor: implications for receptor binding. *Cell* 1994, **77**, p.1101-1116.

259. Metcalf, D. Leukemia inhibitory factor—a puzzling polyfunctional regulator. *Growth Factors* 1992, **7**, p.169-173.

260. McDonald, N., Panayotatos, N. and Hendrickson, W. Crystal structure of dimeric human ciliary neurotrophic factor determined by MAD phasing. *The EMBO journal* 1995, **14**, 2689.

261. Deller, M. C. *et al.* Crystal structure and functional dissection of the cytostatic cytokine oncostatin M. *Structure* 2000, **8**, p.863-874.

262. Grant, S. L. and Begley, C. G. The oncostatin M signalling pathway: reversing the neoplastic phenotype? *Mol Med Today* 1999, **5**, p.406-412.

263. Zhang, F. *et al.* Crystal structure of the obese protein Ieptin-E100. *Nature* 1997, **387**, p.206-209.

264. Pelleymounter, M. *et al.* Effects of the obese gene product on body weight regulation in ob/ob mice. *Science* 1995, **269**, p.540-543.

265. Elkins, P. A. *et al.* Ternary complex between placental lactogen and the extracellular domain of the prolactin receptor. *Nat Struct Mol Biol* 2000, **7**, p.808-815.

266. Bekisz, J. *et al.* Human interferons alpha, beta and omega. *Growth Factors* 2004, **22**, p.243-251.

267. Miller, C. H. T., Maher, S. G. and Young, H. A. Clinical use of interferon-γ. *Ann N Y Acad Sci* 2009, **1182**, p.69-79.

268. Ihle, J. N. Cytokine receptor signalling. *Nature* 1995, **377**, p.591-594.

269. Gadina, M. *et al.* Signaling by Type I and II cytokine receptors: ten years after. *Curr Opin Immunol* 2001, **13**, p.363-373.

270. Ferrao, R. *et al.* The structural basis for class II cytokine receptor recognition by JAK1. *Structure* 2016, **24**, p.897-905.

271. Constantinescu, S. N. *et al.* The erythropoietin receptor cytosolic juxtamembrane domain contains an essential, precisely oriented, hydrophobic motif. *Mol Cell* 2001, **7**, p.377-385.

272. Hilton, D. Negative regulators of cytokine signal transduction. *Cell Mol Life Sci* 1999, **55**, p.1568-1577.

273. Mahlert, F. *et al.* Rational development of lipegfilgrastim, a novel long-acting granulocyte colony-stimulating factor, using glycopegylation technology. *Blood* 2013, **122**, p.4853-4853.

274. Egrie, J. C. and Browne, J. K. Development and characterization of novel erythropoiesis stimulating protein (NESP). *Br J Cancer* 2001, **84 Suppl 1**, p.3-10.

275. To, L. *et al.* Successful mobilization of peripheral blood stem cells after addition of ancestim (stem cell factor) in patients who had failed a prior mobilization with filgrastim (granulocyte colony-stimulating factor) alone or with chemotherapy plus filgrastim. *Bone Marrow Transplant* 2003, **31**, p.371-378.

276. Herman, A. C., Boone, T. C. and Lu, H. S. Characterization, formulation, and stability of Neupogen (Filgrastim), a recombinant human granulocyte-colony stimulating factor. *Pharm Biotechnol* 1996, **9**, p.303-328.

277. Piedmonte, D. M. and Treuheit, M. J. Formulation of Neulasta® (pegfilgrastim). *Adv Drug Deliv Rev* 2008, **60**, p.50-58.

278. Oh-Eda, M. *et al.* O-linked sugar chain of human granulocyte colony-stimulating factor protects it against polymerization and denaturation allowing it to retain its biological activity. *J Biol Chem* 1990, **265**, p.11432-11435.

279. Delorme, E. *et al.* Role of glycosylation on the secretion and biological activity of erythropoietin. *Biochemistry* 1992, **31**, p.9871-9876.

280. Tsuda, E. *et al.* The role of carbohydrate in recombinant human erythropoietin. *Eur J Biochem* 1990, **188**, p.405-411.

281. Souza, L. M. *et al.* Recombinant human granulocyte colony-stimulating factor: effects on normal and leukemic myeloid cells. *Science* 1986, **232**, p.61-65.

