An integrative and predictive model for the influence of protein sequence, structure and excipients on aggregation propensity

A thesis submitted to the University of Manchester for the degree of PhD in the Faculty of Science and Engineering

2016

Spyros Charonis

School of Chemical Engineering and Analytical Science

List of Contents

Word Count: 46,778

List of Figures

List of Tables

List of Equations

List of Acronyms

Abbreviation	Meaning
AA	Amino Acid
APR	Aggregation-Propensity Region
ASP	Atomic Solvation Parameter
AUC	Area Under Curve
CO	Contact Order
FDPB	Finite Difference Poisson Boltzmann
GUI	Graphical User Interface
IDP	Intrinsically Disorder Protein
INS	Insoluble
PDB	Protein Data Bank
RCO	Relative Contact Order
ROC	Receiver Operating Characteristic
scFv	single chain variable fragment
SOL	soluble

Abstract

University: University of Manchester

Candidate: Spyros Charonis

Degree Title: Doctor of Philosophy

Thesis Title: An integrative and predictive model for the influence of protein sequence, structure, and excipients on aggregation propensity

Date: 27/09/2016

Loss of solubility and aggregation of proteins are important bottlenecks in modern bioprocessing pipelines, where formulation and large-scale production of therapeutic proteins such as antibodies is achieved. The mechanistic basis of protein aggregation propensity and solubility are actively investigated using experimental and computational techniques. A significant part of research in this field involves efforts to understand how sequence- and structure-based properties enable proteins to remain functional under conditions and conditions relevant to physiology and delivery of biotherapeutic agents.

Using sequence-based and structural features as well as physico-chemical properties, a model was developed to study how such descriptors can be used in a predictive capacity to separate soluble and insoluble proteins. Therapeutic protein datasets including antibody derivatives and non-antibody biologics were constructed so that their solubility could be studied using the descriptors. Surface charge, polarity, and sequence composition were tested against established thresholds for solubility of *E. coli* proteins. Surface non-polarity was verified as a consistent feature for separating soluble and insoluble therapeutics, in line with its established role as a key player for determining aggregation propensity in the broader scientific community. The ratio of lysine to arginine composition emerged as a novel sequence-based feature that contributes to solubility, where higher lysine composition is favourable for the solubility. There is potential to use this as a method for engineering proteins for higher solubility with minimal disruption to functionality.

The predictive model was subsequently expanded to include a broad array of sequence-based and 3D structural features. Quantitative proteomics studies with high-throughput data for protein solubility, abundance and concentration were used to construct datasets. Web accessible repositories of protein abundance in several species and plasma protein concentration were used to augment the data used to validate the model. Our findings reiterate previously established studies regarding protein length, charge-based properties and surface non-polarity as important descriptors for discriminating soluble and insoluble proteins. The sequence-level lysine/arginine ratio offers a novel perspective on potentially simple ways of protecting proteins against aggregation, which could prove useful for bioprocessing pipelines.

Protein-excipient interactions were studied using a dot product metric to measure the association of a set of crystallisation screen ligands with proteins in the PDB database. Enrichment for predicted small molecule (sugars and buffers) binding sites was observed, although the underlying reasons remain unclear without more sophisticated structure-based techniques.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Acknowledgements

What a trip it has been! The journey to completing a PhD is long and filled with many disappointing moments (and a few rewarding ones). I was fortunate to meet and interact with many bright and talented people along the way to whom I will always be grateful. First and foremost, I would like to deeply thank my supervisors, Dr Jim Warwicker and Dr Robin Curtis, for guiding me throughout the years and for offering me the opportunity to embark on this amazing journey in the first place. Their knowledge and enthusiasm for science were invaluable to my research over the past four years, as was their willingness to provide their scientific insight and mentorship.

I would further like to thank my family in Greece for their unconditional love and support throughout these years. My beloved parents, Aristeidis and Photini, and brother Georgios have always held faith in me through both good and bad times. I could not have completed this journey without them.

I would also like to extend my appreciation to the EPSRC Biochemical Engineering Centre for Doctoral Training at UCL for providing me with the financial support to carry out this research. My frequent visits to London were the source of numerous stimulating scientific conversations and provided me with the opportunity to interact with many great doctoral and postdoctoral researchers. I hope that the CDT continues to flourish and attract great scientific talent.

Finally, a great many thanks to my colleagues and friends at the Manchester Institute for Biotechnology – Max Hebditch, Nick Fowler, Luke Holloway, James Baker, Rose Keeling, Joe Flood, Alejandro Carballo, Ivan Sazanavets and Stefan Ivanov. They filled up the office and made these years a memorable experience, and also helped make Manchester's weather somewhat bearable. Many thanks to Konstantina for our afternoon coffees that brightened up the days!

Rationale for Thesis Submission in Alternative Format

Alternative format is suitable to my PhD because the work that was undertaken featured separate investigations within an umbrella topic, but with a distinct emphasis. Chapter 2 is part of a publication for which the supervising principal investigator (Dr Jim Warwicker) is the first author, but contains results that were a part of this thesis. Two pieces of work (chapters 3 and 4) are currently in preparation for publication. The thesis begins with an introduction covering the topics of protein solubility and aggregation, as well as electrostatic modelling techniques. Results chapters are divided according to sequence-based predction, 3D structure-based prediction and protein-excipient interaction prediction. A conclusion section at the end assesses the aims achieved and those not achieved, and proposes how future work in this direction should proceed.

Chapter 1. Introduction

1.1 Proteins

Proteins are large, highly complex organic molecules exhibiting a remarkable structural and functional versatility that renders them essential in all aspects of cellular life. They are by far the most prolific type of biological macromolecule, assuming vastly greater functions than other classes of the aforementioned (nucleic acids, carbohydrates, and lipids). Indeed, there is no known biological or cellular process that does not involve proteins in some capacity, and it follows that they have immense significance in all areas of theoretical and applied biomolecular science. Being the main components tasked with executing the myriad processes necessary for cellular function, they have evolved over the course of millions of years to accumulate an extremely diverse array of molecular functions.

A fundamental principle underlying all aspects of protein science is that structure dictates function. There is thus an intimate biological relationship between the structure and function of protein molecules, which implies that one cannot be studied independently of the other. Indeed, decades of research were required before a coherent set of principles describing the manner in which structure is utilized to influence function was established (Bourne and Weissig, 2003). These principles have been organized into a four-tiered abstract hierarchy used to describe protein structure in non-physical qualitative terms: primary, secondary, tertiary and quaternary structure.

1.1.1 Primary Structure

Proteins are linear polymers of amino acids, and the sequence of constituent amino acid residues is referred to as primary structure. Amino acids are organic molecules that contain an amino group ($-NH_2$), a carboxyl group (-COOH), and a hydrogen atom bonded to a central carbon atom (α-carbon or C_α). Additionally, all amino acids possess a side chain (R group) bonded to the C_α which distinguishes one amino acid from another. The side chain confers the chemical properties specific to each amino acid, and the set of amino acids used to build all naturally occurring proteins consists of 20 distinct side chains. The structure of a prototypical amino acid is illustrated schematically in figure 1.1.

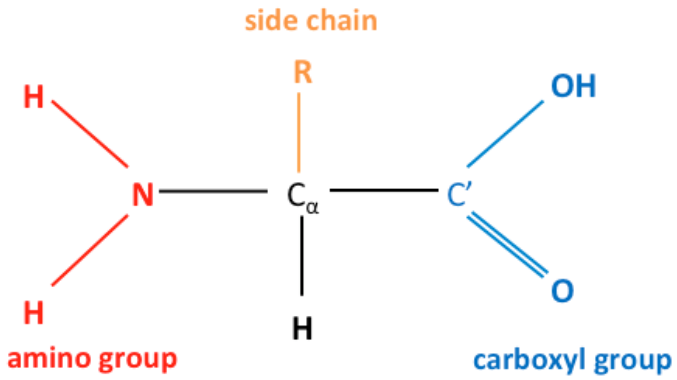


Figure 1.1. **Generic structure of an amino acid molecule**. All 20 naturally occurring have the same main chain (amino and carboxyl groups, C_α, hydrogen) but have distinct side chains (R group) giving them unique chemical properties.

The distinct sequence of amino acid residues determines the three-dimensional structure of a protein. There is an entire field devoted to using computational principles to predict the conformation that a protein will fold into based on its sequence. Amino acids form long polypeptide chains consisting of hundreds of covalently linked molecules. This polymerisation is driven by their ability to form bonds with each other through a reaction of their respective carboxyl and amino groups. The resulting bond is known as the peptide bond, and chains with multiple amino acids covalently linked by the formation of a linear succession of peptide bonds are termed polypeptides. All atoms of an amino acid that are involved in the peptide bond are referred to as the backbone to distinguish them from the side chain atoms.

The specific properties of the peptide bond have important implications for the three-dimensional structure that can be assumed by polypeptide chains. Most importantly, it is planar and rigid and as such a polypeptide chain has rotational freedom only about the bonds involving the C_α atoms. These bonds have been termed the Φ (C_α—N) and Ψ (C_α—C') angles, where rotation is possible but constrained by steric hindrance between the residue side chain and backbone atoms. Consequently, the allowed conformations of any polypeptide are quite limited. Ramachandran plots are used to specify the possible protein conformations corresponding to sterically allowed Φ and Ψ angle combinations (Ramachandran and Sasisekharan, 1968). All naturally occurring proteins are composed from the same set of 20 amino acids linked together in covalent bonds to form polymers known as polypeptides. Each of these amino acids consists of a main backbone made from an amino and a carboxyl functional group, and a side-chain (R group) unique to each residue. The 20 different side-chains give each amino acid different properties, which are commonly classified

according to their capacity to interact with water, the universal solvent. On the basis of this criterion, four classes of amino acids can be distinguished: polar, non-polar, acidic, and basic.

1.1.2 Secondary Structure

The secondary structure of a protein refers to the specific local conformation of the polypeptide chain. Ramachandran constraints corresponding to dihedral angles of the chain backbone as well as hydrogen bonding interactions dictate the secondary structure that can be adopted. Two types of secondary structure have emerged as the most dominant local conformations of polypeptide chains, α-helices and β-sheets. The dominance of these two types of local arrangement can be ascribed to their Φ and Ψ angle combinations falling within the two largest regions of allowed conformational space on a Ramachandran plot. In addition to satisfying Ramachandran criteria, helices and sheets are also stabilized by hydrogen bond interactions between backbone atoms. Furthermore, α-helices and β-sheets exhibit a high degree of regularity, with their specific Φ and Ψ angle combinations in the polypeptide chain being approximately repeated over the entire length of the secondary structure (Bourne and Weissig, 2003).

The α-helix structure is characterised by a curving of the polypeptide backbone so that a regular coil shape is formed. Unlike helices, β-sheets are formed by hydrogen bonds between adjacent polypeptide chains, where sections of the chain participating in the sheet are known as β-strands. Two configurations of β-sheet are possible, parallel and antiparallel. In parallel sheets, the strands are arranged with their backbones aligned in the same direction with respect to their amino (N) and carboxyl (C) terminal ends. In antiparallel sheets, the strands alternate in the opposite orientation, and β-sheets can also form in a mixed configuration containing both parallel and antiparallel sections, although this is less common than the uniform types.

1.1.3 Tertiary Structure

The tertiary structure of a protein is defined as the global three-dimensional structure of its polypeptide chain. The global conformation assumed by a polypeptide depends on the manner in which its constituent helix, sheet, and loop secondary structure elements combine to produce a complete structure. In contrast to secondary structure which is largely dependent on backbone interactions, tertiary structure arises predominantly from interactions between the polypeptide's side chains. One common molecular interaction in protein tertiary structure is the hydrophobic effect

(Tanford, 1979) which describes how residues with hydrophobic side chains are packed into the core of the protein so that they are inaccessible to the solvent, while polar and charged side chains comprise the protein surface and can interact with water molecules and solvated ions.

1.1.4 Quaternary Structure

Tertiary structure describes the three-dimensional spatial organization of a single polypeptide chain, but protein molecules are seldom monomeric (single chain), and instead usually exist as multiple independently folded polypeptide chains (multimers) linked together via non-covalent interactions. Such multimeric proteins are said to possess quaternary structure, and their monomeric subunits may be identical (homomeric) or different (heteromeric).

The formation of multimeric proteins is a highly specific interaction. Quaternary structures are stabilized by the same types of interactions employed in secondary and tertiary structure stabilization. The surface regions involved in monomer subunit interactions generally resemble the cores of globular proteins. The regions are comprised of residues with non-polar side chains and residues that can form hydrogen bonds (Bourne and Weissig, 2003).

1.1.5 Protein Folding

The entire three-dimensional tertiary structure of a protein is commonly referred to as its fold. Accordingly, the physical process by which a polypeptide acquires its native three-dimensional (tertiary) structure, a conformation that is usually associated with a function, is termed folding. Folding is a highly complex process that is not yet completely understood, and is one of the most important topics in all of computational biology. Despite the deterministic nature of protein folding, it is not yet possible to accurately predict the final structure of a protein given only its sequence although it is generally accepted that all information for proper folding is contained within the amino acid sequence. Taking into consideration the number of secondary structure element permutations, it may appear plausible to assume that the number of possible distinct protein folds is almost infinite, *i.e.* the conformational space of polypeptide chains is almost limitless. Interestingly, currently available protein structural data suggest that fold space, the universe of all observed folds, is in fact quite limited. The remarkable propensity of proteins to adopt specific, well-defined folds suggests that evolutionary pressure has imposed specific structural conformations. The concept describing how proteins are able to fold rapidly into their native,

functional state relative to how long it would take them to search the entire set of possible configurations (conformational space) is known as Levinthal's paradox (Zwanzig *et al.*, 1992), and is an important principle for understanding protein folding dynamics.

Protein folding is driven by several forces, one of which is the requirement of a protein to sequester its hydrophobic side chains from water. This sequestering allows the polar groups (amino and carboxyl) of the protein's backbone to form hydrogen bonds. The fold that a protein will adopt depends on the composition of amino acids in its primary sequence and their ordering. As outlined above, two types of SSE (secondary structure elements) occur naturally – the α-helix and the β-sheet. The former is formed via hydrogen bonding between amino and carboxyl groups of the same strand while the latter is formed by hydrogen bonding with other strands. A hydrogen bond is a special case of a dipole interaction driven by an attraction between a hydrogen atom chemically bonded to an electronegative atom of one molecule and an electronegative atom of another molecule. An overview of the secondary, tertiary, and quaternary hierarchies of protein structure and the formation of α-helices and β-sheets via hydrogen bonding is illustrated in figure 1.2. The electronegative atom, which is most often oxygen in proteins, has a partial negative charge. The hydrogen atom hence has a partial positive charge.

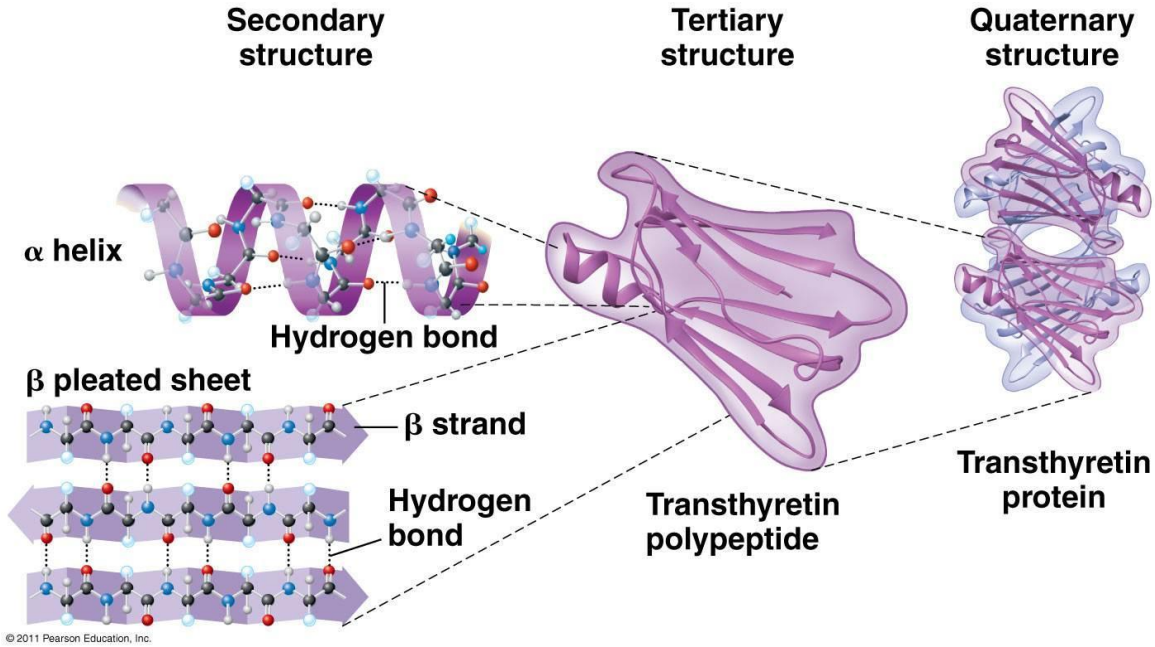


Figure 1.2.

Hierarchy of protein structure. Intramolecular hydrogen bonding gives rise to the α-helix SSE while intermolecular hydrogen bonding gives rise to the β-sheet SSE. Secondary structural elements combine to give a three dimensional configuration which often is repeated in individual subunits comprising the protein. Figure adapted from Campbell.

1.2 Biotechnology and Bioprocessing

Biotechnology is the use of organisms and biological systems to develop useful products. Recombinant genetic technologies, applied immunology, tissue engineering, novel diagnostic techniques, therapeutic proteins used as pharmaceutical agents, as well as industrial protein production processes are all instances of biotechnological applications. Although biotechnology often has health-related objectives and as such makes heavy use of biological sciences (*e.g.* genetics, microbiology, and biochemistry), methodologies and expertise from non-biological fields are frequently employed. Several fields such as chemical engineering, bioinformatics and more generally computational science often lend concepts to biotechnology.

Computational biological concepts such as sequence analysis, protein structure prediction, as well as methodologies from genomics and proteomics are increasingly necessary to cater to the data analysis requirements of biotechnology incentives. These techniques aim to develop fast and efficient algorithms for extracting meaningful information from enormous amounts of data and detecting hidden correlations. The field of computational biology, or bioinformatics, is firmly grounded in statistical theory and aims to develop formal methods to describe complex biological systems as well as validated predictive models for the structure and function of important molecular targets (*e.g.* proteins-ligand systems).

1.2.1 Bioinformatics and Computational Biology

Bioinformatics is the scientific field devoted to the consistent management and extraction of useful information from biological data. Biologically significant data is most commonly found in the form of sequence (*e.g.* proteins, nucleic acids) and structure (proteins), which ideally could be used to deterministically assign function. The inception of bioinformatics came about from a necessity to efficiently organize, analyse, and interpret a hitherto unanticipated surge of biological data. Initiatives such as the human genome project have transformed biology into a data-intensive field, exceeding the traditional experiment-based paradigm. What follows as a consequence of this transformation is the need to adjust the methodologies used to uncover biological principles accordingly. Indeed, a simple yet elegant example of how computational techniques can drive biological knowledge forward is found in the discovery of the APOA5 gene (Pennacchio *et al.*, 2001). This study contributed the discovery of a gene encoding an important apolipoprotein involved in cholesterol and triglyceride homeostasis. In spite of this gene being completely

unknown prior to this study, the discovery was made without employing any experimental methods, simply by aligning human and mouse genomic sequences and observing regions with unusually high similarity. As such, it serves to illustrate how computational techniques (*e.g.* sequence alignment) can provide the basis for revealing hitherto unknown functional genomic elements, and more generally how they can aid in accelerating research endeavors. The bioinformatics boom has had a great impact on biotechnology, and has been driven by the emergence of experimental techniques that generate data in a high-throughput fashion, such as DNA sequencing, microarray expression, mass spectrometry, and more recently next-generation sequencing. Effectively dealing with the complexity and size of datasets that emerge from such techniques is a central theme in bioinformatics.

Several key challenges in biotechnology and bioprocessing involve optimising production processes of therapeutic agents. This requires a comprehensive understanding of both intrinsic (protein-based) properties such as amino acid sequence, folding, and surface properties (charge- and polarity-based) and environmental properties (usually solution-based) such as pH, ionic strength, temperature and pressure that are relevant to therapeutic production pipelines. Efforts to attain an understanding of these properties and the complex interplay between them requires data acquisition and data analysis methodologies. Software applications for comparative sequence analysis, molecular visualisation, as well as protein structure prediction have become invaluable tools in various aspects of biotechnology research. Furthermore, computational methods contribute to building models of biochemical and biophysical processes that can ideally be used in a predictive capacity. This thesis focuses on how sequence- and structure-based properties can be used to predict protein solubility and solubility-related properties, one of the foremost challenges that industrial biotechnology is currently faced with.

1.3 Protein Solubility & Aggregation

Protein aggregation is a molecular phenomenon in which misfolded or unfolded proteins accumulate in either intracellular space (often termed inclusion bodies) or extracellular space. Such aggregates may occur both *in vitro* (such as in bioprocessing pipelines where they are detrimental to the production process involved), as well as *in vivo* (amyloid formations that are usually pathogenic). In the context of bioprocessing, aggregation defines a protein degradation pathway, which is usually made irreversible due to conformational changes in the protein structure and formation of strong inter- and intramolecular interactions (Roberts *et al*., 2015). The study of protein aggregates dates as far back as 80 years ago when the pioneering biophysicist William

Astbury postulated that all proteins might possess a fibrous state in addition to their functional globular state (Astbury *et al.*, 1935). This was the first paper describing the X-ray diffraction pattern of cross-beta fibrils, now accepted as the definitive fingerprint of the amyloid state of proteins. The significance of this study lies in its hinting that all proteins may have an inherent propensity to aggregate, and it has been suggested that this is averted in physiological conditions through evolved chaperoning effects (Dobson, 2004). This finding is of great importance in the context of understanding the evolution of protein folds. The majority of non-membrane proteins have functional forms in globular states, and formation of ordered aggregates (also termed amyloid fibrils) is largely associated with numerous severely debilitating neurodegenerative diseases (*e.g.* Parkinson's, Huntington's, synucleinopathies) in mammals (Dobson, 2003; Chiti and Dobson, 2006). However, functional proteins are also known to possess amyloid states, as is the case with the Pmel17 protein associated with skin pigmentation (Kelly and Balch, 2003), a number of secretory hormones (Maji *et al*., 2009), as well as fungal hydrophobins (De Simone *et al*., 2012).

In addition to the pathophysiological aspects of aggregation and its implications, it also comprises a major issue in the field of industrial protein production. Aggregation of protein-based biopharmaceuticals is the main bottleneck in bioprocessing pipelines, effectively restricting the spectrum of polypeptides obtainable using recombinant techniques (Ventura and Villaverde, 2006). Considerable costs are associated with purification, solubilisation, and renaturation of aggregated protein products (Clark, 2001), so it is desirable to understand this process on an *ab initio* molecular basis as well as to characterise its underlying mechanisms. Large-scale production of protein-based biotherapeutics is a growing industry (Roberts *et al*., 2015). Macromolecular drug compounds include a diverse range of products including growth factors, cytokines, hormones, receptors, enzymes, blood factors, anticoagulants, monoclonal antibodies (mAbs), fusion proteins, recombinant vaccines as well as nucleic acid-based products (Agrawal *et al*., 2011).

Biotherapeutics typically undergo numerous processing steps including production, purification, and formulation before they become clinically available active pharmaceutical agents. Therapeutics also usually spend a considerable amount of time in storage before administration. When stored for extended periods of time drug molecules are subjected to several physical stresses, including high concentrations, variable temperatures, pH extremes, varying ionic strengths, shear stress, and several solid-liquid interfaces (Cromwell *et al*., 2006). These stresses can impact the potency as well as the homogeneity of the active drug product via multiple molecular degradation pathways, the foremost of which is aggregation and loss of solubility (Manning *et al*., 1989). Formation of aggregates is effectively irreversible and a poorly understood degradation pathway at the molecular level. Furthermore, aggregated therapeutics frequently cause immunogenic responses when administered to patients (Agrawal *et al*., 2011). As such, this topic is studied extensively in at

the molecular level in order to characterise the kinetic bottlenecks that give rise to irreversible formation of oligomeric aggregates.

1.3.1 Protein Solubility

Solubility refers to the chemical property of a substance (termed the solute) in any phase of matter (solid, liquid, gas) to dissolve in another substance (termed the solvent) in its own phase so as to form a homogeneous solution of the solute in the solvent. The physical interpretation of solubility relates the chemical potential of a protein in the solution phase to that in the condensed phase. Chemical potential (usually denoted by μ) is often an elusive concept with multiple physical interpretations (Baierlein, 2000). The thermodynamic treatment considers it a form of potential energy that can be absorbed or released during a chemical reaction or phase transition. Under this definition, it can be defined for a molecular species as the mathematical gradient of free energy of a system with respect to a change in the number of moles of that species. Thus, it is the partial derivative of free energy with respect to the molar amount of the species, or $(\partial G/\partial N_i)_{T,P}$ representing the change in energy when one particle is added to a system at constant temperature T and pressure P. Free energy refers to the amount of physical work that a system can perform.

The solubility of a protein is defined in a thermodynamic sense as the concentration of that protein in a saturated solution that is in equilibrium with a condensed phase, either crystalline or amorphous, under a given set of conditions (Arakawa and Timasheff, 1985). It can be mathematically stated as the concentration at which chemical potential of the protein in solution is equal to that in precipitated phases, as given in equation 1.1 below.

$$S = \exp[-\frac{(\mu^S - \mu^C)}{k_B T}] \quad (1.1)$$

In equation 1.1, S represents solubility, μ^S represents the chemical potential of a protein in solubtion at a standard concentration (C = 1 M), μ^C the chemical potential of a protein in condensed phase, k_B the Boltzmann constant and T the absolute temperature. The difference term ($\mu^S - \mu^C$) can be viewed as the free energy of transfer from the condensed phase to the solution phase. The more favourable it is for a protein to transfer from the condensed to the solution phase, the lower the transfer free energy, and correspondingly the higher its solubility (Tjon and Zhou, 2008). In the context of biophysical protein-solvent systems, this definition is most commonly

adopted, *i.e.* the equilibrium concentration of a protein in solution phase when a saturation amount of the protein is present. Less formally, solubility measures the ability of a protein to dissolve in aqueous solutions. Numerous biochemical experiments rely on high levels of protein solubility, such as protein expression and purification, high-resolution structural determination, and quantitative binding assays.

Proteins are increasingly being used as pharmacologic agents, making solubility a primary concern in bioprocessing pipelines. Precipitation refers to the industrial downstream processing of biological protein products in order to concentrate and purify them for clinical and therapeutic purposes. Loss of solubility during overexpression of recombinant proteins is a common problem, leading to inclusion body formation and low protein yield. Although inclusion bodies are a relatively pure form of protein, the protein is not biologically active and therefore must be refolded (Cellmer *et al.*, 2007). Solubility effects are important in both biomedical and industrial contexts, since insolubility is intimately associated with aggregation phenomena, the harmful consequences of which were outlined above.

Because of its importance in biomolecular science, numerous experimental techniques have been developed to improve protein solubility, both *in vivo* during expression and *in vitro* during purification or formulation. These include time consuming solvent screening in order to identify optimal solvent conditions (Howe, 2004), co-expression with molecular chaperones (Machida *et al*., 1998), attachment of a small fusion tag protein (Waugh, 2005) or peptide (Kato *et al*., 2007), and site-directed mutagenesis of surface-exposed side chains (Bianchi *et al*., 1994; Mosavi and Peng, 2003; Trevino *et al*., 2007). Although experimental methods have achieved improvements in protein solubility, it is highly desirable to develop theoretical models for predicting the solubility profile of a protein. Such methods could be used to obtain desired levels of solubility for particular proteins as well as to provide necessary guidance for refining existing experimental techniques to achieve optimal solvent conditions for protein solubility (Tjong and Zhou, 2008). Theoretical models have been developed to predict aqueous solubility for drugs and drug-like compounds (Jorgensen and Duffy, 2000), but developments for larger macromolecules such as proteins remain greatly limited. Of particular interest are methods that would predict how protein solubility varies in tandem with solvent conditions. Efforts in this direction have been taken (Zhou, 2005; Tjong and Zhou, 2008), however progress is slow on account of the complexity associated with protein folding and protein-excipient interactions.

The solubility of proteins in aqueous cellular environments, most commonly the cytoplasm, is influenced by several parameters, including both intrinsic factors (*e.g.* primary sequence, surface properties) and extrinsic factors (*e.g.* pH, temperature, ionic environment). The majority of intracellular proteins are active in the intracellular cytosol, and as such their solubility properties are

crucial to their functioning. Globular proteins normally require high levels of solubility. Achieving high expression levels without concurrent loss of solubility within the congested molecular environment of cells is a central theme in biological systems. Perhaps the most basic intrinsic dependency is the distribution of hydrophilic and hydrophobic amino acids. In globular proteins, hydrophobic residues are predominantly buried in the core but occasionally exist in patches on the surface, whereas hydrophilic residues occur mainly on the surface. Polar and ionisable side chains interact with ionic molecules in the solvent and increase the solubility of a protein. Conversely, high surface hydrophobic amino acid content causes low solubility in aqueous solvents. Proteins are generally more soluble in acidic or alkaline pH ranges on account of excess charges of the same sign, which cause electrostatic repulsion between protein molecules. At their isoelectric point (pI), proteins are usually least soluble due to the lack of an overall net charge and the absence of repulsive intermolecular electrostatic forces. Temperature is also a very important variable. High temperatures are known to denature proteins as excess thermal energy causes destabilisation of non-covalent interactions (*e.g.* hydrogen bonding, hydrophobic interactions and salt bridges), which sustain secondary structure (Pelegrine and Gasparetto, 2005).

There is currently no universally accepted theory of protein solubility, and hence it is very difficult to determine how a specific charge distribution affects or dictates solubility (Garcia-Moreno, 2009). This implies that the bulk of our knowledge on the topic comes from empirical studies. One of the most comprehensive studies examining the relationships between protein solubility and physico-chemical properties as well as tertiary structure was carried out by Niwa and colleagues (2009), in which aggregation analysis for the entire ensemble of *Escherichia coli* proteins was performed using an *in vitro*-reconstituted translation system. A histogram of individual solubilities, based on data from over 3000 translated proteins, exhibited a clear bimodal distribution, indicating that aggregation propensities are non-evenly distributed. Bacterial (*E. coli*-based) protein solubilities were found to correlate well with structural classification of proteins, implying that predictive models require structural information (Niwa *et al*., 2009).

In discussions involving the formal treatment of protein solubility and aggregation, there is an important distinction to be made between the two. Solubility corresponds to the protein concentration in a solution that is in equilibrium with a precipitated phase (equation 1.1) and is primarily controlled via reversible inter-protein interactions. Aggregation refers to the formation of protein oligomers, which are often irreversible due to conformational rearrangements and is dominated by folding-based inter-protein interactions. Hence a key difference between the two is that aggregation pathways occur through partially folded states while proteins remain folded when in equilibrium with a precipitated phase. The corollary of this distinction is that loss of solubility is often reversible whereas the formation of aggregates is effectively irreversible. Although protein

aggregation (also referred to as non-native aggregation) and loss of solubility (insoluble protein product) are occasionally used interchangeably in biomedical and bioprocessing literature, it is important to establish that they comprise distinct thermodynamic processes. In the results chapters of this thesis, the terms soluble and insoluble will often be used to denote aggregation-resistant and aggregation-prone proteins, respectively. The reader is encouraged, however, to keep in mind that there are important physical distinctions between the two.

1.3.2 Molecular Mechanisms of Aggregation

Historically, amyloid-based aggregation has been well studied in medical contexts, as it is one of the defining molecular phenotypes of several neurological disorders. Diseases such as ALS (amyotrophic lateral sclerosis), Alzheimer's, Parkinson's, Huntington's, and prion disease, collectively classified as amyloidoses (De Felice *et al*., 2004), currently lack effective therapies, and offer a primary motivation for rigorous efforts to understand the molecular mechanisms underlying amyloid formation. A further incentive is found in the limitations that aggregation phenomena impose on the throughput yield of industrial bioprocessing pipelines. Novel protein-based therapeutic agents face prolonged and expensive delays in development and manufacturing due to aggregate formation, making the development of *in silico* and experimental screening tools highly desirable.

A range of diverse intrinsic and environmental factors has been reported to influence the aggregation process. The surrounding chemical environment of a protein, including factors such as pH (Su and Chang, 2001), temperature (Kusumoto *et al*., 1998), ionic strength, concentration of co-solutes, concentration of co-solutes and exposure to bulk liquid-fluid and liquid-solid interfactes (Roberts, 2014) all play important roles in aggregation. Furthermore, aggregation has intrinsic protein dependencies such as amino acid sequence (Chiti *et al*., 2003; Conchillo-Solé *et al*., 2007; Goldschmidt *et al*., 2010), charge (Konno, 2001; Tjernberg *et al*., 2002), hydrophobicity (Otzen *et al*., 2000; Schwartz *et al*., 2001), as well as patterns of polar and non-polar residue side chains (West *et al*., 1999). Further to intrinsic and extrinsic factors, a variety of evolved cellular regulatory mechanisms such as molecular chaperones (Hartl *et al*., 2011) and quality control processes which control degradation of partially unfolded proteins (Molinari, 2007) and modulate protein expression levels (Tartaglia *et al*., 2009) act to inhibit aberrant aggregation. The focus of this thesis is primarily on sequence- and structure-based factors, and to a lesser extent on environmental (solution-based) or cellular regulatory mechanisms (chaperone-based), since the former dictate

intrinsic aggregation propensities whereas the latter comprise environment-based (*e.g.* pH, ionic strength) and cellular anti-aggregation "safety" mechanisms.

The primary mechanism allowing proteins to maintain functional and soluble states is strongly believed to be sequence-encoded and hence evolutionarily optimised, involving intrinsic energy barriers that prevent conversion to an aggregation-prone state (Chiti and Dobson, 2006; Tartaglia *et al*., 2008). Aggregation-prone sequences may exist in more than one place within the same protein (Wu *et al*., 2014). These may become exposed transiently via local or partial unfolding of an initially monomer, or if one such monomer forms small, reversibly folded oligomers or "clusters" (Banks *et al*., 2012). Under conditions in which the folded state is favoured over the unfolded states, monomers can initially self-associate reversibly either as folded or partially unfolded species (Roberts, 2014). Aggregation is reversible, at least putatively, at the eearly stages prior to formation of nuclei, due to the existence of kinetic bottlenecks that allow hot spot sequences to sample conformations that enable strong interactions between chains of adjacent proteins (Roberts, 2007). Importantly, the same forces that drive folding of an individual protein molecule are also present when two or more chains interact with one another. Hence the same forces that drive folding also drive aggregation, and as a result, hot spot sequences tend to be stretches of amino acids that are highly hydrophobic, lack charges, and are prone to form β-sheets when paired with adjacent strands (Caflisch, 2006).

Otzen and colleagues (2000) coined the term "structural gatekeeper" to describe residues with charged side chains that impede aggregation by interrupting contiguous stretches of hydrophobic residues in the primary sequence. Richardson and Richardson (2002) further elaborated this phenomenon in a study characterising a variety of secondary structure elements (termed negative design elements) that minimise intermolecular edge-to-edge β-sheet interactions. Both studies reported the specific placement of charged side chains as one strategy of minimising the risk of aggregation for proteins that are β-sheet enriched. Studies aimed at identifying aggregation hot spots (also referred to as high propensity segments) in protein sequences report that although such segments tend to be uniformly distributed throughout the primary sequence, they are often directly flanked by charged residues such as aspartic acid and arginine (Rousseau *et al*., 2006; Goldschmidt *et al*., 2010). These studies illustrate that electrostatic interactions are of key importance in aggregation phenomena. In terms of primary sequence, amino acids with ionisable side chains are crucial in providing protection against aggregation. When considering tertiary structure, this means that the distribution of charge on protein surfaces is a contributing factor to their propensity to self-associate. In order to establish correlations between surface properties and aggregation tendency, and extend such correlations to predictive models, variables such as charge

distribution, detailed surface geometry, as well as polarity distribution (polar and non-polar surface patches) must be considered individually or as a combined metric.

The kinetics of aggregation pathways have also been studied in order to gain insight into the underlying mechanistic basis of amyloid formation. For certain protein classes (mostly non-therapeutic), aggregation is can be modelled as a nucleation polymerisation process, whose reaction rate depends on the protein concentration and can be accelerated via addition of homologous pre-aggregated polypeptides that act as seeds promoting transition from a soluble to an insoluble, aggregated state (Jarrett and Lansbury, 1993; Nielsen *et al*., 2001; Chiti and Dobson, 2006). According to this simplified model, amyloid aggregation can be divided in three main phases: (i) the thermodynamically unfavourable lag phase where soluble polypeptides associate to form nuclei, *i.e*. nucleation, (ii) the population of these transient nuclei induces polymerisation and fibril growth in what is termed the exponential phase, and (iii) depletion of soluble monomeric species leads to the saturation phase in which no further nucleation reactions can occur due to lack of monomers. The lag and exponential growth phases have been suggested to be amenable to drug targeting whereby certain compounds may act to inhibit the process (Invernizzi *et al*., 2012). Kinetic data for amyloid fibril formation exhibit a characteristic sigmoidal appearance (Linse and Linse, 2011) and equilibrium data from experimental studies suggest that the process follows the physico-chemical trends of a phase transition in a highly deterministic, predictable manner (Hellstrand *et al*., 2010). Molecular simulations of aggregation from the native state have further shown that inter-protein association increases as proteins begin to unfold. The partial unfolding of the native state leads to the formation of aggregation precursors. This behaviour can be attributed to the fact that side chains that are buried in the native, folded state become solvent exposed when unfolding occurs and are able to interact with each other (Bratko *et al*., 2006).

1.3.3 Computational Methods for Aggregation Prediction

The current paradigm used to address aggregation involves experimentally screening for excipients and storage conditions that optimise protein stability and minimise aggregation. Sequence and structural information are intrinsic factors and can hence provide insight into the propensity of proteins to aggregate. Advances in hardware speed and power as well as developments in software over the past decades have given computational techniques great scope to contribute to predicting the aggregation propensity of potential protein-based drug candidates.

Algorithms have been developed for predicting the aggregation propensity of a protein given its amino acid sequence. The computational methods that address this issue can be broadly divided

into phenomenological models and molecular simulation techniques (Chennamsetty *et al*., 2009). Phenomenological models rely on correlation of aggregation behaviour with physico-chemical properties, *e.g*. hydrophobicity, polarity, amino acid sequence to identify aggregation-prone regions (APRs) and involve bioinformatics approaches. Molecular simulation methods rely on predicting aggregation by energetic sampling using either coarse-grained models or atomistic models. Although they employ different methodologies and underlying theories, both phenomenological and simulation models aim to develop a systematic way of identifying APRs within a protein sequence. Several efforts have been put forth to develop predictive models from both the phenomenological and the simulation approaches. Some of these initiatives are outlined below and summarised in table 1.1.

TANGO

TANGO (Fernandez-Escamilla *et al*., 2004) is a statistical mechanics-based algorithm that identifies β-aggregating regions of a protein sequence and predicts the sequence-dependent and mutational effects on aggregation. TANGO is based on the physico-chemical principles of β-sheet formation. For each residue, it calculates the energy of structural states derived from statistical and empirical considerations. A partition function of the conformational phase space is computed to predict β-aggregating segments of a polypeptide. The algorithm has been shown to accurately predict the aggregation of a dataset of 179 peptides compiled from literature as well as a dataset of 71 peptides derived from human disease-related proteins. TANGO uses two main assumptions to estimate aggregation propensity of a particular polypeptide sequence, *i.e*. (i) the aggregating regions are fully buried and hence solvated and (ii) electrostatic interactions establish an overall net charge that disfavors aggregation. The success rate of the method to correctly predict a sequence to be aggregation prone is approximately 90%. TANGO has been implemented on a webserver and can be accessed at the URL http://tango.embl.de/.

PAGE

PAGE (Tartaglia *et al*., 2005) develops predictions of absolute protein aggregation rates by identifying segments involved in β-sheet formation using a sliding window of 5 – 9 residues over the length of the input sequence. The aggregation propensity of the queried sequence is computed based on aromaticity, β-sheet formation propensity, charge, and average polar and non-polar accessible surface area of each residue within the selected window. Absolute aggregation rate is related to aggregation propensity using a factor that is a function of concentration and temperature. PAGE predicts APRs for both parallel and antiparallel β-sheets in a large set of natural protein sequences.

PASTA

PASTA (Trovato *et al*., 2006) attempts to identify regions within a protein sequence that may contribute toward the formation of cross-β structural motifs. The algorithm uses a pairwise energy function to compute the propensity of residues found within a β-sheet facing one another on neighboring strands. The energy function is derived from the top500H database (Lovell *et al*., 2003) consisting of a non-redundant set of 500 high resolution X-ray crystallographic structures of globular proteins. PASTA's ability to predict the registry of the intermolecular hydrogen bonds formed between amyloidogenic segments allows it to identify the portions of the sequence forming the cross-β core as well as to discriminate whether the intermolecular β-strands are parallel or antiparallel. As in the case of TANGO, the authors used the data on 179 peptides from literature to benchmark predictions from their methods. PASTA yields approximately 80% correct predictions. The most recent version of the algorithm is available on a webserver at the URL http://protein.bio.unipd.it/pasta2/.

AGGRESCAN

AGGRESCAN (Conchillo-Solé *et al*., 2007) is a webserver implementing an algorithm based on an aggregation-propensity scale for amino acids derived from *in vivo* experiments. For a queried protein sequence, the aggregation propensity value for each residue in the sequence is computed by averaging aggregation propensity value per amino acid over a sliding window of a given length. The relative aggregation propensity value for each of the 20 natural amino acids is derived from the analysis of the intracellular aggregation of mutants in the central position (F19) of the central hydrophobic cluster (L17-V18-F19-F20-A21) in the Aβ peptide. AGGRESCAN also outputs the areas of the profile peaks over a pre-calculated threshold as well as a graphical representation of peak-area values. This method has been validated via comparison of experimental and predicted APRs in 24 fibrillar deposition disease-linked polypeptides. AGGRESCAN uses the key assumption that short and specific sequence stretches modulate protein aggregation. The webserver can be accessed at the URL http://bioinf.uab.es/aggrescan/.

Zyggregator

Zyggregator (Tartaglia *et al*., 2008) predicts relative propensities for folding and aggregation for a given protein sequence, and identifies regions where aggregation propensities are significantly higher. Predictions are based on a Z_{agg} score, which uses a predefined intrinsic propensity of an unfolded polypeptide chain to form amyloid deposits by considering a range of physico-chemical properties. Aggregation propensity per residue is modelled as a combination of four factors intrinsic to the sequence: net charge, hydrophobicity, secondary structure propensity,

and the pattern of alternating hydrophobic and hydrophilic residues. Native propensities to aggregate are also directly influenced by pH, ionic strength, and peptide concentration. Zyggergator is able to correctly identify APRs in the Aβ42 peptide, α-synuclein protein, and tau protein. The predictive method is available on a webserver that can be accessed at the URL http://www.vendruscolo.ch.cam.ac.uk/zyggregator.php/.

BETASCAN

BETASCAN (Bryan *et al*., 2009) calculates likelihood scores for potential β-strands and strand pairs based on correlations observed in parallel β-sheets. For each pair of possible strands, a score is determined to represent the probability of their pairing. This likelihood is based on empirical preferences for each pair of residues in the strands to form hydrogen bonds with each other. Maxima-finding algorithms are employed to detect local maxima of formation propensity across all potentially paired strands. The output of the program is a series of score-ordered lists of all predicted strand pairings. Accurate prediction has been demonstrated for experimentally determined amyloid structures and for a set of known β-aggregates. The BETASCAN webserver is accessible at the URL http://www.groups.csail.mit.edu/cb/betascan.

AMYLPRED

AMYLPRED (Frousios *et al*., 2009) is a consensus prediction tool for amyloidogenic determinants. Five different methods are employed using primary sequence data as input, in order to arrive at the consensus amyloidogenic segment prediction. Among these methods are TANGO (Fernandez-Escamilla *et al*., 2004), SecStr (Hamodrakas, 1988), and average packing density (Galzitskaya *et al*., 2006). SecStr identifies regions which can be both α-helices and β-strands in a given protein sequence. These regions, termed conformational switches, are shown to be good correlators for amyloidogenic propensity. A method that maps hexapeptides of a sequence onto the template microcrystalline structure of *NNQQNY* and calculates the resulting conformational energy using residue-based statistical potentials is also used (see 3D profile method below). AMYLPRED has been implemented on a webserver and is available at the URL http://aias.biol.uoa.gr/AMYLPRED/.

3D Profile Method

The 3D Profile Method (Thompson *et al*., 2006) is an extension of some of the above sequence-based predictive models to incorporate structural features. Crystallographic studies of segments of proteins capable of forming amyloid-like fibrils have revealed that these segments are short (4-10 residues) and self-complementary, stacking into pairs of β-sheets whose side chains

extend and interdigitate (Balbirnie *et al*., 2001; Sawaya *et al*., 2007). Two such segments (*GNNQQNY* and *NNQQNY*) were used to develop a structure-based computational method for identifying other such short segments that fibrillise. In the 3D profile method, the sequences of putative amyloid-forming proteins are scanned by threading them on the backbone of the segment *NNQQNY*. Segments that form similar, self-complementary structures satisfying energy criteria, as assessed by the RosettaDesign potential energy function (Kuhlman and Baker, 2000), are classified as being capable to fibrillise. The method was found to yield correct predictions in a set of fibril-forming segments of amyloid proteins.

The 3D profile method was extended by Goldschmidt and colleagues (2010), who used a set of 16 published hexapeptide fibril-forming crystal structures to derive an energetic threshold of -23 kcal/mol, below which a sequence can be classified as HP (high fibrillation propensity) after being threaded on the *NNQQNY* template backbone. The authors coined the term "amylome" to describe the subset of the proteome encompassing all proteins capable of forming amyloid-like fibrils. The authors further reported that sequence is more important than residue composition as a criterion for determining propensity for formation of amyloid-like fibrils. This was ascertained by shuffling identified exposed and buried HP segments *in silico* and observing a trend for HP segments to be preferentially buried. The shuffling experiments found that when the sequence of an HP segment was shuffled, the rearranged sequence lost its tendency to fibrillise. Conversely, when the sequence of a non-fibrillising segment was rearranged to one below the energetic cutoff, fibril formation was observed.