282. Kubota, N. *et al.* Structural characterization of natural and recombinant human granulocyte colony-stimulating factors. *J Biochem* 1990, **107**, p.486-492.

283. Zink, T. *et al.* Structure and dynamics of the human granulocyte colony-stimulating factor determined by NMR spectroscopy. Loop mobility in a four-helix-bundle protein. *Biochemistry* 1994, **33**, p.8453-8463.

284. Wingfield, P. *et al.* Characterization of recombinant-derived granulocyte-colony stimulating factor (G-CSF). *Biochem J* 1988, **256**, p.213-218.

285. Lu, H. S. *et al.* Folding and oxidation of recombinant human granulocyte colony stimulating factor produced in *Escherichia coli*. Characterization of the disulfide-reduced intermediates and cysteine-serine analogs. *J Biol Chem* 1992, **267**, p.8770-8777.

286. Aritomi, M. *et al.* Atomic structure of the GCSF-receptor complex showing a new cytokine-receptor recognition scheme. *Nature* 1999, **401**, p.713-717.

287. Tamada, T. *et al.* Homodimeric cross-over structure of the human granulocyte colony-stimulating factor (GCSF) receptor signaling complex. *Proc Natl Acad Sci U S A* 2006, **103**, p.3135-3140.

288. Werner, J. M. *et al.* Secondary structure and backbone dynamics of human granulocyte colony-stimulating factor in solution. *Biochemistry* 1994, **33**, p.7184-7192.

289. The PyMOL molecular graphics system, Version 1.7 Schrödinger, LLC.

290. Kawakami, M. *et al.* Levels of serum granulocyte colony-stimulating factor in patients with infections. *Blood* 1990, **76**, p.1962-1964.

291. Nicholson, S. E. *et al.* Tyrosine kinase JAK1 is associated with the granulocyte-colony-stimulating factor receptor and both become tyrosine-phosphorylated after receptor activation. *Proc Natl Acad Sci U S A* 1994, **91**, p.2985-2988.

292. Tian, S. *et al.* Rapid activation of the STAT3 transcription factor by granulocyte colony-stimulating factor. *Blood* 1994, **84**, p.1760-1764.

293. Corey, S. J. *et al.* Granulocyte colony-stimulating factor receptor signaling involves the formation of a three-component complex with Lyn and Syk protein-tyrosine kinases. *Proc Natl Acad Sci U S A* 1994, **91**, p.4683-4687.

294. Ward, A. C. *et al.* The Src-like tyrosine kinase Hck is activated by granulocyte colony-stimulating factor (G-CSF) and docks to the activated G-CSF receptor. *Biochem Biophys Res Commun* 1998, **251**, p.117-123.

295. de Koning, J. *et al.* The membrane-distal cytoplasmic region of human granulocyte colony- stimulating factor receptor is required for STAT3 but not STAT1 homodimer formation. *Blood* 1996, **87**, p.1335-1342.

296. Ward, A. C. *et al.* Tyrosine-dependent and -independent mechanisms of STAT3 activation by the human granulocyte colony-stimulating factor (G-CSF) receptor are differentially utilized depending on G-CSF concentration. *Blood* 1999, **93**, p.113-124.

297. Wang, L. *et al.* Repression of c-Cbl leads to enhanced G-CSF Jak-STAT signaling without increased cell proliferation. *Oncogene* 2002, **21**, p.5346-5355.

298. Dong, F. and Larner, A. C. Activation of Akt kinase by granulocyte colony-stimulating factor (G-CSF): evidence for the role of a tyrosine kinase activity distinct from the Janus kinases. *Blood* 2000, **95**, p.1656-1662.

299. Tian, S.-S. *et al.* Multiple signaling pathways induced by granulocyte colony-stimulating factor involving activation of JAKs, STAT5, and/or STAT3 are required for regulation of three distinct classes of immediate early genes. *Blood* 1996, **88**, p.4435-4444.

300. de Koning, J. P. *et al.* Specific involvement of tyrosine 764 of human granulocyte colony-stimulating factor receptor in signal transduction mediated by p145/Shc/GRB2 or p90/GRB2 complexes. *Blood* 1996, **87**, p.132-140.

301. Hörtner, M. *et al.* Suppressor of cytokine signaling-3 is recruited to the activated granulocyte-colony stimulating factor receptor and modulates its signal transduction. *J Immunol* 2002, **169**, p.1219-1227.