Spatial Aggregation Propensity (SAP)

Spatial aggregation propensity (SAP) is a computational tool based on molecular simulation that predicts APRs (Chennamsetty *et al*., 2009; Chennamsetty *et al*., 2010). Unlike the other tools that have been overviewed, which are largely based on protein sequence, SAP takes into account both the dynamic exposure of residues as well as their spatial proximity in the protein tertiary configuration, and is thus applicable for large proteins such as antibodies. The method does not use data on amyloidogenic peptides and proteins. SAP identifies surface exposed hydrophobic patches that can lead to aggregation. A SAP value over each atom is computed as shown in equation 1.2.

(1.2)

$$\text{Spatial Aggregation Propensity}_{atom,i} = \sum_{S} \left\{ \sum_{R_i} \left(\frac{\text{SAA}_\text{R}}{\text{SAA}_\text{E}} \cdot \text{H} \right) \right\}$$

In equation 1.2, **S** represents the simulation average, $\mathbf{R_i}$ the residue with at least one side chain atom within radius *R* from atom *i*, $\mathbf{SAA_R}$ the solvent accessible area of side chain atoms contain within distance *R* from the central atom, $\mathbf{SAA_E}$ solvent accessible area of side chain of fully exposed residues and **H** the residue hydrophobicity.

A high SAP value indicates an aggregation-prone region. The spatial aggregation propensity is calculated for spherical regions with radius *R* centered on every atom in a protein, yielding a unique SAP value for each atom. SAP for a residue is obtained by averaging the SAP of all its constituent atoms (Chennamsetty *et al*., 2010). The SAP technique has been successfully incorporated in a tool used to predict relative aggregation propensities of IgG1 mAbs (Lauer *et al*., 2009), making it particularly important in bioprocessing applications involving protein-based agents in industrial pipelines. The APRs identified by SAP are not amyloidogenic sequence patterns, but instead are structural motifs consisting of residues, which may or may not be contiguous in sequence. Table 1.1 below provides on overview of each tool.

Table 1.1 Summary of Computational Tools for Protein Aggregation Prediction

Phenomenological/Bioinformatics Methods					
Tool	Input	Parametrisation	APR Prediction Overview	Validation	Applied to Improve Biotherapeutic Developability?
TANGO	Sequence	Short aggregating and non-aggregating peptides	Statistical mechanics based method. Uses physico-chemical principles underlying β-sheet formation	Experimental data on a set of 179 peptides	Yes
PAGE	Sequence	Peptides found in amyloidosis-based diseases	Aromaticity, β-strand propensity, and charge	Experimental data on a number of amyloid-based diseases	Yes
PASTA	Sequence	Protein crystal structures	Pairwise interaction potentials for a pair of residues found facing each other within a cross-β-motif. Interaction potentials were determined from a dataset of 500 high resolution globular crystal structures.	Experimental data on 179 peptides used to validate TANGO (see above)	Unknown
AGGRESCAN	Sequence	Aβ peptide mutants	Intracellular aggregation propensity of mutants of Aβ-42 peptide	Experimental data on 24 fibrillar deposition-linked polypeptides	Unknown
Zyggregator	Sequence	Short peptides	Relative propensities for folding and aggregation in a given sequence region	Numerous known amyloidogenic peptides	Yes
BETASCAN	Sequence	Prion and amyloid proteins	Calculates likely β-strands and strand-pairs for an input sequence using probability tables	Tested against a non-redundant set of crystallised β-strand proteins	Unknown
AMYLPRED	Sequence	Consensus based of five methods parametrized using short peptides and proteins	Uses consensus among five different methods	Experimentally known amyloidogenic short stretches in 18 proteins involved in amyloidoses	Unknown

Molecular Simulation Methods					
Tool	Input	Parametrisation	APR Prediction Overview	Validation	Applied to Improve Biotherapuetic Developability?
3D Profile	Sequence	Potential energy function based on RosettaDesign	Molecular simulation method that evaluates compatibility of a sequence with the crystal structure of hexapeptide *NNQQNY*	Strongly predicted peptides were shown to fibrillize in experimental studies	Unknown
SAP	Structure	Accessible surface area of protein and residue hydrophobicity scale. Data on amyloidogenic peptides/proteins was not used.	SAP computes the dynamically exposed hydrophobicity of a surface patch. High SAP values indicate APR regions.	mAbs (monoclonal antibodies), engineered based on SAP prediction and showed reduced aggregation propensity in experiments.	Yes

Table adapted from (Agrawal *et al*., 2011)

1.3.4 Bioinformatics-based Solubility Prediction

Solubility prediction methods are similar in underlying principles to phenomenological aggregation prediction methods and are based on statistical learning techniques. The development of such models is performed within a probabilistic framework, with machine learning methods being the most common underlying methodology. A wide variety of machine learning methods are employed, including *k*-nearest neighbours, neural networks, Bayesian classification, logistic regression, discriminant analysis, and SVM (support vector machines) (Chang *et al*., 2013). In most cases, a classifier is employed to assign data points to categories (*e.g*. solubility, aggregation propensity).

There is significant interest in developing theoretical models of protein solubility, as this will contribute to accurate predictive methods. Some prominent efforts aimed at developing statistics-based solubility prediction tools include PROSO II (Smialowski *et al*., 2012), CCSOL

(Agostini *et al*., 2012), SOLPro (Magnan *et al*., 2012), PROSO (Smialowski *et al*., 2007), SVM-based (Idicula-Thomas *et al*., 2006), SI-based (**s**olubility **i**ndex) (Idicula-Thomas and Balaji, 2005), and recombinant protein solubility prediction (Wilkinson and Harrison, 1991). Such methods and their datasets comprise an important part of later chapters grounded in sequence-based prediction of solubility.

1.4 Electrostatics

Electrostatics is the branch of physics that is concerned with the phenomena and properties of stationary electric charges. Electrical charge is one of the two main sources of potential energy (the other being gravity), which in turn is one of several known forms of energy and is associated with the force acting on a body. Charge is a fundamental physical concept, similar to mass and spin, and as such cannot be readily defined in terms of other simpler concepts. Hence it is assigned an operational definition, *i.e*. it is a property of matter that causes a body to experience a force when in proximity to another electrically charged body. It follows that two bodies carrying charges will exert a force on each other. This force is termed the electromagnetic force and comprises one of the four fundamental forces of nature. It is repulsive for identical charges and attractive for opposite charges. The force experienced by each body can be described using the notion of an electric field, which is described by the space surrounding a charged body that acts on another charge to produce an electrical force (illustrated in figure 1.3).

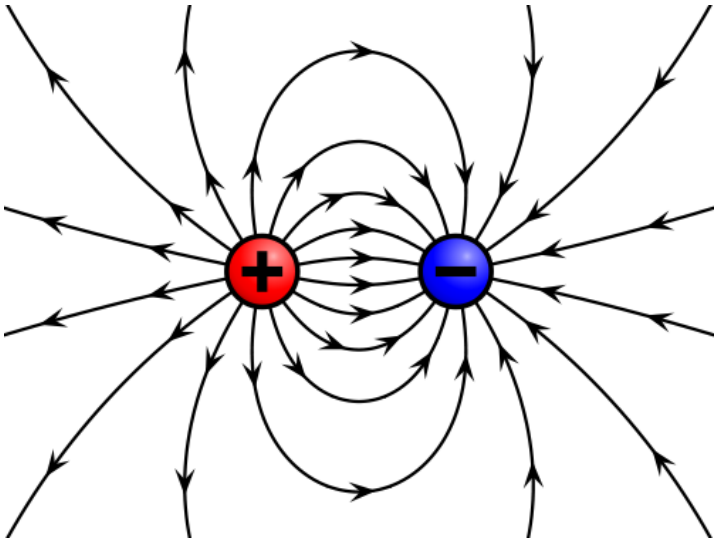


Figure 1.3. **Electric field created by two oppositely charged bodies**. An electric field permeates the space around a positive and negative point charge. Opposite charges are attractive towards each other. Figure adapted from (http://www.phy.ntnu.edu.tw).

A particle can have positive, negative or zero charge (termed neutral). Protons are subatomic particles with positive charge while electrons possess negative charge, and neutrons are uncharged. All atoms are composed of protons, neutrons and electrons (with the exception of hydrogen whose atom contains a single proton and electron). Atoms are neutral, balancing their overall charge by an equal number of protons and electrons. Most atomic properties, including interactions with other atoms, are derived from constituent electrons. When an atom loses or gains an electron, it gains a positive or negative charge, respectively. An atom or molecule with an unequal number of protons and electrons that has acquired a net positive or negative charge is termed an ion.

Electrostatic properties of proteins are conferred by the ionic properties of their constituent amino acids, which have extensive roles in mediating interactions. Proteins, being essentially very large polymeric molecules comprised of thousands of atoms, can be ionic species if they possess an overall nonzero charge. They often contain acidic or basic groups that are ionised depending on the pH (hydronium ion environment). Although a protein can be net neutral, it still has significant electrostatic properties because individual side chains can be ionised. Hence, electrostatic principles maintain a prominent role in the field of biophysics. Electrostatic properties have prominent roles in mediating molecular interactions in protein functional complexes, such as enzyme-substrate and ligand-receptor complexes. Given the accepted biochemical dogma of primary sequence dictating tertiary structure, which in turn determines the function of a protein molecule, it follows that aspects of sequence and structure are key parts of computational studies aimed at better understanding protein function. Electrostatic properties can be studied at both the sequence-level (*e.g.* net sequence charge) and structure-level (*e.g.* surface potential and polarity), and have important functional implications.

Certain amino acids have the ability to become ionised in solution and can therefore possess electrically charged side-chains (termed ionisable or titrable groups). Amino acids with ionisable groups were the focus of the current work, as they are the mediators of electrostatic interactions within and between proteins. Out of the 20 amino acids, seven are known to have ionisable side-chains with five of those being in ionised states at neutral pH – lysine, arginine, histidine, aspartic acid, and glutamic acid (K, R, H, D, and E respectively). Lysine and arginine are basic amino acids with positively charged groups (histidine is generally only partially charged at neutral pH). Aspartic and glutamic acid are acidic amino acids with negatively charged groups and thus are often referred to as aspartate and glutamate, respectively. Cysteine and tyrosine also have ionisable side chains but are usually uncharged at neutral pH.

In order to computationally determine the functionality of ionisable groups, some physical quantity must be used to detect and measure changes in their charge state. The quantities most commonly used to measure electrostatic properties of amino acids (i.e. to characterise their charge states and how they vary) and their polymeric forms (polypeptides and proteins) are pH as well as pK_a, and are discussed in the following section.

1.4.1 Protein Electrostatics

Electrostatic (charge-based) interactions play an important role in molecular biology. The links between the structure and function of a protein and its electrostatic properties have been explored by numerous studies, some of which have been reviewed by Sinha and Smith-Gill (2002). The complex aspects of cellular physiology mean that very large numbers of macromolecules interact with specific binding partners, and these interactions are most often mediated through non-covalent forces. Although significantly weaker than the covalent bonds commonly found in small molecules, non-covalent bonds such as hydrophobic interactions and van der Waals forces have the ability to stabilise protein-protein interaction networks and determine the structure of ternary protein complexes. Such interactions can be both short-range and long-range. At short distances, electrostatic interactions act in conjunction with other factors such as hydrophobic interactions, hydrogen bonding forces, as well as dispersion interactions (Nielsen, 2007). Long-range electrostatic interactions are observed in the binding specificity of proteins that can recognize and discriminate their binding partners in the crowded cellular environment (Schreiber and Keating, 2011). Proteins alter their charge state through protonation and redox reactions as well as via binding charged ligands. The distribution of charged residues is significant when considering protein interactions. Proteins contain charged patches, which are formed by either oppositely charged residues or like charged residues (Karlin, 1995). Interactions between proteins mediated by oppositely charged patches facilitate the formation of multi-protein complexes, whilst those of like charge are thought to keep certain assemblies apart (Karlin, 1995).

In terms of environmental factors, pH has a large effect on protein electrostatic properties. Proteins can become unstable or denature completely under extreme acidic or alkaline pH conditions due to possessing a high excess charge density. Ionisable groups of amino acid side chains are usually localised on the solvent accessible area of a protein on account of the unfavourable solvation energy associated with burying charged residues in the protein core. Receptor-ligand interactions depend heavily on the ionisation state of side chains near the binding interface. Interactions based on electrostatic properties have been used to study structural and functional aspects of several protein systems including enzymes (Greaves and Warwicker, 2005),

mRNA translation factors (Magee and Warwicker, 2005) among others, reflecting that charged regions often provide insight to functional areas. The ionisable groups of amino acid side chains, each with distinct pK_a values (table 1.2), contribute to electrostatic interactions between proteins and molecules. These pK_a values are highly sensitive to local protein conformation and environment, and thus provide a useful way to study electrostatic interactions in proteins.

In macromolecular systems, electrostatic interactions also arise for neutral atomic groups through polarisation effects that occur due to several factors including: (i) the electron density distribution around atoms, (ii) the redistribution of electrons in response to local electrical fields and (iii) the reorientation of polar groups in response to an electrical field (Neves-petersen and Peterson, 2003). The computational modelling of electrostatic interactions and their effects on protein behaviour is usually accounted for using either continuum electrostatics through solving the Poisson-Boltzmann equation or by using molecular simulation techniques. Both these techniques face similar challenges in calculating electrostatic effects. These challenges arise from the fact that interactions between ionisable groups take place in the cellular environment where several factors have to be taken into consideration. These include specific pH levels, specific salt concentrations that may alter binding energies, and surrounding aqueous environment means water molecules possess orientational freedom. Generally, pH can be accounted for using pK_a values, changing salt concentration through changes to ionic strength or screening parameters. Screening refers to the dampening of electric fields caused by the presence of mobile charge carriers such as ionic species. The impact of water is similarly accounted for by using a dielectric constant. However, specific ion effects are not mechanistically well characterised, as it is not always clear how ions interact with proteins.

1.4.2 Continuum Electrostatics

When considering the role of electrostatics in mediating molecular interactions, the most important parameters are the electrostatic potential energy and the pK_a values of ionisable side chains. These parameters are estimated by solving the Poisson-Boltzmann equation (PBE) for a protein-solvent system in a continuum model that treats protein and water as uniformly smooth materials (figure 1.5). The fundamental principle underlying interactions between charged particles of any size in a physical system is Coulomb's law. This relationship describes the force acting between electrically charged particles according to the inverse square law that is stated as follows:

(1.3)

$$\mathbf{F} = \frac{\mathbf{q_i}\,\mathbf{q_j}}{\mathbf{4\,\pi\,\varepsilon_0\,r^2}}$$

Where **F** is the force (*i.e.* the electrostatic energy between point charges q_i and q_j), **r** the distance between the centres of q_i and q_j and $\boldsymbol{\varepsilon_0}$ the Dielectric constant in vacuum

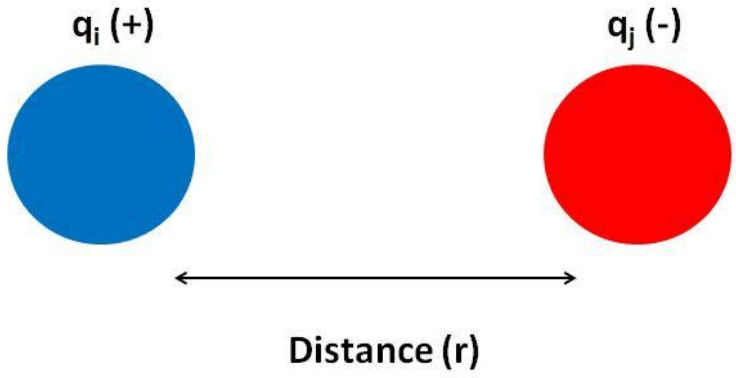


Figure 1.4. Schematic view of Coulomb's Law.

Equation 1.3 coupled with figure 1.4 illustrates how the electrostatic energy *F* between a pair of point charges, q_i and q_j, separated by a distance *r*, is calculated. In the above system, the electrostatic potential φ_i at point *i* in space is given by equation 14:

(1.4)

$$\boldsymbol{\varphi_i} = \frac{\mathbf{F}}{\mathbf{q_i}} = \frac{\mathbf{q_j}}{\mathbf{4\,\pi\,\varepsilon_0\,r^2}}$$

Classical electrostatics offers a quantitative description of the forces that exist between two charged bodies through this relationship. However, this approach is too simplistic for modelling protein-solvent systems where the dielectric properties of each differ substantially. A dielectric refers to any substance that can behave as an electrical insulator and become polarised upon introduction of an electrical field. In order to model protein behaviour, it is necessary to consider how the potential of one point in space varies from multiple point charges. The total potential in this case is the sum of each potential from the source charges to the point. Furthermore, charge-charge interactions in biological systems take place in a medium instead of a vacuum and hence the dielectric constant in Coulomb's equation does not suffice to describe how electrostatic energies between charged molecules vary. Equation 1.3 is accurate for computing electrostatic potentials of point charges only under a uniform dielectric in an infinite medium.

For biological systems with large protein molecules in aqueous environments, a more sophisticated model is necessary. This is found in the linear Poisson-Boltzmann equation (equation 1.3), described independently by Gouy (1910) and Chapman (1913) around a century ago. The PBE was used to equate the chemical potential and the force acting on small adjacent volumes in an ionic solution between two plates at a different voltage. The approach was generalised around a decade later by Debye and Hückel (1923), whose work was applied to the theory of ionic solutions and led to the successful interpretation of experimental thermodynamic data. The linear PBE is formed by the Poisson equation and the Boltzmann probability distribution, and is commonly applied to compute the electrostatic potential at the solvent accessible surface. The PBE can be expressed as follows:

(1.5)

$$\frac{\{\,\nabla\,[\boldsymbol{\varepsilon}(\mathbf{r})\,\nabla\boldsymbol{\varphi}(\mathbf{r})] - \boldsymbol{\varepsilon}(\mathbf{r})\boldsymbol{\kappa}(\mathbf{r})^2\,\sinh[\boldsymbol{\varphi}(\mathbf{r})] + [\,4\pi\boldsymbol{\rho}^{\mathbf{f}}(\mathbf{r}) - \boldsymbol{\kappa}^2\boldsymbol{\varphi}(\mathbf{r})]\,\}}{\boldsymbol{k_B T}} = 0$$

In equation 1.5, ∇ is the del operator, $\varphi(\mathbf{r})$ represents the electrostatic potential in units of kT/q (where k is the Boltzmann constant, T is the absolute temperature and q is the charge on an electron), $\boldsymbol{\varepsilon(r)}$ is the dielectric constant and $\boldsymbol{\rho^f}$ is the fixed charge density (in units of electronic charge). The term $\boldsymbol{\kappa}^2 = 1/\lambda^2$ (where λ represents the Debye length, see section 1.4.4). The term $\boldsymbol{4\pi\rho^f(r)}$ describes the charge density of the protein and $\boldsymbol{\kappa^2\varphi(r)}$ the charge density of the solvent. As in equation 1.1, $\mathbf{k_B}$ and $\boldsymbol{T}$ represent the Boltzmann constant and absolute temperature, respectively.

The variables φ, ε, κ and ρ are all functions of the position vector $\mathbf{r}$. The second term of equation 1.5 accounts for salt effects and is absent when no mobile ions are present in the system ($\kappa = 0$). Under this condition, equation 1.5 reduces to Poisson's equation, which in turn reduces to Coulomb's law when the dielectric constant is uniform throughout space. Water is much more easily polarised by an electric field compared to most solutes contained within (*e.g.* proteins and ions); hence, two dielectric terms are required to encapsulate the underlying physics of polar molecules in aqueous solutions (Honig and Nicholls, 1995).

1.4.3 Finite Difference Poisson Boltzmann

The Poisson-Boltzmann equation (PBE) describes electrostatic interactions between molecules in solutions with ionic species by treating proteins and solvent as uniformly smooth objects. This model was initially used to model charge effects in enzyme active sites (Klapper *et*

al., 1986) and to calculate pK_a values (Bashford and Karplus, 1990). The PBE is useful for modelling simple geometric shapes such as spheres, but for complex protein tertiary configurations a numerical solution must be applied. A variation of this method comes in the form of a numerical FDPB (Finite Difference Poisson-Boltzmann), which calculates the electrostatic potential by applying a discretisation to the PBE and incorporating detailed geometric information so that proteins do not have to be approximated as spherical objects (Warwicker and Watson, 1982). The introduction of finite difference techniques in this context provided a representation of the protein-solvent interface in macroscopic models (Alexov *et al*., 2011). The FDPB method is based on the changes experienced in the dielectric properties of the protein-solvent system which occur due to dipolar reorientation in the presence of an electrical field. Dipoles are separations between positive and negative charges that occur on the atomic level and can be classified into two types: permanent and induced. The former occur when in a covalent bond the shared electron pair between two atoms is permanently pulled closer to the more electronegative atom, while the latter are experienced when the electron cloud around a covalently bonded molecule shifts to one side. Due to the relative freedom of water molecules in aqueous environments, there is extensive dipolar rotation which leads to a high dielectric constant (~80 at 298 K). In contrast, permanent dipoles resulting from the covalently linked backbone in proteins maintain much smaller dielectric constants (~4 at 298 K). The FDPB model therefore describes a protein with a low dielectric constant immersed in a medium with a much higher dielectric. This concept is illustrated in figure 1.5.

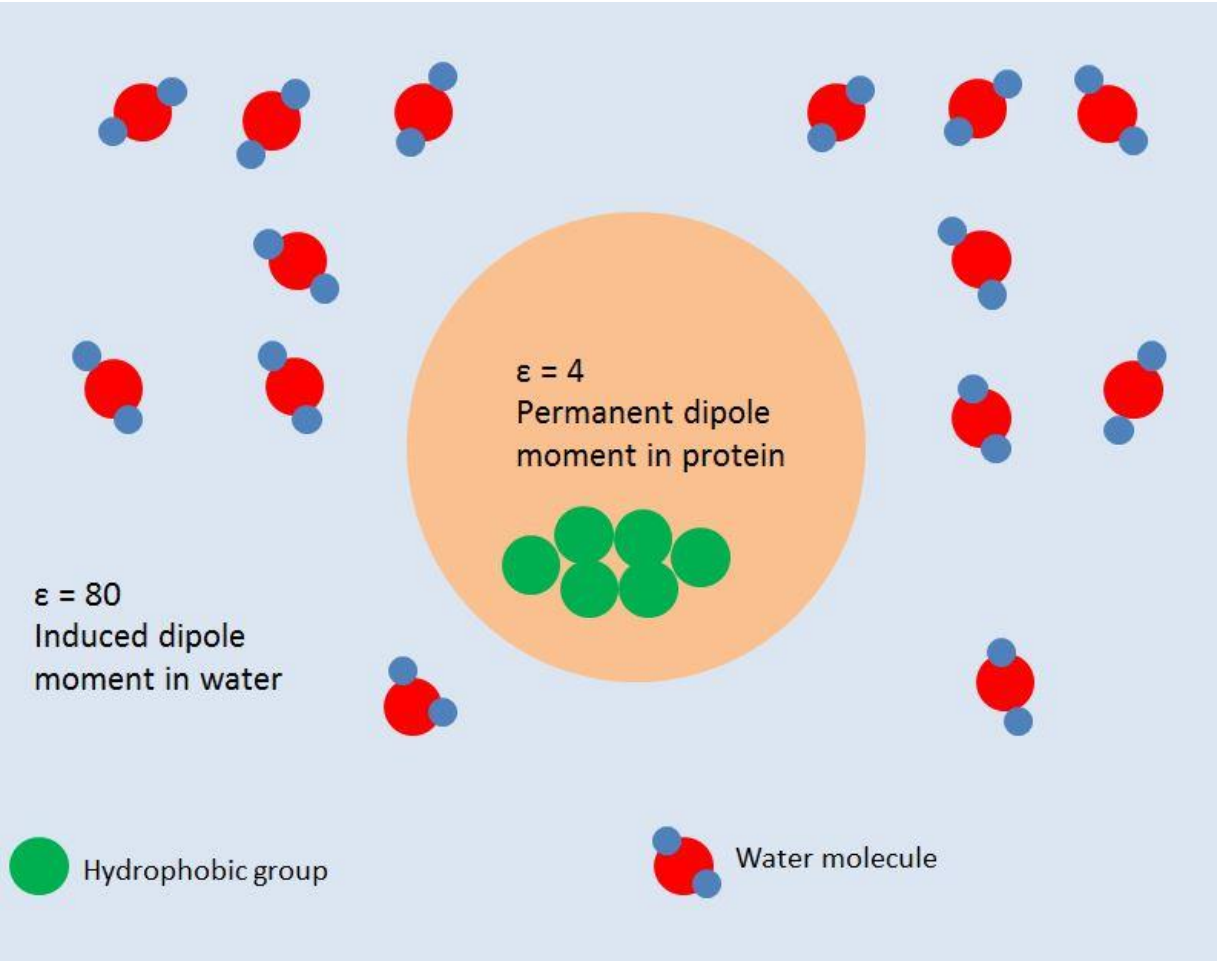


Figure 1.5. **Dielectric Environment of Protein-Solvent Systems.** An overview of differences in dielectric constants within a protein-solvent system. The highly mobile ions in water arise from high dipolar rotation while the polyionic nature of a protein arises from immobile ions with low dipolar rotation. The figure is meant as an illustration only, with the geometry and scale of the actual molecules being ignored for simplicity.

The calculation of electrostatic potential ($\varphi(\mathbf{r})$ in equation 1.5) within the protein-solvent system is achieved by mapping the system onto a Cartesian grid (x, y, z) in order to subdivide the domain in which the PBE is numerically solved. The grid boundary is based on a solvent probe with a radius of 1.4 Å. The electrostatic potential, charge density, and dielectric constant of each charged atom of the protein within a grid cube is replaced by a central grid point. These parameters are influenced by the six nearest grid points. The Poisson-Boltzmann equation is thus integrated and recast in a finite difference form using the gird-based system described. A schematic of the discretisation process is provided in figure 1.6.

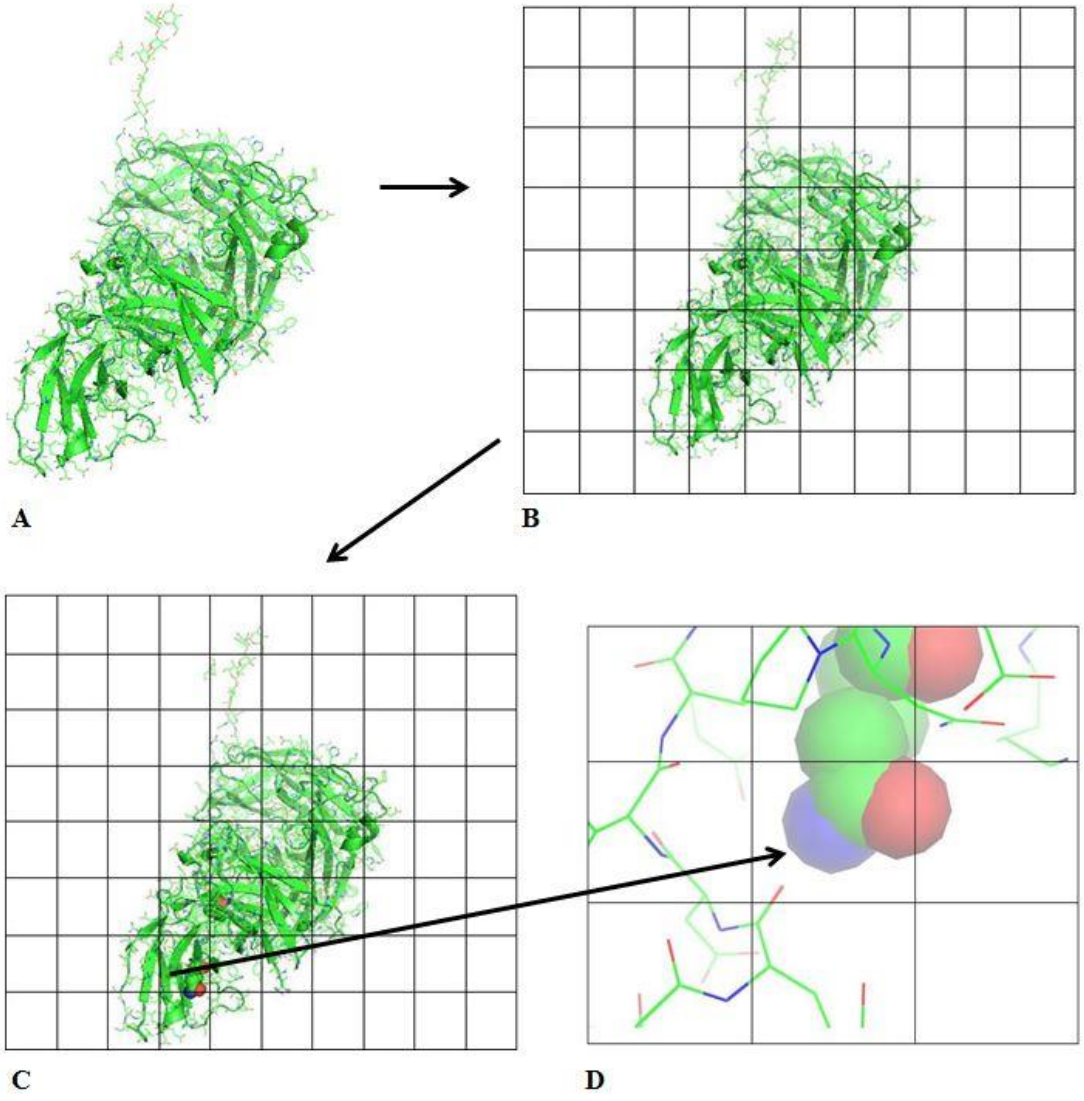


Figure 1.6. Finite Difference Representation of a Protein-Solvent System. (**A**) Protein in aqueous solution (PDB 1A14). (**B**) The protein-solvent system placed into a 3D grid (2D projection of the grid illustrated here). (**C**) A lysine residue highlighted in sphere representation. (**D**) A close up view of the nitrogen atom of the lysine side chain (blue). The actual N atom is the central grid point.

1.4.4 Debye-Hückel Theory

Another method used in electrostatic modelling is based on the Debye-Hückel equation (Debye and Hückel, 1923), which can be used to represent the average charge-charge environments at protein surfaces (Warwicker, 2011). The method is essentially a simplified Poisson-Boltzmann model to account for the non-ideality of electrolyte solutions, especially at low concentrations. The activity of ions in solution is relatively large compared to that of neutral compounds. Although an electrolyte dissolved in water tends to strengthen the solvation of charges, in the context of the

cellular environment this effect is minimal compared to the powerful solvating effects of water itself (Gilson, 2006). For a point charge in solution, the redistribution of mobile ions can be calculated through a screened potential given by a potential *φ(r)* defined as follows:

(1.6)

$$\boldsymbol{\varphi(r) = \frac{q}{4\pi D(\varepsilon 0)\, r}\, e^{-r/r_{Debye}}}$$

In equation 1.6, $\boldsymbol{\varphi(r)}$ is the electorsatic potential, $\boldsymbol{q}$ the point charge, **D** is dielectric, $\boldsymbol{\varepsilon_0}$ is zero permittivity, $\boldsymbol{r}$ the distance from the centre of the point charge and $\boldsymbol{r_{Debye}}$ the Debye length. The Debye length is the distance within which significant Coulombic contributions between charges occur. It is inversely proportional to the square root of ionic strength, *i.e.* it decreases as ionic strength increases. At physiological ionic strength (~150 mM) the Debye length is approximately 8 Å. The underlying principle is that Coulombic interactions experienced between ions in solution are so dominant in contributing to the non-ideality of an ionic solution to the point that other contributions can be neglected. The Debye-Hückel theory postulates that ions in solution are non-uniformly distributed. A solution may be overall electrically neutral, but near any given ion there is a higher concentration of counterions (oppositely charged ions). Over time, a greater amount of counterions than like ions accumulate around a given ion, and their directions are uncorrelated. Averaged over time, this movement appears as a spherically shaped haze with the same net charge as the ion, but with opposite charge. The spherical distribution of this resulting excess of counterions is termed the ionic atmosphere.

When considering the ionic atmosphere around an isolated ion in solution, the concentration of counterions decreases exponentially with distance from the ion. The relationship between the Debye length and counterionic concentration is illustrated in figure 1.7 below.

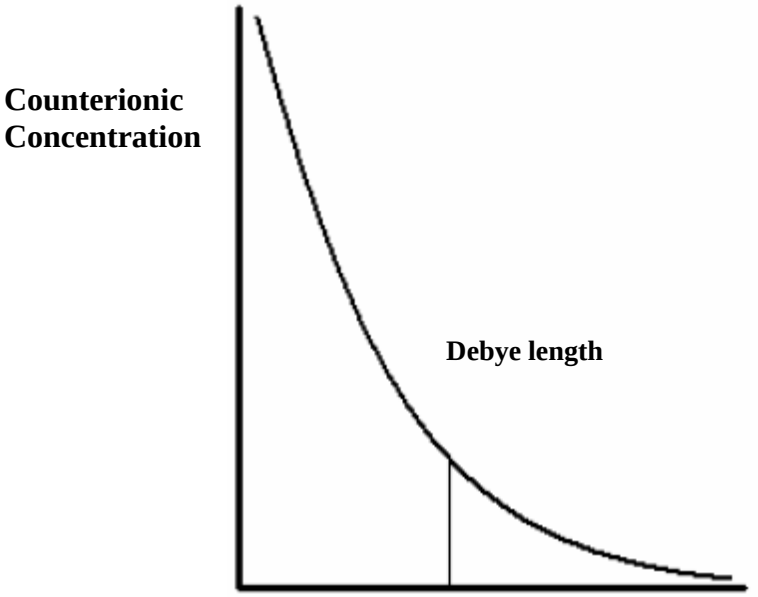


Figure 1.7. Debye length vs counterionic concentration. The counterionic concentration decreases exponentially with Debye length. Physically, the Debye length is the distance at which an ion does not feel the presence of other ions. The concentration of counterions is exponentially proportional to the distance from the ion. Image adapted from (www.uvm.edu).

1.5 pH and pK_a

Two very important concepts in molecular biochemistry and biophysics are pH and pK_a. They describe the propensities of molecules in aqueous solutions to donate or lose protons. Species that tend to lose protons (deprotonation) are termed acids while those that accept protons (protonation) are called bases. Cellular and physiological environments are almost exclusively aqueous, and consequently these concepts are very important in describing electrostatic interactions. pH (power of hydrogen) is one of the most important quantities in biochemistry and physiology, and corresponds to the concentration and activity of solvated hydronium ions (H_3O^+, commonly referred to as hydrogen ions and abbreviated as H^+) in a solution. Solvation refers to the process of dissolving a solute (small molecules, proteins, ionic species, *etc*.) in a solvent (usually water). Thermodynamic interactions between the solvent and the solute can impact the chemical state of the solute, such as the dissociation of a salt compound into its component elements. Protein interactions are considered within biological aqueous environments where acids and bases play key

roles in the underlying solvent-solute chemistry. pH is a measure of the acidity or basicity of a solution and is mathematically expressed as the negative logarithm of hydronium ion concentration:

$$\mathbf{pH} = -\mathbf{log}([\mathbf{H_3O^+}]) \tag{1.7}$$

This relationship means that pH is a function of two variables: (i) the concentration of the solution and (ii) the identity of the acid (given two equally concentrated acids, the solution of the stronger acid will have a lower pH because it is more dissociated than the weaker acid). The pH scale measures dissociation on a logarithmic scale ranging from 1 (extremely acidic) to 14 (extremely alkaline), with 7 being neutral (neither acidic nor basic). The majority of bodily fluids has a pH near neutral and range from 6.5 – 8.0, with that of blood being regulated to remain around a value of 7.4. The process of dissociation can be expressed by the chemical equation:

$$\mathbf{AH + H_2O = A^- + H_3O^+} \tag{1.8}$$

In this reaction equation, AH is a generic acid and A^- is its conjugate base formed by the deprotonation of AH. A special equilibrium constant can be defined to represent the point at which the ratio of reactants and products being exchanged is constant and consequently substances move between both sides of the equation at an equal rate producing no net change, *i.e*. the dissociation constant:

$$\mathbf{K_a} = \frac{[\mathbf{H_3O^+}]\,[\mathbf{A^-}]}{[\mathbf{AH}]} \tag{1.9}$$

The equilibrium constant usually factors into account the concentration of water, but in solutions where acidic compounds are sufficiently diluted, such as the intracellular environment, the concentration of water can be approximated as being constant. Like hydronium ion concentration, the dissociation constant can also be expressed in logarithmic form as follows:

$$\mathbf{pK_a} = -\mathbf{log}(\mathbf{K_a}) \tag{1.10}$$

pK_a, the acid dissociation constant, provides a quantitative measure of the strength of an acid in solution. Because of the negative sign, a low pK_a indicates a high K_a value and therefore a strong acid and vice versa. pH and pK_a are therefore related concepts, the former measuring the acidity of a solution while the latter measures the tendency of an acid to dissociate. There is an important mathematical relationship between the two quantities, *i.e.* pK_a is the pH value at which at which an acid is exactly half dissociated. This is demonstrated by rearranging the equilibrium equation:

(1.11)

$$[\mathbf{H_3O^+}] = \mathbf{K_a}\,\frac{[\mathbf{AH}]}{[\mathbf{A^-}]}$$

$$\mathbf{pH} = \mathbf{pK_a} + \mathbf{log}\frac{[\mathbf{A^-}]}{[\mathbf{AH}]}$$

What follows is that when [AH] = [A^-], it holds that pH = pK_a. This relationship is important because it dictates solubility properties and charge states of amino acids. In general, at a pH above the pK_a an acid exists mostly as A^- in water and will therefore be soluble, whereas at a pH below the pK_a the acid exists mostly as AH in water and will be insoluble. The opposite is true for basic compounds.

The acid-base dynamics apply to proteins in cellular environments as they are polyionic macromolecules with amphoteric units. Amino acids are amphoteric (also termed zwitterions) substances because they react with both acids and bases on account of their α-amino ($-NH_2$) and α-carboxyl (-COOH) groups. All 20 amino acids therefore have two dissociable H_3O^+ ions, but only charged amino acids possessing acidic or basic side chains have more than two. When an amino acid is incorporated into a polypeptide, the charges on the α-amino and α-carboxyl groups disappear. Hence the complex ionic properties of a protein are conferred by the five amino acids whose side chains can possess a charged state under certain pH conditions. The charge state dynamics and their pH dependence for each ionisable side-chain are reviewed in the following section.

1.5.1 pH dependence of charge state for ionisable amino acids

Among the five amino acids with side chains that are generally charged at neutral pH, two become negatively charged (aspartic acid and glutamic acid) and three become positively charged at neutral pH levels (lysine, arginine, and histidine). Despite having slightly different R groups, aspartic and glutamic acid both have a carboxylic group ($-COO^-$) on their respective side chains. At a pH greater than the corresponding pK_a, the carboxylic group is deprotonated (loses a hydronium ion) and becomes negatively charged. Hence in their deprotonated states they become acids. At a pH lower than the corresponding pK_a, both aspartic and glutamic acid are uncharged. The deprotonation occurs as shown in figure 1.8.

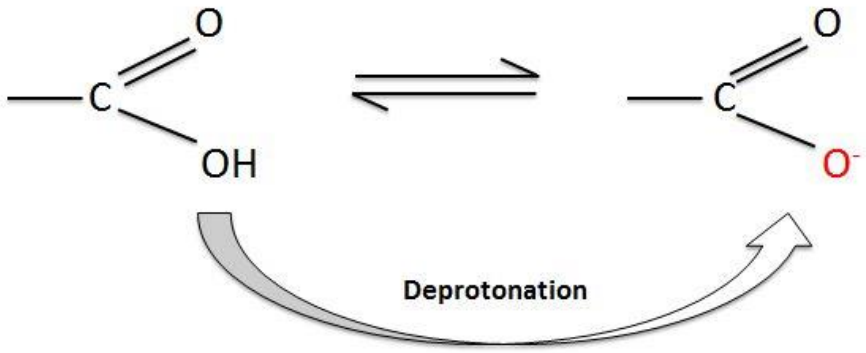


Figure 1.8. **Deprotonation of carboxylic side chain group (Asp/Glu).** Reaction is shown for negatively charged aspartate and glutamate side chains. Diagram is illustrational (structural aspects are not considered).

Lysine, arginine and histidine have distinct side chains. Lysine possesses an amino group ($-NH_3^+$ in protonated form), arginine a guanidinium group in which the charge is delocalised and histidine has an imidazole side chain. At a pH below the corresponding pK_a, the amino group is protonated (gains a hydronium ion) and becomes positively charged. Hence in their protonated states they become basic. At a pH greater than the pK_a, all three side chains are uncharged. The protonation occurs as shown in figure 1.9:

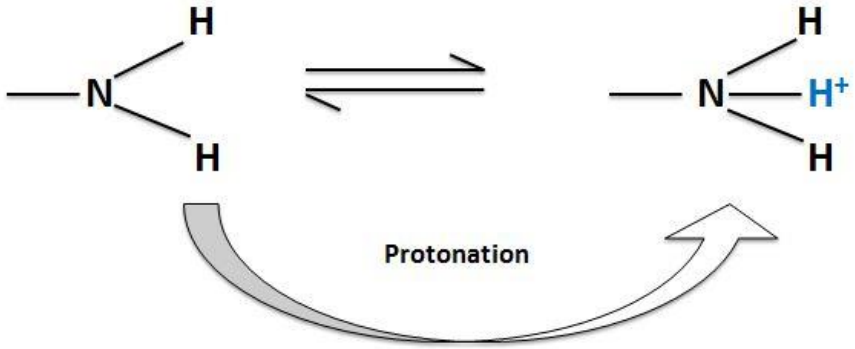


Figure 1.9. **Protonation of amino side chain group (Lys).** Reaction is shown for positively charged lysine side chain specifically (*N.B.* arginine and histidine have different side chains). Diagram is illustrational (structural aspects are not considered).

In this context, the pK_a is the pH value at which half of the carboxylic or amino side chain groups are charged. The pK_a values of the five ionisable side chains, as well as an overview of the pH-dependent charge state dynamics are summarised in tables 1.2 and 1.3 that follow.

Table 1.2 pK_a values for charged side chains

Amino Acid	**Side chain pK_a**
Aspartic Acid (Asp)	3.9
Glutamic Acid (Glu)	4.2
Lysine (Lys)	10.5
Arginine (Arg)	12.5
Histidine (His)	6.0

Table 1.3 pH Dependence of Amino Acid Charge States

Amino Acid	**Ionisable side chain group**	**pH / pK_a Relationship**	**Charge State**
Aspartic Acid (Asp)	Carboxyl (-COOH)	$pH > pK_a$ $pH < pK_a$	Negative (-) Neutral*
Glutamic Acid (Glu)	Carboxyl (-COOH)	$pH > pK_a$ $pH < pK_a$	Negative (-) Neutral
Lysine (Lys)	Amino ($-NH_2$)	$pH > pK_a$ $pH < pK_a$	Neutral Positive (+)
Arginine (Arg)	Guanidinium $-HNC(NH_2)_2$	$pH > pK_a$ $pH < pK_a$	Neutral Positive (+)
Histidine (His)	Imidazole $-(CH)_2N(NH)CH$	$pH > pK_a$ $pH < pK_a$	Neutral Positive (+)

* The pH at which a molecule carries no net charge is termed the isoelectric point (pI)

Electrostatic interactions between proteins and other molecules are based on ionised states of amino acid side-chains. Hence, changes in their pK_a values can be used to probe potential functional relevance of a particular residue or group of residues.

1.6 Aims of the Thesis

The overarching aim of this thesis is to investigate sequence- and structure-based features that influence protein aggregation and solubility. As discussed, understanding the principles underlying non-specific interactions and aggregate formation is crucial in assessing the developability of protein-based therapeutics. Furthermore, extending our understanding of the mechanistic basis of protein stabilisation in solution, most often with excipients, is also pivotal to the field of bioprocessing. Protein aggregation is known to have both intrinsic (*i.e.* native propensity to form unfolded aggregates) and environmental causes. The focus of this work is based largely on the intrinsic (sequence and structural) features of proteins, and specifically on how these can be harnessed in a predictive capacity. The key issues that will be addressed in the following chapters are: (i) how sequence-based and physico-chemical properties can be used to predict solubility and (ii) how studying interactions with excipients known to stabilise protein formulations can be used to develop predictive models. The ultimate goal of research in this area is the development of computational screening methods that may potentially be valuable to upstream production pipelines of protein-based therapeutics. This is certainly the case where protein-based therapeutics suffer aggregation during storage (shelf life), necessitating expensive and time-consuming procedures to re-solubilise the active product. There are three results chapter, briefly outlined here, each focusing on separate albeit overlapping aspects of sequence- and structure-based solubility prediction as well as excipient interaction-based stabilisation prediction.

Chapter 2 describes my contributions to a publication which was co-authored by me and comprises a subset of the presented findings (Warwicker *et al*., 2014). This body of work focuses on protein-based therapeutics and will investigate the performance of three features as binary classifiers where the dichotomy is between soluble and insoluble classes. Class membership is decided based on *E. coli* protein data (Niwa *et al*., 2009). The features considered are both structure- and sequence-based and relate to protein surface potential, surface polarity and lysine/arginine composition. The protein datasets comprise antibody-based therapeutics and non-antibody biologics. The objective of this investigation is to use computationally derived thresholds for each of the features considered (Chen *et al*., 2013) for *E. coli*-based solubility data (Niwa *et al*., 2009) and apply them to therapeutic datasets. The rationale behind this is to apply well performing

solubility classifiers for bacterial proteins (Chen *et al*., 2013) to eukaryotic proteins. Although important differences exist between bacterial proteomes and their eukaryotic counterparts, using the former as a benchmark for analysing solubility trends is reasonable in light of the sparsity of high-throughput "–omics" studies of protein solubilities for proteomes of higher organisms (Obrezanova *et al*., 2015). Furthermore, the choice of *E. coli* as a popular expression system for recombinant proteins (Rosano and Ceccarelli, 2014) renders the details of its cellular physiology relevant to studies investigating the developability of protein-based therapeutics.