302. Lee, C.-k. *et al.* STAT3 is a negative regulator of granulopoiesis but is not required for G-CSF-dependent differentiation. *Immunity* 2002, **17**, p.63-72.

303. Chi, E. Y. *et al.* Roles of conformational stability and colloidal stability in the aggregation of recombinant human granulocyte colony-stimulating factor. *Protein Sci* 2003, **12**, p.903-913.

304. Brems, D. N. The kinetics of G-CSF folding. *Protein Sci* 2002, **11**, p.2504-2511.

305. Arakawa, T. *et al.* Cysteine 17 of recombinant human granulocyte-colony stimulating factor is partially solvent-exposed. *J Protein Chem* 1993, **12**, p.525-531.

306. Kolvenbach, C. G. *et al.* Characterization of two fluorescent tryptophans in recombinant human granulocyte-colony stimulating factor: Comparison of native sequence protein and tryptophan-deficient mutants. *J Protein Chem* 1993, **12**, p.229-236.

307. Banks, D. D., Zhang, J. and Siska, C. C. Relationship between native-state solubility and non-native aggregation of recombinant human granulocyte colony stimulating factor: practical implications for protein therapeutic development. *Mol Pharm* 2014, **11**, p.3431-3442.

308. Krishnan, S. *et al.* Aggregation of granulocyte colony stimulating factor under physiological conditions: characterization and thermodynamic inhibition. *Biochemistry* 2002, **41**, p.6422-6431.
309. Zhang, J. *et al.* Effects of sucrose and benzyl alcohol on GCSF conformational dynamics revealed by hydrogen deuterium exchange mass spectrometry. *J Pharm Sci* 2015, **104**, p.1592-1600.
310. Sambrook, J., Fritsch, E. F. and Maniatis, T. Molecular cloning. Cold spring harbor laboratory press, New York, 2nd Edition 1989.
311. Gill, S. C. and von Hippel, P. H. Calculation of protein extinction coefficients from amino acid sequence data. *Anal Biochem* 1989, **182**, p.319-326.
312. Pace, C. Determination and analysis of urea and guanidine hydrochloride denaturation curves. *Methods Enzymol* 1986, **131**, p.266-280.
313. Kelly, S. M. and Price, N. C. The application of circular dichroism to studies of protein folding and unfolding. *Biochim Biophys Acta* 1997, **1338**, p.161-185.
314. Greenfield, N. J. and Fasman, G. D. Computed circular dichroism spectra for the evaluation of protein conformation. *Biochemistry* 1969, **8**, p.4108-4116.
315. Morrisett, J. D. *et al.* Interaction of an apolipoprotein (apoLP-alanine) with phosphatidylcholine. *Biochemistry* 1973, **12**, p.1290-1299.
316. Chu, B. Laser light scattering: basic principles and practice. Academic Press, ISBN, 2nd Edition 1991.
317. Koppel, D. E. Analysis of macromolecular polydispersity in intensity correlation spectroscopy: the method of cumulants. *J Chem Phys* 1972, **57**, p.4814-4820.
318. Provencher, S. W. Inverse problems in polymer characterization: Direct analysis of polydispersity with photon correlation spectroscopy. *Die Makromolekulare Chemie* 1979, **180**, p.201-209.
319. Cavanagh, J. *et al.* Protein NMR spectroscopy: principles and practice. Academic Press, Massachusetts, 2nd Edition 1995.
320. Delaglio, F. *et al.* NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* 1995, **6**, p.277-293.
321. Vranken, W. F. *et al.* The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins* 2005, **59**, p.687-696.
322. Morris, G. A. and Freeman, R. Enhancement of nuclear magnetic resonance signals by polarization transfer. *J Am Chem Soc* 1979, **101**, p.760-762.

323. Bodenhausen, G. and Ruben, D. J. Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. *Chem Phys Lett* 1980, **69**, p.185-189.

324. Keeler, J. Understanding NMR Spectroscopy. John Wiley & Sons, New Jersey, 1st Edition 2005.

325. Kay, L. E. *et al.* Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. *J Magn Reson* 1990, **89**, p.496-514.

326. Grzesiek, S. and Bax, A. Improved 3D triple-resonance NMR techniques applied to a 31 kDa protein. *J Magn Reson* 1992, **96**, p.432-440.

327. Farmer, B. T. *et al.* A refocused and optimized HNCA: increased sensitivity and resolution in large macromolecules. *J Biomol NMR* 1992, **2**, p.195-202.

328. Bax, A. and Ikura, M. An efficient 3D NMR technique for correlating the proton and ^{15}N backbone amide resonances with the α-carbon of the preceding residue in uniformly ^{15}N/^{13}C enriched proteins. *J Biomol NMR* 1991, **1**, p.99-104.

329. Muhandiram, D. R. and Kay, L. E. Gradient-enhanced triple-resonance three-dimensional NMR experiments with improved sensitivity. *J Magn Reson* 1994, **103**, p.203-216.

330. Clubb, R. T., Thanabal, V. and Wagner, G. A constant-time three-dimensional triple-resonance pulse scheme to correlate intraresidue ^{1}HN, ^{15}N and ^{13}C' chemical shifts in ^{15}N/^{13}C-labelled proteins. *J Magn Reson* 1992, **97**, p.213-217.

331. Grzesiek, S. and Bax, A. Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR. *J Am Chem Soc* 1992, **114**, p.6291-6293.

332. Williamson, M. P. Using chemical shift perturbation to characterise ligand binding. *Prog Nucl Magn Reson Spectrosc* 2013, **73**, p.1-16.

333. Hvidt, A. A discussion of the pH dependence of the hydrogen-deuterium exchange of proteins. *C R Trav Lab Carlsberg* 1964, **34**, 299.

334. Hvidt, A. and Nielsen, S. O. Hydrogen exchange in proteins. *Adv Protein Chem* 1966, **21**, p.287-386.

335. Englander, S. W. and Kallenbach, N. R. Hydrogen exchange and structural dynamics of proteins and nucleic acids. *Q Rev Biophys* 1983, **16**, p.521-655.

336. Bai, Y. *et al.* Primary structure effects on peptide group hydrogen exchange. *Proteins* 1993, **17**, p.75-86.

337. Richa, T. and Sivaraman, T. CIntX: A software tool for calculating the intrinsic exchange rates of labile protons in proteins. *J Pharm Sci & Res* 2012, **4**, p.1852-1858.

338. Glasoe, P. K. and Long, F. Use of glass electrodes to measure acidities in deuterium oxide. *J Phys Chem* 1960, **64**, p.188-190.

339. Schanda, P., Kupče, Ē. and Brutscher, B. SOFAST-HMQC experiments for recording two-dimensional deteronuclear correlation spectra of proteins within a few seconds. *J Biomol NMR* 2005, **33**, p.199-211.

340. Allerhand, A., Doddrell, D. and Komoroski, R. Natural abundance carbon-13 partially relaxed fourier transform nuclear magnetic resonance spectra of complex molecules. *J Chem Phys* 1971, **55**, p.189-198.

341. Kay, L. E., Torchia, D. A. and Bax, A. Backbone dynamics of proteins as studied by nitrogen-15 inverse detected heteronuclear NMR spectroscopy: application to staphylococcal nuclease. *Biochemistry* 1989, **28**, p.8972-8979.

342. Carr, H. Y. and Purcell, E. M. Effects of diffusion on free precession in nuclear magnetic resonance experiments. *Phys Rev* 1954, **94**, 630.

343. Meiboom, S. and Gill, D. Modified spin-echo method for measuring nuclear relaxation times. *Rev Sci Instrum* 1958, **29**, p.688-691.

344. Studier, F. W. and Moffatt, B. A. Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J Mol Biol* 1986, **189**, p.113-130.

345. Rizzi, C. *et al.* DAHP synthase from *Mycobacterium tuberculosis* H37Rv: cloning, expression, and purification of functional enzyme. *Protein Expr Purif* 2005, **40**, p.23-30.

346. de Mendonca, J. D. *et al.* Functional characterization by genetic complementation of aroB-encoded dehydroquinate synthase from *Mycobacterium tuberculosis* H37Rv and its heterologous expression and purification. *J Bacteriol* 2007, **189**, p.6246-6252.

347. Magalhaes, M. L. *et al.* Cloning and expression of functional shikimate dehydrogenase (EC 1.1.1.25) from *Mycobacterium tuberculosis* H37Rv. *Protein Expr Purif* 2002, **26**, p.59-64.