Chapter 3 extends the investigation of chapter 2 to include a board range of physico-chemical properties used as features to discriminate expression levels of proteins. There are several important divergences from the first results chapter in this context. As opposed to focusing explicitly on structure-based features that perform well in classifying *E. coli* protein solubility, a more global approach is taken, with properties relating to charge, folding, β-strand propensity and amino acid composition being considered. However, large-scale studies reporting data on proteome-wide solubilities are sparse. This issue is managed by using more commonly quantified properties such as protein abundance and expression levels (usually measured as number of protein copies per cell) as a proxy for solubility. The rationale behind this study was two-fold. First, it was desired to explore how the novel lysine/arginine composition feature investigated in chapter 2 could be extended to non-therapeutic protein datasets given the observed trends related to enrichment in the human proteome (Warwicker *et al*., 2014). Furthermore, there was interest in studying how charge-based features scale against broader physico-chemical (both sequence- and structure-based) features when used to separate high- and low-abundance proteins, both in bacterial and eukaryotic environments. Equally important was to investigate how the separation capabilities of these features compares to established findings regarding properties with well-characterised contributions to protein solubility prediction such as protein length, surface charge (Kramer *et al*., 2012), surface polarity (Chennamsetty *et al*., 2009; Chennamsetty *et al*., 2010) and amino acid sequence (Thompson *et al*., 2006).

Hence, relevant data from literature were used to establish if the features that performed well as binary classifiers of solubility in *E. coli* and therapeutic datasets would replicate their discrimination ability in abundance/expression level data. Statistical *z*-scores and Pearson correlation coefficients are used to compare features enriched in high- and low-solubility/abundance proteins and displayed on coloured heatmaps. The overarching goal of this chapter's work was to identify sequence-based and 3D structural features that perform well in classifying proteins as soluble/insoluble and high-/low-abundance/expression. Features that perform well against both *E. coli* proteome solubility data and high-throughput abundance data are integrated into a prediction model that has been implemented on a University-hosted webserver, which will hopefully be

expanded upon substantially based on future work. A brief overview of the GUI is presented in this results chapter. The findings of this work have not yet been published but have been written in manuscript format. An abstract of the intended manuscript to be submitted is included in the beginning of the chapter.

Chapter 4 investigates protein-excipient interactions using structural annotations from the PDB database. A set of PDB-derived compounds used as crystallisation agents (Gorrec, 2009) is used in this capacity as a set of excipients that bind to proteins. Each of these excipient compounds is searched against the entire PDB using a web tool that retrieves structural annotations for a queried ligand. This information is used to compare the side chain interaction environment of between each excipient and matching PDB entries. A cosine metric of similarity between vectors encoding excipient contact information and amino acid side chain interactions is used to compare. The work carried out in this chapter comprises a low-resolution study of protein-excipient contacts. The aim was to establish any underlying patterns in the contact environment of proteins with structural annotations and the considered set of excipients. Finally, a conclusion chapter assesses what aims were achieved and not achieved and discusses how future work in this direction should proceed.

Chapter 2. Protein Therapeutics: A Novel Sequence-based Solubility Prediction Tool

---

Publication

Lysine and Arginine Content of Proteins: Computational Analysis Suggests a New Tool for Solubility Design[1]

ABSTRACT: Prediction and engineering of protein solubility is an important but imprecise area. While some features are routinely used, such as the avoidance of extensive non-polar surface area, scope remains for benchmarking of sequence and structural features with experimental data. We study properties in the context of experimental solubilities, protein gene expression levels, and families of abundant proteins (serum albumin and myoglobin) and their less abundant paralogues. A common feature that emerges for proteins with elevated solubility and at higher expression and abundance levels is an increased ratio of lysine content to arginine content. We suggest that the same properties of arginine that give rise to its recorded propensity for specific interaction surfaces also lead to favourable interactions at non-specific contacts, and thus lysine is favoured for proteins at relatively high concentration. A survey of protein therapeutics shows that a significant subset possesses a relatively low lysine to arginine ratio, and therefore may not be favoured for high protein concentration. We conclude that modulation of lysine and arginine content could prove a useful and relatively simple addition to the toolkit available for engineering protein solubility in biotechnological applications.

KEYWORDS: *protein aggregation, bioinformatics, solubility prediction, biologics, amino acid side chain charge*

---

[1]Warwicker, J., Charonis, S., Curtis, R.A. (2014). Lysine and Arginine Content of Proteins: Computational Analysis Suggests a New Tool for Solubility Design. *Mol Pharm*. **11**(1), 294 - 303

2.1 Objectives

This chapter focuses on therapeutic proteins, and specifically engineered antibody fragments as well as non-antibody biologics. Sequence-based and 3D structural features that perform well in predicting protein solubility for the *E. coli* proteome are used on constructed datasets of therapeutic proteins. Three properties are investigated: (i) KR-ratio (*i.e.* the ratio of lysine to arginine content of a protein's primary sequence), (ii) positively charged surface patch size, and (iii) non-polar surface patch size. The reason for using such specific features as binary classifiers is based on their high classification performance for cell-free *E. coli* proteins (Niwa *et al*., 2009). Performance in discriminating soluble from insoluble proteins has been measured by ROC analysis in a study reporting the emergence of positive surface charge and non-polarity as the best binary classifiers for solubility in the *E. coli* proteome (Chan *et al*., 2013). Hence, the empirically derived thresholds (see table 2.1) for each of these features in a bacterial system are tested on therapeutic proteins. The differences between the *E. coli* and human proteomes as well as their cellular physiologies imply that using solubility data from one template to separate proteins from the other may impose significant limitations. However, as discussed in the previous chapter, the sparsity of large-scale solubility datasets for eukaryotic proteomes means that until fruitful high-throughput studies are undertaken, benchmark data will necessarily be constrained to non-eukaryotic proteomes.

Using this rationale, this chapter investigates how features that discriminate *E. coli*-based proteins as soluble or insoluble scale to antibody-based and therapeutic proteins native to eukaryotes. Two of the three considered features are structure-based (surface positive charge and surface non-polarity) and hence require structural annotations, whereas analysis of the lysine/arginine composition ratio can be extended to all relevant proteins with known primary sequences. Although surface charge (Kramer *et al*., 2012) and non-polarity are features with well-established contributions to protein solubility (Chennamsetty *et al*., 2009), the ratio of lysine to arginine composition (henceforth referred to as KR-ratio) is potentially a hitherto uncharacterised solubility prediction feature. Hence, another focus of this chapter is to assess the ability of KR-ratio to separate soluble and insoluble therapeutic proteins, and to compare its discrimination performance to that of patch-based surface charge and non-polarity. The findings presented in this chapter comprise a subset of those described in a publication detailing and extending the current investigation to non-therapeutic human proteins present in high concentrations (*i.e.* myoglobins and serum albumins) and their less abundant paralogues (Warwicker *et al*., 2014).

2.2 Protein-based Therapeutics

Protein-based therapeutics (also termed biologics), with their ability to deliver high affinity binding to targeted molecules, are gaining increasing importance in pharmacological intervention (Roberts *et al*., 2015). The number of modified and unmodified proteins approved for clinical use by regulatory authorities of the USA and the EU runs into the hundreds, with several more in the pipeline, and monoclonal antibodies (mAbs) accounting for almost one half (48%) of all sales revenue published in 2010 (Dimitrov, 2012).

Despite the rapid growth of interest in biologics, the process of bringing a protein-based pharmacologic agent into the market is a challenging task. Establishing a therapeutic requirement and a specific molecular target are first steps and rely on input from genomic, transcriptomic, and proteomic datasets from studies. Once a candidate protein has been identified, early in the process the amino acid sequence is locked in, so that there are no changes during the prolonged clinical trial stages (Roberts *et al*., 2015). This presents a major challenge for avoiding the solubility and aggregation problems that may become evident only in later stages of development such as formulation. Hence, screening for and designing protein sequences having optimised solubility properties is a very important objective in the field of bioprocessing.

Antibodies and their derivatives comprise one of the most successful paradigms for the design of high-affinity, protein-based binding reagents (Holliger and Hudson, 2005). In the realm of biopharmaceuticals, they have arguably the most streamlined solubility behaviour, perhaps on account of being subjected to evolutionary pressures to circulate at relatively high concentrations *in vivo* (Roberts *et al*., 2015). Protein solubility as a generic principle has been a crucial property up to and through the advent of recombinant protein production. Prior to the era of recombinant protein expression, the natural abundance of proteins largely determined which were studied in detail. When protein overexpression techniques were introduced, along with whole genome sequencing, proteins could be chosen and studied on the basis of the underlying science in question instead of their abundance levels. However, overexpression did not guarantee that a soluble product would result (Esposito and Chatterjee, 2006). This is related to the problem of avoiding protein aggregates when developing and formulating proteins in the increasing important area of biopharmaceuticals (den Engelsman *et al*., 2011).

2.3 Immunoglobulins

Immunoglobulins are large, complex proteins that act as the main agents of the adaptive immune system, protecting against foreign pathogenic agents known as antigens. Immunoglobulins of the same antigen specificity are secreted as antibodies by specialized B-lymphocytes, and as mediators of adaptive immunity they are one of the most important classes of proteins in the pharmaceutical industry. Bioprocessing pipelines which aim to manufacture high-concentration antibody solutions for healthcare applications frequently run into aggregation bottlenecks. Hence, understanding these issues mechanistically is pivotal in the development of more stable therapeutics. Indeed, monoclonal antibodies are successful drugs because they exhibit significant therapeutic potential with few side effects. However, they are less stable than low molecular weight chemical compounds and are prone to chemical and physical degradation (Wang *et al*., 2007). Aggregated antibody is a degraded product that is often generated during the manufacturing process (Vázquez-Rey and Lang, 2011) and can exhibit low efficacy and trigger immunogenic responses (Rosenberg, 2006).

Antibodies are roughly Y-shaped immunoglobulin molecules consisting of two light chains and two heavy chains. Each of these chains is in turn comprised of a variable region at the N-terminus and a constant region at the C-terminus. Being the secreted form of the B-lymphocyte receptor, they are soluble and readily obtainable from blood serum, and hence have been well studied. The generic structure of an antibody is shown schematically in figure 2.1.

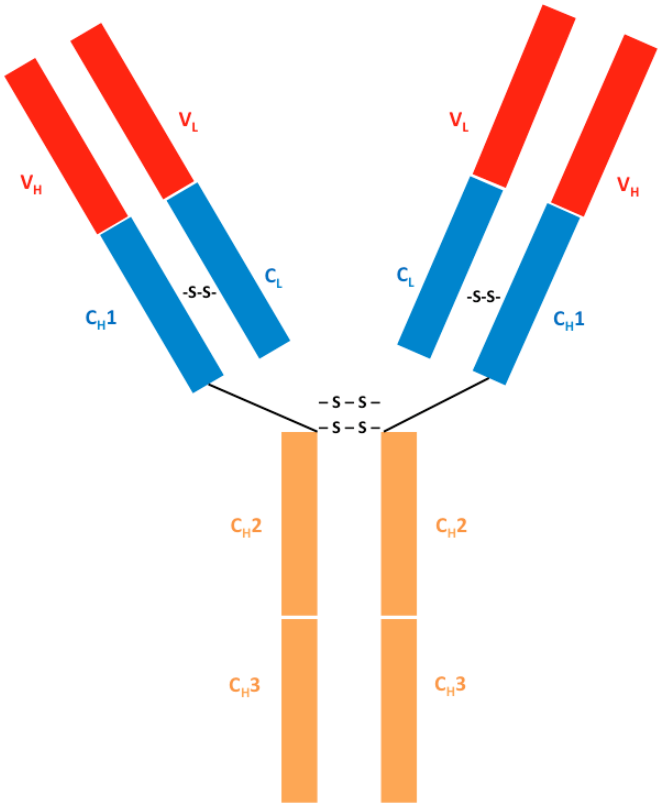


Figure 2.1. Generic Structure of an Immunoglobulin Molecule. Antibodies contain two immunoglobulin (Ig) heavy chains and two Ig light chains.

Two types of light chains are found in antibodies – κ and λ, and the constant regions of heavy chains have one of five different sequence types – γ, α, μ, δ, or ε. The sequence type of the heavy chains determines the class that the immunoglobulin belongs to; the five classes in humans are IgG, IgA, IgM, IgD, and IgE. The IgG class is by far the most abundant immunoglobulin and has several subclasses (1, 2, 3, and 4 in humans). The variable region of an antibody (F_v) consists of two identical light and heavy chain components, one on each branch of the molecule. The variable regions are collectively known as the CDR (complementarity determining region) and contain the site of interaction between antibody and antigen. The CDR possesses by far the largest sequence diversity, reflecting its need to accommodate binding of vast numbers of antigens.

2.3.1 Recombinant Antibody Fragments

The modular structure of antibodies described above means that properly folded and assembled antibodies consist of three equal-sized globular portions, rendering them readily

cleavable into functionally distinct fragments. Proteolytic enzymes (proteases) such as papain have been used to dissect antibodies into their constituent elements – Fab fragments (**F**ragment **a**ntigen) and Fc fragments (**F**ragment **c**rystallisable). Fab fragments consist of a complete light chain paired with the V_H and C_H1 domains of a heavy chain and thus correspond to the two identical branches of an antibody containing the antigen-binding domains. Hence they are non-synthetic proteolytic fragments of immunoglobulins, essentially wild-type antibodies lacking the Fc region. Fc fragments consist of paired C_H2 and C_H3 domains and correspond to the part of an antibody that interacts with effector molecules. Fc-mediated effects are not required and often undesirable for a range of applications; this has given rise to genetically engineered monovalent and bivalent fragments (Hollinger and Hudson, 2005).

Single chain variable fragments (scFv) are popular recombinant monovalent formats in which the V domain of a heavy chain (V_H) and the V domain of a light chain (V_L) are joined together with a flexible synthetic polypeptide linker preventing dissociation. Antibody Fab and scFv fragments retain the specific, monovalent, antigen-binding affinity of their parent IgG molecules, while exhibiting improved pharmacokinetic properties (Holliger and Hudson, 2005). Schematic diagrams of the Fab and scFv fragments of an antibody are illustrated in figure 2.2.

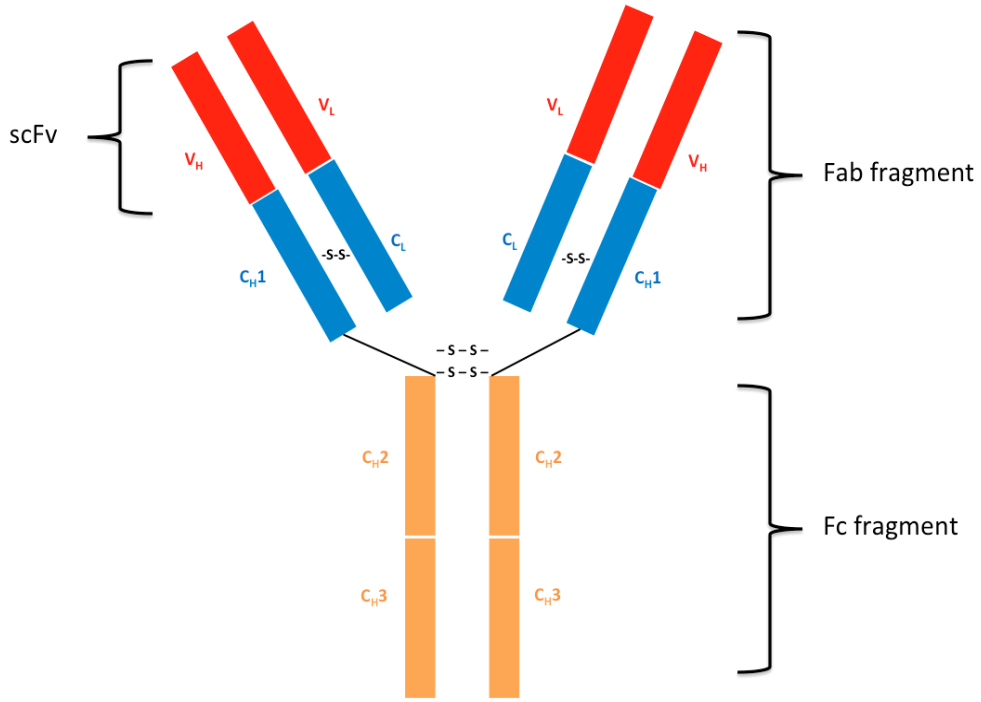


Figure 2.2. Recombinant Antibody Fragments. The single variable chain fragment (scFv) Fab fragment (fragment antigen), and Fc (fragment crystallisable) constituents of an immunoglobulin molecule.

2.4 Sequence- and Structure-based Features for Solubility Prediction

Several properties pertaining to both sequence and 3D structure, some of which were overviewed in section 1.3.3, have been used for correlating and predicting protein solubility. A corollary of the oil drop model for protein folding (Kauzmann, 1959) is that non-polar regions could lead to protein-protein interactions. This is a common theme throughout the analysis of protein structure and function (Jones *et al*., 2002). One approach involves ranking non-polar patches (Cole and Warwicker, 2002), which are expected to mediate non-specific interactions and influence colloidal stability. In this context, a patch is defined as a localised region of a protein's solvent accessible surface area with chosen dimensions. Such patches are a key component of the spatial aggregation propensity (SAP) method that has been successfully implemented to redesign antibodies for improved solubility (Chennamsetty *et al*., 2009; Chennamsetty *et al*., 2010). Whereas non-polar patch analysis requires a 3D structure or model in order to make predictions, several of the previously reviewed sequence-based methods (TANGO, PAGE, PASTA, AGGRESCAN, Zyggregator, and BETASCAN) focus on identifying amyloidogenic sequence

stretches of a protein. As discussed, the propensity of a sequence to form amyloid structures is commonly assessed by the likelihood of β-sheet formation.

Returning to the colloidal stability properties probed by 3D-based patch analysis, non-polar surface reduction is complemented by changing the surface charge properties of a protein. Several reports indicate that negative charge is preferred over positive charge for properties related to protein solubility, such as aggregation resistance (Arbabi-Ghahroudi *et al*., 2009; Perchiacca *et al*., 2011; Dudgeon *et al*., 2012). Furthermore, certain studies have found protein solubility to be correlated with negative surface charge (Kramer *et al*., 2012). The detail of protein solubility modification by charge is likely to be a case- and condition-dependent combination of factors that include reduction of non-polar regions (Ho and Middelberg, 2004), overall net charge repulsion (Chi *et al*., 2003; Olsen *et al*., 2009), and attraction from positive and negative regions in an anisotropic charge distribution (Saluja *et al*., 2007; Yadav *et al*., 2012). There is strong evidence from supercharging protein surfaces (Lawrence *et al*., 2007), as well from the tradition of handling proteins away from their pI values to prevent aggregation, that net charge can be a major factor in determining solubility. It is believed that at least a part of the effect of supercharging lies in preventing aggregation of partially unfolded states (Der *et al*., 2013), analogous to the avoidance of β-sheet forming regions.

The motivations for the work undertaken in this chapter were grounded in findings relevant to the correlation of surface non-polarity and surface charge properties with protein solubility. An experimental solubility dataset from a large-scale study of protein aggregation for *Escherichia coli* in a cell-free environment (Niwa *et al*., 2009) was used, in which 3173 translated proteins were quantified using PCR and classified as either soluble or insoluble using centrifugation assays. The experiments in this study were performed in a chaperone-less cellular environment, so that the native aggregation propensity of the proteome could be measured. An interesting finding was that the distribution of protein solubilities exhibited a clear bimodal shape, leading to the conclusion that *E. coli* proteins can be classified in terms of native aggregation propensity as either soluble or insoluble.

The authors of this work concluded that factors correlating to some degree with solubility include charge and structural class. This genome-wide dataset has been used to study structural features that correlate with solubility. Interestingly, the structural feature that correlated best with solubility was a lack of large positively charged surface patches, where the difference in positive patch signatures for the separation of soluble and insoluble datasets was similar to that seen for DNA-binding versus non-DNA-binding proteins (Chan *et al*., 2013). As a result, it was speculated that interactions between expressed proteins and nucleic acid (mRNAs, tRNAs) may lead to formation of insoluble protein/nucleic acid aggregates via an unknown mechanism. Consistent with

this hypothesis is the fact that there was no equivalent separation observed for negatively charged patches. Although this observation may be specific for solubility in expression systems, and not necessarily relevant for concentrated protein solutions with low levels of nucleic acid, the difference between positive and negative charge is intriguing. Because of the bimodality observed in protein solubility data, *i.e.* separated into subsets of soluble and aggregation-prone proteins, it makes sense, from a prediction standpoint, to treat properties such as charge and polarity as binary classifiers.

2.4.1 Binary Classification and Receiver Operating Characteristic Analysis

Binary classification is the task of classifying elements of a given set into two groups based on a classification rule. In the context of the current work, the set is composed of protein sequences and structures, and the classification rule is some property (*e.g.* surface charge, polarity) that will have the ability to discriminate between soluble and insoluble proteins with some level of accuracy. Surface polarity and surface charge have been shown to perform best among a series of numerous properties in binary classification of solubility (Chan *et al.*, 2013) for the cell-free *E. coli* study described above. These findings are illustrated as ROC plots in figure 2.3 adapted from the relevant publication.

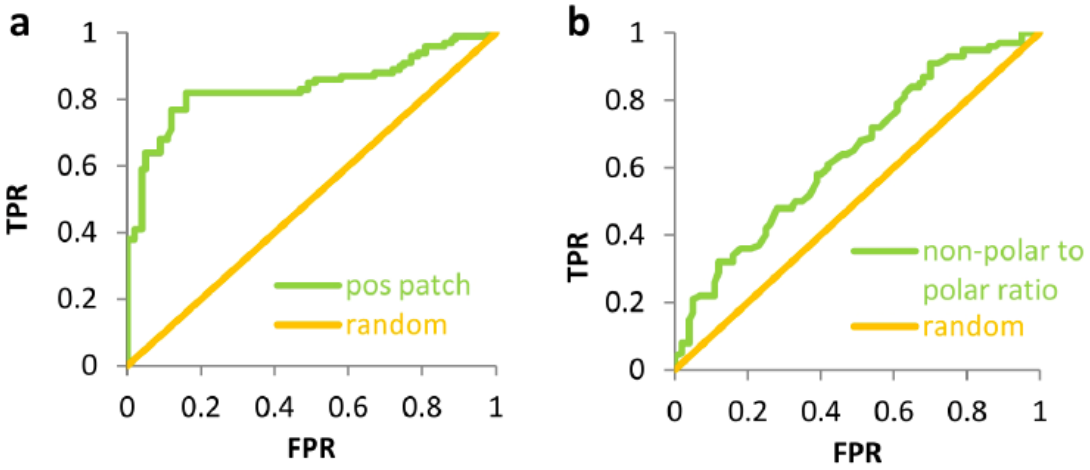


Figure 2.3. **Surface Charge and Polarity as Binary Classifiers.** ROC plots for soluble and insoluble subset separation. (**A**) ROC plot (AUC = 0.85) showing separation by positive surface potential. (**B**) ROC plot (AUC = 0.62) quantifying the separation by non-polar to polar surface ratio. TPR = True Positive Rate, FPR = False Positive Rate. Figure adapted from (Chan *et al.*, 2013).

ROC (receiver operating characteristic) analysis is a technique used frequently in machine learning to graphically illustrate the performance of a binary classifier as its discrimination threshold is varied. Specifically, ROC curves plot sensitivity (ordinate, y-axis) versus [1 –

specificity] (abscissa, x-axis). Sensitivity quantifies how well false negatives are avoided whereas specificity quantifies how well false positives are avoided.

Sensitivity (also termed the true positive rate or TPR) measures the proportion of positives that are correctly identified as such, *i.e.* TPR = TP / (TP + FN) where TP and FN are true positives and false negatives, respectively. Specificity (also termed the true negative rate or TNR) measures the proportion of negatives that are correctly identified as such. The value used in ROC plots on the abscissa corresponds to [1 – specificity], referred to as the false positive rate, or FPR. It is expressed as FPR = FP / (FP + TN) where FP and TN are false positives and true negatives, respectively. Accuracy measures the overall performance of a predictor and is expressed as [(TPR + 1 – FPR) / 2]. Hence, a ROC plot assesses the performance of a binary classifier by plotting the true positive rate versus the false positive rate. FPR, TPR, and accuracy all take values between 0 and 1, with TPR and accuracy being equal to 1 and FPR being equal to 0 in the case of perfect performance (classifying all members into their respective subset correctly). In each of the ROC plots in figure 2.3, a predictor (surface charge in 2.3a, surface polarity in 2.3b) is tested against a threshold value (linear curve titled random). The area under the curve (AUC) measures the ability of the discriminator to correctly predict proteins as either soluble or insoluble, with 0.5 indicating random and 1 indicating perfect discrimination.

The performance of large positively charged surface regions as a discriminator (figure 2.3a from Chan *et al*., 2013) between soluble and insoluble proteins means surface potential is an important feature for predicting solubility in *E. coli* proteins. It is important to acknowledge that the cell-free environment from which this data was obtained implies that this may not be the case for other systems. Nonetheless, the finding that surface charge is a better performing binary classifier than non-polarity (AUC = 0.85 vs. AUC = 0.62) for this system is an interesting one. This leads to the question of whether amino acid side chains that are normally positively charged at physiological pH (lysine and arginine) contribute equally to the observed correlation with solubility. A limited analysis indicated that they do not, with lysine being favoured over arginine for soluble proteins. Specifically, the positively charged surface patches that performed better than any other feature in terms of separation (soluble vs. insoluble) were enriched for arginine (Chan *et al*., 2013). This chapter explores this interesting observation further and the extent to which it can be applied to protein-based therapeutics. The incentive for this is that the KR-ratio, being a purely sequence-based and hence readily adjustable property, may potentially prove to be a simple method for designing proteins with improved solubility. If lysine is preferred over arginine in proteins that have evolved to function at high concentration levels, it might be expected that protein abundance or mRNA levels display a correlation with KR-ratio. Additionally, extracellular proteins at high abundance would have a high KR-ratio, while paralogues of these proteins, themselves present at

lower concentration, would be expected to have a lower KR-ratio. Paralogues are genes that share a common ancestor but have undergone functional divergence, encoding proteins with different functions.

2.5 Methods and Data

2.5.1 Sequence- and Structure-based Datasets

Three sets of protein structures were obtained from the Protein Data Bank (Berman *et al*., 2000), for scFv's, Fab fragments, and non-antibody protein-based therapeutics (biologics). Complete antibody structures are sparse in the PDB, but there are numerous scFv and Fab fragment structures. Sets of single chain variable fragments, Fab fragments, and biologics were constructed from the PDB and UniProtKB databases in order to perform sequence- and structure-based calculations.

Single Chain Variable Fragments (scFv's)

In order to compile a dataset of scFv structures, an advanced search of the RCSB PDB website was performed using "text search" and "macromolecule name" criteria, with the sequence similarity cutoff set at 100%. This set of criteria was chosen so that structure matches would correspond to true scFv fragments, since both a term in the form of a text string and a structural annotation would have to be present for an entry to be considered a hit. The sequence similarity cutoff was specified so that only sequences with 100% similarity (*i.e.* identical sequences) would be rejected. It is most often the case that scFv's are solved in complex with another protein. In order to isolate the relevant antibody chains from non-antibody polypeptide chains, a graphics screen was employed to ascertain the chains associated with genuine scFv structures. This resulted in a dataset of 24 scFv structure matches.

Fab fragments

The search criteria for Fab fragments were relaxed in comparison to those for scFv structures since they are naturally occurring substructures of wild type immunoglobulins. A basic text search using the term "Fab" was performed and subsequently all entries with 100% sequence identity (*i.e*. redundant sequences) were removed. This search yielded slightly over 1400 structures. The structures of many Fab fragments are obtained in combination with an antigen, making it necessary to isolate the antibody component. Rather than analysing this number of coordinate sets

using molecular graphics, in order to limit this set to true Fab fragments, only entries with H and L chains were selected using the custom report filtering option of the PDB website. This was followed by inspecting for chains of the expected size and a coordinate file header that contains reference to an antibody Fab fragment, resulting in 408 extracted H and L chains. H and L are part of the nomenclature traditionally used to denote antibody chains (since antibody chains can be classified as either heavy or light), but this notation is not universal and it is likely that certain fragment structures were excluded. Although selecting only H and L chains from the initial search may likely exclude a number of true Fab fragments, searching through the entire set of 1400 structures in a non-automated manner was not feasible. The subset of 408 structures with H and L chains was considered adequate in size for the scope of this investigation.

Non-Antibody Protein-based Therapeutics (Biologics)

Biologics encompass all therapeutic products used in clinical applications, and in this case literature searches were performed to identify such recombinant proteins. Dimitrov (2012) provides an excellent summary of biologic products currently on the market, while Holliger and Hudson (2005) review recombinant and engineered antibody fragments that are increasingly being developed and used as biopharmaceuticals. In order to compare computed structural features for therapeutic proteins with a background set of human proteins, a search for human proteins in the PDB was made, followed by a filter with the PISCES sequence culling tool (Wang and Dunbrack, 2005) for crystal structures with sequence identity less than 30% (non-redundancy) and sequence length within 40-10,000 amino acids. In this case, the dataset is too large to reliably screen for biological units, and calculations are performed on single protein chains (2073). A dataset of structures for therapeutic proteins that have reached the market was prepared from queries in the PDB database for entries corresponding to products reviews by Dimitrov (2012). A single coordinate entry was used for each biologic, and the analysis was again based on chains (62) rather than biological units, maintaining consistency with the background set of human protein structures. For estimation of the concentration of a marketed biologic at the point of delivery, the DailyMed resource (http://dailymed.nlm.nigh.gov) was searched for preparation and delivery guidelines of each therapeutic protein. The final dataset consisted of proteins such as erythropoietin, interferon, insulin, as well as monoclonal antibodies such as rituximab, pertuzumab, and herceptin. The full list of proteins and chains, along with information such as commercial name, concentration, and mode of administration where available, can be found in Appendix 2.A.

Sequence-based Datasets

With the focus on sequence-based KR-ratio, there is no need to restrict analysis to those proteins for which 3D structural annotations exist, when comparing with proteome-wide solubility data for cell-free expression in *E. coli*. For all three datasets (scFv's, Fab fragments, biologics), computing the KR-ratio was a straightforward task as only sequence annotations were required.

Hence, 3173 data points (open reading frames and their protein products) from experimental studies of solubility in this bacterial system were used as a benchmark, where subsets of low solubility (<30%) and high solubility (>70%). Solubility is defined in terms of inherent aggregation propensities and assessed by quantifying them within a reconstituted system containing only essential *E. coli* factors responsible for protein synthesis (Niwa *et al*., 2009). The histogram of solubilities exhibited a bimodal shape, indicating that *E. coli* proteins could clearly be dividing into either soluble or insoluble under the applied experimental conditions.

2.5.2 Sequence- and Structure-based Calculations

Three properties were chosen as binary predictors of solubility: (i) surface potential, (ii) surface polarity (both structure-based), and (iii) KR-ratio (sequence-based). These properties were tested on the therapeutic protein datasets described in section 2.5.1 to ascertain how well they performed in discriminating soluble from insoluble proteins.

Surface Potential

Surface electrostatic potential describes the distribution of charge on the surface of a protein. In this instance, the maximum positive contoured patch size was the feature of interest, as it has been demonstrated to be a well performing binary classifier (AUC = 0.85, figure 2.3a) (Chan *et al*., 2013) for cell-free *E. coli* expression data (Niwa *et al*., 2009). This 3D structure-based feature is referred to as the **maximum positive surface patch size** (table 2.1).

Electrostatic potential was calculated around each protein using a finite difference Poisson-Boltzmann method (Warwicker, 1986), as solving the Poisson-Boltzmann equation analytically is not tractable for irregular protein geometries. The linear PBE (equation 1.5, section 1.4.2) is solved numerically in terms of electrostatic potential $\varphi(\mathbf{r})$, where φ is a function of the position vector $\mathbf{r}$. Charges are assigned in a protein-solvent counterion system within a continuum electrostatics framework (the Poisson-Boltzmann equation) that is discretised via inscription onto a Cartesian grid and solved for electrostatic potential using the numerical method of finite differences. Negatively

charged aspartic and glutamic acid side chains and C-termini, as well as positively charged lysine and arginine side chains and N-termini were included. PBE solution calculations use existing code in the group based on earlier work (Warwicker, 1986), but similar results could also be obtained using one of several available PBE solvers implemented since then.

The resulting potential map was contoured at thresholds of + kT/e on a shell around the protein (k is the Boltzmann constant, T is absolute temperature 300 K, and e is electronic charge), with the ionic strength fixed at 0.15 M. The size of the largest contoured positive patch is recorded for each protein. Surface potential was calculated using a Cartesian grid-based approach as illustrated in figure 1.6 in which the linear PBE is solved for a protein-solvent system. Structural annotations for all therapeutic protein calculations were retrieved from the PDB database as described in section 2.5.1. The parameters used to construct electrostatic potential grids are fixed, so that a number of grid points always represent the same surface area for all proteins. A grid step of 0.6 Å was used, in which a two-dimensional grid element corresponds to 0.36 $Å^2$ and the threshold value of 3000 grid points is approximately equivalent to 1000 $Å^2$ surface area.

Surface Polarity

Surface polarity describes the distribution of polar and non-polar patches on the surface of a protein. Specifically, the maximum ratio of non-polar to polar solvent accessible surface area (SASA) for a patch was of interest, as it was also shown to perform well as a discriminator (AUC = 0.62, figure 2.3b) (Chan *et al*., 2013) for *E. coli* protein solubility. This structure-based feature is referred to as the **maximum ratio of non-polar to polar SASA** (table 2.1).

Solvent accessible surface area (SASA) refers to the area over which the centre of a water molecule can move while maintaining unobstructed contact with the side chain atoms. The ratio of non-polar SASA to polar SASA is calculated using a patch-based approach, in which patches are drawn out by spheres with radius 13 Å centered on all non-hydrogen atoms (Cole and Warwicker, 2002), with the maximum ratio for each protein recorded. A 1.4 Å radius solvent probe, approximately equivalent to the atomic radius of an H_2O molecule, was used to generate the solvent accessible surfaces. The ratio of non-polar to polar surface per patch is calculated using equation 2.1:

(2.1)

$$\textbf{patch ratio} = \frac{\textbf{SASA}_{np}}{\textbf{SASA}_{p}}$$

Here, $SASA_{np}$ refers to the non-polar surface area value for a single patch, drawn with 13 Å radius sphere, and $SASA_p$ is the equivalent value for the polar surface area of the patch. Carbon atoms are considered to contribute to non-polar SASA while nitrogen, oxygen and sulfur contribute to polar SASA. Atoms that are buried from solvent have zero contribution, as in the case of solvent inaccessibility it holds by definition that SASA = 0.

Graphical illustrations of surface electrostatic potential and surface polarity can be drawn using visualisation software such as PyMOL, an open source molecular graphics package. Figure 2.4 illustrates these concepts using one of the Fab fragments (PDB 7AB) as a template.

KR-ratio

KR-ratio is simply the ratio of the number of lysine residues to the number of arginine residues in a protein sequence, and is therefore readily computable for any protein with sequence annotations. All lysine and arginine residues of a protein sequence are counted without considering localisation in tertiary structure, as residues with ionisable side chains are most often solvent exposed and would only rarely be expected to be located in the buried protein core. In the rare case where there are no arginine residues in the sequence (*i.e.* zero denominator), a value of 1 is recorded in order to avoid a divide-by-zero error.

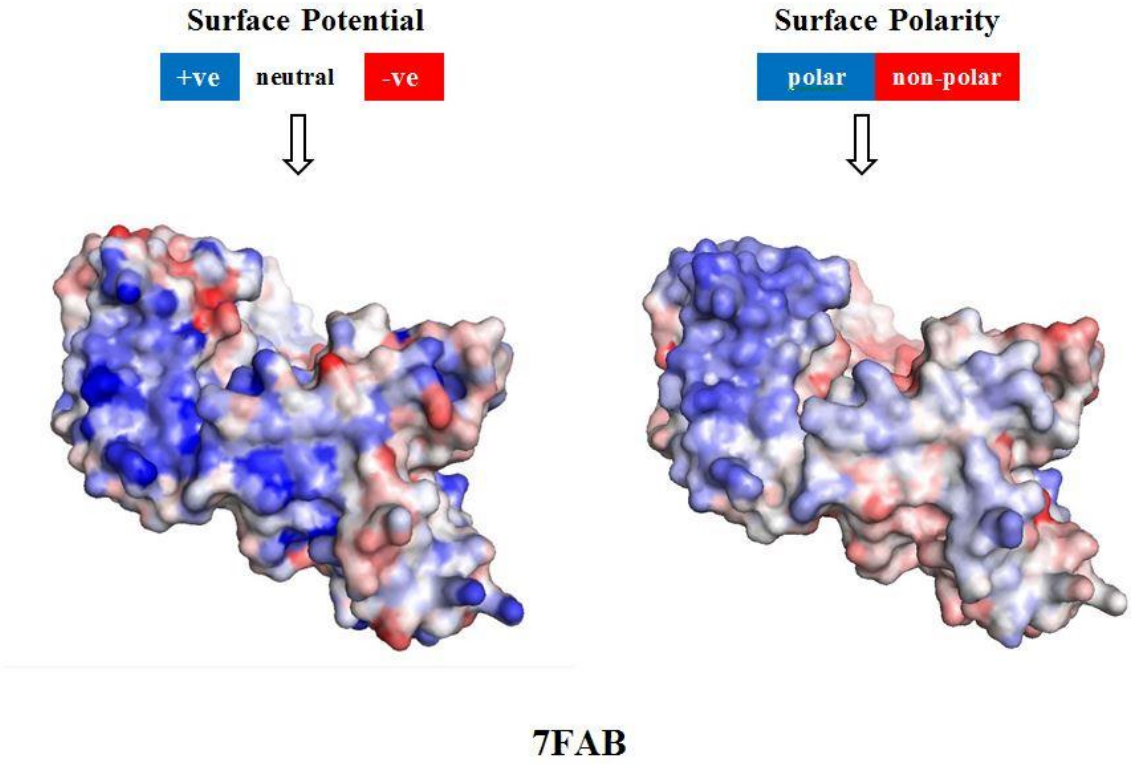


Figure 2.4. Surface Potential and Polarity Maps. The surface of the Fab fragment structure 7FAB, drawn using a charge map (left hand side) and a polarity map (right hand side). Illustrations were rendered using PyMOL.

Threshold values for these three solubility properties were derived in previous work (Chan *et al*., 2013). In the case of maximum positive patch size and ratio of maximum non-polar to polar patch size, these thresholds could only be calculated based on proteins of the soluble (>70%) and insoluble (<30%) subsets of the *E. coli* solubility dataset (Niwa *et al*., 2009) for which 3D structures have been solved. Following cross-referencing PDB structures to the UniProtKB sequence repository (Bairoch *et al*., 2005) and further processing to remove redundant sequences, final subsets of 111 (soluble) and 56 (insoluble) *E. coli* proteins were available for processing (167 PDB structures total). The thresholds represent the values that best separate soluble and insoluble subsets for the cell-free expression data, when these subsets are plotted as cumulative distributions. The thresholds obtained in this manner are 3000 grid points (~1000 $Å^2$) for maximum positive patch size, 4.5 for the ratio of non-polar to polar SASA, and 1.2 for the ratio of lysine to arginine residues a sequence. These thresholds are depicted as dotted red lines in figures 2.5 – 2.7.

Table 2.1 Threshold Values for Solubility Determining Properties

Binary Classifier for Solubility Prediction	Calculation Method	Threshold Value*
Maximum positive surface patch size	Numerical solution of linear PBE (equation 1.5) solved for electrostatic potential $\varphi(\mathbf{r})$ using a finite difference Cartesian 3D grid-based method (figure 1.6)	3000 grid points (~1000 Å^2)
Maximum ratio non-polar to polar SASA	Patch-based ratio of non-polar to polar surface area (equation 2.1)	4.5
KR-ratio	Ratio of lysine to arginine residues	1.2

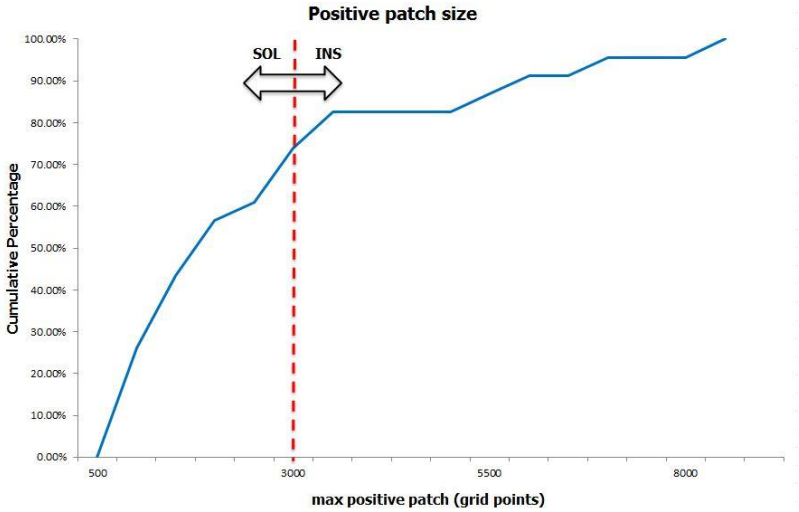
* Thresholds for all three features are empirically derived from cell-free *E. coli* expression data (Niwa *et al*., 2009) using ROC analysis from (Chan *et al*., 2013).

Unix shell scripts were written to automate various data cleaning tasks, such parsing protein sequences from PDB files of scFv and Fab fragment structures. A Python script was written to extract sequence-based information from PDB structures, *e.g*. KR-ratio as well as amino acid composition and output them in Excel-readable format. Excel spreadsheets were used for data presentation and analysis. An in-house computational pipeline implemented in Perl (code contributed by Jim Warwicker) was used to perform all structure-based calculations.

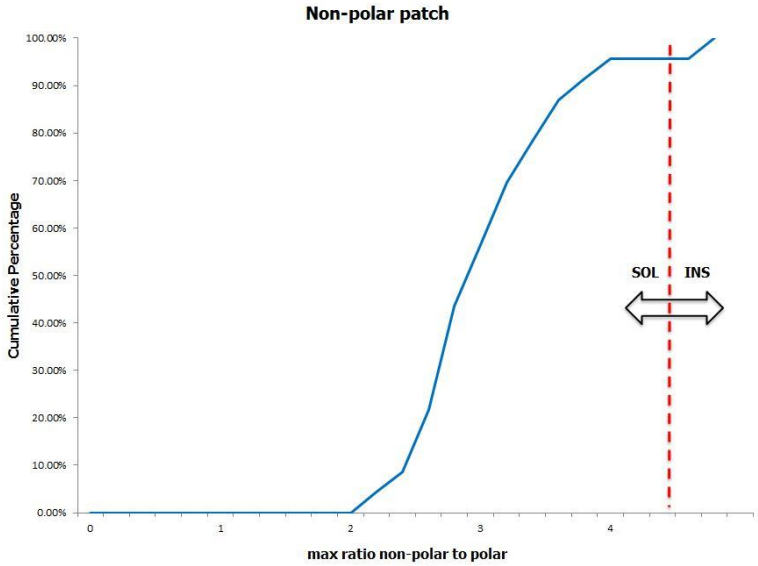
2.6 Distributions of Charge and Non-Polar Features for Therapeutic Proteins

The three described features were calculatxed individually for each protein dataset (scFv's, Fab fragments, and biologics). The findings are presented in figures 2.5 – 2.7 that follow, where cumulative frequency plots are used to show the separation of soluble and insoluble proteins according to each feature based on *E. coli* protein solubility data. In each of the three figures, plots A, B and C refer to maximum positive surface patch size, maximum ratio of non-polar to polar SASA and KR-ratio, respectively. The proportion of proteins classified as soluble/insoluble is illustrated in each cumulative frequency plot where the point of separation is indicated by a dotted line with an arrow at the top. For all protein datasets, these lines correspond to the *E. coli*-based thresholds reported in table 2.1 and each feature is calculated as described in section 2.5.2 (summarised in the second column of the table).

A



B



C

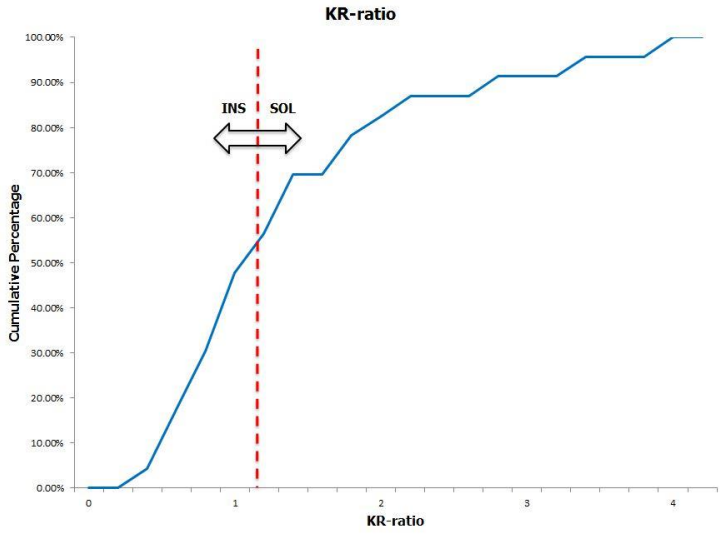
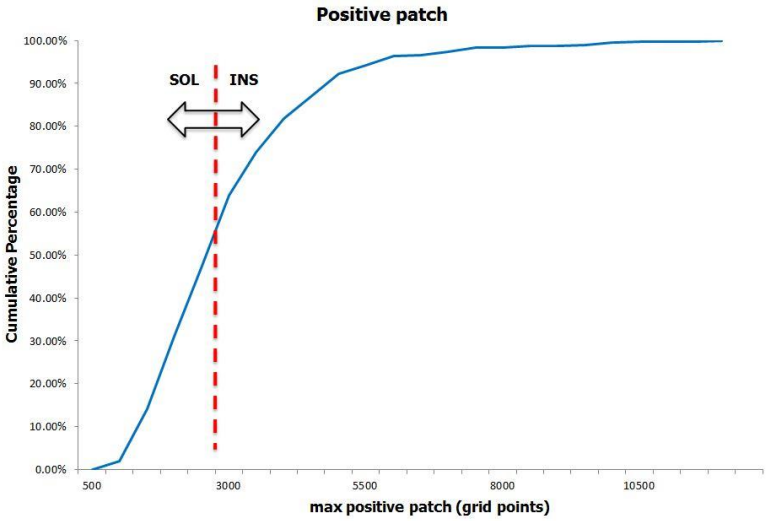
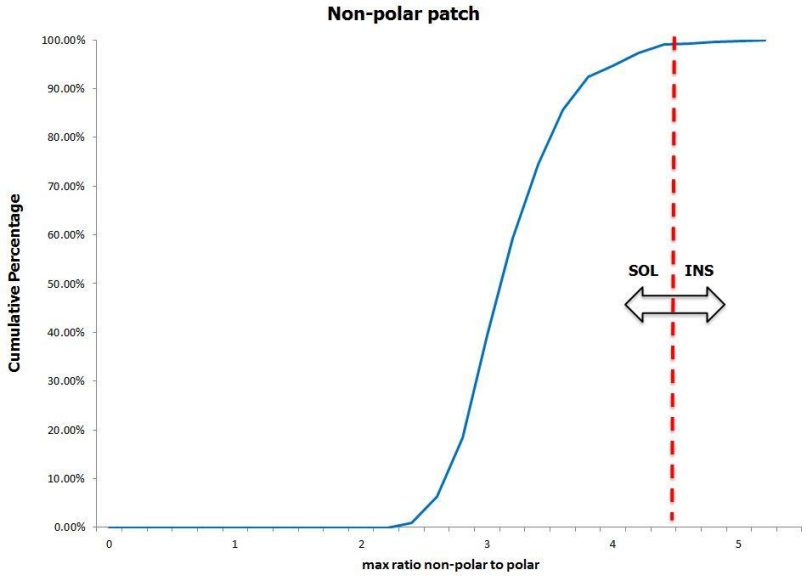


Figure 2.5. scFv dataset. Calculated features plotted as cumulative percentages for scFv's compared with the thresholds that best separate soluble and insoluble *E. coli* proteins in cell-free expression. (**A**) Maximum positive potential surface patch (threshold is 3000 points, values less than which are more soluble). (**B**) Maximum ratio of non-polar to polar SASA for a patch (threshold is 4.5, values less than which are more soluble. (**C**) Ratio of Lys to Arg content (KR-ratio) for each protein (threshold is 1.2, values above which are more soluble). The dotted red line represents an *E.coli*-based empirically deduced threshold.

A



B



C

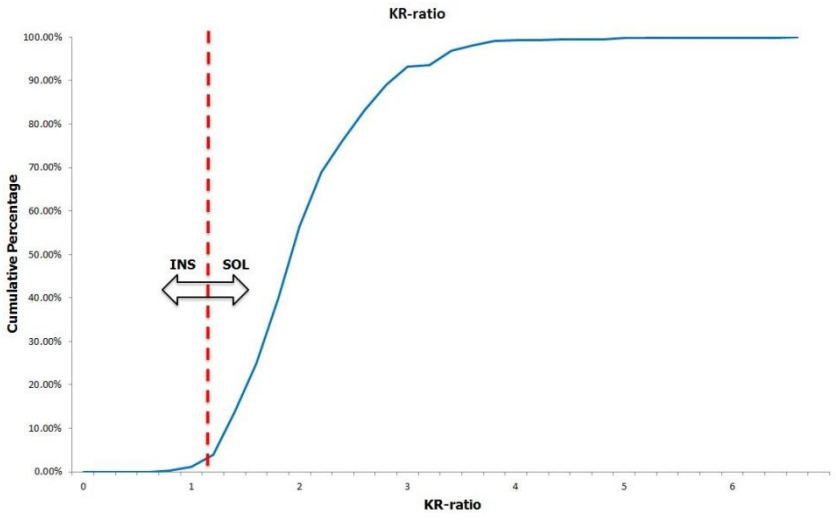
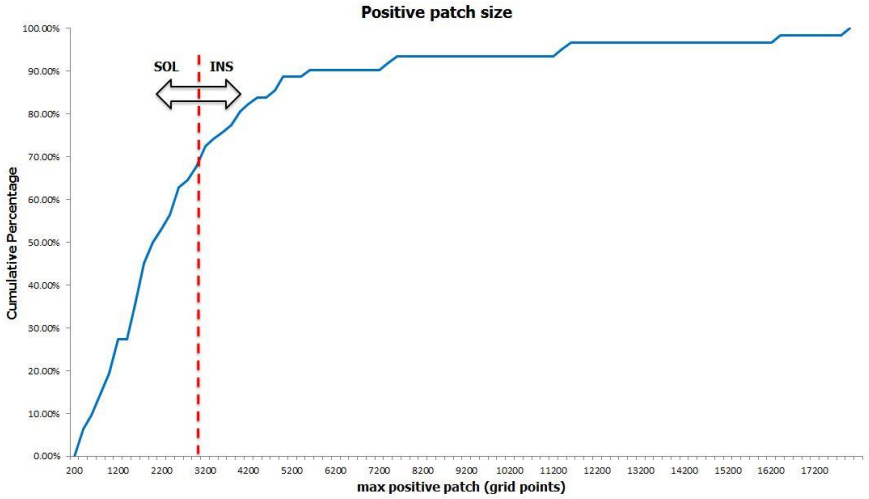
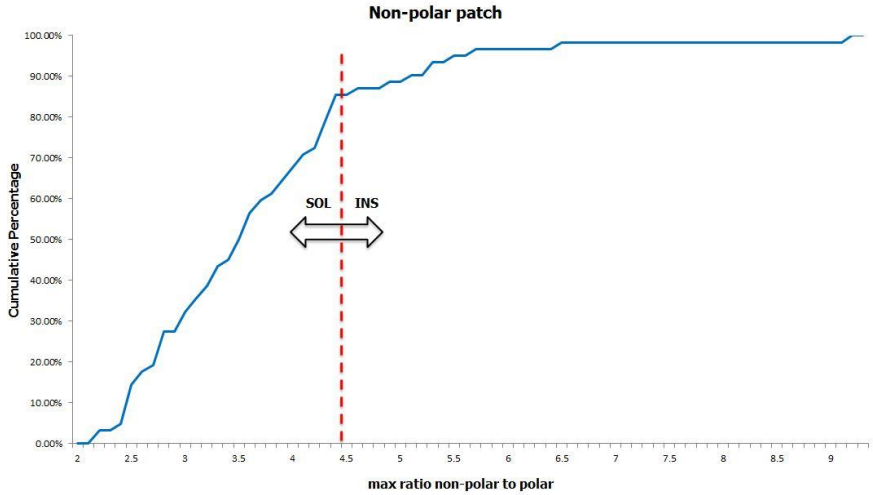


Figure 2.6. Fab Dataset. Calculated features plotted as cumulative percentages for the Fab fragment protein set (408 structures) compared with the thresholds that best separate soluble and insoluble *E. coli* proteins in cell-free expression. (**A**) Maximum positive potential surface patch (threshold is 3000 points, values less than which are more soluble). (**B**) Maximum ratio of non-polar to polar SASA for a patch (threshold is 4.5, values less than which are more soluble. (**C**) Ratio of Lys to Arg content (KR-ratio) for each protein (threshold is 1.2, values above which are more soluble). The dotted red line represents an *E.coli*-based empirically deduced threshold.

A



B



C

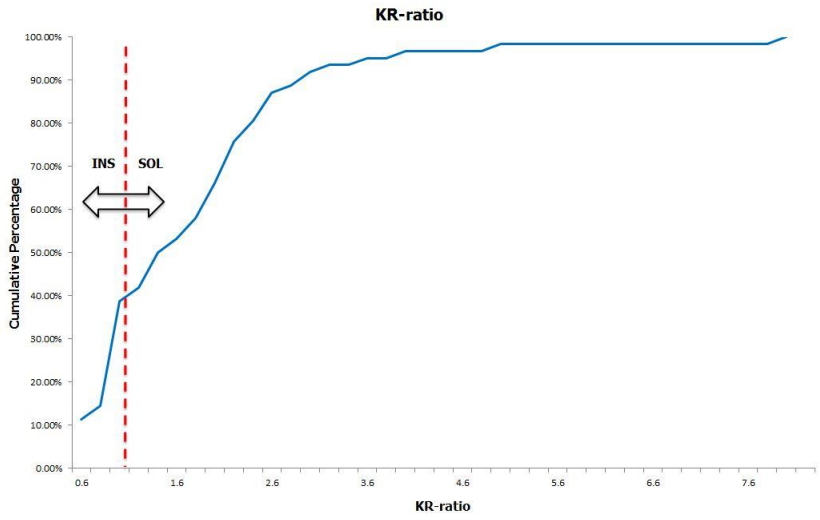


Figure 2.7. Biologics Dataset. Calculated features plotted as cumulative percentages for the biologic protein set (31 structures, 62 chains) compared with the thresholds that best separate soluble and insoluble *E. coli* proteins in cell-free expression. (**A**) Maximum positive potential surface patch (threshold is 3000 points, values less than which are more soluble). (**B**) Maximum ratio of non-polar to polar SASA for a patch (threshold is 4.5, values less than which are more soluble. (**C**) Ratio of Lys to Arg content (KR-ratio) for each protein (threshold is 1.2, values above which are more soluble). The dotted red line represents an *E.coli*-based empirically deduced threshold.

Table 2.2 summarises the separation of soluble and insoluble proteins in each of three datasets for each feature used to classify soluble and insoluble proteins. The proportion of soluble and insoluble (SOL/INS) proteins in each set is listed as separated according to the property used as a classifier.

Table 2.2 Soluble/Insoluble Separation of Therapeutic Datasets

Feature	Dataset (SOL/INS Proportion)		
	scFv fragments	**Fab fragments**	**Biologics**
Max positive patch size	0.74 / 0.26	0.64 / 0.36	0.67 / 0.33
Max non-polar to polar SASA ratio	0.96 / 0.04	0.99 / 0.01	0.87 / 0.13
KR-ratio	0.57 / 0.43	0.96 / 0.04	0.58 / 0.42

For the structure-based positive patch size, none of the three therapeutic datasets show a clear preference for either side of the threshold, although they do separate to a certain extent. All three datasets are slightly skewed toward the soluble side. It is not clear that large positively charged surface patches are directly relevant for the solubility of proteins (Fab fragments of antibodies) at high levels of circulatory concentration. Otherwise, evolutionary pressure would have pushed Fab fragments closer toward the region lower than the threshold in figure 2.6A. This argument arguably applies less for scFv's, as they are excised from Fab fragments and re-engineered (Demarest and Glaser, 2008). The case for non-polar SASA to polar SASA is vastly different, where the largest disparity in separation of soluble/insoluble subsets is observed. All three therapeutic protein datasets are located almost exclusively on the soluble side of the threshold. This is consistent with the role of non-polar patches in protein insolubility, whether in expression or at high concentration of secreted protein.

Interestingly, the sequence-based KR-ratio is largely above the solubility threshold for Fab fragments, but only slightly so in the case of scFv's and biologics. Again, this may indicate an evolutionary pressure on antibody sequences (which Fab fragments most closely resemble) to maintain a relatively high ratio of lysine to arginine content. Taking into consideration the evolutionary pressure that Fab fragments have undergone as components of antibodies, their sequence and structural properties will likely have been streamlined on much larger timescales than their synthetic counterparts. Similarly, at least some of the preference exhibited by biologics to tend to the soluble side of the threshold under all three features (although markedly less so for positively charged patches and KR-ratio compared to non-polar patches) is likely due to

evolutionary forces at work, as several of the included structures are either humanised antibodies or modified naturally occurring proteins (Appendix 2A).

2.7 KR-ratio and Solubility in Cell-Free Expression

For a sequence-based property, it is possible to extend calculations for the cell-free expression dataset beyond proteins that can be annotated with 3D structure to all proteins that can be cross-referenced. Solubility and KR-ratio for 2931 *E. coli* proteins correlate with R = 0.22 (Pearson correlation coefficient), $p > 1 \times 10^{-8}$. The least and most soluble groups are clearly separated ($p = 1.98 \times 10^{-35}$), where higher KR-ratio associates with higher solubility. This is illustrated in figure 2.8 below. Of the two systems, *i.e. E. coli* cell-free expression and Fab fragments (representing proteins that circulate at relatively high concentration), some properties may relate to solubility in both. This appears to be the case for KR-ratio and the maximum of the ratio of non-polar to polar surface area, whereas the maximum positive patch appears to be more relevant to cell-free expression environments.

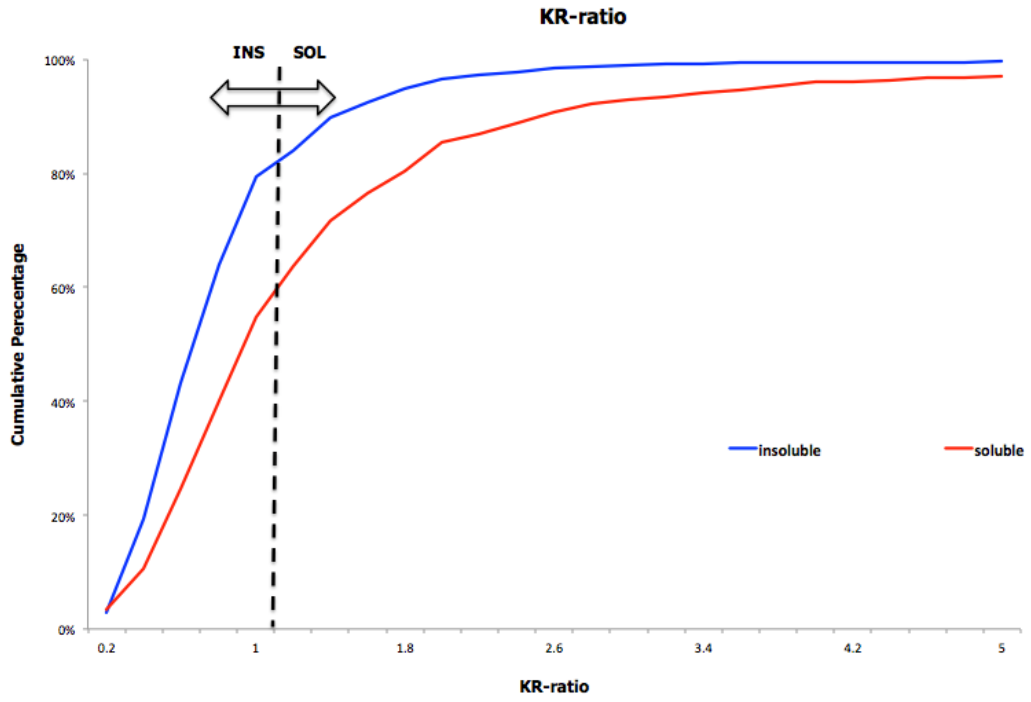


Figure 2.8. Separation of KR-ratio for soluble and insoluble subsets of proteins, from cell-free expression of *E. coli* proteins (Niwa *et al.*, 2009). Predicted soluble and insoluble regions are indicated relative to the threshold value of KR- ratio (1.2). The dotted black line is drawn at the value of the threshold.

2.8 Conclusions

The finding that large positively charged surface patches correlate with insolubility (Chan *et al*., 2013) in a cell-free expression system (Niwa *et al*., 2009) led to the question of whether these relationships were more ubiquitous. This was investigated with threshold values obtained from analysis of low and high solubility subsets of the cell-free expression data. The availability of structural representatives of antibody Fab fragments deposited in the PDB, in combination with the fact that they typically circulate at high concentration, presented a good case with which to examine protein solubility in a physiological environment. While separation based on non-polar surface patches appears to apply both for synthetic proteins (scFv's and certain biologics) and natural proteins (Fab fragments) (figures 2.5B – 2.7B), separation based on positive surface charge does not (figures 2.5A – 2.7A). Hence, surface non-polarity for therapeutic datasets conforms to the established mantra, appearing to be a general property that should be considered when designing natively structured proteins for high solubility.

Of particular interest in the current work is the emergence of KR-ratio as a potential novel feature that correlates to solubility in physiological environments. The interest is twofold, as a sequence-based property offers more scope for investigation against high-throughput protein expression data compared to a structure-based feature (where structural data is necessary), and furthermore offers a relatively simple solubility improvement tool in protein engineering efforts. If the correlation of KR-ratio with solubility is found to be more generic, similar to that of non-polar surface patches, then several members of the scFv set could be improved in this respect through experimental protocols such as rational mutagenesis.

KR-ratio has been compared with expected values for random KR assignment given the proteome lysine and arginine content, with equivalent analysis for DE-ratio (fraction of aspartic/glutamic acid sequence composition) in other proteins existing *in vivo* in high concentrations (Warwicker *et al*., 2014). In the case of serum albumins, which also have clinical applications (Mendez *et al*., 2005) and myoglobins, the reported findings showed DE-ratios being in line with the expected distributions across species whereas KR-ratios were almost uniformly high, strengthening the results ascertained from the therapeutic datasets described above. Furthermore, comparison of KR-ratios in serum albumin and myoglobin proteins with those in paralogues existing at lower concentrations (neuroglobin and cytoglobin) demonstrated a marked difference in which high concentration proteins possess higher KR-ratios (Warwicker *et al*., 2014). A refined model of the relationship between KR-ratio and protein solubility can be obtained by identifying proteins that are soluble at high concentration but have low KR-ratio. One notable example is that of γ-crystallins, the subfamily of eye lens proteins, which have KR-ratios as low as 0.0048 for

human γD-crystallin. Crystallins maintain solubility at high concentration and without protein turnover, and have been extensively studied (Bloemendal *et al*., 2004). An interesting observation is that, in the human γD-crystallin crystal structure, arginine side chains are involved in extensive charge networks with acidic side chains (Basak *et al.*, 2003), presumably reinforcing folded state stability. It may be the case that selection for arginine over lysine increases intramolecular interactions, and reduces arginine influence on intermolecular interactions through sequestering the side chains into charge networks. The case of crystallins as a counterexample to therapeutics and other high concentration proteins with regards to KR-ratio demonstrates that there is more to understand about the relative contributions of ionisable side chains to stability and solubility.

For the charge-based properties presented in this work, it is important to consider confounding factors relevant to electrostatic interactions. The correlation reported between maximal positively charged surface patches and solubility in cell-free expression (Chan *et al*., 2013) does not account for how different counterions and their binding to proteins could influence solubility. Calculations of positive patches were performed using a simple model for 0.15 M monovalent counterions, and do not incorporate specific ion binding or varying ionic strengths. The value of 0.15 M was chosen as this resembles physiological ionic strength, and therapeutic formulations generally aim to maintain protein-based products at conditions similar to those encountered upon administration to patients. Clearly, this remains an area to develop computationally, and could be particularly important for cases such as biologics where formulation and solution conditions are highly variable.

Perhaps one of the most important considerations of such computational models would be incorporate the effects of modulating the ionic strength of the underlying aqueous environment. In general terms, a higher ionic strength is associated with a decrease in the screening length, *i.e*. the quantity measuring the length over which charged species can "sense" electrostatic attractions or repulsions (Roberts, 2014). Hence, interactions between like-charged molecules attenuate in this case. Conversely, at low ionic strengths the screening length increases and repulsion between proteins is stronger. Although beyond the scope of this thesis, a model that could dynamically adjust PBE-based calculations of electrostatic potential to varying levels of parameters such as ionic strength and pH would be highly valuable as a pre-screening tool in the design of protein-based therapeutics.

In a more bioinformatics-based context, a study investigating correlations of sequence properties with solubility has shown that the net charge of myoglobin has evolved to higher values in animals with greater diving ability, suggesting that prevention of aggregation at high myoglobin concentrations underlies this observation (Mirceta *et al*., 2013). This finding offers an intriguing parallel to the current work concerned with therapeutic proteins, as well as to further work

demonstrating that myoglobins have statistically very high KR-ratios (Warwicker *et al*., 2014), although no investigation of charged side chains such as arginine and lysine was reported.

The current analysis of therapeutic proteins in addition to other high concentration protein families (Warwicker *et al*., 2014) suggest that, when assessed over large datasets, KR-ratio may comprise a previously uncharacterised correlate of physiological protein concentration and solubility. Differences in arginine and lysine side chain interactions have been reported. Specifically, arginine is known to be over-represented in functional protein-protein and protein-nucleic acid interfaces (Jones *et al*., 2000). A preference of arginine over lysine for cation-pi interactions with aromatic groups has been observed (Martis *et al*., 2008; Shah *et al*., 2012). Furthermore, it has been observed that lysine to arginine mutations can increase crystallisation propensity, thought to be on account of differential conformational mobility (Czepas *et al*., 2004). Furthermore, arginine is commonly used as a component of additive solutions that stabilise against protein aggregation (Golovanov *et al*., 2004). It will be important to establish the mechanistic basis by which argninine can be used as a stabilizing additive in protein formulations (Arakawa *et al*., 2007; Shukla and Trout, 2011). In this sense, it is interesting that lysine is enriched in high-abundance human proteins and high-solubility therapeutics, as the arginine side chain has a more complex structure enabling a broader range of interactions such as cation-π and strong salt bridges (White et al., 2013). The delocalisation of the charge in the guanidinium side chain group may render arginine more versatile in chemical interactions, and thus more capable of contributing to non-specific protein-protein interactions, some of which may induce growth of monomeric aggregate precursors. There is currently not a clear mechanistic basis for lysine to be more effective at inhibiting loss of solubility than arginine, although the results presented here suggest that there could be evolutionary pressures at play.

Several studies have reported that supercharging proteins, either positively or negatively, offers a tool to improve solubility. It has been reported that prevention of aggregation from partially unfolded states contributes to the effect of supercharging (Der *et al*., 2013). The current work, which comprises a part of the study published by Warwicker and colleagues (2014), suggests that in terms of positive charge, lysine should be more effective than arginine in promoting solubility via supercharging. Comprehensive comparisons of charge roles in solubility should be tractable by experimentally swapping arginine to side chains to lysine. A more formally established correlation between KR-ratio and solubility could have implications for protein expression in biotechnology, and for modulation of therapeutic proteins. An experimentally confirmed, ubiquitous correlation between KR-ratio and solubility could offer a relatively simple method for increasing solubility in therapeutic formulation, where it could be implemented using some mutagenesis-based protocol (*e.g*. substituting a subset of non-essential arginine residues with

lysine). KR-ratio is a sequence-based property, without reference to the 3D conformation of amino acids in tertiary structure. It may therefore parallel the study of amyloidogenic sequence regions, and their suggested role in promoting aggregation from partially unfolded protein states.

In summary, protein solubility is a key property for biotechnological and biopharmaceutical applications, where a protein may be required to be soluble when removed from its native, physiological environment. Certain properties associated with optimal solubility such as absence of non-polar interactions and the importance of charged groups and net charge are well established. Information is emerging regarding the roles of positive and negative charge. In the current work, the preference of lysine over arginine in terms of an aggregation inhibiting mechanism in biotherapeutics was investigated, and possibly as a sequence-based mechanism encoding high solubility in other protein families (Warwicker *et al*., 2014). The statistically significant enrichment of lysine over arginine in high-concentration serum albumins and myoglobins establishes that KR-ratio is a proteome-wide feature not limited to small-scale therapeutic datasets (figures 2.5 – 2.7) or cell-free *E.coli* data (figure 2.8). The underlying principles for lysine preference can investigated using mechanistic studies, and high-throughput "-omics" studies on eukaryotic proteomes measuring inherent aggregation propensities should offer insight into how widespread this correlation is in the proteomes of higher organisms. The main conclusion from this body of work is that scope exists for simple adjustment of lysine and arginine content to enhance protein solubility, although how generic this feature is in determining solubility compared to well-established ones such as net charge and non-polarity remains unclear. Experimental studies in this direction will establish whether the correlation is general, and if so, the underlying mechanistic basis of this phenomenon.

Appendix 2A. Biologics Dataset

PDB	Chains	Description	Commercial Name	Concentration	Administration Mode
2OSL	HL-AB	IgG-Fab	Rituximab	375 mg/m2*	Solution
4G3Y	HL	IgG-Fab-fragment	Infliximab	5 mg/kg	Lyophilized Powder
1AU1	A-B	Interferon-beta-1	Avonex-Rebif	30 µg/kg per week	Solution
2R7E	AB	Factor-VIII	Octocog-alfa		
1YY8	AB-CD	IgG-Fab	Cetuximab	400 mg/m2	Solution
1W7X	HL	Factor-VIIa	Eptacog-alfa		
4GBC	AB-CD	Insulin-modified	Insulin-aspart		
3V0A	A	Botulinum-toxin-A	Ona-botoxA		
1EER	A	Erythropoietin	EPO-R		
4F1D	AB-CD	Insulin	Humulin		
2OSL	HL-AB	IgG1-Fab	MabThera	375 mg/m2*	Solution
3IU3	HL-ABCD	IgG1-Fab	Basiliximab		
3NFS	HL	IgG1-Fab	Daclizumab		
3EO9	HL	IgG1-Fab	Efalizumab		
2XA8	HL	IgG1-Fab	Omalizumab	30 mg/mL	Lyophilized Power
4G5Z	HL	IgG1-Fab	Canakinumab		
3HMW	HL	IgG1-Fab	Ustekinamab	90 mg/mL	Solution
3GIZ	HL	IgG1-Fab	Ofatumumab	20 mg/mL	Solution
1A22	A	HGH	HGH		
3LC3	AB	Factor-IX	Factor-IX		
1AUT	CL	Protein-C	Protein-C		
1QLP	A	anti-trypsin-inhibitor	a1-anti-trypsin		
3CXE	B	G-CSF	Filgrastim	20 drops/day**	Solution
1XWD	AB	FSH	FSH		
1HRP	AB	Chorionic-gonadotropin	HCG		
1M47	A	Interleukin-2	IL-2		
3SE3	B	Interferon-alpha2beta	IFN-a2b		
1FG9	A-B	Interferon-gamma	IFN-g		
3BMP	A	BMP-2	Dibotermin-a		
1M4U	L	BMP-7	BMP-7		
3MJG	A	PDGF	PDGF		
3KE0	A	Glucocerebrosidase	hydrolase		
4JXP	A	alpha-L-iduronidase	Laronidase		
1FSU	A	4-sulfatase-(ARSB)	Galsulfase		
1R46	A-B	human-galactosidase-A	Agalsidase-beta		
3IAR	A	Adenosine-deaminase	A-deaminase		
1BDA	A	tPA-catalytic-domain	tPA-protease		

1TRN	A	human-trypsin	Trypsin
1F31	A	Botulinum-toxin-typeB	Bot-Toxin-B
4AWN	A	human-DNase-I	DNase-I
2PE4	A	Hyaluronidase-1	Hyaluronidase
9PAP	A	Papain	Papain
4GDT	A	L-Asparaginase	L-Asparaginase
1BML	C	Streptokinase	Streptokinase
2B4X	I-L	Antithrombin-III	AntithrombinIII

* Administered dosage can vary throughout different stages of therapy
** Non-quantitative unit

Chapter 3. Sequence-based and Structure-based Solubility Prediction in Proteins

---

Manuscript

Protein solubility correlates with protein abundance, and lysine/arginine content is a major factor[2]

ABSTRACT: Protein solubility is an important property that spans multiple areas of biotechnology, from over-expression for synthetic biology to maintenance of stable solutions of biotherapeutics at high concentration. It has previously been found that lysine is preferred to arginine in proteins of higher solubility (Warwicker *et al*., 2014). The current work shows that lysine to arginine content remains a key factor in distinguishing high and low solubility sets of proteins, when a range of other features are included. Since the availability of high-throughput data for protein solubility is low, additional analysis is made of protein abundance from quantitative mass spectrometry data. It is found that determinants of higher protein solubility correlate with features that associate with higher cytoplasmic abundance. Protein solubility therefore appears to be a general property that is selected for to different levels. Thus, many proteins of biotechnological interest and for which soluble expression at high levels is required will be problematic since they have evolved to 'fit' a lower abundance in the cell. Expression engineering of various types becomes important in these cases. With regards to amino sequence alterations, the current work reinforces the potential for arginine substitution with lysine as a possible contributory factor to improve solubility.

---

[2]Charonis, S., Curtis, R.A, Warwicker, J. (*in writing*).

3.1 Objectives

This chapter focuses on extending sequence- and structure-based analysis to determine features that separate high- and low-solubility proteins. A significantly more diverse range of both sequence- and structure- based features is considered here, and protein datasets analysed span several organisms, both prokaryotic and eukaryotic. In contrast to the previous chapter that used empirically derived thresholds for features that separated *E. coli* proteins well, this chapter will examine a broad range of properties relating to charge, folding, β-strand propensity and amino acid composition. Both sequence-level and 3D structure-based features are investigated in terms of their predictive capacity, *i.e*. how well they can classify proteins as high-/low-abundance based on data from proteomic studies. The main objectives behind this work can be summarised as follows: (i) to compare KR-ratio to other charge-based and generic sequence properties and (ii) to compare positive surface charge and non-polarity to other charge-based and non-electrostatic structural properties. Thus the three features focused on in chapter 2 (KR-ratio, surface charge/non-polarity) are extended significantly in both the sequence and 3D structural domains to determine how well they extend to larger-scale proteomic data and how they compare to other similar features in terms of solubility/abundance prediction. The enrichment of selected features in abundant/soluble and non-abundant/insoluble proteins are compared across proteomes spanning several organisms using *z*-scores (sequence-based features) and Pearson correlation coefficients (structure-based features) and are illustrated using heatmaps. A qualitative study of sequence-based features enriched in soluble/insoluble and high-/low-expression proteins obtained from literature is presented first (section 3.4.1, figures 3.1 – 3.4). Quantitative analyses are subsequently presented for both publication-based datasets as well as data from repositories of protein abundance measurements covering a range of proteomes from both prokaryotic and eukaryotic organisms.

3.2 Protein Solubility Prediction

Protein solubility and aggregation are important properties and have been extensively investigated and were discussed in previous chapters. They are becoming increasingly important in the growing areas of bioprocessing and biopharmaceuticals. These issues are exemplified by biologics, where dosing schedule and drug delivery are limited by the ability to maintain stable, high concentrations of the therapeutics, typically up to 100 – 150 mg/mL (Kayser *et al*., 2011). For proteins that have evolved to maintain high concentrations *in vivo*, such as circulating antibodies,

this may not be a problem, although it is common for changes in the complementarity determining regions (CDRs) of antibodies to cause large alterations in solubility (Ducancel and Muller, 2012).

Despite extensive efforts made in this direction (Niwa *et al*., 2009), there is no clear consensus model for predicting protein solubility. This is presumably due in part to the complexity of the processes involved, but also to the lack of large-scale benchmark datasets in the public domain (Obrezanova *et al*., 2015). When considering relevant properties, it is appropriate to include those based on protein sequence alone, and those derived from 3D structure where available. In addition, there is coupling between properties that influence protein association in the folded form, relating to colloidal stability and those that relate to stability of the folded form (conformational stability). In the Lumry-Eyring model for nucleated protein polymerisation, irreversible aggregation is associated with partial, or full, unfolding and subsequent aggregate growth (Li and Roberts, 2009). Predictive studies of aggregation have included sequence and structure-based properties, from aggregation that occurs close to the pI to the role of β-strands in aggregation-prone regions. More recently, prediction tools have focused on a few prominent features, particularly net charge (Kramer *et al*., 2012) and non-polar surface patches (Voynov *et al*., 2009).

A major obstacle to the progress of predictive methods is the lack of benchmark datasets for validation of the underlying algorithm. Acquisition of solubility and aggregation propensity data is generally labour intensive and is often associated with issues of commercial sensitivity, *e.g*. pharmaceutical companies being reticent to share in-house platforms and other proprietary technologies. Use of data generated by structural genomics and high-throughput proteomics studies can be useful, but may be too broad to focus on properties that are effective in separating proteins of high and low solubility. Despite the scarcity of large-scale solubility and abundance data obtained from proteomics technology platforms, high throughput screens of relevant data are becoming available. A study of solubility in cell-free expression for over 3000 *E. coli* proteins demonstrated a bimodal distribution of solubilites, and an indication that negative charge associates, on average, with soluble proteins (Niwa *et al*., 2009). This dataset was used in the previous chapter to show additionally that therapeutic proteins of high solubility are enriched in lysine relative to arginine (Warwicker *et al*., 2014). This leads to the question of what other properties, either sequence-based or structure-based, might be enriched in a set of soluble proteins, and how they compare with features used in existing prediction methods.

3.3 Protein Abundance and Aggregation Propensity

Protein solubility is an important property that spans multiple areas of biotechnology, from overexpression for synthetic biology applications to stabilisation of biotherapeutic solutions at high concentration. It has previously been found that lysine is preferred to arginine in proteins of higher solubility (Chapter 2; Warwicker *et al*., 2014). As discussed in chapter 2, the mechanistic basis of solubility modulation by KR-ratio is still unclear. At the molecular level, the ability of the arginine guanidinium group to form multiple strong interactions (Sokalingam *et al*., 2012) and strong salt bridges (White *et al*., 2013) is thought to underpin a role in increasing structural stability and the protein-protein interactions that in turn reduce solubility. A similar distinction is seen in the enrichment of arginine at interacting sites in proteins that bind glycosaminoglycans (Hileman, 1998). Furthermore, a study of nucleic and amino acid sequences across many species found that lysine and arginine abundance is a trade-off between higher protein thermal stability (*e.g*. higher arginine content in extremophiles) and higher lysine content in the absence of a requirement for increased folding stability (Goncearenco *et al*., 2014).

The caveat of largely lacking high-throughput data measuring protein solubility in a manner similar to the *E. coli* data used in chapter 2 is compensated by using related but distinct physical quantities that are more commonly measured in "-omics" studies. Often these quantities include protein abundance and expression levels. Although such quantities are proxies for estimating solubility, underlying correlations have been explored with important findings. It has been shown that the *in vitro* aggregation rates of several human proteins are not comparable with their gene expression levels *in vivo*, as estimated from measurements of cellular mRNA concentrations (Tartaglia *et al*., 2007). Computational analysis of predicted aggregation propensities of large sets of human (Tartaglia and Vendruscolo, 2009) and bacterial (Tartaglia *et al*., 2009; Castillo *et al*., 2011) proteins from their primary sequences indicates an inverse correlation between theoretical aggregation propensity values and experimentally determined cellular concentrations of the corresponding mRNAs, suggesting that gene expression levels may be linked to the solubility of the encoded protein. In this respect, although protein expression is regulated both temporally and spatially, most proteins have an intrinsic range of functionally effective abundance levels (Wang *et al*., 2011). The range of abundance levels is extremely diverse, measuring from a few molecules per cell for signalling proteins to millions of molecules for structural proteins. If indeed evolutionary forces have provided a mechanism for regulating gene expression and cellular abundance based on the solubility of encoded gene product, it is likely that this will be reflected to a certain extent in the protein sequences.

3.4 Quantitative Proteomics Studies

In the previous chapter, three protein-based datasets (scFv's, Fab fragments, and biologics) were used to assess the ability of three properties (surface charge/polarity and KR-ratio) to discriminate soluble from insoluble proteins. The focus here is on investigating sequence- and structure-based relationships that can be used to predict solubility/abundance. As discussed, large-scale quantitative proteomics studies directly measuring solubility are considerably limited, necessitating the use of proxies such as protein abundance or mRNA expression data. Nonetheless, certain studies have been published with the underlying data made publicly available. Such studies usually make use of protein sequence data, and their underlying datasets comprise gene identifiers and their corresponding sequences. Using data from relevant studies where available, a range of sequence-based properties was investigated for separating proteins based on solubility or expression levels. Qualitative analyses were performed for studies on human, bacterial (*E. coli*), yeast (*S. cerevisiae*) and fungi (*A. niger*) proteomes.

3.4.1 Qualitative Comparison of Sequence-based Features

A simple median-based quantity was used to compare datasets in terms of sequence composition with the aim of visualising disparities between high-/low- solubility and expression subsets of proteins. Protein sequences for each dataset were obtained from the supplementary material section of relevant publications. A Perl pipeline was used to calculate sequence composition statistics (code contributed by Jim Warwicker). The method for comparing sequence-based properties uses a normalised median statistic that is outlined below. The median was used as a measure of central tendency of the distribution of sequence-based features as it is less sensitive to outliers than the mean. No significance tests are performed on these results, as the aim was simply to observe any patterns between features that stand out when comparing solubility and expression levels from different proteomes. Furthermore, lysine and aspartic acid have been validated as being statistically enriched over arginine and glutamic acid (KR-ratio and DE-ratio respectively) in *E. coli*, yeast and human proteomes (Warwicker *et al*., 2014) as discussed in chapter 2. For each protein dataset plotted in figures 3.1 – 3.4, a set of features including KR-ratio (fraction of lysine/arginine), DE-ratio (fraction of aspartic/glutamic acid), protein length (*n*umber of *a*mino *a*cids *naa*), and the percentage composition of each standard amino acid (single-letter abbreviation) is used. A normalised median is calculated as shown in equations 3.1:

(3.1)

(i) For a sequence feature *i* (*e.g*. KR-ratio), the unweighted average of the mean of the two protein subsets (soluble and insoluble subsets) is calculated as:

$$\bar{\iota} = \frac{\boldsymbol{\mu_{i(SOL)}} + \boldsymbol{\mu_{i(INS)}}}{\mathbf{2}}$$

(ii) Subsequently, a normalised median is calculated by dividing the median value of feature *i* by the unweighted average of means as determined in step (i):

$$\widetilde{\boldsymbol{m}}_{\boldsymbol{i\_norm}} = \frac{\widetilde{\boldsymbol{m}_{\boldsymbol{\iota}}}}{\bar{\iota}}$$

In equations 3.1, the mean (μ_i) and median ($\widetilde{m_i}$) of a sequence feature *i* refer to statistics taken over all data points (protein sequences in this context) of each subset (high-/low-solubility or abundance) of proteins. Normalised median values are plotted for each sequence-based feature so that disparities between subsets can be readily observed (figures 3.1 – 3.4). The rationale behind this was to establish in an informal (non-quantitative) manner whether certain features produce better separation than others (*i.e*. have wider gaps between the red and blue lines in each plot) and if there was any recurrence in the features that do so. Although statistical significance is not tested for here, features that produce good separation across multiple proteomes of different organisms would be interesting, as they would hint towards being generally applicable to a certain extent for separating soluble/insoluble proteins. In figures 3.1 – 3.4, the statistic used (*y*-axis) has been normalised around the value of 1 meaning that the absolute values of each sequence feature (*x*-axis) have been shifted.

There have been several high-throughput "-omics" initiatives to carry out large-scale, proteome-wide studies of protein solubility and proxies such as abundance and expression levels. Such studies often have different scopes with the latter type (cellular abundance of proteins) not always being directly interested in the underlying solubility trends of the proteome. The datasets that are used in this chapter to establish sequence- and structure-based features that can be used in predictive models can be loosely divided into two categories: (i) those measuring native aggregation propensities to determine aggregation-related quantities such as solubility and (ii) those measuring mRNA expression and protein abundance levels. The datasets for which the previously described qualitative comparison was carried out are outlined below.

***Escherichia coli* Dataset (Niwa *et al*., 2009)**

This study discusses one of the most comprehensive large-scale solubility datasets to date and has been mentioned extensively in the previous chapter. The authors employ a cell-free *E. coli* system to express all known ORF proteins and quantify their solubility levels. Because of the chaperone-free cellular environment that was employed, the intrinsic aggregation propensities of thousands of proteins were measured in a translation-coupled manner. Approximately 70% of *E. coli* open reading frames were successfully quantified using PCR (3173 translated proteins). Solubility was quantified using autoradiography-measured band intensities. A histogram on individual solubilities, based on data from the translated proteins, revealed a bimodal distribution. The bimodality of the solubility distribution indicates that intrinsic aggregation propensities are not evenly distributed across a continuum and that cytoplasmic proteins can be categorised into an aggregation-prone (insoluble) group and a highly soluble one.

Interestingly, subtraction of integral membrane proteins (IMPs) did not alter the bimodality of the solubility distribution. Furthermore, the authors found that protein solubility is not correlated with the rates of conversion between unfolded and aggregated states of proteins. The factors found to correlate to some degree with solubility included protein charge and structural class. Figure 3.1 below illustrates the results of the above statistical analysis protocol for the *E. coli* dataset. The authors point out that a significant caveat regarding their solubility data is the complete dependence of observed aggregation rates on the centrifugation method used. Other conditions such as a higher-speed centrifugation might have revealed a histogram with a different shape, and there is a possibility that soluble fractions might include oligomers that are aggregation precursors. A median-based comparison of sequence feature enrichment in soluble and insoluble subsets is illustrated below in figure 3.1.

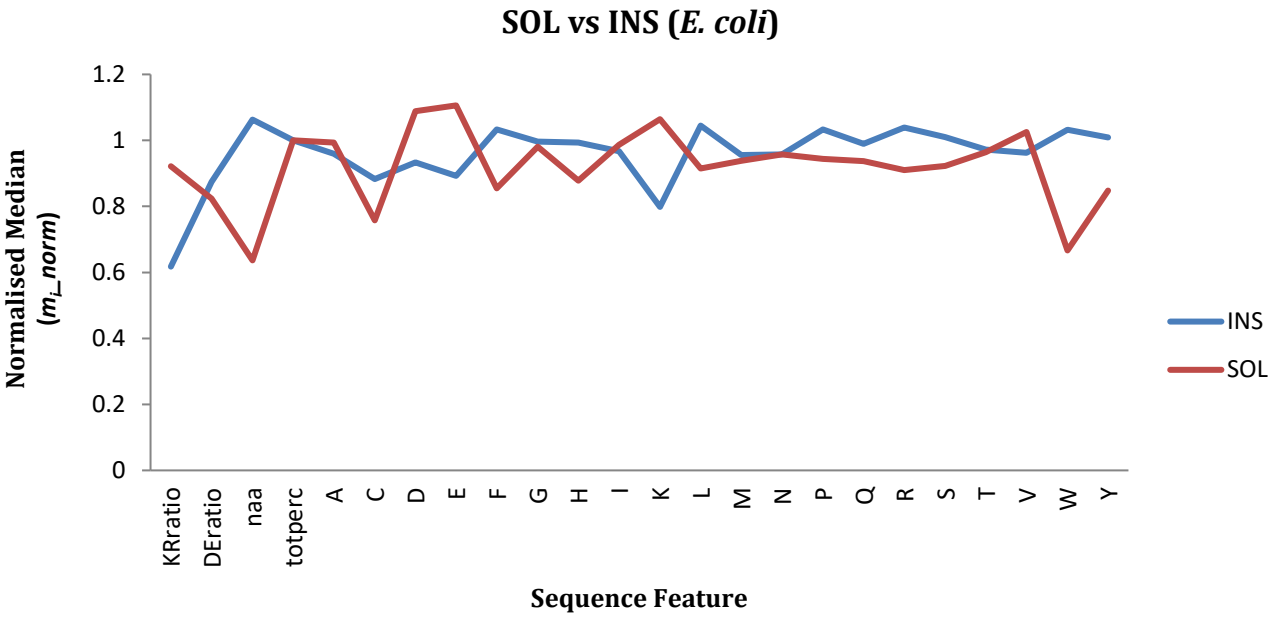


Figure 3.1. Separation of soluble and insoluble *E. coli* proteins based on sequence features. Separation of soluble (SOL) and insoluble (INS) subsets of cell-free *E. coli* expression (Niwa *et al*., 2009) using a normalised median value for each sequence property.

SOLP Dataset (Magnan *et al*., 2009)

The SOLP dataset is a large, non-redundant set of proteins expressed in *E. coli* that combines data from the PDB, SwissProt (The UniProt Consortium, 2007) and TargetDB (Chen *et al*., 2004) databases. These databases were merged with the protein dataset used in Idicula-Thomas and Balaji (2005) with a rigorous threshold (25% sequence similarity) being applied to reduce the redundancy of the sequences. To discriminate sequences between soluble and insoluble, annotations in the primary datasets are used to filter out those that are labelled as employing an *E. coli* expression system. The dataset is further refined by computational means to filter out sequences (i) belonging to membrane proteins, (ii) having several unknown amino acids, or (iii) that are extremely short (<10) or extremely long (>10000). SOLP is a very large dataset (17408 proteins), but it does not apply sequence-based or structural criteria to assess solubility. Rather, it relies on annotations of the primary datasets from which it aggregates protein sequences. A median-based comparison of sequence feature enrichment in soluble and insoluble subsets is illustrated below in figure 3.1.

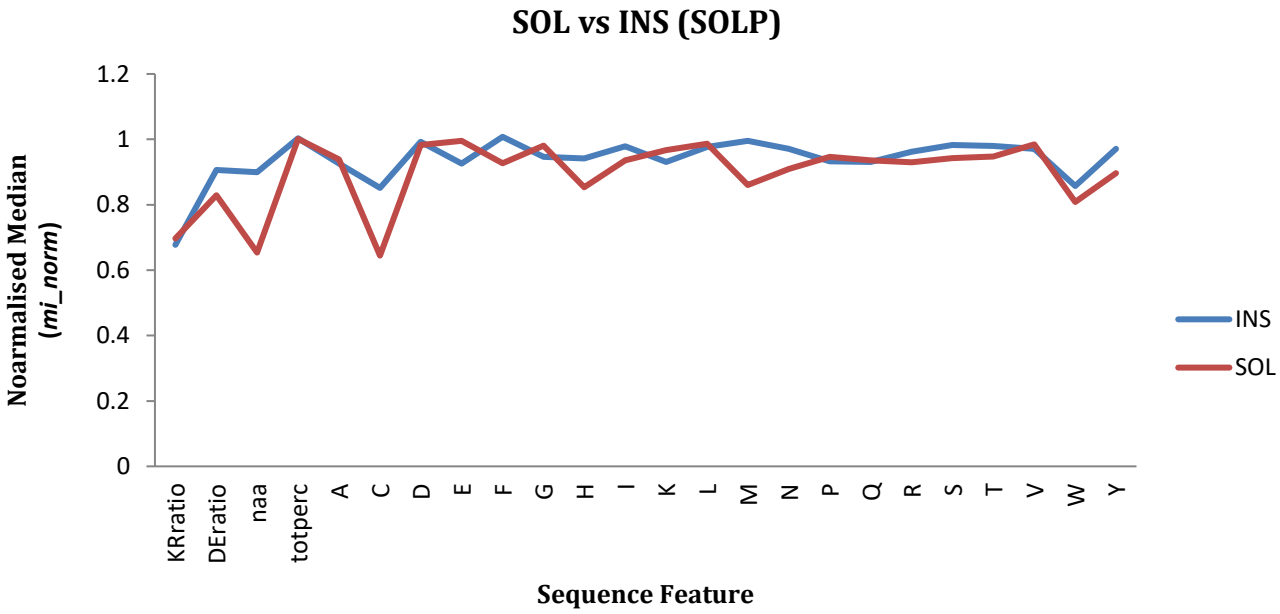


Figure 3.2. Separation of soluble and insoluble protein subsets based on sequence features for SOLP dataset. Separation of soluble (SOL) and insoluble (INS) subsets of SOLP (Mangan *et al*., 2009) using a normalised median value for each sequence property.