348. Vanz, A. L. *et al.* Human granulocyte colony stimulating factor (hG-CSF): cloning, overexpression, purification and characterization. *Microb Cell Fact* 2008, **7**, 13.

349. Grossman, T. H. *et al.* Spontaneous cAMP-dependent derepression of gene expression in stationary phase plays a role in recombinant expression instability. *Gene* 1998, **209**, p.95-103.

350. Moffatt, B. A. and Studier, F. W. T7 lysozyme inhibits transcription by T7 RNA polymerase. *Cell* 1987, **49**, p.221-227.

351. Abolghasemi, D. S. *et al.* An efficient purification method for high recovery of recombinant human Granulocyte Colony Stimulating Factor from recombinant E. coli. *International J Environ Sci Dev* 2010, **1**.

352. Ramya, M. and Selvarajan, E. Purification of human recombinant granulocyte colony stimulating factor from *Escherichia coli*. *Afr J Biotechnol* 2012, **11**, p.11104-11109.

353. Kolvenbach, C. G. *et al.* Granulocyte-colony stimulating factor maintains a thermally stable, compact, partially folded structure at pH 2. *J Pept Res* 1997, **50**, p.310-318.

354. Narhi, L. O., Kenney, W. C. and Arakawa, T. Conformational changes of recombinant human granulocyte-colony stimulating factor induced by pH and guanidine hydrochloride. *J Protein Chem* 1991, **10**, p.359-367.

355. Royer, C. A. Probing protein folding and conformational transitions with fluorescence. *Chem Rev* 2006, **106**, p.1769-1784.

356. Pierce, D. W. and Boxer, S. G. Stark effect spectroscopy of tryptophan. *Biophys J* 1995, **68**, 1583.

357. Royer, C. A. *et al.* Effects of amino acid substitutions on the pressure denaturation of staphylococcal nuclease as monitored by fluorescence and nuclear magnetic resonance spectroscopy. *Biochemistry* 1993, **32**, p.5222-5232.

358. Colucci, W. J. *et al.* Conformational studies of a constrained tryptophan derivative: implications for the fluorescence quenching mechanism. *J Am Chem Soc* 1990, **112**, p.9182-9190.

359. Chen, Y. and Barkley, M. D. Toward understanding tryptophan fluorescence in proteins. *Biochemistry* 1998, **37**, p.9976-9982.

360. Shinitzky, M. and Fridkin, M. Spectral studies with histidyl-tryptophan peptides. *Eur J Biochem* 1969, **9**, p.176-181.

361. Shizuka, H. *et al.* Excited-state behavior of tryptamine and related indoles. Remarkably efficient intramolecular proton-induced quenching. *J Am Chem Soc* 1988, **110**, p.1726-1732.

362. Saito, I. *et al.* Photochemical hydrogen-deuterium exchange reaction of tryptophan: the role in nonradiative decay of singlet tryptophan. *J Am Chem Soc* 1984, **106**.

363. Chen, Y., Liu, B. and Barkley, M. D. Trifluoroethanol quenches indole fluorescence by excited-state proton transfer. *J Am Chem Soc* 1995, **117**, p.5608-5609.

364. Yu, H. T. *et al.* Fluorescence quenching in indoles by excited-state proton transfer. *J Am Chem Soc* 1992, **114**, p.8449-8454.

365. Ricci, R. W. and Kilichowski, K. B. Fluorescence quenching of the indole ring system by lanthanide ions. *J Phys Chem* 1974, **78**, p.1953-1956.

366. Petrich, J. W. *et al.* On the origin of nonexponential fluorescence decay in tryptophan and its derivatives. *J Am Chem Soc* 1983, **105**, p.3824-3832.

367. Munishkina, L. A. and Fink, A. L. Fluorescence as a method to reveal structures and membrane-interactions of amyloidogenic proteins. *Biochim Biophys Acta* 2007, **1768**, p.1862-1885.

368. Kane, J. F. Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. *Curr Opin Biotechnol* 1995, **6**, p.494-500.

369. Kurland, C. and Gallant, J. Errors of heterologous protein expression. *Curr Opin Biotechnol* 1996, **7**, p.489-493.