Saccharomyces cerevisae Dataset (Ghaemmaghami *et al*., 2003)

This study features a yeast protein expression dataset in which *S. cerevisiae* open reading frames were tagged with a high-affinity epitope and expressed from their natural chromosomal location. Subsequently, protein abundances were measured during log-phase growth by immunodetection of the tag and the yeast proteins were split into high-level and low-level expression. A total of 3853 out of 6234 tagged open reading frames were successfully quantified (expression level). This dataset was extracted and the proteins were ranked by expression level. Subsequently, the top and bottom 1000 proteins were used to create high-expression and low-expression subsets. A median-based comparison of sequence feature enrichment in high- and low-expression subsets is illustrated below in figure 3.1.

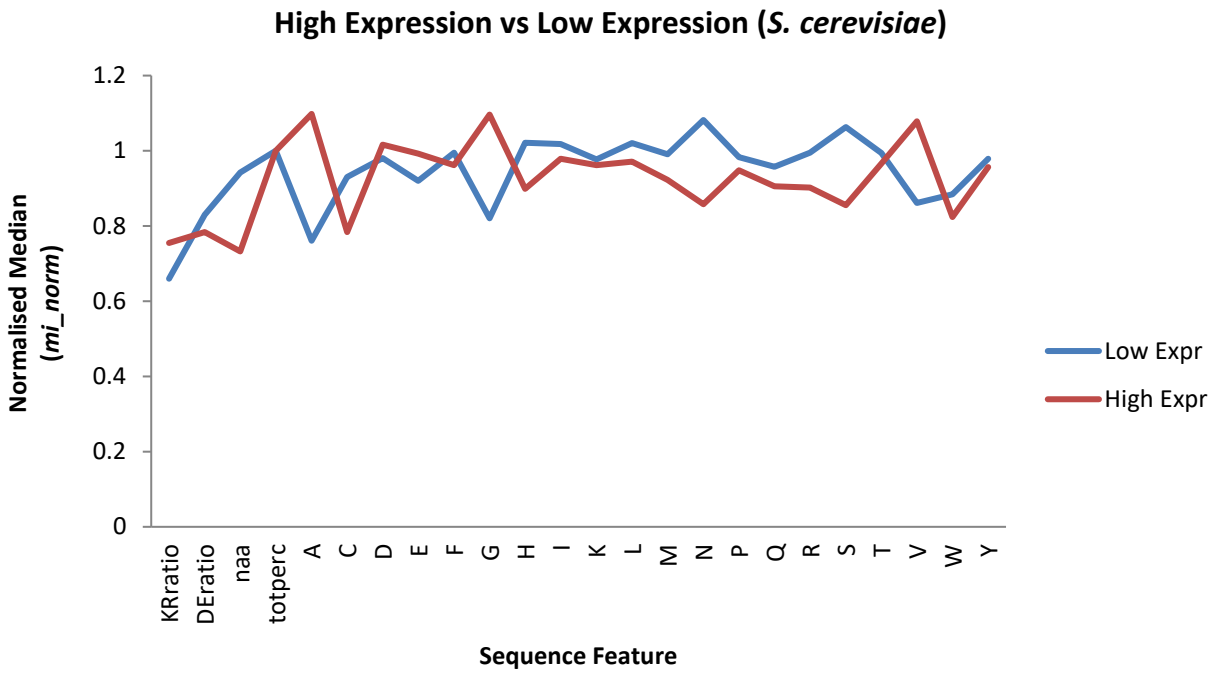


Figure 3.3. Separation of soluble and insoluble *S. cerevisiae* proteins based on sequence features. Separation of high expression and low expression subsets of yeast proteins (Ghaemmaghami *et al*., 2003) using a normalised median value for each sequence property.

Aspergillus niger Dataset (van den Berg *et al*., 2012)

In this study, a library of over 600 homologous and nearly 2000 heterologous fungal genes was constructed and overexpressed in *A. niger* using a standardised expression cassette and scored for high versus zero production. Machine learning techniques were subsequently applied for identifying sequence-based predictors of expression. The amino acid composition of each protein was reported to be highly predictive of expression levels and for both homologous and heterologous genes, the same features were important.

Two protein datasets were tested for homologous and heterologous gene expression. After removing redundant sequences and using cluster analysis, the final dataset consisted of 345 secretory proteins that were overexpressed in *A. niger* and tested for detectable extracellular concentrations by placing the obtained extracellular medium on a gel after growing the culture in shake flask. Proteins for which a band on the gel was observed were labelled as the high-production subset (167 proteins) whereas those with no observed bands were labelled as low-production subset (178). A median-based comparison of sequence feature enrichment in high- and low-production subsets is illustrated below in figure 3.1.

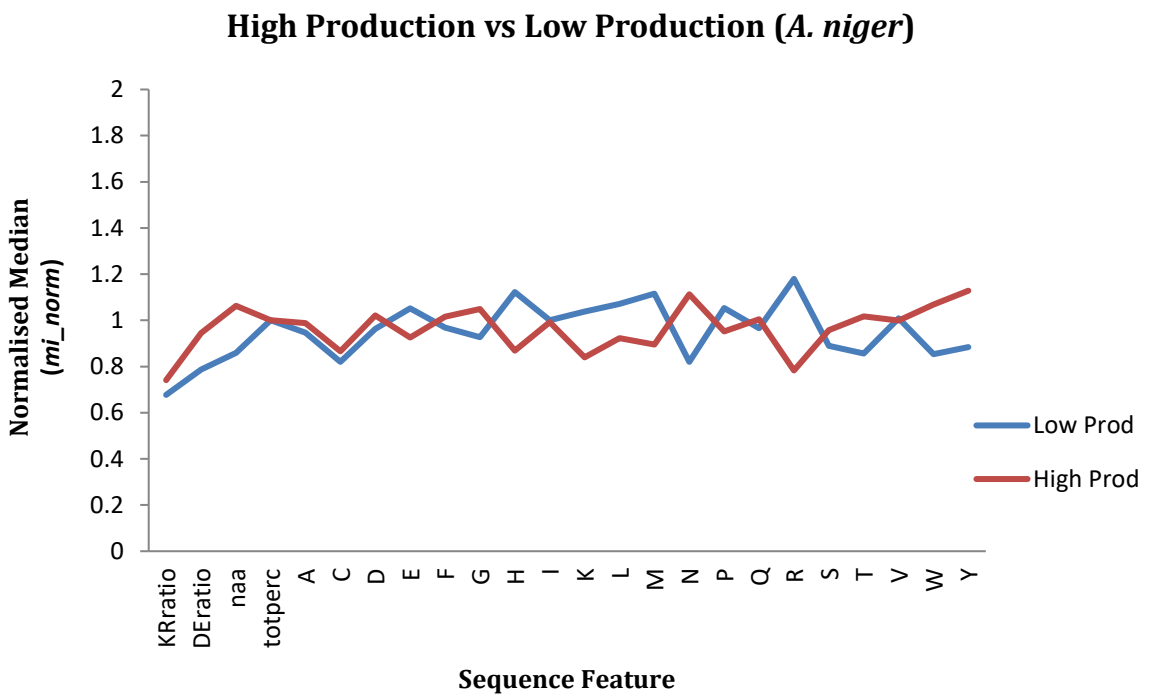


Figure 3.4. Separation of soluble and insoluble *A. niger* proteins based on sequence features. Separation of high production and low production subsets of secretion proteins (van Den Berg *et al*., 2012) using a normalised median value for each sequence property.

Figure 3.1 illustrates that as expected, KR-ratio is lower for insoluble proteins than soluble ones in *E. coli*. Interestingly, protein length (*naa*, *n*umber of *a*mino *a*cids) shows an even better separation of the soluble and insoluble subsets. A similar separation is observed in tryptophan (W) enrichment for insoluble *E. coli* proteins. Lesser, albeit notable differences, are observed in the fraction composition of certain amino acids: aspartic and glutamic acid (both enriched in soluble proteins) and lysine (enriched in soluble proteins).

Figure 3.2 depicts that, in contrast to the *E. coli* dataset, KR-ratio does not separate SOLP proteins as soluble and insoluble, as both subsets have nearly equal median values. However, protein length (*naa*) appears to separate the subsets just as in *E. coli*. The only other notable difference is observed in cysteine composition (enriched in insoluble proteins). Histidine and methionine are marginally enriched in insoluble proteins. The SOLP dataset shows the least separation between subsets, with both soluble and insoluble features largely overlapping. This suggests that the SOLP dataset is noisier than experimental studies despite being by far the largest of four sets, with less power to separate proteins based on solubility.

Figure 3.3 shows only a slight enrichment of KR-ratio in high-expression proteins in *S. cerevisiae*. Alanine (high-expression enriched), asparagine (low-expression enriched), serine (low-expression enriched) and valine (high-expression enriched) appear to separate subsets. Protein

length is again enriched in low-expressed proteins, suggesting that this could be a feature useful for prediction solubility/expression since proteins from three different organisms (*E. coli*, *H. sapiens* and *S. cerevisiae*) share this trend.

Figure 3.4 shows several small differences in sequence features between high-production and low-production proteins. The largest of these is observed in arginine, which contributes to low-expression. Interestingly, the trend for protein length is reversed in this dataset, with longer proteins associating with high-expression.

North East Structural Genomics Dataset (Price *et al*., 2011)

This study features statistical analyses of results from a high-throughput protein-production pipeline of the NESG (Northeast Structural Genomics Consortium). Proteins expressed in *E. coli* and subsequently purified were scored independently for expression and solubility levels. A range of primary sequence features that influence expression levels and solubility were analysed in order to uncover underlying correlations. The dataset consisted of 9644 randomly selected protein target sequences expressed using the NESG uniform *E. coli* protein expression and purification pipeline. Sequences with significant transmembrane α-helical components or low-complexity components (>20%) were excluded from the pipeline. The dataset was composed of 94% bacterial sequences (82% eubacterial and 12% archaebacterial), 5.7% human sequences, and 0.3 % miscellaneous eukaryotic sequences. Price and colleagues (2011) assigned integer scores (0-5) for expression level and solubility, placing them on ordinal scales. A strong correlation was observed between expression level and solubility score. Higher expression correlated strongly with higher solubility. This finding corroborates other studies mentioned which report inverse correlation between aggregation propensities and mRNA concentration levels.

Logistic regression analysis was performed to evaluate the dependence of expression and solubility on primary sequence parameters. Interestingly, it was reported that fractional lysine content correlated positively with solubility while fractional arginine content showed a slight negative correlation with solubility, a finding that converges interestingly with the KR-ratio hypothesis of the previous chapter. This effect was attributed partly to the fact that arginine is encoded by rare codons, which are known to impede expression in certain cases. The authors further report that the fractional content of the two negatively charged amino acids (Asp, Glu) strongly correlates with higher expression level and solubility score. This finding presents an interesting parallel to the observations of Kramer and colleagues (2012), who report that negative charge is correlated with *in vitro* solubility. The sequence properties that correlated most strongly with high expression levels and solubility scores were charge-based, *i.e.* total charge and net charge.

Unfortunately, the protein dataset generated in this study was not made publicly available and could not be used in the predictive model described in this chapter. Although the protein datasets from these experiments could not be used, total charge and net charge were added to the sequence-based feature set used in this chapter for statistical analysis on the basis of the reported findings.

3.5 Analysis of Sequence-based Features in Multiple Datasets

The manual process of retrieving solubility and abundance data from publications is laborious and error-prone. Furthermore, data picked from a few specific studies is insufficient to discover underlying trends in sequence patterns that truly separate soluble and insoluble protein subsets. Visualising such data in an informative manner is also important. Line plots are useful for individual datasets (figures 3.1 – 3.4), where there are only two subsets whose sequence properties are being compared. However, to observe trends in such properties throughout multiple datasets, heatmaps were preferred. Furthermore, a robust statistical analysis of the trends that emerge as separating high-/low-solubility and abundance is required for a predictive model to be implemented.

The set of sequence-based features was expanded considerably (section 3.5.1). For statistical analyses to be carried out, *z*-scores were calculated for the full set of selected sequence-based features (table 3.1). A *z*-score measures how many standard deviations from the mean a data point is located. For each feature *i*, the *z*-score is calculated using the equation:

$$z_i = \frac{X_i - \mu_i}{\sigma_i} \tag{3.2}$$

where X_i is the raw value of the feature, μ_i is the feature's mean and σ_i the feature's standard deviation. The difference of *z*-score values between soluble and insoluble (or high- and low-abundance) protein subsets was calculated and heatmaps were constructed to visualise how these varied between the different sets. Hence feature enrichment in either subset could be observed based on the sign (+/-) of the difference between *z*-scores, (*i.e.* $z_{i(\mathrm{SOL})} - z_{i(\mathrm{INS})}$), where a positive sign indicates the feature is above the mean value (enriched in soluble or high-abundance proteins) and a negative sign indicates it is below the mean (enriched in insoluble or low-abundance proteins). Figure 3.5 illustrates a heatmap of *z*-score differences of the full sequence-based feature set (table 3.1) between soluble and insoluble proteins for the *E. coli* dataset (Niwa *et al*., 2009). In each of the heatmaps presented, the raw data (*z*-score differences) have been removed for readability but can be found in appendices 3A – D at the end of the chapter.

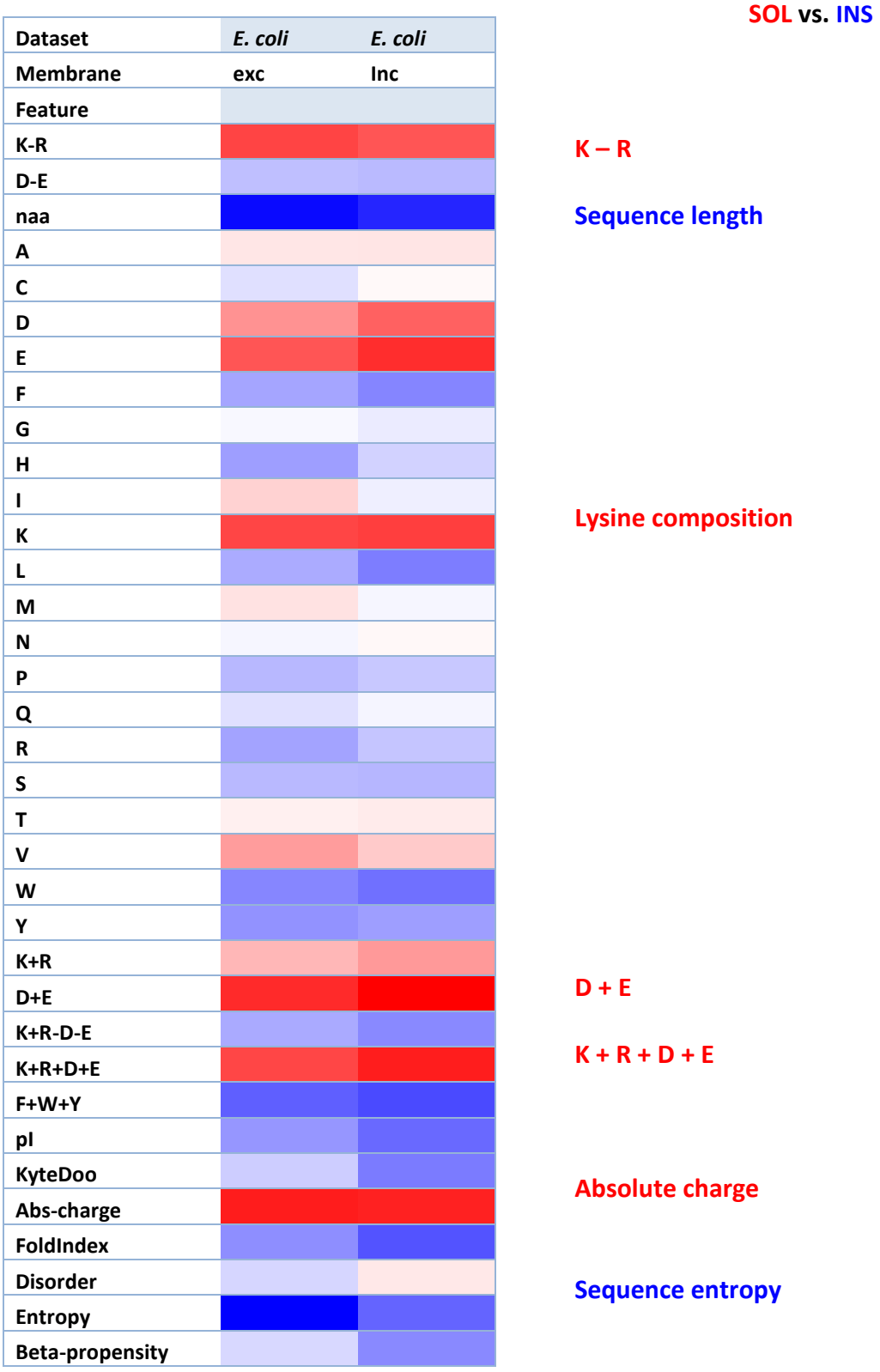


Figure 3.5. Heatmap of Sequence Properties in *E. coli* cell-free expression data. Calculations of *z*-scores were performed under two different conditions: (i) including membrane proteins and (ii) excluding membrane proteins (labelled *inc* and *exc*). Red denotes that a sequence feature is favoured for solublily; blue denotes that one is favoured for insolubility. The lighter the colour, smaller the z-score value (and vice versa), so that lightly coloured cells indicate sequence features that separate SOL/INS less than darker ones. Raw z-score differences are listed in appendix 3A.

All heatmaps use a three-colour scheme where red indicates a higher *z*-score for the sequence feature in the soluble dataset, blue indicates a higher *z*-score in the insoluble dataset, and white indicates parity. A description of the sequence-based features is presented in section 3.5.1 and summarised in table 3.1. As observed in the heatmap and in the figure 3.5, the sequence features that have a higher *z*-score in soluble proteins are mostly charge-based. Those more prevalent in insoluble proteins include sequence length, aromatic residues and interestingly, sequence entropy.

Although sequence entropy is not a property that can be readily manipulated from a protein engineering perspective, higher sequence entropy in aggregation-prone *E. coli* proteins is intriguing since low complexity sequences are more common in disordered proteins. Figure 3.6 plots the mean residue composition for each of the twenty amino acids for the soluble and insoluble datasets of the *E. coli* dataset, as well as over the entire *E. coli* proteome, obtained from the UniProtKB database. Although this method does not measure information entropy, it compares amino acid compositions of soluble and insoluble *E. coli* proteins (Niwa *et al*., 2009) to the full *E. coli* proteome and shows that certain amino acids are enriched in soluble proteins.

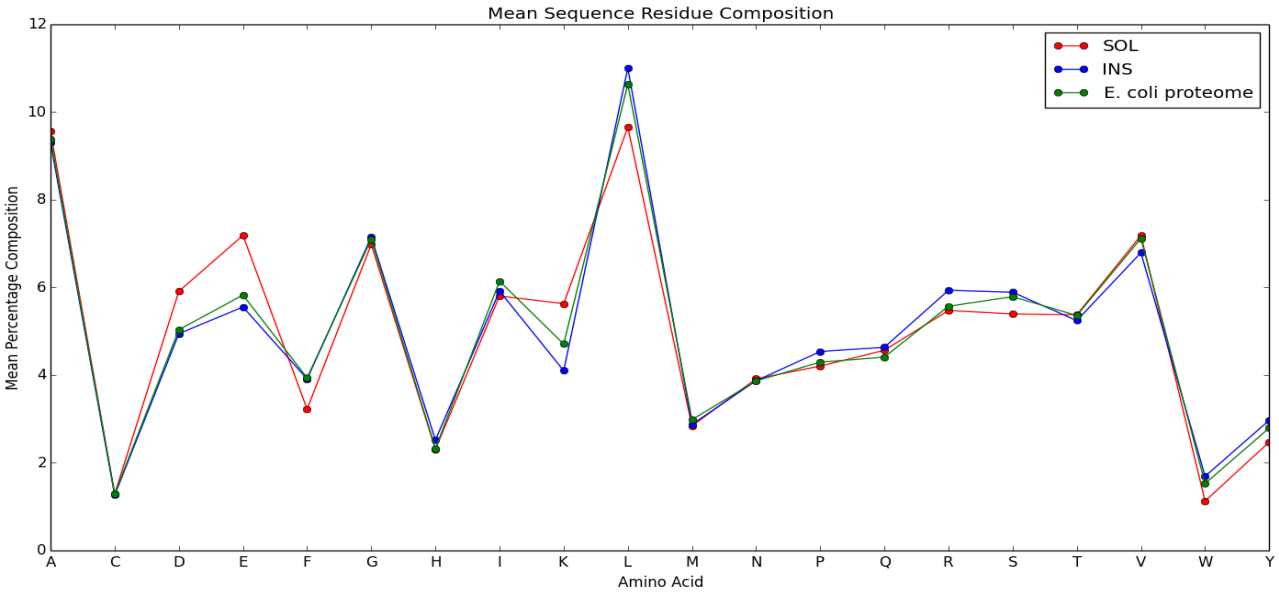


Figure 3.6. Mean Sequence Residue Composition (*E. coli* dataset). Amino acid compositions from soluble and insoluble protein sets (Niwa *et al*., 2009) are compared to that of the *E. coli* proteome.

Mean amino acid compositions were calculated using a Python script to count the occurrence of each residue in each of the sequences for soluble and insoluble datasets and to calculate the arithmetic mean, which is plotted above. Mean residue compositions for protein solubilities obtained from cell-free *E.coli* (Niwa *et al*., 2009) and the entire *E. coli* K12 proteome are considerably similar. Soluble proteins display enrichment for ionisable side chains, notably lysine, aspartic acid and glutamic acid. Insoluble proteins show a slightly higher content of arginine as well as aromatic side chains, specifically tryptophan and valine. The peaks at leucine and alanine most likely occur due to the fact that these are the two most commonly occurring amino acids in globular proteins (Trinquier and Sanejouand, 1998; Tompa, 2002).

3.5.1 Sequence-based Calculations

Thirty-five sequence features are computed in the pipeline. These include the 20 individual amino acid percentage compositions, and several combinations relating to specific physico-chemical properties. In a slight variation to the previous chapter's work, the combined lysine/arginine and aspartate/glutamate properties are now calculated as differences of percentage compositions rather than as ratios (K-R and D-E respectively). This is done in order to avoid numerical pitfalls encountered when there are zeros in the denominator (*e.g*. no arginine/glutamate residues). In calculating charge properties, it is assumed that only lysine, arginine, aspartate, and glutamate contribute significantly at neutral pH that is common physiologically. Hence histidine, cysteine and tyrosine are excluded from charge considerations, although each is tested as an individual amino acid composition. Other combined charge properties (all as percentage compositions) include K+R (positive charge summed), D+E (negative charge summed), K+R+D+E (sum over all charged groups, *i.e*. total charge) and K+R-D-E (net charge). Absolute charge is defined as the modulus of net charge so that it always assumes positive values. Additionally, the pI of each protein is estimated, purely from sequence with no charge-charge interactions and pK_a shifts taken into account. The amino acid side chain pK_as used for pI calculation are as follows: aspartic acid 4.0; glutamic acid 4.4; histidine 6.4; lysine 10.4; arginine 12.0; tyrosine 10.1 (Warwicker, 1999).

Several non-charge properties are computed, including the percentage of amino acids with aromatic side chains (*i.e*. F+W+Y). The Kyte-Doolittle measure of hydrophobicity (Kyte and Doolitle, 1982) is used in several ways. It is recorded as an overall measure for a protein sequence and also included in a measure of folding propensity (Uversky *et al*., 2000). The fold propensity

score is calculated as 2.785<H> - <R> - 1.151, where the range (-4.5 to 4.5) of the Kyte-Doolittle hydrophobicity is scaled (0 to 1). <H> is the mean (per amino acid) scaled hydrophobicity and <R> is the mean (per amino acid) net charge, over the protein sequence. Uversky and colleagues (2000) established that the relationship <R> = 2.785 – 1.151 defines a threshold separating proteins known to be folded and those known to be intrinsically disordered. Thus, the fold propensity score predicts a folded protein when taking a positive value and intrinsically disordered when assuming a negative value. The Kyte-Doolittle hydrophobicity (unscaled) is also used as a screen to reduce the number of membrane proteins included in the analysis. A protein is excluded if any 21 amino acid window has an average hydrophobicity of greater than 1.6, a threshold developed in the original study (Kyte and Doolittle, 1982). Another measure of disorder that is used in the predictive model employs amino acid propensities for intrinsic disorder from the GlobPlot scheme (Linding *et al*., 2003).

Given the importance of β-strands in amyloid formation, and the consideration of β-forming propensity in previous reports of solubility prediction, β-strand propensity has been included, averaged over each protein, using propensities reported by Costatini and colleagues (2006). Finally, a measure of sequence entropy was incorporated. The Shannon entropy *H* for amino acid diversity within a protein sequence is computed using the definition for information entropy as follows:

(3.3)

$$\boldsymbol{H = -\sum_{i}[f_i \log_2(f_i)]}$$

where f_i is the fractional amino acid content for each residue type *i*, and the sum runs over the 20 standard amino acids. Table 3.1 below summarises the sequence properties used to compare solubility and protein abundance data throughout multiple datasets.

Table 3.1 Sequence Features used in Predictive Model

Feature	Description
K – R	Difference of lysine and arginine percentage compositions
D – E	Difference of aspartate and glutamate percentage compositions
Naa	Sequence length (*n*umber of *a*mino *a*cids)
A, C, D, E, F, G, H, I, K, L, M, N, O, P, Q, R, S, T, V, W, Y	Amino acid percentage composition for each residue type
K + R	Positive charge summed
D + E	Negative charge summed
K + R – D – E	Net charge
K + R + D + E	Total charge (sum over all charged groups)
F + W + Y	Percentage of aromatic groups
pI	Isoelectric point
KyteDoo	Kyte-Doolittle hydrophobicity
Abs-charge	Absolute charge, *i.e.* |K + R – D – E|
FoldIndex	Uversky fold
Disorder	Propensity to form disordered folds
Entropy	Shannon entropy
Beta-propensity	Propensity to form β-strands (based on Chou-Fasman helix prediction)

3.5.2 Protein Abundance Datasets

PaxDb

A particularly useful resource in the scope of large scale such studies is PaxDb (Wang *et al*., 2012), a web-accessible centralised repository of experimentally determined protein abundance levels that integrates such data across a broad spectrum of organisms. PaxDb is a meta-resource, drawing data exclusively from published experimental studies and from the often labourious work performed at the primary proteomics data repositories. Protein abundance values are calculated by converting tandem mass spectrometry (MS-MS) data into units of "parts per million" (ppm). Abundances in ppm describe each protein with reference to the entire expressed proteome and in particular to the most abundant proteins therein; the latter are usually confined to the cellular translation apparatus and to key proteins in metabolism or structural maintenance (Wang *et al*., 2012). It currently covers 56 species from all three domains of life (eukaryotes, bacteria, and archaea) and aggregates quantitative proteomics data on organism-wide averages and organ/tissue-

wide averages. For each organism, individual, tissue-level datasets are provided as well as a single, consolidated abundance estimate of all detectable proteins.

Plasma Proteome Database (PPD)

The plasma proteome database (Nanjappa *et al*., 2014) was developed as a part of the Human Proteome Organisation's (HUPO) initial effort to characterise the human plasma proteome. The database houses information on approximately 1300 proteins detected in serum and plasma. The plasma proteome represents an important subproteome, as it contains proteins secreted in all tissues (Anderson and Anderson, 2002). Furthermore, plasma comprises one of the most extensively investigated body fluid in clinical diagnostics, making it is highly relevant to the therapeutics field.

In addition to classical blood proteins, plasma contains proteins secreted by various cells, glands and tissues along with proteins derived from infectious organisms residing inside the body. The plasma proteome comprises 22 highly abundant proteins including albumin, immunoglobulins, transferrin and haptoglobin, which make up 99% of total protein abundance in plasma. The remaining fraction is composed of proteins of much lower abundance including proteolytically cleaved protein fragments (Tirumalai *et al*., 2005). The dynamic range of protein abundance, exceeding 10 orders of magnitude (Nanjappa *et al*., 2014), renders the plasma proteome highly relevant to the current investigation, as it represents a dataset containing proteins existing at high and low concentrations. Data from the PPD was used in the sequence-based feature analysis.

3.5.3 Heatmap Analysis of Sequence Features

Quantitative proteomics datasets were obtained from the PaxDb and plasma proteome databases and combined with those from individual publications. As discussed, the rationale behind this comparison was that using heatmaps, sequence-based features enriched in soluble datasets would become apparent, as would any potentially underlying trends discriminating soluble and insoluble proteins. A Perl pipeline was implemented to perform statistical calculations (code contributed by Jim Warwicker) and Python scripts were written for data cleaning and formatting where necessary. Calculations for comparative analysis of high/low protein abundance, concentration and solubility datasets were performed under two conditions: including membrane and membrane excluding.

For the protein datasets retrieved from PaxDb and PPD, further to comparing sequence properties in humans (PPD) and across multiple organisms (PaxDB), it was considered appropriate to compare the high and low abundance/concentration proteins within a single dataset, as was done for the proteomics studies in section 3.4. In order to visualise disparities in sequence-based features within a single dataset, PaxDb and PPD were compartmentalised into subsets of high abundance/concentration and low solubility/concentration. The cutoff (percentage of data points included in each subset) for high and low solubility or concentration proteins was decided based on frequency plots. Table 3.2 summarises the datasets used in the current work, including the organism from which protein data was obtained and the environment of the proteins under consideration, *e.g*. intracellular or extracellular, as well as the number of protein sequences in the dataset and each of the subsets. The final two columns (cutoffs and separation) refer to how datasets were divided into high and low values (solubility/abundance/concentration). Cutoffs are used in datasets where there is no mention of the proportion proteins belong to high and low subsets of the quantity being measured, whereas separation is used when the studies report the proportion of proteins comprising each subset.

Table 3.2 Summary of Datasets for Sequence-based Features

Quantitative Proteomics Datasets						
Dataset	**Reference**	**Biological Environment**	**Number of sequences/ORFs**	**Cutoffs (Upper/Lower)**	**Separation**	
E. coli Bacterium	Niwa *et al*., 2009	Intracellular (cell-free expression)	3173	Not used	70% SOL	30% INS
S. cerevisiae Yeast[1]*	Ghaemmaghami *et al*., 2003	Intracellular	3853	33%	None	
S. cerevisiae Yeast[2]	Lu *et al*., 2007	Intracellular	592	10%	None	
S. cerevisiae Yeast[3]	Lee *et al*., 2012	Intracellular	6530	33%	None	
A. niger Fungus	van den Berg *et al*., 2012	Extracellular/ Secreted	345	Not used	48% High	52% Low
SOLP	Magnan *et al*., 2009	Intracellular & Extracellular	17408	Not used	50% SOL	50% INS

Protein Databases							
Repository/ Web server	**Reference**	**Biological Environment**	**Number of sequences/ORFs**	**Cutoffs (Upper/Lower)**		**Number of sequences in upper/lower subsets**	
PaxDb** *Multiple Species*	Wang *et al*., 2012	Intracellular & Extracellular	53963 (total)	5%	33%	Varies across species	Varies across species
PPD *H. sapiens*	Nanjappa *et al*., 2014	Plasma (Extracellular)	1278	8%	33%	100 (8% cutoff)	421 (33% cutoff)

* Superscripts 1, 2, 3 refer to columns 3 – 10 of the relevant heatmap (figure 3.7)
** Refer to table 3.3 for a detailed breakdown of the PaxDb datasets

Quantitative Proteomics Datasets

Four proteomics studies were overviewed in section 3.4, each one from a different organism and SOLP, which pooled together protein sequences from three primary repositories. The *E. coli* and SOLP datasets were by design, *a priori* divided into soluble and insoluble protein subsets. For *E. coli*, the separation was based on the bimodal distribution of solubilities, whereby 70% of quantified ORFs had soluble protein products and 30% encoded aggregation-prone proteins. Similarly, the fungal dataset (table 3.2) was divided into high production and low production subsets based on experimental results. The yeast datasets did not have this attribute, as they were studies involving high throughput proteomics and transcriptomics techniques to quantify gene expression and mRNA abundance levels. For these datasets, whenever there were more than 1000

data points, the highest and lowest 1/3 (~33%) values was used. These tails were chosen to form the lower and upper subsets, given that this both gives a large separation of solubility or abundance in the distributions, and maintains large enough numbers of sequences for processing.

In addition to these four studies, two similar studies were included in the heatmap analysis, in which further sequence properties were considered. In the first, absolute profiling expression (APEX), a method for large-scale absolute protein expression measurements, was applied to estimate the relative contributions of transcriptional- and translational-level gene regulation in the *S. cerevisae* proteome (Lu *et al*., 2007). APEX was applied to yeast growing in rich and minimal medium to measure the absolute abundance of 454 proteins. The authors reported good correlation when the APEX-derived protein abundances were compared with published measurements of absolute expression of the corresponding mRNAs. Subsequently, measurements of 626 proteins observed from yeast grown rich and minimal media (annotated appropriately on figure 3.7) were compared. It was reported that sensitivity of expression levels changed under different conditions, with changes in expression predominantly reflecting differential expression of metabolic enzymes (Lu *et al*., 2007). The abundance levels of 592 verified ORFs were measured in units of molecules per cell. The highest and lowest 10% abundance tails were used to generate subsets of high and low abundance in order to analyse the differences in their sequence properties (columns 6 – 9, figure 3.7).

The final study included in the heatmap analysis used quantitative transcriptomics data acquired from *S. cerevisiae* cultures grown under two conditions to predict genome-scale metabolic flux patterns. Gene expression data was generated using RNA sequencing, which provides expression levels in terms of counts of expressed transcripts that can be related to transcripts per cell (Lee *et al*., 2012). A total of 6717 transcripts were quantified, 6530 of which had non-zero values. The highest and lowest 33% transcripts were used to generate the relevant subsets for heatmap visualisation (columns 10 – 11, figure 3.7).

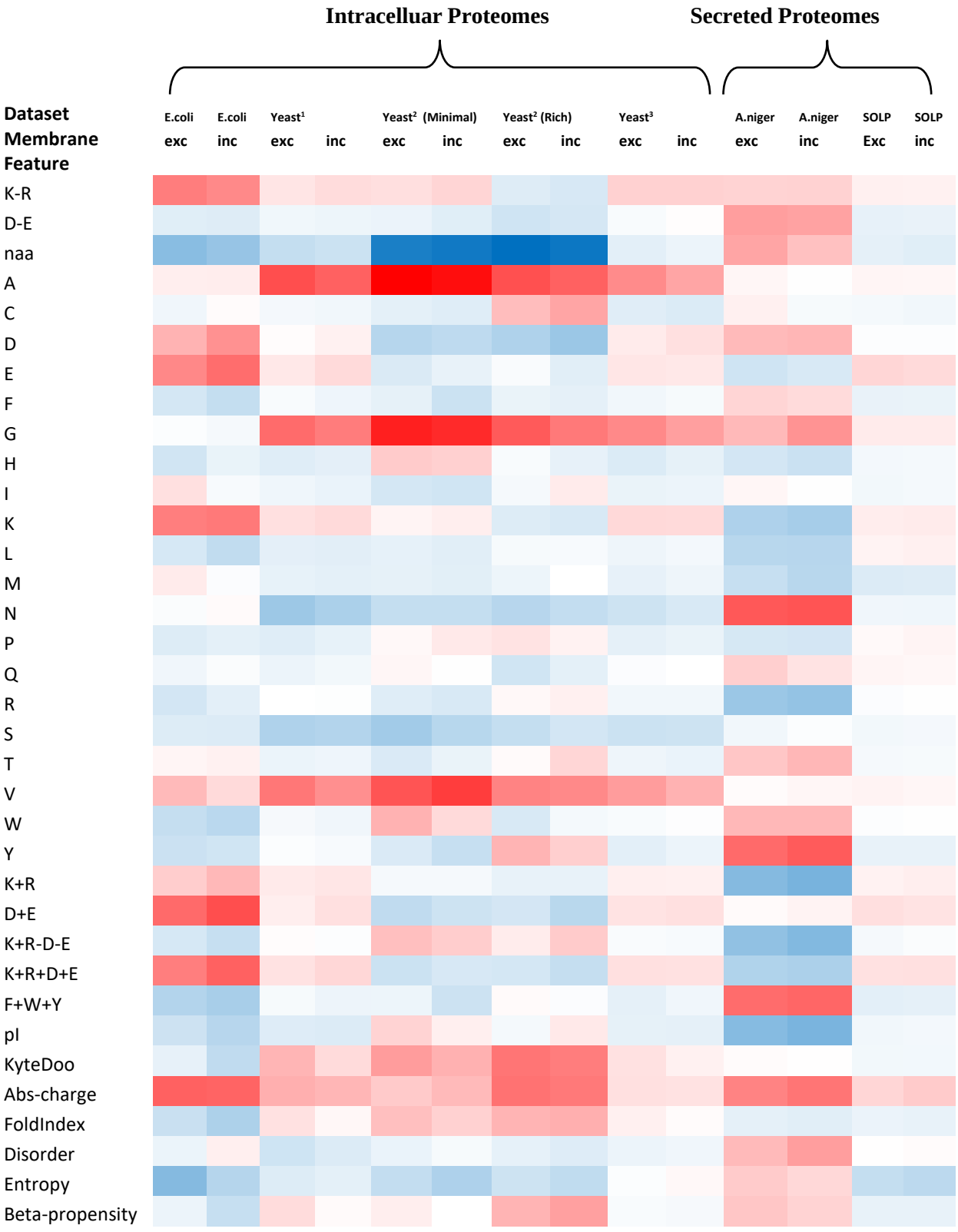


Figure 3.7. Heatmap of Quantitative Proteomics Datasets. Calculations of *z*-scores were performed under two different conditions: (i) including membrane proteins and (ii) excluding membrane proteins (labelled *inc* and *exc*). Red denotes that a sequence feature is favoured for high solubility/abundance, while blue denotes that one is favoured for low solubility/abundance. Raw *z*-score differences are listed in appendix 3B.

The heatmap for quantitative proteomics-based data is arranged so that cell-free expression (*E. coli*) and intracellular (*S. cerevisiae*) data are located on the left-hand side and data based on secreted proteins (*A. niger*) is situated on the right-hand side. The two rightmost columns refer to the SOLP dataset. Several properties appear to be consistent throughout datasets. K-R is elevated in almost all cases for high solubility, as are other charge-based properties. Sequence entropy and aromatic residues (F+W+Y) appear to be generally enriched in low solubility subsets, and the same holds true for sequence length as well as serine/leucine composition. The SOLP dataset exhibits by far the poorest separation between soluble and insoluble proteins, as there is almost a complete lack of red/blue bands.

Within the intracellular proteomes, strong enrichment bands are observed for alanine, glycine and valine (high-abundance proteins) as well as protein length (low-abundance proteins) in yeast. For yeast proteins, the data from APEX (Lu *et al*., 2007) studies follow the patterns of protein abundance levels during log-phase growth (Ghammaghami *et al*., 2003) as seen by dark red bands in the relevant columns (also presented qualitatively in figure 3.3). Larger proteins are strongly associated with low-abundance in rich and minimal medium-grown yeast, but less so for measurments made during log-phase growth and metabolix flux studies (Lee *et al*., 2012) as seen by dark blue bands.

Importantly, there appears to be a general divergence in enriched features between cell-free and intracellular proteins and extracellular/secreted proteins. Indeed, for several sequence properties, the columns corresponding to the *A. niger* experimental study (van den Berg *et al*., 2012) show trends opposite to those of other experimental studies. This is observed most explicitly for asparagine (N), valine (V) and aromatic residues (F, W, Y). It is worth noting that the fungal dataset was the smallest (345 proteins), but this marked divergence remains interesting nonetheless.

PaxDb Datasets

The PaxDb repository is a web accessible resource (www.pax-db.org) storing protein abundance data across multiple organisms. Protein sequences from four species were considered in the current investigation – *E. coli* K12, mouse (*M. musculus*), yeast (*S. cerevisiae*) and human (*H. sapiens*). For these datasets, only membrane-excluded calculations were performed using the Kyte-Doolittle criterion to test for hydrophobicity common in transmembrane regions (cutoff at 1.6), since proteins with membrane parts are expected to have different solubility properties, as they possess extensive hydrophobic interaction regions due to their transmembrane components. Table

3.3 provides details of species' datasets and figure 3.9 presents *z*-score differences for sequence features throughout *E. coli*, yeast, mouse and human proteomes.

Table 3.3 Summary of PaxDb Datasets

PaxDb Protein Abundance Datasets					
Species	**Number of sequences/ORFs**	**Cutoffs (Upper/Lower)**		**Number of sequences in upper/lower subsets**	
				5%	33%
E. coli	3119	5%	33%	136	1040
S. cerevisiae	5942			297	1981
M. musculus	13930			696	4643
H. sapiens	16578			829	5526
	Total 39569				

Figure 3.8 shows the cumulative frequency of protein abundance values, which was used to determine thresholds for dividing the dataset into high and low abundance subsets. There are two curves for human proteins, one corresponding to the entire PaxDb dataset and the other corresponding to non-membrane proteins. The plot shows that for human proteins, a threshold corresponding to approximately five percent of data points is reasonable. The same cutoff value was used for mouse and yeast protein datasets for consistency. Hence the highest and lowest 5% of concentration values were used in one subset. The other subset consisted of the highest and lowest 33% of concentration values. The heatmap for PaxDb-based proteomes is shown in figure 3.9.

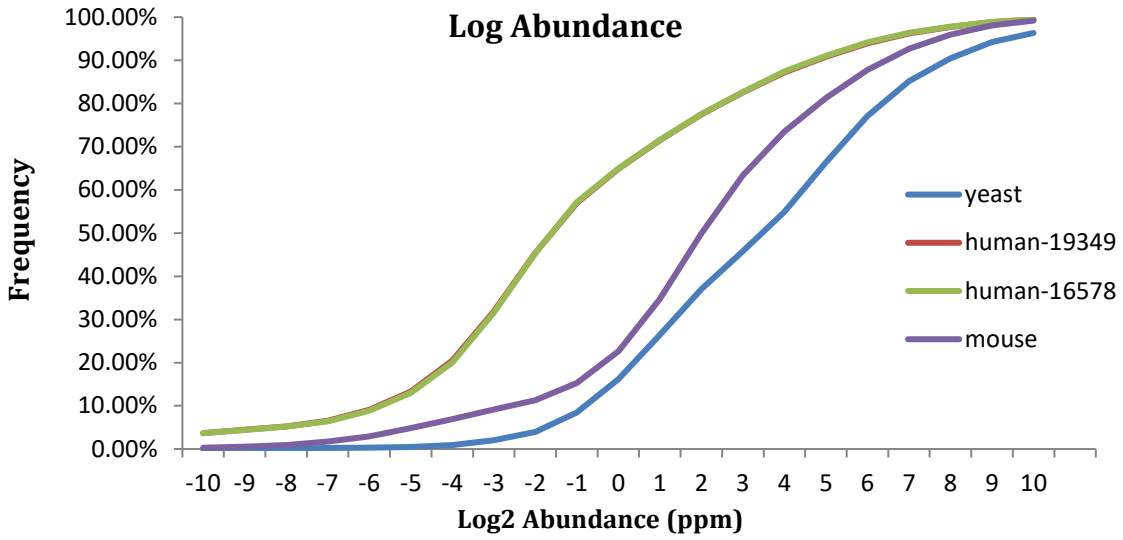


Figure 3.8. Cumulative Frequency of PaxDb abundance levels. The 5% tail observed for the lowest abundance levels of human proteins was used as a cutoff for constructing small high/low concentration level subsets. The large subsets were constructed using the 33% cutoff.

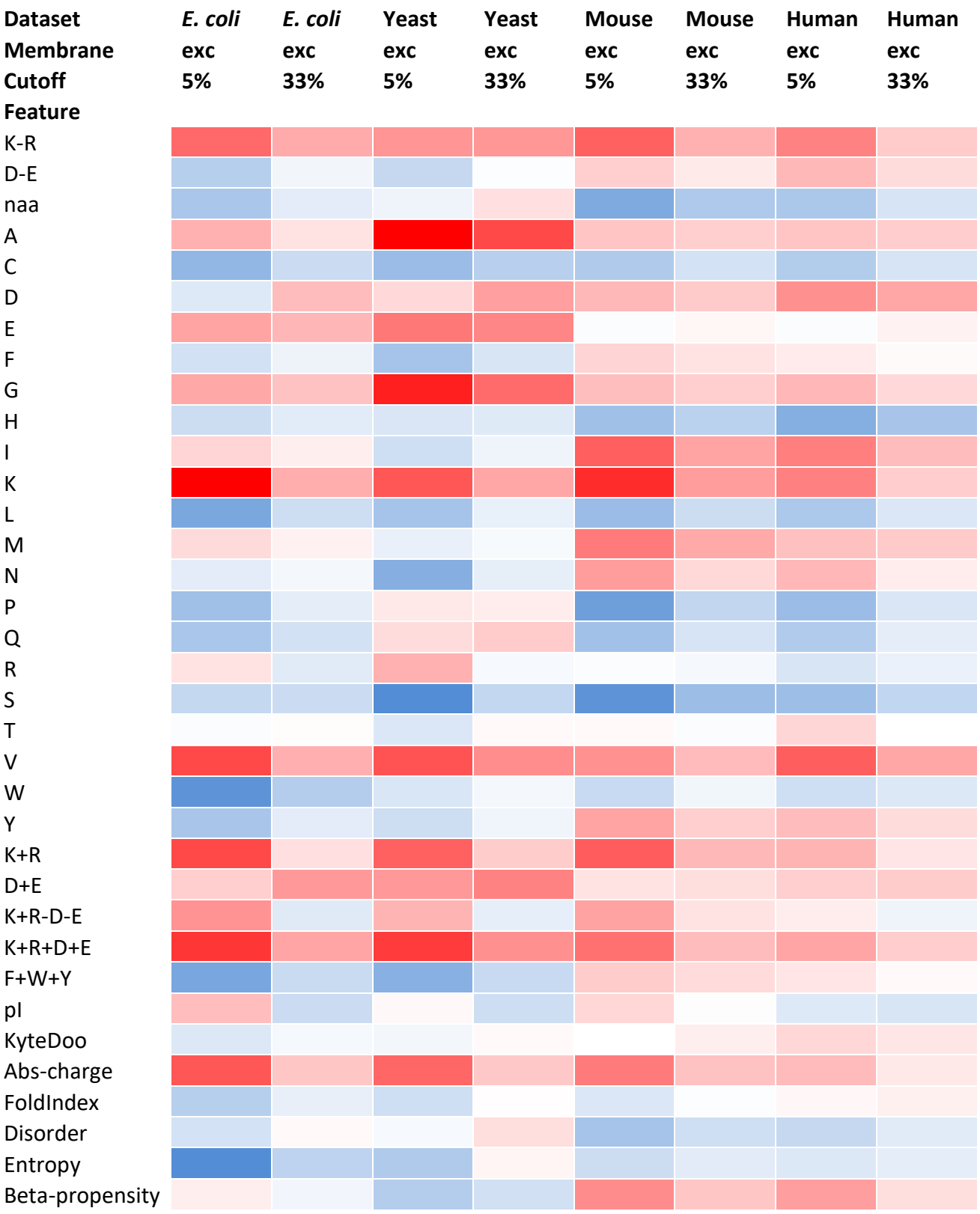


Figure 3.9. Heatmap of PaxDb Datasets. Calculations of *z*-scores were performed solely for non-membrane proteins (labelled *exc*). Red denotes that a sequence feature is favoured for high protein abundance levels, while blue denotes that one is favoured for low protein abundance levels. Raw *z*-score differences are listed in appendix 3C.

As shown in the figure 3.9, charge-based properties are favoured in proteins with high abundance levels. Enrichment in high-abundance proteins is observed across all proteomes for several features, including K-R, lysine (K), valine (V), K+R, D+E and charge-based properties. This reinforces the dogma that charge is important in aggregation-related phenomena. In this context, increased levels of charge may serve as an evolutionary mechanism enabling cells to express proteins needed in large cytosolic concentrations whilst inhibiting non-specific interations leading to aggregate formation. Enrichment in low-abundance proteins is observed uniformly throughout all organisms in histidine (H), leucine (L), serine (S), tryptophan (W) and sequence entropy. Encrichment of aromatic amino acid composition is split between high-abundance (*E. coli* and yeast) and low-abundance (mouse and humans) proteins. A similar divergence is also observed for phenylalanine composition (F), which is elevated in high-abundance subsets in the proteomes of mouse and humans but elevated in low-abundance subsets in those of *E. coli* and yeast.

It is worth noting that there are more dark red bands than blue bands. The colour intensity indicates the difference in *z*-score values, meaning that features that associate with high-abundance proteins (mostly charge-based) do so more strongly than do features that associate with low-abundance proteins.

PPD Datasets

Figure 3.10 shows the cumulative frequency of protein concentration values, which was again used to determine thresholds for dividing the dataset into high and low abundance subsets. A logarithmic scale was used to account for the multiple orders of magnitude of protein concentrations found in human plasma (Nanjappa *et al*., 2014). The plot shows that distribution has a "tail" corresponding to approximately 8 percent of data points. Hence the highest and lowest 8% of concentration values (100 protein sequences) were used in one subset. The other subset consisted of the highest and lowest 33% of concentration values (421 sequences).

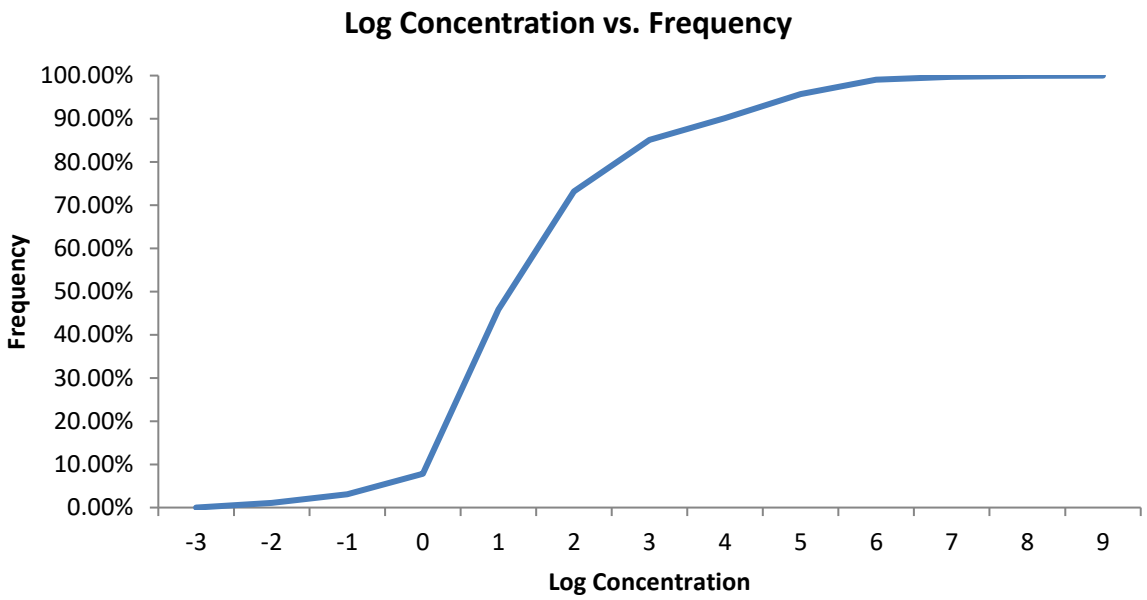


Figure 3.10. Cumulative Frequency of Plasma Protein Concentrations. A logarithmic scale was used to account for the multiple orders of magnitude of varying protein concentrations in human plasma (Nanjappa *et al*., 2014). The 8% tail justifies using the equivalent percentage as a cutoff for making high/low concentration level subsets.

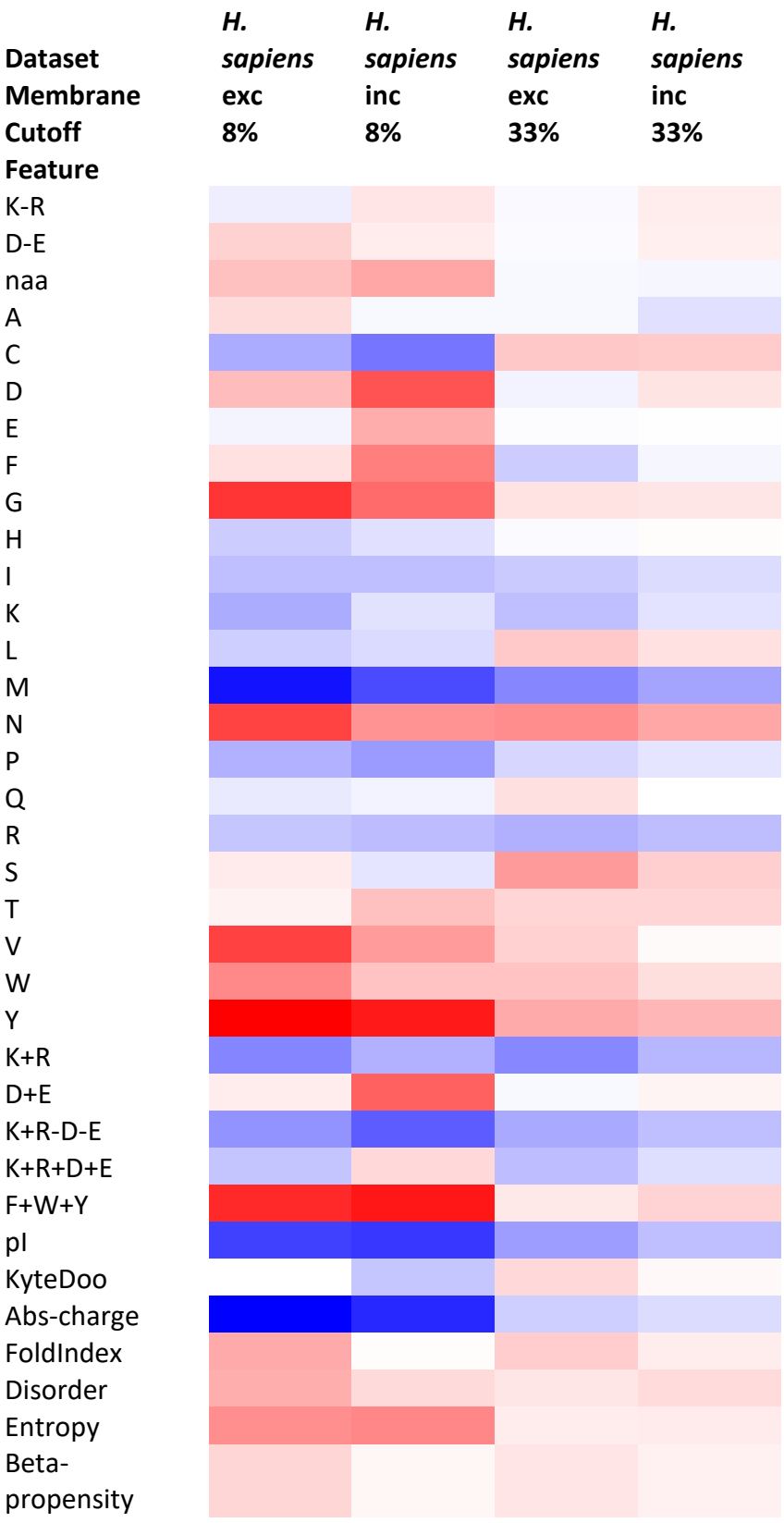


Figure 3.11. Heatmap of PPD Datasets. Calculations of *z*-scores were performed under two different conditions: (i) including membrane proteins and (ii) excluding membrane proteins (labelled *inc* and *exc*). Red denotes that a sequence feature is favoured for high protein concentration levels, while blue denotes that one is favoured for low protein concentration levels. Raw *z*-score differences are listed in appendix 3D.

Interestingly, a reversal of the prevalence of charge-based properties in proteins existing at high concentrations is observed in the plasma proteome dataset. Most charge-related features are enriched in proteins found at low concentrations. Entropy and aromatic residues also appear to follow an opposite trend to the quantitative proteomics (figure 3.7) and PaxDb datasets (figure 3.9), as they are enriched in proteins found at high concentrations. Protein length is associated with high abundance in small subsets (8% cutoff) and with low-abundance in large subsets (33% cutoff). K-R and D-E, corresponding to the KR-ratio and DE-ratio discussed in chapter 2 have irregular enrichment patterns. Enrichment in low-abundance proteins is observed uniformly for isoleucine (I), lysine (K) and methionine (M) and in high-abundance proteins for asparagine (N), tyrosine (T), valine (V), tryptophan (W) and threonine (T). Importanntly, the observed trends match those of the secreted protein dataset (*A. niger*) in the first heatmap. This may suggest that protein localisation (intracellular vs. extracellular) may play a key role in aggregation propensity.

3.5.4 Protein-Sol: Web Server Development

The sequence-based feature analysis is being integrated into a web server currently under development (Hebditch *et al*., in writing) that is currently hosted on a University of Manchester virtual machine server (accessible at the URL www.protein-sol.manchester.ac.uk). Although currently in the primary stages of development, it provides basic functionality such as receiving a user-specified sequence and providing a detailed solubility prediction output. The output provides a summary of sequence-based features, including deviations from the population averages of *E. coli* cell-free solubility data (Niwa *et al*., 2009) for each of the 35 calculated properties (table 3.1). Furthermore, a charge score (per amino acid) and an Uversky-based fold score (per amino acid) are calculated across the sequence using a window of 21 amino acids. Finally, a scaled solubility score (0-1) is calculated reporting that the queried sequence is predicted as folding into an either soluble (>0.5) or insoluble (<0.5) protein. The server allows the user to download the results as a comma-separated (.csv) file. The graphical user interface (GUI) of Protein-Sol is displayed in figure 3.12 below, which shows the output of a query run on a human myoglobin sequence (UniProtKB P02144).

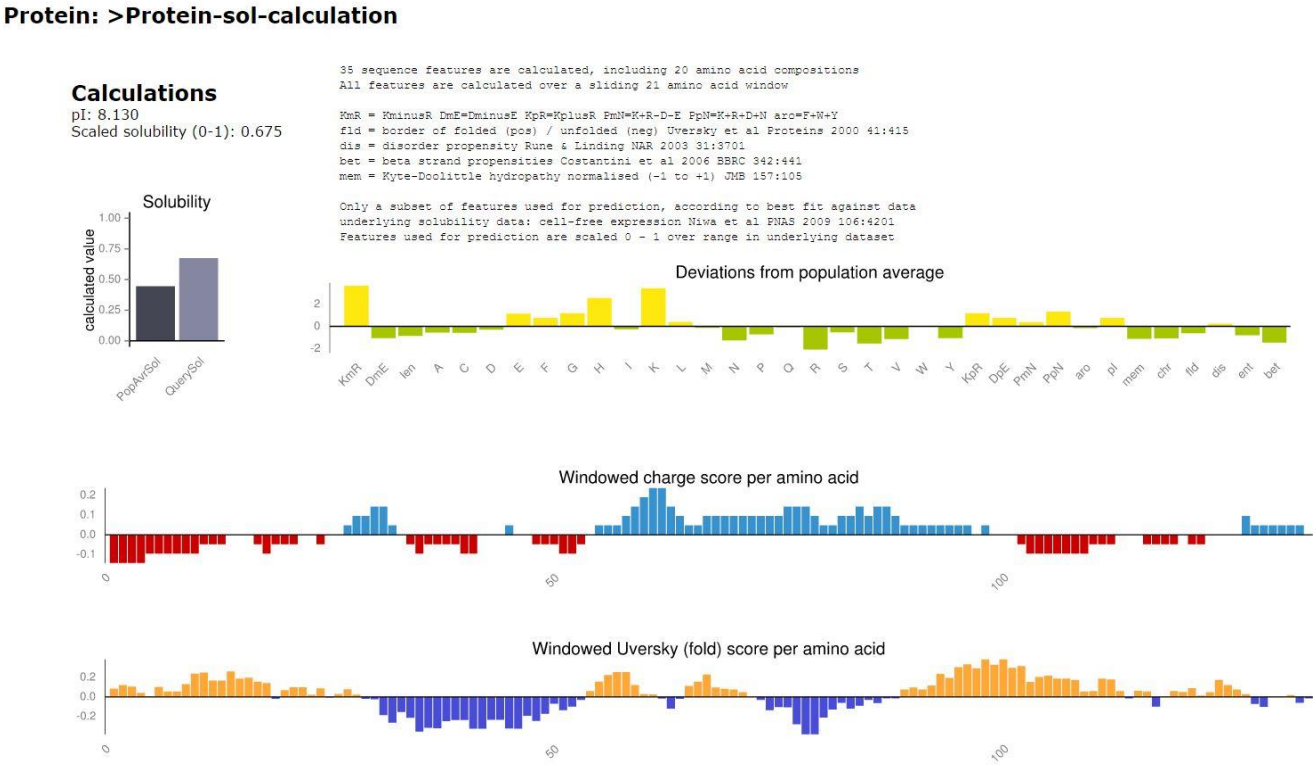


Figure 3.12. Protein-Sol Solubility Prediction. The solubility prediction output is organised into four panels: (i) deviation from the average of *E.coli*-based solubility data for thirty-five computed sequence features (upper right panel). (ii) Windowed (21 residues) charge score per amino acid (middle panel). (iii) Windowed (21 residues) Uversky (fold) score per amino acid (lower panel). (iv) Calculated pI and scaled solubility score (0-1) (upper left panel).

3.5.5 Sequence-based Solubility Trends

Sequence-based disparities in *E. coli* solubility data

For sequence-level analysis, comparison between computed features and experimental properties (solubility/abundance/concentration) is made by assessment of how well each feature distinguishes between tail-end subsets (PaxDb and plasma proteins) and *a priori* separated subsets (quantitative proteomics studies). Expanding the analysis of sequence features and *E. coli* protein solubility from previous work (Chapter 2; Warwicker *et al*., 2014), it was found that several properties directly related to primary sequence provide some distinction between the lower and higher solubility subsets. Separation was plotted as a *z*-score difference of lower subsets (average) subtracted from higher subsets (average).

Protein length tends to be greater for less soluble proteins. Length has the largest deviation, presumably since longer proteins can be more complex with more scope for containing elements that lower solubility. In this dataset, charge is important in several ways, but less so in terms of net charge (K+R-D-E), rather in terms of the number of charge groups (assuming pH 7.0), *i.e.* K+R+D+E. Both positive charge (K+R) and negative charge (D+E) summed have positive values, and this is favourable for solubility. However, for lysine and arginine, this actually masks an even higher positive difference for lysine, accompanied by a small negative difference for arginine. Amino acids with aromatic side chains are enriched in less soluble proteins; F+W+Y and each of F, W and Y have negative differences. This is reasonable in the sense that the aromatics are amongst the most non-polar amino acids, which could in principle lead to protein-protein interactions and aggregation. Aromatic side chains may contribute to the negative differences for Kyte-Doolittle hydrophobicity and fold propensity. Interestingly, the disorder prediction from the GlobProt parameters gives little difference, so that the fold propensity and disorder metrics used here are not equivalent. Overall, the analysis suggests that less soluble proteins have elevated fold propensity. Although this may seem counterintuitive, it is possible that proteins calculated with lower fold propensity (greater charge and less hydrophobicity) tend towards intrinsically disordered protein (IDP) properties, in which case solubility would be retained in the absence of a folded structure. This idea is partially supported by the large negative difference for sequence entropy, *i.e.* there is more sequence variation, and hence more equal sampling over the 20 amino acids, in the lower solubility subset. This could be the case if the insoluble subset were enriched for folded domains, with sampling of both polar and non-polar amino acids, with the soluble subset enriched for IDP-like proteins, sampling more from polar amino acids with a commensurate reduction of sequence entropy.

In this view, proteins exhibiting low solubility will be those that contain folded domains that can presumably lead to aggregation through partial unfolding and subsequent non-specific protein-protein interactions. Importantly, of the sequence features that best discriminate soluble and insoluble proteins in *E. coli*, manipulation in the context of protein engineering would entail altering either the net charge of a protein or the balance of hydrophobic and polar groups (fold propensity). Perhaps the feature with the greatest scope for manipulation with minimal disruption with regard to tertiary structure and purification properties of a protein is the KR-ratio, as practically this would entail the relatively simple procedure of swapping non-essential arginine residues with lysine.

Sequence-based disparities in Intracellular and Secreted Proteomes

Perhaps the most interesting observation made from the comparison of *E .coli* solubility data to quantitative proteomics sets from other organisms (figure 3.7) is the divergence of intracellular and secreted proteome trends. Comparing the intracellular proteomes to the secreted (*A. niger*) one, K-R remains largely positive. The lysine row (K) maintains positive *z*-score differences across intracellular protein datasets but reverses sign in the secreted dataset. Overall, K-R deviation is maintained due to an enrichment of arginine in lower abundance proteins of the secreted dataset. Charge-based properties such as total positive charge (K+R), net charge (K+R-D-E) and total charge (K+R+D+E) follow generally opposite trends in intracellular proteomes (enriched in more soluble/abundant proteins) and secreted proteomes (enriched in less abundant proteins), perhaps indicative of a lesser role for charged surfaces when proteins are free of the crowded cytoplasmic environment. Furthermore, asparagine (N) is substantially elevated in the more abundant secreted subset, possibly as a result of the stabilising effects of N-linked glycosylation. This effect is also observed, albeit to a lesser extent, in tyrosine (T). Aromatic side chains are also enriched in the more abundant secreted proteins, again possibly indicating a role for folded state stability.

Interestingly, the reversal of both charge-based properties and aromatics between intracellular and secreted proteomic datasets carries over to the plasma subproteome (figure 3.11). In the heatmap covering sequence disparities between high and low concentration proteins in human plasma, charge-based properties follow the same trend as those in the fungal secreted protein dataset, *i.e*. elevation in low concentration proteins. These observations further support the hypothesis that charged surfaces may not be as important for maintaining solubility in non-cytoplasmic environments. Furthermore, plasma proteins and secreted proteins are the only datasets in which greater protein length appears to be favourable for high concentrations (although this is only observed in the upper/lower 8% of plasma protein dataset), hinting that in higher order organisms, extracellular proteins may have experienced different evolutionary pressures than their cytoplasmic counterparts.

It must be the case that the highly abundant, secreted proteins such as immunoglobulins are relatively soluble, but it remains to be established precisely how this is achieved. Some features (K-R in particular) are uniformly enriched in more soluble and more abundant subsets, and therefore could be a focus for solubility engineering efforts. Finally, SOLP (final right-hand columns of figure 3.7) has poor discrimination ability, clearly producing the weakest signal of all analysed datasets. This is possibly due to the fact that it uses annotation-based criteria to classify a protein as either soluble or insoluble (Magnan *et al*., 2009), as opposed to experimental solubility or

abundance data. Presumably, classification criteria of protein behaviour used in structural genomics projects are too broad to yield large deviations between proteins expressed at high and low levels.

Sequence-based disparities in PaxDb datasets

Z-score difference calculations were further made for the lower and upper abundance subsets of protein abundance levels based on data from PaxDb (table 3.3) for *E. coli*, yeast, mouse, and human proteomes. Figure 3.9 shows a heatmap comparison of the results, matched to an average of z-score differences over the four species' abundance datasets. It is clear that features in general vary with respect to species. The positive differences (red) for charge-related properties such as K, K-R, K+R, D+E and K+R+D+E are reproduced across all datasets. Similarly, protein length, fold propensity and sequence entropy are mostly negative (blue). There is some variation in fold propensity across the datasets. Although it was suggested that IDP characteristics could be enriched for more soluble proteins, this cannot be the only contributing factor to abundance, since there will likely be evolutionary constraints on the levels of disordered proteins that a cell can produce.

An interesting trend that arises in the PaxDb heatmap is the opposite enrichment pattern of closely related amino acids. Valine and leucine, both non-polar molecules, present one example of this trend. Valine is consistently enriched in more abundant proteins (red), whereas leucine is enriched in less abundant proteins (blue). Similarly, serine and threonine, both possessing polar side chains, have intriguingly divergent trends in abundance data. Serine is consistently enriched in less abundant proteins whilst threonine shows little to no enrichment in any of the upper and lower abundance subsets (white), *i.e.* the *z*-score difference is close to zero.

3.6 Analysis of Structure-based Features

Further to sequence-based features, structural features were also considered. Structural annotation was accomplished using the SIFTS cross-referencing tool (Velankar *et al.*, 2013). The computational pipeline was extended so that the SIFTS tool was integrated for downloading UniprotKB IDs mapped to PDB IDs. In cases where multiple matches were encountered, the PDB structure with the best resolution was chosen, and only X-ray structures were considered.

3.6.1 Structure-based Calculations

Seventeen structure-based features are calculated, including both whole protein properties and features calculated as maxima or minima over patches on the protein surface. Charge-based features are included via calculations of electrostatic potential maps, with in-house code to handle calculation of patch features (Chan *et al*., 2013). Electrostatic potential is calculated at pH 7 with ionisable group charges adjusted according to the standard pK_a values given for the sequence-based pI calculation, and with no adjustment for charge interactions. A Finite Difference Poisson-Boltzmann (FDPB) method is used (Warwicker and Watson, 1982; Warwicker, 1986) to solve the PBE (equation 1.5) for the vector $\varphi(\mathbf{r})$. Ionic strength is set at 0.15 M. Contours are constructed for potentials at +/- *kT*/e (*k* = Boltzmann constant, *T* = 300 K, and *e* = electronic charge) on the surface of proteins inscribed on a 3D finite difference grid (figure 1.6). Numbers of grid points are then summed in each contour patch, and the largest patch sizes recorded (for each structure) for positive (*posQ*) and negative (*negQ*) potentials, and for a patch uniformly below the *kT*/*e* threshold (*nonQ*). Summing the patches in each of the three classes (positive, negative, neither positive nor negative), gives overall values for each protein. These are recorded as: (i) [*non/*(*pos*+*neg*)] (a measure of non-charged versus charged surface), (ii) *pos/neg* (the degree of positive versus negative surface), and (iii) *non*+*pos*+*neg* (a measure of protein size via surface area). Protein size is defined simply as the number of amino acids (*numAA*).

Further charge-based properties considered are the net charge per amino acid (*QperAA*), and the estimated charge-charge interactions, normalised to a per amino acid value (*QQperAA*). These interactions are calculated with a Debye-Hückel model (equation 1.6) in a uniform relative dielectric of 78.4, an ionic strength of 0.15 M, and with the amino acid side chain charges used in electrostatic field and patch calculations. A measure of the degree of non-polarity of a surface was included as a ratio of non-polar to polar (*nonpol/pol*) solvent accessible surface area (SASA), calculated both over the whole protein surface, and taken as the maximum value for all patches. Following previous work, patches were calculated for 13 Å spheres centred on all non-hydrogen atoms (Chapter 2; Warwicker *et al*., 2014). Contact order provides a measure of the extent of packing within a protein (Ivankov *et al*., 2003), and contact order as well as relative contact order were calculated and averaged over protein length to give *COperAA* and *RCOperAA*. Contact order was calculated based on C_α atoms only, with a contact radius of 7.3 Å (Bahar *et al*., 1997), and relative contact order included the residue distance (number of amino acids) between contacting C_α atoms. Relative contact order measures the average number of contacts (with other residues) around an amino acid within a given radius. Each contact is scaled by its distance from the central amino acid along the sequence. Contact order and relative contact order were also recorded as

minimum values for each protein over spherical patches with a 13 Å radius, given that localised loose packing could mediate partial unfolding and reduce solubility. The minimum is used in this case as it is anticipated that lower contact order would imply looser packing, increasing the scope of the protein or localised patch to unfold.

Furthermore, an atomic solvation parameter (ASP) based approach was used, following the work of Eisenberg and McLachlan (1986), averaging over the protein (*ASPperAA*). Here, atomic solvation parameters for each amino acid are multiplied by the SASA of that amino acid to obtain its contribution to solvation. More negative ASP values relate to more hydrophilic amino acids, and more positive to less polar and less favourable for water exposure. Again, a value for the whole protein was taken, along with one for the maximum of a 13 Å patch/sphere. The patch maximum records the least favourable solvent surface exposure of amino acids, in the ASP model, on the basis that this could lead to folded form conformational instability. The charge-based and contact-based structural features that were computed over entire protein structures as well as over defined patches are summarised in table 3.4 below.

Typically, a whole protein value for is used for structural properties (see table 3.4), although a patch value is added for instances in which there may be localised values that could correlate with measured solubilities, *e.g*. perhaps loose packing in a patch could lead to transient local unfolding and lead to aggregation with available monomeric species.

Table 3.4 Structural Features used in Predictive Model

Feature	Whole protein/patch	Description
posQ	patch max	Largest recorded positively charged patch
negQ	patch max	Largest recorded negatively charged patch
nonQ	patch max	Largest recorded uncharged patch
non/(pos+neg)	whole protein	Non-charged vs. charged surface
pos/neg	whole protein	Positively vs. negatively charged surface
non+pos+neg	whole protein	Surface area, *i.e.* $(\sum \text{positve, negative, uncharged})$
numAA	whole protein	Protein size
QperAA	whole protein	Net charge per amino acid
QQperAA	whole protein	Charge-charge interactions per amino acid
nonpol/pol	patch max whole protein	Surface non-polarity (calculated over entire surface)
COperAA	patch min whole protein	Contact order per amino acid
RCOperAA	patch min whole protein	Relative contact order per amino acid
ASPperAA	patch max whole protein	Solvation per amino acid

3.6.2 Heatmap Analysis of Structural Features

Comparison of sequence-based feature enrichment between soluble/high-abundance and insoluble/low-abundance proteins was done using a *z*-score difference. For structural features, a correlation value was used as a metric to compare high and low solubility or abundance proteins. The Pearson correlation coefficient R (-1 to 1) of a protein's solubility value (*E. coli*) or abundance value (PaxDb) with each structural feature (table 3.4) was computed. Two datasets were studied, including *E. coli* protein solubility (Niwa *et al*., 2009) and *E. coli* protein abundance (PaxDb). For each dataset, protein sequences having structural annotation were collected and calculations were performed for single protein chains from the PDB entry (single chain) and with the biological unit listed (biological unit). The results are displayed in the heatmap in figure 3.13). Raw data (Pearson coefficients) have been removed for readability but can be found in appendix 3E at the end of the chapter.

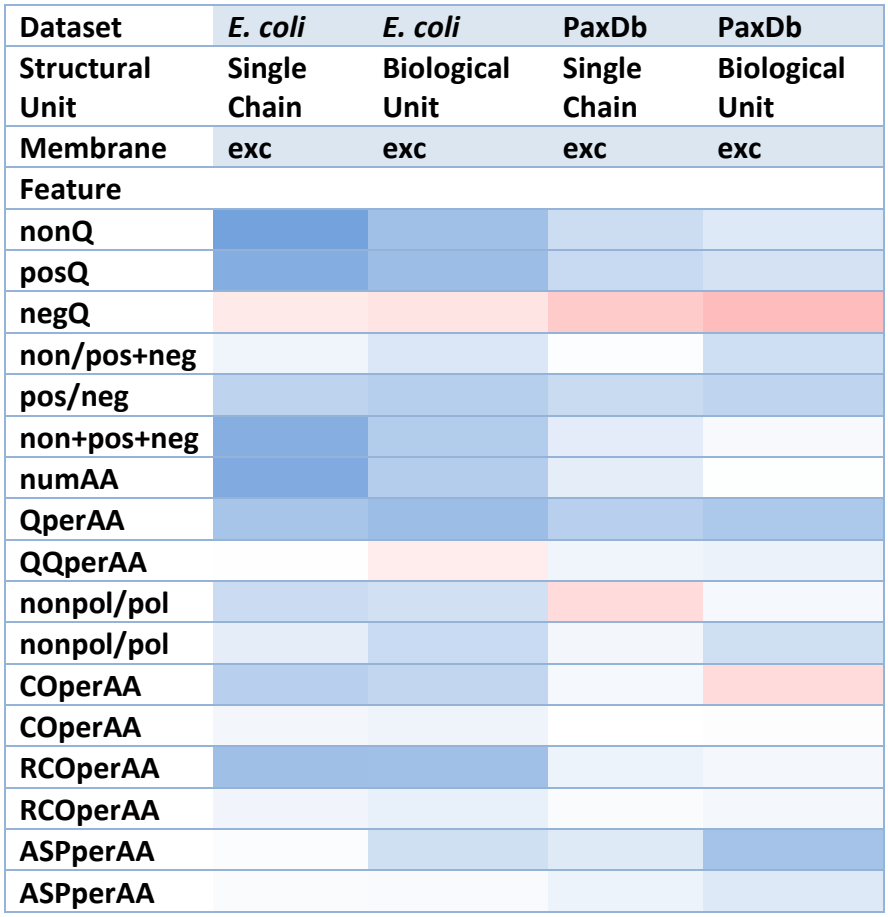


Figure 3.13. Heatmap of structure-based properties in *E. coli*. Pearson correlation coefficient R for structure-based properties and protein solubility/abundance scores. Red denotes that a structural feature is favoured for high solubility/abundance proteins, while blue denotes that one is favoured for low solubility/abundance proteins. Pearson correlation values are listed in appendix 3E.

Structural feature analysis was limited to *E. coli* proteins mainly due to the fact that solubility data was only available for this species (compared with abundance or concentration data), rendering this as perhaps the most reliable dataset to use as a benchmark. Importantly, there is significant overlap between the *E. coli* cell-free and PaxDb datasets (2434 sequences in common), so structural annotations (PDB entries) would largely overlap. Although the caveats of using *E. coli*-derived solubility data to benchmark non-bacterial protein abundance have been discussed, the findings presented in figure 3.13 suggest that enrichment patterns of *E. coli* surface charge- and polarity-based structural features (columns labelled ***E. coli***) are largely replicated in non-*E. coli* proteins (columns labelled **PaxDb**). This implies that using protein abundance data, which is currently far more accessible in the public domain than large-scale measurements relevant to solubility or native aggregation propensity, is reasonable to a certain extent until more refined data becomes commonplace.

3.6.3 Structure-based Solubility Trends

As is the case for sequence-based features, correlations are generally consistent between *E. coli*-based solubility and abundance data from multiple proteomes. Charge-based features and protein size give the best correlations. Similar results have been reported in a comprehensive study of properties that affect high-throughput analysis in structural genomics pipelines (Goh *et al*., 2004). Additionally, more sophisticated models employing molecular dynamics to identify structure-based descriptors have reported net charge as well as dipole moments to be key players in mediating aggregation propensities (Brunsteiner *et al*., 2013).

Large positively charged patches (*posQ*) appear enriched in insoluble (Chan *et al*., 2013) and less abundant proteins, as do large uncharged patches (*nonQ*). Conversely, large negatively charged patches are enriched in soluble and high abundance proteins. Patch observations are consistent with whole protein correlations. Non-charged vs. charged surface (*non*/*pos*+*neg*) has a weak negative correlation with solubility and abundance, showing that charged surface is preferentially negative for soluble and abundant proteins. Protein size (*numAA* and *non*+*pos*+*neg*) shows the same inverse correlation with solubility and abundance observed in sequence-based feature heatmaps. The net protein charge (*QperAA*) yields a slight preference for negative charge, whilst the predicted contribution of charge-charge interactions (*QQperAA*) shows little correlation with solubility or abundance.

Non-polar to polar surface ratio (either over whole protein or a patch maximum) is weakly inversely correlated with solubility, as expected, but shows no consistent variation with abundance. For the contact order (*COperAA* and *RCOperAA*) and atomic solvation (*ASP*) properties, patch extremes (minimum for contact order and maximum for solvation parameter) generally show less correlation with solubility and abundance than do whole protein values. The strongest and most consistent correlation for these features is with whole protein relative contact order (inverse correlation). A positive correlation might be expected for proteins that are more closely packed, since on average they could be more resistant to aggregation (*e.g*. via partial unfolding). As displayed in figure 3.14, it is likely that the observed inverse correlation for whole protein relative contact order is related to protein size, with larger proteins reducing their surface area to volume ratio and incorporating higher side chain separation.

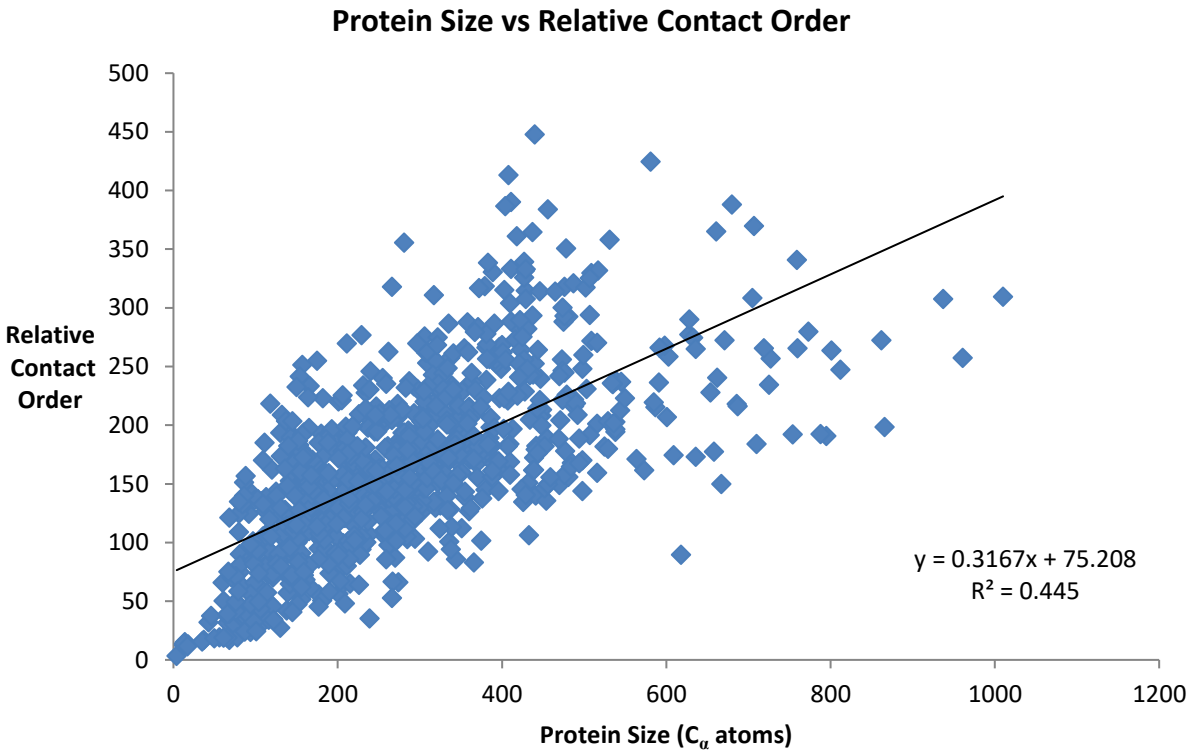


Figure 3.14. Protein size (C_α atoms) vs. Relative Contact Order. Whole protein relative contact order enrichment in insoluble *E. coli* proteins is likely due to the relative effect of RCO increasing with increasing protein size.

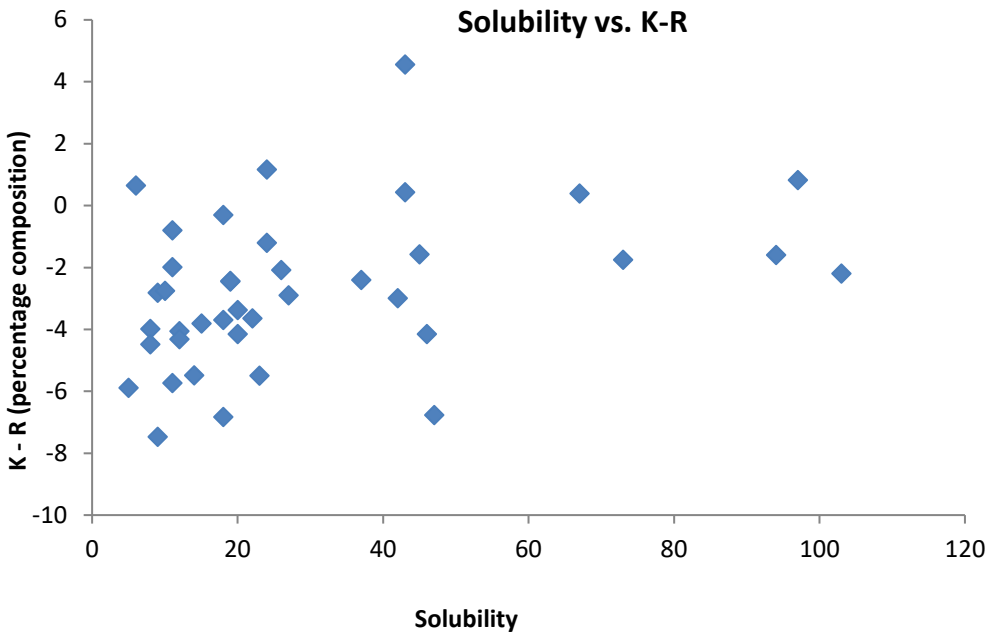
Whole protein solvation gives a moderate correlation for three of the four sets, with the expected negative sign where more negative solvation energies relative, on average, to more soluble/abundant proteins. However, the magnitude of these correlations is generally inferior to those for directed charge surface measures seen in the heatmap (figure 3.13). Mirroring the case for sequence analysis, charge features have the most distinguishable correlations with solubility. Furthermore, consistency between datasets containing solubility data and protein abundance is largely reproduced in structure-based features.

3.6.4 Variation of Sequence-/Structure-based Features in *E. coli* Paralogues

It has been observed that lysine to arginine ratios vary significantly between more and less abundant proteins in the serum albumin and myoglobin eukaryotic families of paralogous proteins (Warwicker *et al.*, 2014), with lysine being enriched in the more abundant ones. Quantitative proteomics techniques offer scope for larger scale studies focusing on paralogue families. In this context, a brief investigation of the largest family of paralogues based on a study of gene families in *E. coli* (Pushker *et al.*, 2004) was carried out. In this study, the relationship between genes belonging to paralogous families and bacterial genome size was investigated. The authors report

that the relative content of paralogues increases with genome size and that the size of a given gene family is remarkably similar in strains of the same species and in closely related species. The gene family having the greatest number of paralogues (52 members) encodes ATP-binding proteins, of which 39 can be cross-referenced with solubility data in *E. coli* (Niwa *et al*., 2009). These sequences were analysed to determine if there was a correlation between solubility and K-R (difference of lysine and arginine composition). Figure 3.15A shows that they are positively correlated.

A



B

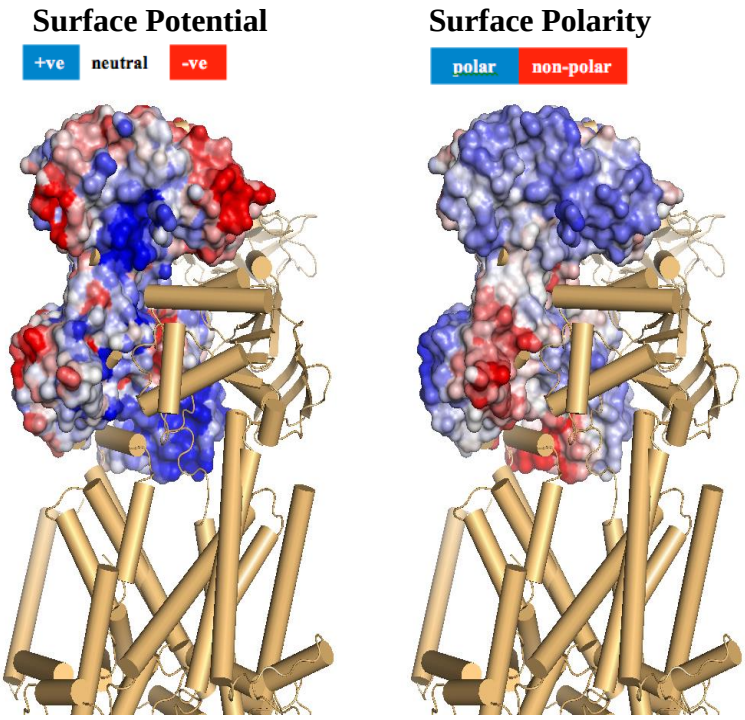


Figure 3.15. Sequence-based and Structure-based Features in *E. coli* Paralogues. (**A**). K-R vs. solubility as measured in (Niwa et al., 2009) for the *E. coli* family with the most paralogues (Pushker *et al*., 2004). Pearson correlation coefficient R = 0.37 (p = 0.02). (**B**) 3RLF, the crystal structure of a protein in complex with membrane is depicted here with electrostatic potential (left-hand side) and polarity (right-hand side).

As shown in the current work, several factors can contribute to solubility, but enrichment of arginine over lysine at low solubility for the ATP-binding family reinforces the view that relatively simple charge engineering could provide one route to improve solubility. The proteins encoded by this family exist in protein complexes (ATP-binding proteins bound to membrane transporters). A structural perspective is given in figure 3.15B. Very few structures for this proteins family are available, and they relate to low solubility variants. The structure depicted is that of a maltose-binding transporter complex (PDB 3RLF, UniProtKB P68187). This protein has been quantified in the Niwa study (19% solubility) and is shown with both electrostatic potential and non-polar surface. Both the largest positive surface region and the largest non-polar surface region are located at the protein-protein interfaces and at the protein-membrane interface in the naturally occurring complex. Either of these features could contribute to the low solubility of this protein. Additionally, it also has an excess of arginine over lysine (K-R = -2.43%). The extent to which engineering of an orthogonal feature (swapping arginine for lysine) could improve solubility is an interesting question for biotechnological and biopharmaceutical systems alike.

The idea of using relatively simple experimental protocols such as mutagenesis to improve solubility (*e.g.* swapping non-essential arginine residues to lysine) was introduced in the previous chapter, where the effects of KR-ratio on solubility were explored extensively. This principle of using "non-invasive" techniques (lysine to arginine mutations) as opposed to structurally disruptive interventions (altering surface polarity/hydrophobicity) to modify protein chemistry in terms of solubility is illustrated schematically in figure 3.16. The figure is meant for purely illustrative purposes.

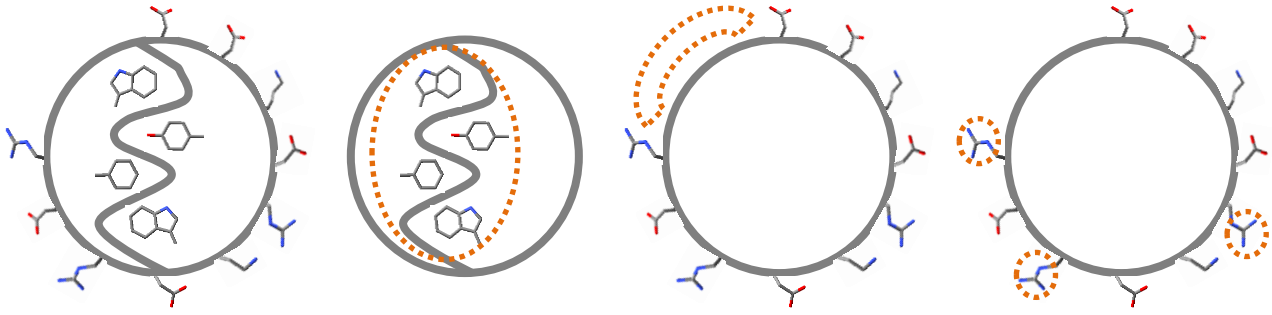


Figure 3.16. Modifying protein properties to optimise solubiliy. From left to right: (**A**) wild-type (aromatic side chains packed within protein core), (**B**) the protein core repacked with smaller non-polar side chains, (**C**) adding surface charge to a non-polar region, and (**D**) swapping (non-essential) arginine with lysine.