370. Kudla, G. *et al.* Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 2009, **324**, p.255-258.

371. Devlin, P. E. *et al.* Alteration of amino-terminal codons of human granulocyte-colony-stimulating factor increases expression levels and allows efficient processing by methionine aminopeptidase in *Escherichia coli*. *Gene* 1988, **65**, p.13-22.

372. Hajnsdorf, E. and Boni, I. V. Multiple activities of RNA-binding proteins S1 and Hfq. *Biochimie* 2012, **94**, p.1544-1553.

373. Qu, X. *et al.* Ribosomal protein S1 unwinds double-stranded RNA in multiple steps. *Proc Natl Acad Sci U S A* 2012, **109**, p.14458-14463.

374. Duval, M. *et al. Escherichia coli* ribosomal protein S1 unfolds structured mRNAs onto the ribosome for active translation initiation. *PLoS Biol* 2013, **11**, e1001731.

375. Allert, M., Cox, J. C. and Hellinga, H. W. Multifactorial determinants of protein expression in prokaryotic open reading frames. *J Mol Biol* 2010, **402**, p.905-918.

376. Jeong, K. J. and Lee, S. Y. Secretory production of human granulocyte colony-stimulating factor in *Escherichia coli*. *Protein Expr Purif* 2001, **23**, p.311-318.

377. Alrokayan, S. Chemical synthesis and improved expression of recombinant human granulocyte colony-stimulating factor cDNA. *Genet Mol Res* 2011, **10**, p.2671-2678.

378. Henaut, A. and Danchin, A. Analysis and predictions from *Escherichia coli* sequences, or *E. coli in silico*. *Escherichia coli and Salmonella: cellular and molecular biology* ASM Press, Washington DC 1996, p.2047-2066.

379. Wohlgemuth, I. *et al.* Evolutionary optimization of speed and accuracy of decoding on the ribosome. *Philos Trans R Soc Lond B Biol Sci* 2011, **366**, p.2979-2986.

380. Pavlov, M. Y. *et al.* Slow peptide bond formation by proline and other N-alkylamino acids in translation. *Proc Natl Acad Sci U S A* 2009, **106**, p.50-54.

381. Johansson, M. *et al.* pH-sensitivity of the ribosomal peptidyl transfer reaction dependent on the identity of the A-site aminoacyl-tRNA. *Proc Natl Acad Sci U S A* 2011, **108**, p.79-84.

382. Muto, H. and Ito, K. Peptidyl-prolyl-tRNA at the ribosomal P-site reacts poorly with puromycin. *Biochem Biophys Res Commun* 2008, **366**, p.1043-1047.

383. Wohlgemuth, I. *et al.* Modulation of the rate of peptidyl transfer on the ribosome by the nature of substrates. *J Biol Chem* 2008, **283**, p.32229-32235.

384. Doerfel, L. K. *et al.* Entropic contribution of elongation factor P to proline positioning at the catalytic center of the ribosome. *J Am Chem Soc* 2015, **137**, p.12997-13006.

385. Charneski, C. A. and Hurst, L. D. Positively charged residues are the major determinants of ribosomal velocity. *PLoS Biol* 2013, **11**, e1001508.

386. Li, G.-W., Oh, E. and Weissman, J. S. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 2012, **484**, p.538-541.

387. Holtkamp, W. *et al.* Cotranslational protein folding on the ribosome monitored in real time. *Science* 2015, **350**, p.1104-1107.

388. Nissley, D. A. *et al.* Accurate prediction of cellular co-translational folding indicates proteins can switch from post- to co-translational folding. *Nat Commun* 2016, **7**.

389. Buhr, F. *et al.* Synonymous codons direct cotranslational folding toward different protein conformations. *Mol Cell* 2016, **61**, p.341-351.

390. Vassar, P. S. and Culling, C. Fluorescent stains, with special reference to amyloid and connective tissues. *Arch Pathol* 1959, **68**, 487.

391. Natalello, A. *et al.* Biophysical characterization of Met-G-CSF: effects of different site-specific mono-pegylations on protein stability and aggregation. *PLoS One* 2012, **7**, e42511.

392. Pawar, A. P. *et al.* Prediction of “aggregation-prone” and “aggregation-susceptible” regions in proteins associated with neurodegenerative diseases. *J Mol Biol* 2005, **350**, p.379-392.