3.7 Conclusions

The findings discussed in this chapter suggest that features that discriminate protein solubility also correlate with protein abundance. This reinforces the view that naturally occurring abundance is an important feature in the evolutionary landscape, with proteins being under pressure to modulate their solubility in accordance with their functionally required concentration/abundance levels (Tartaglia *et al*., 2007; Tartaglia *et al*., 2009). This observation is particularly important in the bioprocessing field, where therapeutic proteins are often required to be produced, stored and used at concentrations well above those at which they occur naturally. For antibody-derived scaffolds, this may normally not be an issue, as immunoglobulins have evolved to exist at high concentration levels in plasma. However, other protein-based therapeutics such as engineered fragments (scFv's) and biologics have not undergone the same evolutionary pressures and hence would not be expected to have adapted their sequence and structure features to enable them to exist at artificially high concentrations. In this respect, two encouraging results are reported. First, protein abundance data can be used to seed generation of predictive models, since large-scale solubility data remains scarce in the public domain. Second, certain features appear to have increased importance in discriminating solubility; in particular, the findings from this chapter suggest that lysine/arginine engineering could offer an aid to improve protein solubility and reduce aggregation propensities.

Of the sequence and structure-based features that correlate best with protein solubility and abundance in this work, two (protein length and sequence entropy) cannot be changed without fundamentally altering the protein. However, protein length is effectively altered when partial constructs or individual domains are used in a popular reductionist approach for studying protein structure/function relationships (Vogel *et al*., 2004). It was hypothesised earlier that sequence entropy may be flagged due to an underlying higher average solubility of intrinsically disordered proteins. Decreased sequence entropy in IDPs might be due to unequal sampling of amino acids in their primary sequence, in which polar groups would be enriched non-polar groups depleted. Interestingly, a strong inverse correlation ($R = 0.97$) between the IDP content of a proteome and its aggregation load has been observed by Rousseau and colleagues (2006), where the aggregation load is defined as the percentage of residues having a high probability to aggregate (Fernandez-Escamilla *et al*., 2004). However, even if this holds true, it cannot be used practically in a solubility engineering context as altering a protein of interest towards IDP (with possibly increased solubility) would likely disrupt functional domains.

Figure 3.16 depicts some of the most important features related to solubility and abundance that are highlighted in the current work. Amino acids possessing aromatic side chains could be replaced with other hydrophobic residues (*e.g.* valine, which is enriched in high solubility/abundance proteins in almost all datasets that were studied), although such re-packing of a protein core would be a disruptive design strategy (figure 3.16B). Protein surface regions lacking charged amino acids could be modified to incorporate negative charge (altering pI) or they could be designed to balance positive and negative charges (maintaining pI) (figure 3.16C). The most conservative and least invasive modification would be mutating non-essential arginine residues to lysine residues (figure 3.16D). In practice, it will be of interest to compare the effects of such modifications, how they influence each other and how they influence partial unfolding, widely considered to be necessary for irreversible aggregation (Chaudhuri *et al*., 2014).

One key aspect could involve reducing the strength of non-specific protein-protein interactions, whether mediated in folded or partially unfolded states. The preliminary analysis of structural stability presented here, using simple contact order measures, has not revealed a clear relationship with solubility and abundance. This can perhaps be partially attributed to the fact that the current model was limited to *E. coli* solubility and abundance data. Structure-based prediction of protein solubility is more demanding than using sequence-based features, as it requires experimentally obtained high-resolution 3D structures, which are often hard to acquire for aggregation-prone proteins. In more formal structure-based prediction models, the free energy difference between aggregation and solution phases is computed (Habibi *et al*., 2014). This arguably offers greater scope for understanding how important features such as non-polar and charged protein surfaces contribute to non-specific interactions that promote aggregation and loss of solubility. Given that under currently available technologies the rate of structure determination will significantly lag that of sequence acquisition, there is scope in using both sequence- and 3D structure-based descriptors.

Although several interesting sequence-based and structure-based correlations were observed, it is important to establish that protein abundance as a physical quantity provides a proxy to solubility, which has a more formal thermodynamic definition. This stands as perhaps the most salient shortcoming of the current predictive model. Given that protein abundance cannot fully supplant solubility, any predictive effort relying on the former will be constrained in this respect. Nonetheless, using solubility data to the extent possible and comparing with abundance or expression measurements appears to be the only model validation method currently available. Another important constraint when seeding predictive models with protein abundance and mRNA expression data is the spatial and temporal regulation of gene expression. Cells express different proteins at different times throughout their life cycle, and this means that measurements of protein

or mRNA abundance levels are closely related to the cell cycle phase. In this respect, with the possible exception of housekeeping genes, protein and mRNA abundance measurements would provide a "snapshot" of the proteome, depending on which proteins are functionally required, which would likely be different if measured at another cell cycle phase. Further to this, higher organisms such as eukaryotes have evolved an enormous diversity of cells with specialised functions (*e.g.* tissues, organs, organ systems), with a commensurate differentiation in terms of proteome composition.

Perhaps one way to mitigate this issue is to focus high-throughput proteomics data on biochemical pathways that are ubiquitous throughout all domains of life, *e.g.* glycolysis. The metabolic pathway for converting glucose into pyruvate whilst releasing free energy is a universal pathway used for synthesis of high-energy compounds (ATP and NADH). Biochemical pathways related to vital processes often require cellular machinery regulated by housekeeping genes. Proteins encoded by such genes are required constantly by a cell, and hence could offer a useful benchmark against which to compare protein abundance levels on a proteome-wide scale. Even within the confines of non-solubility data such as protein and mRNA abundance, it is worth augmenting the current predictive model to include more species. This would not present a major challenge, as public domain proteomic data is constantly increasing in volume. Of particular interest would be to incorporate extremophile proteomes into the model, as these species would be expected to have evolved housekeeping proteins adapted to function under environments of extreme temperature, pressure or pH (*e.g.* thermophiles, piezophiles, acidophiles and alkalophiles). High-throughput studies of such organisms would provide useful insight for investigating the sequence-based and structural modifications adopted by proteins sustaining life under extreme conditions, *e.g.* thermostable proteins. Moreover, comparing the findings of our model with other similar efforts in the field can aid in validating the methodology. The SPiCE web tool (van den Berg *et al.*, 2014) offers sequence-based prediction of properties such as subcellular localisation and solubility using classification algorithms, and uses the same fungal (van den Berg *et al.*, 2012) and yeast (Ghaemmghami *et al.*, 2003) datasets that were used in the current work for validation.

Whilst it is true that the increasing amount of quantitative proteomics data provides an important new benchmark for computational studies of protein solubility, it is clear that there is also a requirement for benchmark datasets of solubility data, such as that reported by Niwa and colleagues (2009), and biophysical characterisation of aggregation, particularly for proteins used in therapeutic capacities.

Appendix 3A. Heatmap 1: Standard *z*-score Differences for *E. coli* Dataset

Dataset	*E. coli*	*E. coli*
Membrane	MEM exc	MEM inc
Feature		
K-R	0.654	0.595
D-E	-0.229	-0.246
naa	-0.886	-0.786
A	0.089	0.091
C	-0.109	0.022
D	0.384	0.553
E	0.597	0.736
F	-0.323	-0.44
G	-0.025	-0.07
H	-0.348	-0.159
I	0.159	-0.056
K	0.651	0.675
L	-0.303	-0.467
M	0.104	-0.03
N	-0.031	0.027
P	-0.255	-0.198
Q	-0.109	-0.035
R	-0.331	-0.208
S	-0.251	-0.262
T	0.054	0.071
V	0.348	0.187
W	-0.436	-0.515
Y	-0.392	-0.35
K+R	0.255	0.358
D+E	0.747	0.891
K+R-D-E	-0.306	-0.426
K+R+D+E	0.649	0.793
F+W+Y	-0.575	-0.651
pI	-0.378	-0.541
KyteDoo	-0.18	-0.476
abs-charge	0.794	0.778
FoldIndex	-0.407	-0.619
disorder	-0.148	0.082
entropy	-0.921	-0.559
betapropensity	-0.139	-0.425

Appendix 3B. Heatmap 2: Standard *z*-score Differences for Proteomics Datasets

Dataset	E.coli	E.coli	Yeast[1]	Yeast[1]	Yeast[2]	Yeast[2]	Yeast[2]	Yeast[2]	Yeast[3]	Yeast[3]	A.niger	A.niger	SOLP	SOLP
Membrane	exc	inc	exc	inc	exc	inc	exc	inc	exc	inc	exc	inc	exc	inc
Feature														
K-R	0.654	0.595	0.134	0.177	0.161	0.214	-0.25	-0.3	0.233	0.234	0.224	0.229	0.08	0.072
D-E	-0.23	-0.25	-0.106	-0.13	-0.15	-0.23	-0.36	-0.32	-0.05	0.011	0.488	0.469	-0.18	-0.16
naa	-0.89	-0.79	-0.443	-0.39	-1.73	-1.8	-1.93	-1.84	-0.21	-0.139	0.454	0.313	-0.2	-0.23
A	0.089	0.091	0.888	0.795	1.277	1.213	0.879	0.792	0.582	0.453	0.047	-0.004	0.055	0.048
C	-0.11	0.022	-0.082	-0.1	-0.2	-0.24	0.335	0.453	-0.23	-0.269	0.08	-0.062	-0.08	-0.1
D	0.384	0.553	0.017	0.073	-0.55	-0.49	-0.6	-0.75	0.105	0.157	0.347	0.368	-0.02	-0.02
E	0.597	0.736	0.119	0.187	-0.28	-0.17	-0.04	-0.23	0.13	0.116	-0.356	-0.291	0.209	0.188
F	-0.32	-0.44	-0.047	-0.12	-0.18	-0.39	-0.15	-0.2	-0.1	-0.063	0.215	0.184	-0.17	-0.16
G	-0.03	-0.07	0.744	0.657	1.123	1.068	0.83	0.676	0.595	0.484	0.354	0.541	0.102	0.102
H	-0.35	-0.16	-0.236	-0.2	0.262	0.238	-0.05	-0.18	-0.27	-0.184	-0.332	-0.396	-0.09	-0.08
I	0.159	-0.06	-0.117	-0.15	-0.32	-0.35	-0.07	0.105	-0.16	-0.138	0.05	-0.002	-0.09	-0.08
K	0.651	0.675	0.157	0.19	0.059	0.087	-0.26	-0.29	0.194	0.186	-0.608	-0.668	0.095	0.102
L	-0.3	-0.47	-0.203	-0.22	-0.18	-0.22	-0.06	-0.05	-0.13	-0.084	-0.533	-0.543	0.062	0.076
M	0.104	-0.03	-0.175	-0.2	-0.19	-0.22	-0.13	0.001	-0.18	-0.123	-0.426	-0.536	-0.26	-0.25
N	-0.03	0.027	-0.737	-0.63	-0.44	-0.44	-0.54	-0.45	-0.38	-0.288	0.837	0.858	-0.11	-0.12
P	-0.26	-0.2	-0.235	-0.18	0.036	0.111	0.142	0.067	-0.19	-0.156	-0.308	-0.325	0.034	0.06
Q	-0.11	-0.04	-0.141	-0.09	0.05	-0.01	-0.35	-0.2	-0.03	0.001	0.243	0.144	0.051	0.044
R	-0.33	-0.21	0.004	-0.01	-0.24	-0.3	0.036	0.079	-0.1	-0.1	-0.75	-0.805	-0.03	-0
S	-0.25	-0.26	-0.6	-0.58	-0.7	-0.54	-0.44	-0.33	-0.39	-0.375	-0.101	-0.024	-0.09	-0.08
T	0.054	0.071	-0.148	-0.12	-0.27	-0.16	0.029	0.207	-0.13	-0.152	0.29	0.362	-0.07	-0.07
V	0.348	0.187	0.686	0.565	0.857	0.972	0.622	0.595	0.498	0.388	0.024	0.05	0.062	0.049
W	-0.44	-0.52	-0.068	-0.11	0.386	0.187	-0.29	-0.08	-0.05	-0.015	0.357	0.358	-0.02	-0
Y	-0.39	-0.35	-0.024	-0.06	-0.27	-0.42	0.378	0.244	-0.21	-0.143	0.748	0.825	-0.17	-0.17
K+R	0.255	0.358	0.109	0.126	-0.07	-0.07	-0.17	-0.17	0.082	0.08	-0.922	-1.016	0.068	0.09
D+E	0.747	0.891	0.089	0.16	-0.47	-0.37	-0.33	-0.53	0.142	0.157	0.027	0.061	0.163	0.141
K+R-D-E	-0.31	-0.43	0.019	-0.02	0.324	0.253	0.099	0.265	-0.04	-0.06	-0.831	-0.943	-0.07	-0.04
K+R+D+E	0.649	0.793	0.147	0.201	-0.39	-0.3	-0.32	-0.43	0.157	0.155	-0.591	-0.627	0.154	0.156
F+W+Y	-0.58	-0.65	-0.062	-0.14	-0.13	-0.38	0.027	-0.03	-0.19	-0.112	0.741	0.768	-0.21	-0.2
pI	-0.38	-0.54	-0.243	-0.26	0.223	0.084	-0.08	0.114	-0.19	-0.194	-0.914	-1.001	-0.11	-0.09
KyteDoo	-0.18	-0.48	0.372	0.185	0.498	0.395	0.693	0.656	0.155	0.071	0.022	0.01	-0.09	-0.09
abs-charge	0.794	0.778	0.404	0.366	0.269	0.347	0.709	0.676	0.156	0.146	0.622	0.694	0.206	0.264
FoldIndex	-0.41	-0.62	0.154	0.042	0.325	0.234	0.378	0.402	0.08	0.023	-0.192	-0.22	-0.15	-0.17
disorder	-0.15	0.082	-0.364	-0.26	-0.15	-0.06	-0.18	-0.25	-0.14	-0.113	0.349	0.489	0.009	0.021
entropy	-0.92	-0.56	-0.253	-0.2	-0.45	-0.61	-0.38	-0.47	-0.02	0.038	0.261	0.2	-0.44	-0.5
betapropensity	-0.14	-0.43	0.179	0.029	0.086	0.004	0.38	0.479	-0.05	-0.068	0.286	0.223	-0.17	-0.17

Appendix 3C. Heatmap 3: Standard *z*-score Differences for PaxDb Datasets

Dataset	*E. coli*	*E. coli*	Yeast	Yeast	Mouse	Mouse	Human	Human
Membrane	exc	exc	exc	exc	exc	exc	exc	exc
Feature								
K-R	0.864	0.486	0.417	0.401	0.788	0.392	0.623	0.26
D-E	-0.574	-0.103	-0.525	-0.209	0.248	0.106	0.353	0.174
naa	-0.673	-0.216	-0.282	-0.01	-0.88	-0.552	-0.572	-0.27
A	0.453	0.167	1.266	0.851	0.288	0.243	0.29	0.249
C	-0.852	-0.42	-0.769	-0.601	-0.54	-0.297	-0.526	-0.278
D	-0.258	0.386	0.025	0.357	0.354	0.266	0.553	0.438
E	0.533	0.422	0.581	0.501	-0.028	0.049	-0.022	0.069
Feature	-0.357	-0.132	-0.705	-0.419	0.216	0.141	0.1	0.026
G	0.503	0.351	1.089	0.664	0.326	0.239	0.362	0.193
H	-0.399	-0.224	-0.403	-0.373	-0.646	-0.466	-0.835	-0.597
I	0.244	0.099	-0.467	-0.283	0.798	0.461	0.644	0.334
K-R	1.466	0.476	0.777	0.312	1.049	0.489	0.632	0.252
L	-1.049	-0.39	-0.703	-0.322	-0.69	-0.349	-0.567	-0.245
Membrane	0.213	0.082	-0.321	-0.237	0.657	0.427	0.316	0.266
naa	-0.219	-0.082	-0.888	-0.338	0.496	0.194	0.361	0.091
P	-0.75	-0.212	-0.061	-0.088	-0.992	-0.419	-0.685	-0.251
Q	-0.673	-0.349	0.01	0.105	-0.636	-0.271	-0.534	-0.183
R	0.165	-0.237	0.258	-0.242	-0.021	-0.063	-0.265	-0.135
S	-0.462	-0.412	-1.182	-0.538	-1.102	-0.68	-0.672	-0.421
T	-0.026	0.019	-0.399	-0.156	0.034	-0.023	0.204	0.002
V	1.053	0.462	0.791	0.462	0.548	0.34	0.802	0.438
W	-1.271	-0.595	-0.409	-0.257	-0.381	-0.09	-0.334	-0.229
Y	-0.677	-0.22	-0.481	-0.278	0.46	0.245	0.335	0.174
K+R	1.055	0.185	0.714	0.099	0.809	0.353	0.376	0.13
D+E	0.279	0.589	0.399	0.526	0.141	0.164	0.245	0.257
K+R-D-E	0.632	-0.242	0.238	-0.333	0.462	0.142	0.099	-0.108
K+R+D+E	1.161	0.523	0.93	0.441	0.709	0.334	0.448	0.253
F+W+Y	-1.058	-0.426	-0.872	-0.507	0.254	0.183	0.13	0.03
pI	0.382	-0.42	-0.152	-0.477	0.199	-0.012	-0.22	-0.265
KyteDoo	-0.273	-0.074	-0.26	-0.157	0.001	0.086	0.2	0.125
Abs-charge	0.975	0.331	0.683	0.121	0.658	0.304	0.339	0.115
FoldIndex	-0.574	-0.182	-0.473	-0.181	-0.247	-0.016	0.041	0.075
Disorder	-0.345	0.037	-0.242	-0.004	-0.611	-0.333	-0.398	-0.207
Rntropy	-1.366	-0.516	-0.651	-0.135	-0.345	-0.198	-0.23	-0.176
Beta-propensity	0.1	-0.097	-0.625	-0.456	0.575	0.284	0.482	0.165

Appendix 3D. Heatmap 4: Standard *z*-score Differences for PPD Datasets

Dataset	*H. sapiens*	*H. sapiens*	*H. sapiens*	*H. sapiens*
Membrane	exc	inc	exc	inc
Cutoff	8%	8%	33%	33%
Feature				
K-R	-0.055	0.064	-0.019	0.045
D-E	0.111	0.045	-0.012	0.041
naa	0.153	0.218	-0.021	-0.029
A	0.085	-0.021	-0.02	-0.096
C	-0.278	-0.455	0.134	0.128
D	0.165	0.418	-0.038	0.066
E	-0.036	0.203	-0.008	-0.003
F	0.074	0.313	-0.168	-0.027
G	0.495	0.364	0.071	0.063
H	-0.167	-0.097	-0.01	0.008
I	-0.21	-0.209	-0.175	-0.115
K	-0.274	-0.094	-0.209	-0.091
L	-0.157	-0.117	0.133	0.074
M	-0.783	-0.596	-0.399	-0.301
N	0.459	0.264	0.278	0.217
P	-0.256	-0.333	-0.133	-0.085
Q	-0.067	-0.039	0.076	0.001
R	-0.186	-0.22	-0.26	-0.212
S	0.049	-0.084	0.247	0.12
T	0.033	0.153	0.104	0.102
V	0.463	0.244	0.113	0.014
W	0.288	0.147	0.147	0.082
Y	0.619	0.56	0.207	0.179
K+R	-0.402	-0.258	-0.397	-0.236
D+E	0.047	0.384	-0.023	0.029
K+R-D-E	-0.356	-0.537	-0.282	-0.211
K+R+D+E	-0.193	0.094	-0.218	-0.107
F+W+Y	0.52	0.568	0.056	0.107
pI	-0.63	-0.66	-0.323	-0.211
KyteDoo	0	-0.188	0.093	0.019
abs-charge	-0.844	-0.714	-0.161	-0.115
FoldIndex	0.207	0.009	0.123	0.047
disorder	0.199	0.09	0.061	0.088
entropy	0.275	0.292	0.047	0.05
beta-propensity	0.1	0.024	0.065	0.036

Appendix 3E. Heatmap 5: Pearson Correlation Coefficients for *E. coli* Datasets

Dataset	*E. coli*	*E. coli* Biological	PaxDb	PaxDb Biological
Structural Unit	Single Chain	Unit	Single Chain	Unit
Membrane	exc	exc	exc	exc
Feature				
posQ	-0.28562401	-0.230103024	-0.12669922	-0.098586198
negQ	0.033660047	0.043722025	0.084173496	0.106633169
non/pos+neg	-0.03401098	-0.083598346	-0.00614793	-0.1106799
pos/neg	-0.15267312	-0.167754345	-0.12527562	-0.149969665
non+pos+neg	-0.28093812	-0.17861242	-0.0644884	-0.018360734
numAA	-0.29507513	-0.172341341	-0.06016326	-0.002709778
QperAA	-0.20256557	-0.230492561	-0.16459798	-0.191633732
QQperAA	0.002202934	0.028966891	-0.0331623	-0.043450172
nonpol/pol	-0.120866	-0.105924292	0.05763193	-0.022515023
nonpol/pol	-0.05913202	-0.123628732	-0.02703408	-0.108519578
COperAA	-0.16445505	-0.141443646	-0.02177168	0.056599656
COperAA	-0.02568647	-0.03702365	0.000597628	-0.003624388
RCOperAA	-0.22205272	-0.219227161	-0.04113504	-0.023592388
RCOperAA	-0.03164078	-0.05137861	-0.01301226	-0.025353277
ASPperAA	-0.01030292	-0.106998482	-0.07297629	-0.207036063
ASPperAA	-0.01391885	-0.015455806	-0.03962052	-0.075185396

Chapter 4. Protein-Excipient Interactions

4.1 Objectives

This chapter serves as a preliminary study of protein-excipient interactions in aqueous solutions. A set of crystallisation-specific chemical compounds is used an excipient dataset for which interactions with PDB-annotated proteins is investigated. Protein structures were used to carry out a low-resolution analysis on interactions with excipient for compounds spanning several chemical classes.

An overview of the mechanistic basis of excipient stabilisation effects is first presented for several compounds used as excipients in protein-based formulations. An EBI tool (PDBeXpress) is then used to collect information on the contact environment between each excipient of the crystallisation set and proteins with structural representation in the PDB. This information is provided in the form of side chain contacts. A statistical analysis of the protein-excipient contact environment is undertaken by using a dot product metric to measure the similarity between the PDBeXpress environment and the surface of a protein that is found to be in contact with. Hence similarity is measured in terms of side chain contacts of naturally occurring amino acids. Statistical significance of the similarity score is assigned using a brute force sampling method with a Bonferroni correction applied to compensate for multiple sampling. Finally, contact information for each excipient of the crystallisation set is tested against the bulk of the PDB database.

The rationale behind this work was to introduce a structural aspect to the scope of predicting solubility or aggregation propensity in the context of therapeutic formulation development. Although the work undertaken here is only a cursory study, the goal was to determine whether any type of chemical reproducibility exists in the interaction of different excipient compounds with proteins. Hence the main objective was to establish a foundation upon which future work could build in order to expand the Protein-Sol prediction server (section 3.5.4).

4.2 Excipients

Protein-based therapeutics and engineered antibody fragments in particular, have established a prominent role in the biopharmaceutical industry as an effective method for treatment of a wide spectrum of diseases (Holliger and Hudson, 2005). Despite their increasing importance in the field, stabilisation of protein-based therapeutics remains a challenge. Indeed, proteins used as standalone

therapeutic agents in clinical applications are only marginally stable and highly prone to physical degradation (Kamerzell *et al*., 2011). To address this problem, a wide range of excipients is often used as additives in formulations in order to stabilise proteins by inhibiting aggregation. The term excipient refers to both specific and non-specific interaction partners of a protein, often associated with conferring a stabilising effect under certain known storage conditions. Understanding protein-excipient interactions is a crucial step in the development of formulations containing active protein-based therapeutic agents. The overall goal of protein formulation development is to transform a purified protein solution into a stable and efficacious biopharmaceutical agent (active drug or vaccine) for administration to patients.

The use of targeted pharmaceutical excipients to enhance factors such as protein stability, solubility, and biological activity at varying concentrations is vital to increasing formulation shelf life. Excipients have a wide array of uses beyond stabilising protein structures; they are used to aid in manufacture of the dosage form, as part of drug delivery systems in the body, as well as to provide tonicity for injected formulations (Kamerzell *et al*., 2011). Although it is usually the case that pharmaceutical excipients are biologically inert, certain additives may possess toxicological activity. Hence the safety profiles of a novel protein product (active drug) and stabilising excipient must be considered together, since the formulation that is tested in clinical trials consists of the combination of drug and excipient. The result is that excipients used in a pharmaceutical capacity must abide to safety requirements typically established by regulatory agencies. One such classification system is the GRAS (**G**enerally **R**ecognized **a**s **S**afe), an FDA (Food and Drug Administration) designation that a substance is recognized as safe for use as a food additive (Gaynor *et al*., 2006).

Numerous formulation additives, termed osmolytes, have been shown to enhance protein stability and as a consequence mitigate aggregation of stable proteins. Osmosis is the process by which solvent molecules in a solution experience a net movement across a semi-permeable membrane (such as the cellular plasma membrane) into a region of higher solute concentration. In biological systems, osmosis is vital as it defines how certain molecules can traverse the membrane while others are obstructed. Osmotic pressure is the external force required so that there is no net movement of solvent molecules across the membrane. Osmolytes are solutes utilized in nature to raise the osmotic pressure of cellular environments and to ensure macromolecular function and viability (Yancey *et al*., 1982) by protecting proteins against inactivation (Domenico and Levvacia, 2000). In cellular environments, protein unfolding precedes aggregation, and the structure-stabilising excipients reduce aggregation by stabilising the native structure (Ohtake *et al*., 2011). Excipients encompass a wide array of chemical compounds, including sugars, salts, polymers, buffers, surfactants, and amino acids (most notably arginine). They are classified as either structure

stabilising or solubility enhancing, depending on the mechanistic basis of their interaction with proteins.

For each class of compound, the underlying mechanisms that mediate non-specific interactions can be considered separately for solution (aqueous) and freeze-dried (post-lyophilisation) state. Liquid and dried states of protein products differ markedly. In the absence of water, fundamentally different mechanisms are at play, as any excipient-water interactions are not applicable. Lyophilisation (also known as freeze-drying) is a dehydration process used in the preservation of protein products (with applications to other perishable materials), which consists of freezing a material and consequently reducing the pressure in the surrounding medium. This allows frozen water to undergo sublimation and transition from solid phase into gas phase. Lyophilisation is commonly used in the manufacture of protein products, particularly as they are frequently unstable in the aqueous phase (Pikal, 1990). Freeze-dried formulations are less prone to shear-induced denaturation and precipitation than their aqueous counterparts, and have been reported to undergo less pH-induced or temperature-induced hydrolysis reactions (Ohtake *et al*., 2011). However, excipients must be employed to stabilise the protein effectively against the stress associated with freezing and drying processes.

In the current work, which required structural data in the form of X-ray crystals, focus was placed on crystalline phase interactions. However, as noted previously, freeze-drying is widely used in formulation processes, and as such protein-excipient interaction mechanisms are studied extensively in non-aqueous states as well.

4.2.1 Protein Stabilisers

Protein-stabilising co-solvents encompass saccharides, salts, amino acids, amines, and buffering agents. Each class of compounds has an extended history of use in the scope of enhancing stability, more recently for therapeutic protein manufacturing. This renders it difficult to select one compound class over the other, especially as there is some overlap between the constituents of each class; rather, it is perhaps more useful to acknowledge that they can be employed somewhat interchangeably (Ohtake *et al*., 2011). This section focuses on sugars and salts, some of the earliest compounds known and applied as excipient additives.

Saccharides and Carbohydrates

Sugars were among the first compounds whose stabilising effects were observed when present at high concentrations during purification (Lee and Timasheff, 1981). Among saccharide

compounds, sucrose and trehalose have been most frequently used, as they have been shown in certain applications to be highly effective in increasing the melting temperature of proteins (Tiwari and Bhat, 2006). These compounds have prominent roles in nature as osmolytes in extremophiles living in severe environments, *e.g.* high temperatures and high pressures. In a more bioprocessing-relevant context, sorbitol has been shown to increase the melting temperature of human IgG and reduce its aggregation during the heating process, which is used for viral inactivation (Gonzalez *et al.*, 1995). Polyols (alcohols containing multiple hydroxyl groups) have been shown to increase the unfolding temperature of several antibody molecules where extent of stabilisation increased with higher polyol concentration (Sek, 2008). Another important observation was made in hemoglobin, a tetrameric protein whose subunits readily dissociate and aggregate under thermal stress (Antonini and Brunori, 1971). Sorbitol and sarcosine stabilise hemoglobin against heat-induced dissociation and subsequently reduce aggregation (Domenico and Lavecchia, 2000). In Figure 4.1 below a plot of the log aggregation rate constant against osmolyte concentration is shown, adapted from Domenico and Lavecchia (2000).

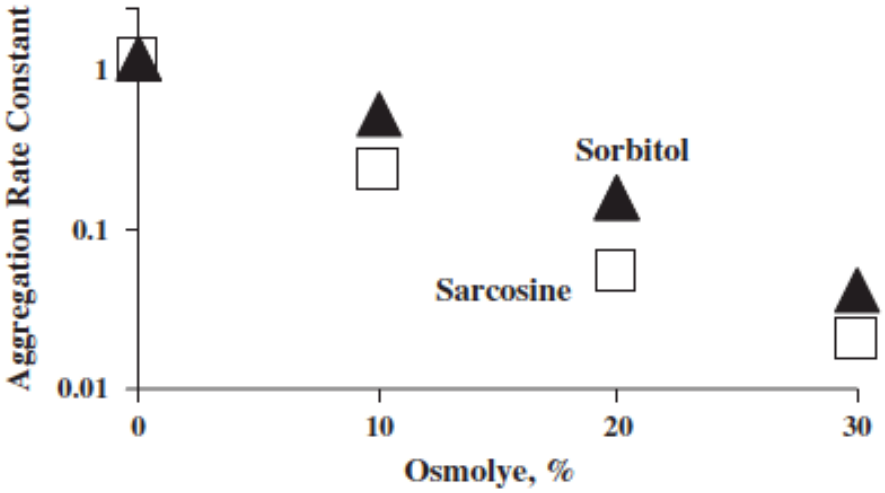


Figure 4.1. **Aggregation Rate vs. Osmolyte Concentration.** The effect of osmolyte concentration on the aggregation rate constant (log) of hemoglobin illustrated on a semi-logarithmic plot. Sorbitol and sarcosine were examined in a concentration range between 0 – 30%. Figure adapted from (Domenico and Lavecchia, 2000)

It is clear that the aggregation rate was greatly reduced in the presence of both osmolytes, with sarcosine being slightly more effective than sorbitol. The aggregation rate was measured in terms of protein monomer half-life, which corresponds to the time required to reach half the initial monomer concentration. More specifically, it was defined as the ratio of protein monomer half-life in presence of additives to that half-life in pure buffer under the same temperatures. These results are in line with stabilisation effects observed in other globular proteins (Back *et al.*, 1979; Gekko, 1982; Xie and Timasheff, 1997; Cioci and Lavecchia, 1998).

Salts

The effect of salts on protein stability has been studied as early as the nineteenth century, and extensively thereafter ever since the discovery of the Hofmeister effect (Hofmeister, 1888). The Hofmeister series is a classification of ions based on their ability to either "salt in" or "salt out" proteins. The origin of this series has been attributed to the structural changes that ions impose on the network of surrounding water molecules. Salting in refers to the effect where increasing salt concentration increases protein solubility; conversely, salting out reflects a reduction in protein solubility with increasing salt concentration. Anions appear to have a greater effect than cations, although the mechanism of the Hofmeister phenomenon is not well understood. The interacting anions and cations are referred to as kosmotropes or chaotropes depending on their effect on proteins. Kosmotropes stabilise proteins and hydrophobic aggregates in solution and reduce the solubility of hydrophobes. They are strong salting out agents and are effective at increasing protein melting temperature and folding stability. Chaotropes unfold proteins, destabilise hydrophobic aggregates and increase the solubility of hydrophobes. Salts are commonly present in protein formulations, typically in the form of buffers, and enhance protein stability by increasing the surface tension of interacting water molecules (Ohtake *et al*., 2011).

A notable example of salt excipients having a stabilising effect on a protein product is that of KGF (Keratinocyte Growth Factor), an approved treatment for oral mucositis. KGF has a strong tendency to aggregate in solution due to its inherent instability (Chen and Arakawa, 1996), as it begins to melt at ~ 40°C in 10 mM phosphate at pH 7.0. Various protein stabilisers were tested in this study to enhance the thermal stability of the protein in the phosphate buffer. Dramatic improvement in thermal stability and shelf life were conferred by ammonium sulfate and sodium phosphate (both salting out salts). The two salts were highly effective thermal stabilisers, raising the melting temperature (T_o) by ~ 10°C and ~ 12°C at 0.2 and 0.5 M concentrations, respectively. In the context of protein solutions, melting temperature is defined as the temperature at which the free energy of unfolded and folded states is equal, *i.e*. half the population is folded and the other half is unfolded. The enhancement was even more pronounced when considering shelf life (Ohtake *et al*., 2011).

Protein-Excipient Interaction Mechanisms

The protein-solvent interaction interface has been studied extensively, and is crucial for understanding the mechanisms of protein stabilisation by co-solvents. The structure-stabilising

properties of co-solvents with regards to proteins in solution have been studied and characterized by Timasheff and colleagues (1998, 2002). Protein stabilisation is conferred by a non-binding mechanism which plays a fundamental role in biological systems with high osmotic pressure (Yancey *et al*., 1982). Broadly speaking, four distinct but inter-related mechanisms have been postulated to describe co-solvent stabilisation effects: the (i) surface tension effect, (ii) excluded volume effect, (iii) peptide bond interaction effect, and (iv) preferential interaction effect. All mechanisms involve interactions with water molecules in some manner (Ohtake *et al*., 2011). These mechanisms are overviewed below.

Surface Tension Effect

Excipients increase the surface tension of water molecules, exerting a cohesive force on the solvent. Surface tension is the elastic tendency of a fluid that constrains it to acquire the least surface area possible. This cohesion was observed by Traube (1910) and was termed attraction pressure. The effect of this phenomenon is that salts to be preferentially excluded (discussed in the following sections) from the protein surface.

Excluded Volume Effect

The phenomenon of excluded volume originates from the very simple consideration that two molecules cannot occupy the same space in solution. As a result of steric hindrance, each macromolecule is expcted to exclude other molecules from its neighbourhood (Ralston, 1990). In the simplest case of spherical particles whose spatial positions are completely specified by the position of their centres, the closest two such molecules can approach each other is a distance equal to the sum of their radii (Kuznetsova *et al*., 2015). This has the effect of generating a spherical excluded volume around each molecule that is inaccessible to centres of all other molecules in solution. This mechanism has been predominantly used to explain the influence of large, polymeric molecules on protein stability and solubility. Molecules larger than water possess a region of "excluded volume", *i.e.* a volume that is inaccessible to other molecules in the system as the result of the presence of a macromolecule. This concept is illustrated in figure 4.2 below.

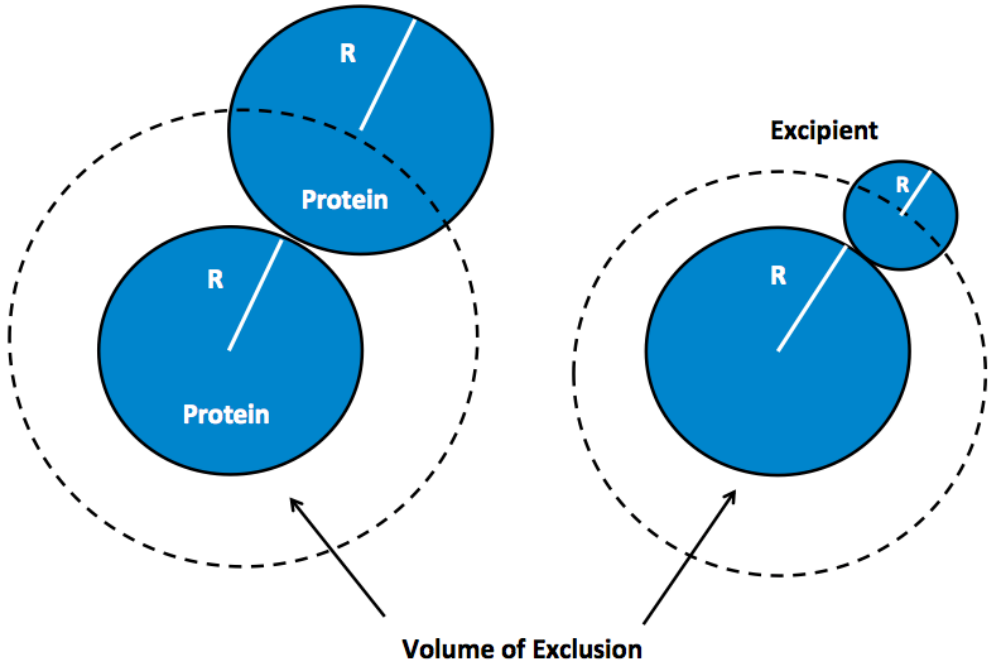


Figure 4.2. **Schematic illustration of the excluded volume effect.** The same protein is represented by a circle coming into contact with another protein (left-hand side) and with a co-solvent such as an excipient (right-hand side). Although the molecules are not drawn to scale, the white area enclosed in the dotted line represents the layer of water that is free from the excluded excipient/protein molecule. *R* represents the hydrodynamic radius in the protein-protein (left-hand side) and protein-excipient (right-hand side) system.

No solute molecule in a solution may occupy the same space as other molecules in a solvent. This volume of occupation is referred to as the excluded volume. The size and conformation of solute molecules determines the volume of exclusion. The centres of two identical spherical solutes cannot be located closer to each other than a distance twice as great as their radius, as illustrated above. The consequence is that the centre of mass for large molecules will be excluded from a much larger region around the surface of the protein and hence will exist at a lower concentration.

In the case of proteins in solution with excipients, there is a layer of water surrounding the protein surface from which the excipient molecule is excluded due to its hydrodynamic radius (the length of the excipient "sphere" that does not come into contact with the protein surface). This exclusion is thermodynamically unfavourable and repulsive, as repulsion increases in proportion with protein accessible surface area (Ohtake *et al*., 2011). Hence, the excipient constrains the protein to assume an equilibrium structure possessing the smallest possible solvent accessible surface area (SASA). This acts to stabilise the native structure of the protein and therefore acts to protect against denaturation and aggregation. More generally, any time an excipient is excluded from around the protein, it will increase the free energy of the protein surface. This exclusion can

be driven by either the surface tension effect or excluded volume. The excluded volume effect is quite small when compared to direct interactions

Peptide Bond Interaction

The protein surface is usually highly heterogeneous in terms of charge and polarity, and consequently may have affinity for specific excipient molecules. Affinity for a particular chemical structure implies specificity in amino acid side chains interactions, and has been investigated using solubility measurements (Nozaki and Tanford, 1963). These studies reported a number of important conclusions for the mechanism of protein denaturation by urea, guanidine hydrochloride (GdnHCl), and other organic solvents. Other experimental studies focusing on interacting mechanisms between amino acids and protein stabilisers (Gekko, 1981; Liu and Bolen, 1995) established that unfavourable interactions between peptide bonds of the protein backbone and stabilising excipients are the driving force of protein stabilisation. Such an unfavourable interaction may be closely related to both the surface tension effect and excluded volume effect described above; both mechanisms should in principle favor excipients to remain in bulk water. Because of the repulsive nature of all three mechanisms it is currently not known which one plays the most dominant role in stabilising proteins (Ohtake *et al*., 2011).

Preferential Interaction/Exclusion

Various interactions (both specific and non-specific) contribute to the to the overall protein/co-solvent interface in solutions. A technique known as equilibrium dialysis can be used to quantify the binding of specific and non-specific ligands to macromolecules (Nimmo *et al*., 1977). In the context of protein-excipient interactions, these interactions can be grouped into two modes, preferential binding and preferential exclusion. They are illustrated in figure 4.3.

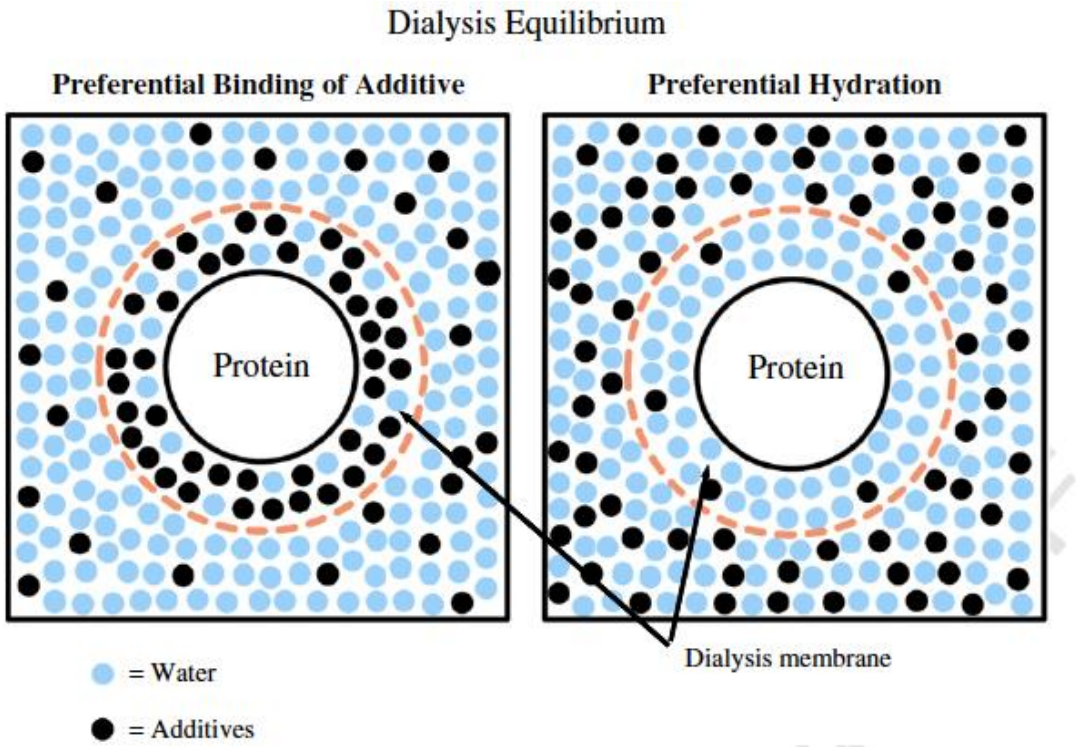


Figure 4.3. **Schematic illustration of the preferential interaction mechanism between co-solvents and proteins.** Preferential binding (left panel) is observed when high concentrations of excipient are present close to the protein surface relative to the bulk solvent; preferential exclusion (right panel) occurs when the opposite takes place. Figure adapted from (Ohtake *et al*., 2011).

In the case of preferential binding (left panel), co-solvents (represented as black circles) are present in higher concentration in the vicinity of the protein surface as opposed to their concentration in the bulk phase (separated by the dotted orange circle). This effect is known as preferential binding (also termed preferential interaction). The opposite case, known as preferential hydration (also termed preferential exclusion), is characterised by a lower concentration of co-solvents close to the protein surface relative to that in the bulk solvent. The space within which excipient concentrations vary from the bulk values is known as the zone of exclusion and is also referred to as the hydration shell. Many sugar and salt-based excipients, which are known to stabilise proteins in the aggregated state and decrease their solubility, are preferentially excluded from the protein surface vicinity (Arakawa and Timasheff, 1985). Preferential exclusion of excipient additives is in line with the repulsive interaction they experience with proteins.

The mechanism by which salt and sugar excipients stabilise proteins and decrease their solubility is grounded in the thermodynamic incompatibility of interactions between stabilising osmolytes and proteins, resulting in them being preferentially excluded (as shown in the right panel of figure 4.3) from the protein surface. Specifically, there is an energetic penalty in bringing osmolytes in the vicinity of the protein, which causes an increase in the free energy of the native state (Ohtake *et al*., 2011). Although experimentally unverified, a greater exclusion of excipients

would be expected from the unfolded (non-native) structure, as it possesses a greater surface area compared to the folded, native state. The energetic penalty would be even greater for the unfolded state in the presence of co-solvent, and this leads to a greater free energy difference between the native and unfolded structures in the presence of stabilising excipients. This means that a greater amount of energy is required to unfold proteins in the presence of preferentially excluded co-solvents, therefore their presence stabilises proteins in their native, folded state. As preferential exclusion (and hence unfavourable interactions) increases with excipient/osmolyte concentration, the native protein structure is stabilised to a greater extent at higher excipient concentrations. This concept is illustrated in figure 4.4.