393. Strickland, E. H. and Beychok, S. Aromatic contributions to circular dichroism spectra of protein. *CRC Crit Rev Biochem* 1974, **2**, p.113-175.

394. Hawe, A., Sutter, M. and Jiskoot, W. Extrinsic fluorescent dyes as tools for protein characterization. *Pharm Res* 2008, **25**, p.1487-1499.

395. Weber, G. and Laurence, D. Fluorescent indicators of adsorption in aqueous solution and on the solid phase. *Biochem J* 1954, **56**.

396. Stryer, L. The interaction of a naphthalene dye with apomyoglobin and apohemoglobin: a fluorescent probe of non-polar binding sites. *J Mol Biol* 1965, **13**, p.482-495.

397. Matulis, D. and Lovrien, R. 1-anilino-8-naphthalene sulfonate anion-protein binding depends primarily on ion pair formation. *Biophys J* 1998, **74**, p.422-429.

398. Gasymov, O. K. and Glasgow, B. J. ANS fluorescence: Potential to augment the identification of the external binding sites of proteins. *Biochim Biophys Acta* 2007, **1774**, p.403-411.

399. Latypov, R. F. *et al.* Structural and thermodynamic effects of ANS binding to human interleukin-1 receptor antagonist. *Protein Sci* 2008, **17**, p.652-663.

400. Leichtling, B. H. and Klotz, I. M. Catalysis of hydrogen-deuterium exchange in polypeptides. *Biochemistry* 1966, **5**, p.4026-4037.

401. Vehlow, C. *et al.* CMView: Interactive contact map visualization and analysis. *Bioinformatics* 2011, **27**, p.1573-1574.

402. Bai, Y. *et al.* Protein folding intermediates: native-state hydrogen exchange. *Science* 1995, **269**, p.192-197.

403. Lou, S.-C. *et al.* The knotted protein UCH-L1 exhibits partially unfolded forms under native conditions that share common structural features with its kinetic folding intermediates. *J Mol Biol* 2016, **428**, p.2507-2520.

404. Feng, H., Vu, N.-D. and Bai, Y. Detection of a hidden folding intermediate of the third domain of PDZ. *J Mol Biol* 2005, **346**, p.345-353.

405. Chu, R. *et al.* Relationship between the native-state hydrogen exchange and folding pathways of a four-helix bundle protein. *Biochemistry* 2002, **41**, p.7998-8003.

406. Llinás, M. *et al.* The energetics of T4 lysozyme reveal a hierarchy of conformations. *Nat Struct Biol* 1999, **6**, p.1072-1078.

407. Kato, H. *et al.* The folding pathway of T4 lysozyme: An on-pathway hidden folding intermediate. *J Mol Biol* 2007, **365**, p.881-891.

408. Vu, N.-D., Feng, H. and Bai, Y. The folding pathway of barnase: The rate-limiting transition state and a hidden intermediate under native conditions. *Biochemistry* 2004, **43**, p.3346-3356.

409. Milkovic, N. M. *et al.* Transient sampling of aggregation-prone conformations causes pathogenic instability of a parkinsonian mutant of DJ-1 at physiological temperature. *Protein Sci* 2015, **24**, p.1671-1685.

410. Tobler, S. A. and Fernandez, E. J. Structural features of interferon-γ aggregation revealed by hydrogen exchange. *Protein Sci* 2002, **11**, p.1340-1352.

411. Kabsch, W. and Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983, **22**, p.2577-2637.

412. Hekkelman, M. L. and Vriend, G. MRS: a fast and compact retrieval system for biological data. *Nucleic Acids Res* 2005, **33**, p.W766-W769.

413. Jenik, M. *et al.* Protein frustratometer: a tool to localize energetic frustration in protein molecules. *Nucleic Acids Res* 2012, **40**, p.W348-W351.

414. Kendrick, B. S. *et al.* A transient expansion of the native state precedes aggregation of recombinant human interferon-γ. *Proc Natl Acad Sci U S A* 1998, **95**, p.14142-14146.

415. Kelly, J. W. Alternative conformations of amyloidogenic proteins govern their behavior. *Curr Opin Struct Biol* 1996, **6**, p.11-17.

416. Roberts, C. J. Therapeutic protein aggregation: mechanisms, design, and control. *Trends Biotechnol* 2014, **32**, p.372-380.