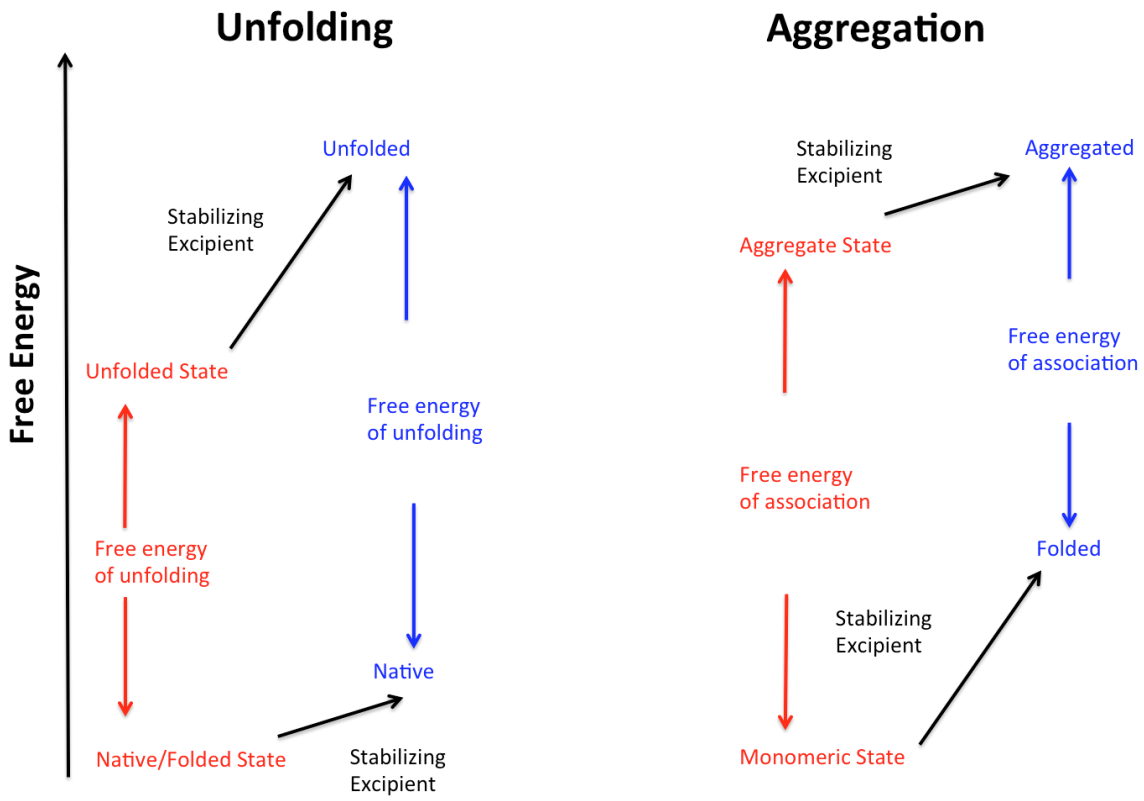


Figure 4.4. Free energy diagram of protein unfolding/aggregation and the effect of excipient interactions. Figure adapted from (Ohtake *et al.*, 2011).

The free energy difference idea can be extended to the situation in which self-association between proteins is taking place. The aggregated state is stable in the presence of osmolytes, which increase aggregation. However, since protein unfolding is the key catalyst for aggregation processes, preferentially excluded excipients reduce aggregation by stabilising the native structure. This mechanism is depicted in figure 4.4 as a free energy diagram. In the left panel, the free energy difference of co-solvents interacting with proteins is greater in the unfolded state than that of the

native state; this is illustrated as the gradient of the black arrows, which is noticeably steeper for unfolded structures compared to properly folded ones. In the case of protein denaturation, followed by aggregation (right panel), the presence of excipients would promote aggregation since the energetic penalty for reverting from self-associated to native state is greater than that of remaining aggregated.

4.2.2 Polymers

Polymeric excipients refer to high molecular weight compounds that stabilise proteins via both specific and non-specific interactions. Common polymers used to stabilise protein formulation include PEG (polyethylene glycols), HSA (human serum albumin), PVP (polyvinylpyrrolidone), dextran and gelatin. Polymers have demonstrated effective stabilising properties in both lyophilized formulations (Costatino *et al*., 1998) and protein solutions (Ping *et al*., 2006).

Polymeric excipients, in contrast to low molecular weight additive compounds, are often charged (*i.e*. polyelectrolytes) and can stabilise proteins via electrostatic interactions through their multiple charged binding sites (Zhang *et al*., 2007). Although this effect tends to be protein specific, it suggests one mechanism by which polymers may stabilise proteins, namely charge-charge interactions between polymers and proteins possessing large regions of opposite surface charge. This effect has been demonstrated for the aFGF (acidic fibroblast growth factor) protein, which possesses a cluster of positively charged ionizable groups on its surface (Chavan *et al*., 1994). The only requirement for aFGF stabilisation was shown to be the presence of one or more regions of high negative charge density (Tsai *et al*., 1993). A multitude of sulphated and phosphorylated anionic polymers (*e.g*. heparin, dextran sulfate, pentosane sulfate, enoxaparin, and phosvitin) were found to be effective at stabilising aFGF (Won *et al*., 1998). Negatively charged biopolymers, most notably nucleic acids, were found to be effective in stabilising aFGF. Furthermore, negatively charged dextran sulfate was found to be effective in preventing aggregation of ribonuclease A (Tsai *et al*., 1998). Cationic polymers have also been used to stabilise negatively charged proteins; one example is PEI (polyethyleneimine), whose effects on lactate dehydrogenase were investigated and found to improve storage stability (Andersson and Hatti-Kaul, 1999). Attractive electrostatic interactions between densely and oppositely charged macromolecular surfaces are non-covalent and act at long range. Protein stabilisation is thus achieved by strong long-range intermolecular interactions between the active product and its polymer additive under this mechanism.

In addition to the electrostatically mediated, protein-specific stabilising mode of charged (polyanionic and polycationic) polymers, both polar and hydrophilic polymers can stabilise proteins

irrespective of physico-chemical properties. The most prominent mechanism in this case is the macromolecular crowding effect, as has been observed for protein transport in polymer gels and concentrated protein solutions (Laurent, 1963). The crowding effect is a phenomenon in which high concentrations of macromolecules (*e.g*. proteins and nucleic acids) in a cellular system reduces the volume of available solvent for other molecules in the environment, leading to a decrease in their effective concentrations (Ellis, 2001). Macromolecular crowding is known to affect protein folding, binding of small molecules, interaction with nucleic acids, enzymatic activity, protein-protein interactions and protein aggregation (Kuznetsova *et al*., 2015). Furthermore, proteins in crowded cellular environments experience volume restrictions due to surrounding macromolecules and this constrains the energetically allowed conformation and also suggests that the dynamics of proteins are different *in vivo* from that *in vitro* (dilute solutions in test tubes). In this capacity, polymeric excipients can act as "crowding agents" to emulate the highly concentrated conditions that are commonplace in biological systems (Homouz *et al*., 2009). Effectively, this is simply one example of the excluded volume effect discussed above, applied to high concentration additives (polymers).

When the population of protein molecules in solution reaches equilibrium between the native, folded state and the unfolded, pre-aggregated state, geometric constraints render the folded state more favourable. Retention of the native structure is preferred as in this state the protein possesses a smaller radius and hence a smaller surface area for exclusion. This mechanism of steric exclusion is an example of volume exclusion (described in the previous section). Essentially, molecular crowding produces a repulsive interaction between a protein and a polymer as they both compete for hydration by surrounding water molecules. Polymer exclusion increases with size, rendering larger polymers generally more effective in stabilising proteins in solution (Minton, 2005).

This "crowding agent" mechanism is used by polymers to stabilise proteins in solution, although the manner in which transient protein conformations and folding mechanisms are affected by macromolecular crowding remains unclear. One commonly used polymer for protein stabilisation in both solution and lyophilized states is PEG, a polyether compound formed by the polymerization of ethylene oxide. Thermodynamic interaction measurements indicate PEG polymers are preferentially excluded from protein surface (Arakawa and Timasheff, 1985). The volume exclusion increases with molecular weight of the polymer, as several PEGs were investigated (PEG-200, -400, -600, -800, and -1000). However, polymers such as PEG can bind to proteins through hydrophobic interactions as well; specifically, interaction of PEG with aromatic side chains has been demonstrated (Hirano *et al*., 2012). The stabilising effect of PEG is thus a delicate balance between its capacity as a crowding agent in order to thermodynamically restrict

proteins to their native state and its favourable interaction with aromatic groups whereby it acts as a weak organic solvent. More specifically, the native structure of a protein would have hydrophobic sequences buried in its core and away from the solvent accessible area whereas in the unfolded state these regions would be exposed to solvent. Hence, in the former case the dominant interaction between PEG and protein would be the excluded volume effect (whereby polymers compete with proteins for hydration and thus stabilise their native states). In the latter case PEG would stabilise the unfolded protein structure via hydrophobic interactions.

4.2.3 Arginine

Arginine is not a protein-stabilising excipient in the sense that has thus far been discussed. However, it is highly effective in suppressing protein aggregation. The aggregation suppression effect of arginine was first observed by Rudolph and Fischer (1989), where it was reported that inclusion of arginine during refolding of plasminogen activator led to increased recovery of the protein. Arginine has been shown to inhibit the aggregation of lysozyme during refolding following thermal denaturation (Kudou *et al*., 2003), and also to reduce the aggregation of heat- or urea-induced denaturation of lysozyme (Shiraki *et al*., 2002). Furthermore, it has been shown to stabilise monoclonal antibodies used in pharmaceutical formulations in solution when it is combined with L-glutamate (Arg-Glu) in equimolar mixtures (Kheddo *et al*., 2014). Arginine is more effective at higher concentrations, which is indicative of weak interaction mechanisms with protein molecules (Arakawa *et al*., 2007).

The mechanistic basis of arginine's stabilising effects is the subject of extensive investigation, as there is no universally accepted theory. However, it is widely accepted that the stabilising effects are in part related to cation-π interactions in aqueous environments (Shukla and Trout, 2010). The cation-π interaction is a non-covalent bond between an aromatic group (containing multiple electrons) and a cation (the positively charged arginine side chain). The bonding energies and strength of such interactions is considerably high, with solution-phase values being on the same order of magnitude as hydrogen bonds and salt bridges (Ma and Dougherty, 1997). Detailed studies of protein structures in the Protein Data Base (PDB) have revealed that cation-π interactions involving arginine are common at protein-complex interfaces, and of all such interactions, those between the side chains of arginine and tyrosine are the most common (Gallivan and Dougherty, 1999). The aggregation suppression characteristics exhibited by arginine in conjunction with its enrichment in protein-complex interfaces in PDB structures, suggest the guanidinium functional group may have an important role in conferring its stabilisation properties

(Ishibashi *et al*., 2005). Indeed, Shukla and Trout have also demonstrated that arginine interacts with both aromatic and charged residues due to the cation-π bond and electrostatic forces.

Preferential interaction and hydration between protein molecules and arginine have also been suggested as interaction mechanisms (Arakawa *et al*., 2007). A concentration dependent interaction effect has been observed with BSA (bovine serum albumin), but not with ribonuclease or lysozyme, suggesting that arginine binds to BSA but is excluded from the surface of both of the latter (Kamerzell *et al*., 2011). Shukla and Trout (2010, 2011) have suggested that preferential interaction depends on the protein surface characteristics. They demonstrated that arginine interacts weakly at low arginine concentration (positive interaction coefficient), while at high co-solute concentrations arginine is excluded (negative interaction coefficient). The studies also focused on hydrogen bonding, reporting that the number of such interactions between a protein and arginine increases with increasing co-solute concentration before reaching a plateau, hinting that saturation of binding sites may be occurring.

4.2.4 Summary of Protein-Excipient Interaction Mechanisms

The preceding sections discussed the interaction mechanisms of several classes of excipients used in pharmaceutical formulations with proteins in liquid phase. Their effects on stability and aggregation were considered. There are principally two different categories of stabilising interactions: (i) those that result in thermodynamically unfavourable interactions and (ii) those that suppress aggregation. These concepts are summarised in figure 4.4, which illustrates that unfavourable interactions are a means of stabilising proteins by forcing them to retain their folded state. Exclusion causes an increase in the free energy of folded and unfolded states. As a result of the unfolded states having a larger surface area, there is a larger increase in free energy for the unfolded versus the folded state leading to the stabilisation effect.

4.3 Non-specific Protein Interactions

Proteins interactions lie at the heart of almost every biological process. Thus, understanding the underlying mechanisms and principles of these interactions comprises an important field of bioprocessing as well as computational biology. A wide range of compounds are used as formulation excipients that contribute to protein stabilisation. The interactions forces that exist between proteins and other molecules can be broadly classified as either: (i) specific interactions which occur at discrete interfaces and evolutionarily conserved binding sites and (ii) non-specific interactions which occur between all proteins (Bahadur *et al*., 2004). Specific interactions are well

studied as they tend to be involved in important biological fucntions such as enzyme catalysis and ligand recoginition. In contrast, non-specific interactions are less well characterised, as they tend to be transient and have little biological function. The interactions between proteins and excipients described in section 4.2 include both specific (*e.g*. hydrogen bonding) and non-specific (*e.g*. long-range electrostatic) types. This chapter does not focus on either specific or non-specific interactions, although extensive use of PDB structural data is made meaning that at least in some cases, results may be biased towards biological specificity due to the fact that all excipients analysed are found in solved protein structures.

A dataset of excipients spanning several compound classes was desired in order to investigate any patterns or trends in their interactions with proteins, and to determine if a "chemical signature" effect established. Research was undertaken to ascertain if low-resolution structure-based analyses can provide useful information relevant to ligand binding. High-resolution methods such as molecular docking and scoring function-based techniques were not considered due to time limitations and sophistication, although it is worth mentioning that they are more precise techniques for computational prediction of protein interactions. A prerequisite for such studies is structural representation and the PDB database was used as a source of high-resolution data. Cursory studies of lysozyme interactions with arginine produced encouraging results, as a certain degree of reproducibility was observed. Arginine was chosen as a starting point for analysis of protein-excipient interactions as it is known to have aggregation inhibiting as well as refolding and stabilising effects on protein mixtures (Golovanov *et al*., 2004; Tsumoto *et al*., 2004; Arakawa *et al*., 2007).

The PDB database was queried using advanced search parameters (specifying that the chemical name under the "chemical component" category should contain the term "arginine"). In order to remove highly similar structures and form a non-redundant set, the 90% sequence identity cutoff was employed, yielding a total of 108 structures. Manual inspection of a customized Excel-readable file provided by the PDB server was subsequently used to identify the subset of those structures that contained arginine bound in a non-specific manner. Although this is not straightforward to deduce, in certain cases the titles of the PDB structures themselves immediately reveal the type of interaction (biological/substrate vs. non-specific) that the protein experiences with the excipient. In the case of the lysozyme data, there were only five structures that could be unequivocally classified as having arginine bound via non-biological interactions. Figure 4.5 illustrates these five PDB files structurally aligned and displayed as a single lysozyme molecule with excipients from all five structures.

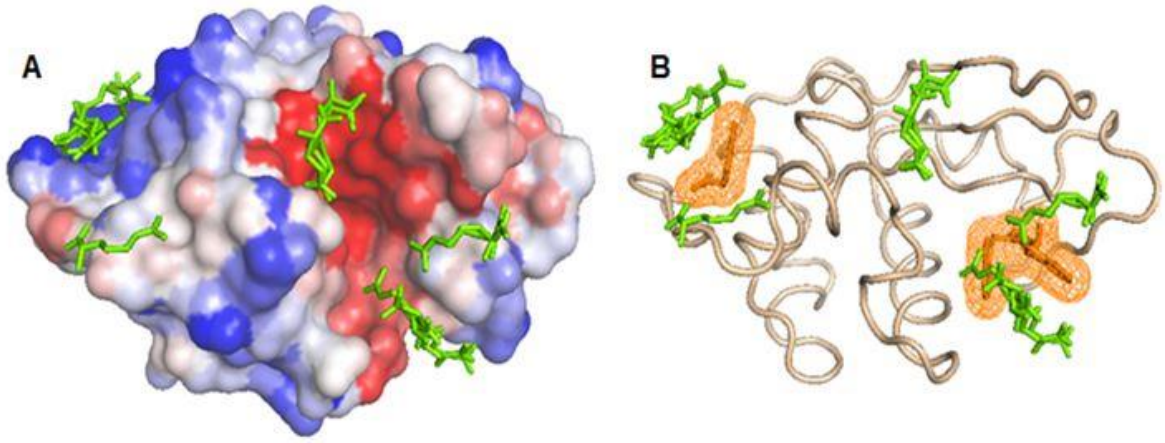


Figure 4.5. Superposed Lysozyme-bound Arginine. Arginine interacts with lysozyme reproducibly, according to charge and charge-aromatic interactions. Binding sites are superposed for 5 lysozyme structures (3AGI, 4EOF, 4HSF, 3A34, 3EMS) containing arginine as an additive. **A**. The lysozyme electrostatic potential (blue/positive, red/negative) is illustrated. Although largely positive at neutral pH, arginine interactions occur on the face containing the negatively charged active site cleft. **B**. Several of the arginine molecules interact with the 3 surface exposed tryptophan (Trp) side chains.

The superposition of arginine excipients on the lysozyme molecule suggests the presence of chemical reproducibility, which may extend to other protein-excipient interactions. If a generalization can be established in the mechanism of protein-excipient binding, this could potentially be useful in bioprocessing pipelines, where protein formulations could be prepared with excipients selected to optimize stability and protect against aggregation.

In this context, several compounds commonly used as excipients were probed to investigate whether chemical reproducibility patterns could be detected. The first task in this approach was to establish which small molecules are most commonly used as non-specific binders. Initially, this issue was addressed computationally, by using text filters on PDB files that would check the occurrence of a specified term (*e.g.* an excipient's name such as "sucrose") throughout several text fields of a PDB file, including the title of the file and over several further annotations. This method of specifically filtering out only non-biologically interacting excipients from PDB searches was not fruitful, as no consistency with manual inspection of the same data could be established.

The only feasible alternative was to use literature-based searches in order to find publications where protein-excipient solutions have been studied. One such study reported a set of screened optimal conditions for promoting successful crystallisation to generate the MORPHEUS screen (Gorrec, 2009). This screen comprises of 47 PDB-derived ligands that could be integrated into numerous crystallisation conditions (a full list of the compounds is provided in appendix 4A). The incentive behind this study was to formulate a screen that would maximise the chances of successful crystallisation, taking into account ligands and additives that can promote crystal

formation via both specific and non-specific interactions. The author reports performing an extensive search of the PDB in order to identify small molecules and ions that bind to biological macromolecules. The author's rationale for using the PDB as a source of data is that compounds found abundantly are potentially good crystallisation agents (the vast majority of the structures deposited in PDB are X-ray crystals). This is due to their stabilisation effects as well as their ability to create variants by changing possible interactions of the molecular surface, hence increasing the chances of obtaining different crystals (Gorrec, 2009). Hence, the MORPHEUS screen is comprised of various chemical compounds including sugars, counter-anions, organic salts containing carboxylic acid groups, halides, and polyethylene glycols (PEGs). Sugars such as sucrose are well known for their thermodynamic stabilisation of macromolecules (Arakawa and Timasheff, 1982). Small counter-anions such as nitrate, phosphate and sulfate possess a multitude of possible binding modes enabled via varying spatial arrangements of oxygen atoms (Gorrec, 2009), rendering them ideal crystallisation components. Small organic salts facilitate crystal growth for the same reason (McPherson, 2001), and halides that promote different crystal forms can help crystal phase determination (Lim *et al*., 1998). Finally, PEGs are known for their tendency to form linear binding patterns in clefts on protein surfaces (Hasek, 2006). Although arginine was not used in the screen, it has been included in the current work because of its well-established role in stabilising protein solutions against aggregation (Arakawa and Tsumoto, 2003).

4.4 Methods

4.4.1 The PDBeXpress Tool

The set of ligands integrated in the MORPHEUS crystallisation screen provided a reasonably robust dataset to use as excipients in order to study SASA (solvent accessible surface area) interactions and to establish any observable patterns. Although a web-based resource reporting such specific information is not readily available, a useful alternative was found in the EBI PDBeXpress suite of tools. PDBeXpress is a web available service containing a set of tools to extract and present useful statistics from the PDB. The tool of relevance in this context retrieves all residues for which there are documented interactions with a queried ligand, *e.g.* arginine. This information is presented on an interactive plot, reporting all tabulated interactions for a queried compound. The graph is enables the user to view individual PDB entries in which any interactions occur, as well as generate reports with this data. An example of the output of PDBeXpress is shown below in figure 4.6, in which arginine was used as the search query. Interaction propensities are

displayed graphically as percentages categorised by amino acid residue, although there is no detailed explanation provided as to how an interaction is precisely defined by the underlying search algorithm.

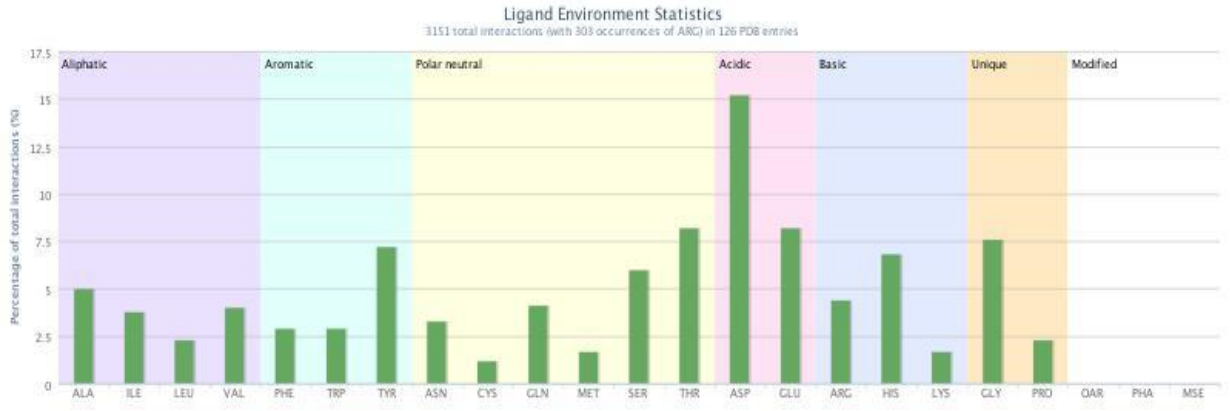


Figure 4.6. PDBeXpress Output: Arginine Interactions in the PDB Database. The output of a search for documented arginine interactions on PDBeXpress. Each bar represents the percentage of interactions with arginine for a particular amino acid as documented in the PDB database.

Nonetheless, this resource was useful for the purposes of the research undertaken. It is perhaps reasonable to assume that a contact would be classified as an interaction when a ligand is within a certain specified distance from any atom of a side chain, but the technical details of any such criteria (*e.g.* Euclidean distance metric) cannot be verified in the absence of a relevant publication. Another shortcoming worth mentioning is that minor variations in the number of identified ligand interactions between different searches for the same query were observed. More specifically, there seems to be some imposed upper limit on the number of retrieved hits (*i.e.* true interactions) at 5500, as this value is never exceeded. Although the exact workings of the algorithm cannot be known without referring to relevant publications, which unfortunately do not seem to be available, it is possible that different search paths are allowed. In spite of the caveats described, PDBeXpress currently appears to be the only web-accessible resource providing this type of protein-ligand interaction information and was therefore used for the current research.

The output of each search is made available for download in the JSON (JavaScript Object Notation) format. JSON is a lightweight format used for data interchange that uses hash tables to store information. JavaScript objects are stored in a format that makes their string representation easy to interpret in Python. A Python script was written to extract the relevant data fields from the output and convert them into an Excel-readable format. For each ligand, the number of interactions, percentage of interactions, and number of occurrences in PDB entries for each amino acid residue were collected.

4.4.2 Computing PDB-based Protein-Excipient Interactions

Ligand interaction data was obtained from the PDBeXpress resource and written into an Excel spreadsheet file. As shown in table 4.1, for each ligand the number of interactions, percentage of interactions, and the number of occurrences in PDB entries per amino acid residue was tabulated. The data for arginine are displayed in table 4.1 below.

Table 4.1 Arginine PDBeXpress Ligand Interaction Data

LIGAND	Residue	Num Inter	Pcnt Inter	PDB Occurrences	Num PDB Entries
Arginine	ALA	159	5	93	37
	ARG	140	4.4	98	51
	ASN	107	3.3	97	55
	ASP	481	15.2	208	73
	CYS	39	1.2	38	20
	GLN	131	4.1	128	60
	GLU	261	8.2	185	87
	GLY	241	7.6	147	56
	HIS	217	6.8	106	39
	ILE	120	3.8	87	37
	LEU	74	2.3	50	27
	LYS	55	1.7	47	21
	MET	55	1.7	51	25
	MSE	2	0	0	1
	OAR	1	0	0	1
	PHA	1	0	0	1
	PHE	93	2.9	73	37
	PRO	74	2.3	72	35
	SER	191	6	131	64
	THR	261	8.2	132	46
	TRP	92	2.9	82	42
	TYR	227	7.2	155	79
	VAL	129	4	92	47

*italics indicate non-standard amino acids

This data was compiled for each of the 47 ligands in the MORPHEUS dataset as well as arginine (shown above). Subsequently, the percentage of interactions per side chain for each of the ligands was tabulated as shown in table 4.2. Each column in table 4.2 represents the percentage of interactions of arginine with a particular amino acid side chain that have been annotated in PDB

structures. A truncated sample of the full twenty amino acids is shown below. The full list of PDB contacts in all MORPHEUS ligands predicted by PBDeXpress can be found in Appendix 4B.

Table 4.2 Arginine PDBeXpress Side Chains Interactions

Ligand	ALA	ARG	ASN	ASP	CYS
Arginine	5	4.4	3.3	15.2	1.2
(RS)-Tartaric Acid	4.4	17.2	7.3	4	0.9
1,2-RS-Propanediol	7.8	5.8	4.4	8.5	0.7
1,3-Propanediol	4.7	2.8	7.5	13.2	0
1,4-Butanediol	2.9	6.3	0.9	5.9	0.5
1,6-Hexanediol	5.6	6	3	4	1.8
1-Butanol	10.8	7.2	3.6	4.5	2.7
2-Propanol	4.1	6.7	4.8	4.2	1.8
Acetate ion	4.2	10.3	4	5.3	1.4
Ammonium cation	2.7	4.1	13.6	11.3	0.8
Bicine	4.5	5.5	3.7	8.7	0.5
Bromide anion	4.7	11.9	6.2	3.5	1.2
Calcium cation	1.3	1.2	9.3	38.5	0.3
Chloride anion	4.4	14	7.5	3.2	1.4
Citrate anion	4.1	17.7	4.1	4.1	0.9
D-Galactose	1.8	6.4	10.3	15.3	2.2
D-Glucose	3.6	7.3	6	9.9	3.4
Diethylene glycol	5.3	8.1	4.7	6.9	0.9
DL-Alanine	4.6	6.3	7	15.2	3.9
DL-Lysine	4.2	4.8	8.2	6	1.5
DL-Serine	5.1	10.8	7.2	6.4	4.2
D-Mannose	4.6	7	8.6	10.6	0.2
D-Xylose	1.3	5.6	12.8	5.3	0.3

The PDBeXpress interaction data was used as a proxy for true protein-excipient interactions to perform patch calculations in order to determine the interaction profile between proteins (as matched from PDBeXpress data) and excipients (each of the MORPHEUS ligands). A Perl pipeline was implemented for performing a low-resolution level analysis of the contact environment between excipients and each of the matched PDB structures as per PDBeXpress (code contributed by Jim Warwicker). The first program was used for calculating SASA per CNOS atom (heavy or non-hydrogen atoms, *i.e*. carbon nitrogen, oxygen, and sulfur) for each PDB structure. A surface patch centred at each CNOS atom of the protein (>5 $Å^2$ accessible surface area) was used to define the area within which a contact would be considered as a true interaction. A series of patch radii ranging from 1 to 16 $Å^2$ were run separately at increments of 1 Å. The range of patch radii are loosely associated with excipient size, so that each one is assigned a single patch size proportional

to its own size in terms of molecular weight. The cutoff of 5 $Å^2$ was chosen in order to define sufficient SASA of an atom for inclusion in a surface patch. Buried atoms were excluded, as they are not able to form contacts with any ligand molecule present in solution. Each amino acid is added once in a patch, so that multiple atom hits for a single amino acid are rendered redundant. Patch radius is estimated empirically so that it is proportional to its number of CNOS atoms. Effectively, each MORPHEUS excipient is being placed at the surface CNOS of each of the proteins that it forms contacts with (PDBeXpress query matches), with patches being subsequently generated.

Mathematically, each patch can be represented as a 20-dimensional vector in which the vector elements correspond to the amino acid compositions for contacts with an excipient. This constitutes a raw form of conformational sampling, as excipient contact information is stored in each surface patch. In order to assess the interaction propensity between excipients and proteins, a proxy interaction parameter was defined. This was calculated as the dot product of the vector for an excipient and the average vector for that excipient (averaging over locations where it is found). Algebraically, the dot product of two vectors is the sum of their component-wise product. The geometric interpretation of the dot product between a pair of vectors **A** and **B** is a measure of the distance that vector **A** extends in the direction of vector **B** and is given by the equation 4.1.

$$\mathbf{A} \cdot \mathbf{B} = |\mathbf{A}| * |\mathbf{B}| \cos \boldsymbol{\theta} \tag{4.1}$$

Equivalently, this can be described as the length of the vector that would result if vector **A** were projected onto **B**. In this context, the cosine metric is used to measure the similarity, or in mathematical terms the co-linearity, of two 20-dimensional vectors. The first vector encodes excipient contacts queried in PDBeXpress and the second vector encodes amino acid composition within patch-defined surface regions of interacting proteins with structural annotations (PDB). The dot product of a surface patch vector for each CNOS atom of an interacting protein is calculated. Subsequently, the PDB-excipient pair that yields the maximum dot product value is recorded as the angle in the calculation of the cosine. This idea is summarised in figure 4.7 below, using arginine as the queried excipient.

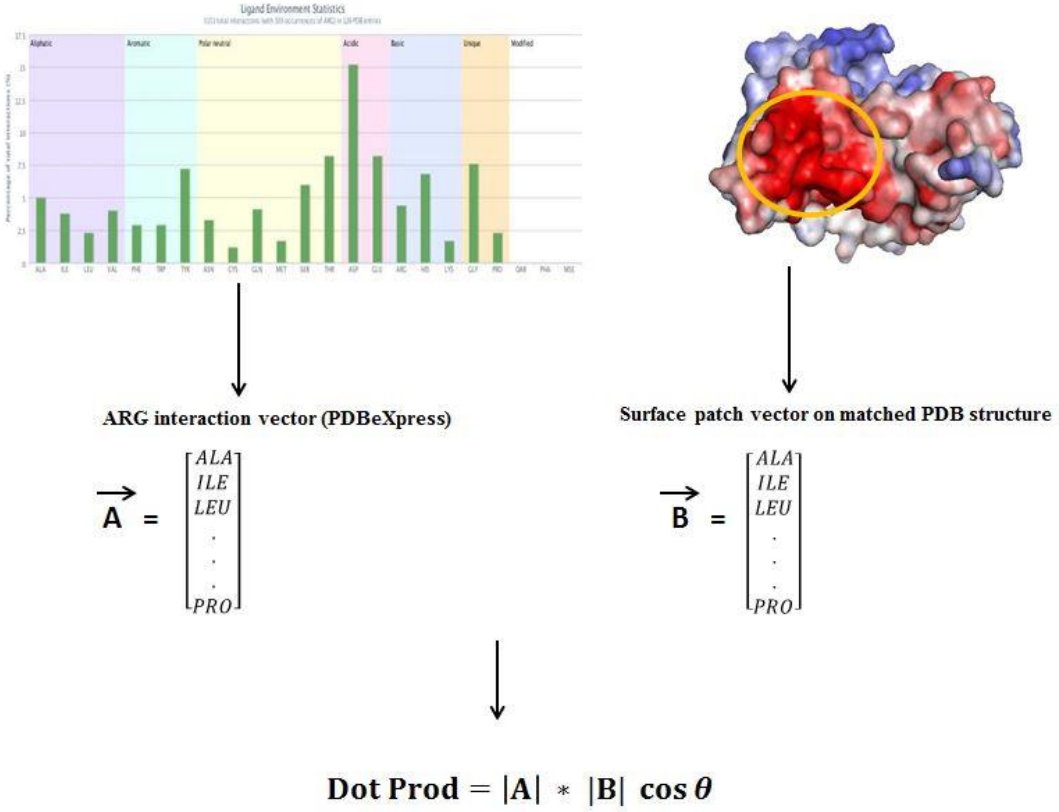


Figure 4.7. Dot Product calculation for measuring protein-excipient interaaction. Diagram summarising the dot product calculation between a queried excipient's (arginine above) interaction environment, obtained via the PDBeXpress web tool (vector **A**) and the surface patches of a PDB structure containing the excipient (vector **B**). Vectors **A** and **B** are populated by integers representing the number of contacts of each amino acid with the excipient. If no side chain contacts are reported on PDBeXpress or none are retrieved on the surface patch, the position of those amino acids is given a value of zero in the vector. The encoding of the surface patch colours on the right hand figure is not directly relevant in this case. Rather the size of the patch is important.

The underlying assumption of this approach is that for a queried ligand, PDBeXpress offers a fairly accurate description of protein-ligand interactions, although as mentioned previously no details regarding the stringency of the criteria for classifying a match as a contact are provided. Non-standard amino acid interaction matches for any of the MORPHEUS ligands were disregarded, so that interactions with only naturally occurring side chains were considered. Since both vectors in the dot product have positive integer magnitudes, the cosine of the angle θ between them will take values ranging from 0 to 1 (0 meaning co-directional, *i.e*. perfect similarity to the PDBeXpress environment, and 1 meaning orthogonal, *i.e.* no similarity to PDBeXpress). Thus the cosine value provides a metric of similarity and is used as a proxy for interaction propensity. Normalisation is applied such that vector magnitudes are divided by the dot product so that the resulting vectors are unit vectors. The pipeline requires as input the set of PDB files that are retrieved from the

PDBeXpress and outputs various ranking information for each PDB/excipient combination. Five different parameters comprise the ranking scheme for a protein-excipient interaction, summarised in table 4.3.

Table 4.3 Parameters Defining PDB-MORPHEUS Interaction Criteria

Parameter	Description
COSINE	Measures the angle between normalised vectors **A** and **B** as described above. This indicates the similarity of protein surface patch environment to that described by the PDBeXpress tool for that protein
COS_MEAN	Measures the mean COSINE value for all patch/excipient combinations in one PDB
COS_ABOVE	Measures the fraction of patches for one excipient/PDB combination for which COSINE >0.75, *i.e.* the number of patches in the third quartile
DOTPRD	A non-normalised version of the COSINE metric adjusted to retain size information of the patch vectors
DPRD_MEAN	Measures the mean DOTPRD value for all patch/excipient combinations in one PDB (analogous to COS_MEAN)

The output of the program is a summary of the above properties for each PDB structure, a sample of which is provided in figure 4.7 below. The value of each of the five interaction properties is shown for eight excipients across two PDB files.

PDB	PROPERTY	ARG	TLA	PGR	PDO	BU1	HEZ	1BO	IOH
1b4b	COSINE	0.815	0.750	0.737	0.733	0.688	0.711	0.664	0.716
1b4b	DOTPRD	1.934	1.732	1.011	1.080	0.931	0.960	0.783	0.727
1b4b	COS_MEAN	0.622	0.495	0.578	0.524	0.503	0.531	0.470	0.482
1b4b	COS_ABOVE	0.141	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1b4b	DPRD_MEAN	1.178	0.888	0.560	0.550	0.496	0.606	0.484	0.394
1m15	COSINE	0.863	0.894	0.840	0.828	0.812	0.778	0.707	0.784
1m15	DOTPRD	2.159	2.573	0.969	1.023	0.973	1.066	0.963	0.776
1m15	COS_MEAN	0.666	0.576	0.624	0.520	0.563	0.550	0.467	0.485
1m15	COS_ABOVE	0.241	0.096	0.051	0.016	0.012	0.006	0.000	0.001
1m15	DPRD_MEAN	1.190	0.976	0.569	0.514	0.521	0.588	0.455	0.377

Figure 4.8. Parameters for PDB/Excipient Combinations. Each PDB-ligand combination is recorded for each of the five parameters in table 4.3.

Another Perl program in the pipeline was used to assess the statistical significance of the PDBeXpress data for reported contacts of MORPHEUS ligands with proteins (PDB files). The input to this program was the file containing ranking information for PDB/excipient combinations (a sample of which is shown in figure 4.8 above) along with a text file containing a list of PDB structures, cross-referenced to an excipient, obtained using the customisable report query from the RCSB website.

A simple brute force sampling method was used to investigate if PDB-excipient interactions determined by the dot product method described above were statistically significant or random. For each dot product calculation between surface patches and excipient, the maximum cosine value, *i.e.* $max(cos\,\theta)$, is recorded. The random sampling procedure consists of generating random integers (using a pseudorandom number generator) between 1 and 48 (the number of excipients). For each excipient, the total number of PDBs containing that excipient was recorded and random number generation trials were repeated for a number of times equal to the number of occurrences of that excipient. For instance, in the case of tartaric acid (TLA), 238 of the processed PDB files have it in complex (table 4.4A, second column), so random integer generation (1 – 48) is repeated 238 times. The mean value of each of these samples is computed and recorded, and this process is repeated 1000 times. This method of sampling the underlying population, which has a distribution of values from 1 – 48, yields a sample distribution that is approximately normal, as illustrated in figure 4.9. Tables 4.5 and 4.6 summarise the statistical properties used to evaluate protein-excipient interaction, while tables 4.4A – D record the values of all parameters used in the sampling procedures as well as their values for each of the 48 excipients, split across four tables.

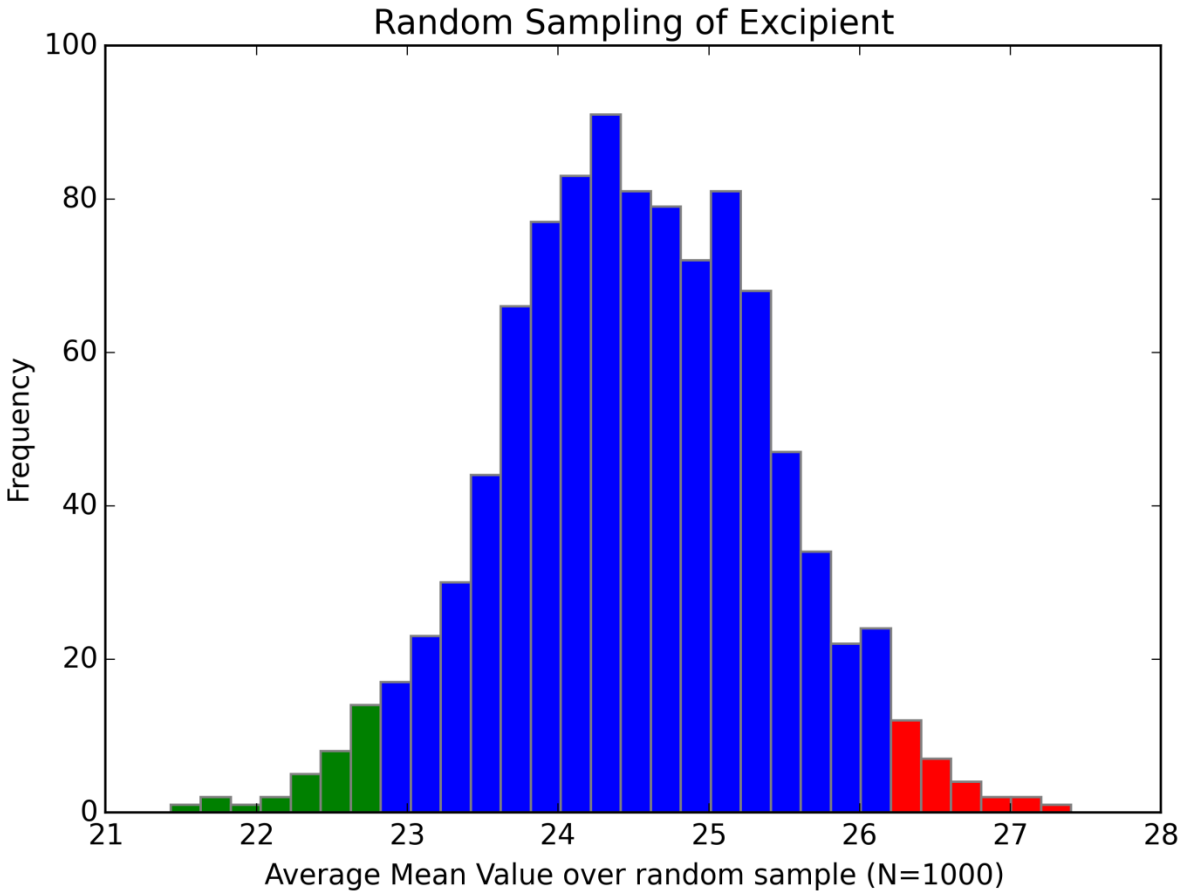


Figure 4.9. Distribution of Brute Force Random Sampling. The brute force sampling method is illustrated using tartaric acid (TLA) as an example. The green and red coloured regions indicate 5% and 95% percentiles for an excipient in the sample, respectively. In this context, they refer to the range of values for an excipient in which it is considered significant, whereas values within the blue region are considered to be random. The same process was carried out on each of the 48 ligands of the MORPHEUS set. A bell shaped curve for the distribution of the random sample is noticeable.

The distribution generated by brute force sampling was displayed using a Python script. The ranking of excipients is normally shaped, and this is most likely due to the central limit theorem. This theorem states that a random variable that is defined as the average of a large number of independent and identically distributed random variables is itself approximately normally distributed.

Table 4.4 Statistics for Ranking PDB-MORPHEUS Contacts

A

Excipient	ARG	TLA	PGR	PDO	BU1	HEZ	1BO	IOH	ACT	NH4	BCN	BR
Statistic												
average_whole	24.9	25.0	24.5	24.4	24.4	24.3	24.2	24.0	24.2	24.3	25.0	24.0
average_nolig	24.5	25.3	25.0	25.1	25.3	25.4	25.1	24.5	24.6	24.1	25.4	23.8
average_onelig	22.8	24.0	21.1	24.0	23.7	28.9	30.8	26.4	25.5	24.7	31.5	24.7
number_whole	84207	84207	84207	84207	84207	84207	84207	84207	84207	84207	84207	84207
number_one	117	238	20	14	44	72	11	395	2976	129	44	270
perc50_numONE	24.5	24.6	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5
No Bonferroni Correction												
perc05_numONE	22.4	23.1	19.4	18.7	21.1	21.9	17.7	23.4	24.1	22.5	21.1	23.1
perc95_numONE	26.6	25.9	29.4	30.2	28.1	27.3	31.5	25.7	24.9	26.5	28.1	25.8
tstONE_negBonf	ins	ins	ins	ins	ins	PSIG	ins	PSIG	PSIG	ins	PSIG	ins
Bonferroni Correction Applied												
percBL_numONE	21.0	21.7	15.4	14.1	17.3	18.7	12.5	22.5	23.7	21.2	17.3	22.2
percBT_numeONE	28.2	21.7	33.1	34.9	30.0	29.2	35.6	26.4	25.2	28.3	30.0	27.2
tstONE_posBonf	ins	ins	ins	ins	ins	ins	ins	ins	PSIG	ins	PSIG	ins

B

Excipient	CA	CL	FLC	GLA	GLC	PEG	ALA	LYS	SER	MAN	XYP	EDO
Statistic												
average_whole	24.9	24.0	25.2	24.6	24.4	24.5	24.7	24.6	24.5	24.9	23.8	24.1
average_nolig	25.1	23.0	25.2	23.5	23.6	25.3	24.9	23.8	25.1	23.6	22.8	24.4
average_onelig	26.4	24.2	27.9	36.1	32.3	24.4	25.5	27.4	24.7	29.1	40.1	24.0
number_whole	84207	84207	84207	84207	84207	84207	84207	84207	84207	84207	84207	84207
number_one	6723	7236	245	119	430	1344	58	68	49	1154	174	4528
perc50_numONE	24.5	24.5	24.6	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5
No Bonferroni Correction												
perc05_numONE	24.2	24.2	23.1	22.4	23.5	23.9	21.7	22.0	21.3	23.9	22.7	24.2
perc95_numONE	24.8	24.8	25.9	26.6	25.6	25.2	27.6	27.3	27.7	25.2	26.3	24.8
tstONE_negBonf	PSIG	NSIG	PSIG	PSIG	PSIG	ins	Ins	PSIG	ins	PSIG	PSIG	NSIG
Bonferroni Correction Applied												
percBL_numONE	24.0	24.0	22.1	21.0	22.4	23.4	18.3	18.8	17.6	23.2	21.1	23.9
percBT_numeONE	25.0	24.9	27.3	28.3	26.3	25.4	29.4	29.2	29.7	25.7	27.6	25.1
tstONE_posBonf	PSIG	ins	PSIG	PSIG	PSIG	ins	ins	ins	ins	Ins	PSIG	ins

C

Excipient	F	FMT	GOL	GLY	EPE	IMD	IOD	FUC	GLU	MG	MES	MPO
Statistic												
average_whole	23.6	24.2	24.4	24.4	25.1	24.3	24.1	24.8	24.8	25.0	24.9	24.5
average_nolig	23.6	24.5	24.4	23.8	24.8	25.2	24.5	23.2	23.8	25.2	25.6	25.4
average_onelig	26.1	23.2	24.3	28.2	26.9	23.4	23.3	31.5	24.5	27.1	26.1	31.1
number_whole	84207	84207	84207	84207	84207	84207	84207	84207	84207	84207	84207	84207
number_one	28	706	9399	133	539	472	529	592	172	6978	650	45
perc50_numONE	24.4	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5
No Bonferroni Correction												
perc05_numONE	20.2	23.6	24.3	22.5	23.6	23.5	23.6	23.6	22.7	24.2	23.7	21.1
perc95_numONE	29.0	25.4	24.7	26.4	25.5	25.6	25.5	25.4	26.3	24.8	25.4	28.2
tstONE_negBonf	ins	NSIG	ins	PSIG	PSIG	NSIG	NSIG	PSIG	ins	PSIG	PSIG	PSIG
Bonferroni Correction Applied												
percBL_numONE	15.5	22.8	24.1	21.1	22.5	22.4	22.5	22.7	21.2	24.0	22.6	17.2
percBT_numeONE	31.9	26.2	24.9	27.9	26.1	26.2	26.2	26.1	27.6	25.0	26.1	29.8
tstONE_posBonf	ins	ins	ins	PSIG	PSIG	ins	ins	PSIG	ins	PSIG	PSIG	PSIG

D

Excipient	MRD	NAG	NO3	OXM	1PE	PO4	K	NA	SO4	PG4	PGE	TRS
Statistic												
average_whole	24.5	240	24.3	24.6	25.2	24.8	23.6	23.9	24.8	25.2	24.8	24.7
average_nolig	26.0	22.0	24.1	24.3	25.7	24.3	22.5	23.2	24.4	26.7	25.7	24.9
average_onelig	25.4	33.9	27.6	29.6	27.0	26.5	21.6	24.0	26.0	24.3	25.2	25.5
number_whole	84207	84207	84207	84207	84207	84207	84207	84207	84207	84207	84207	84207
number_one	967	3443	402	18	331	3437	1242	4511	12165	660	514	724
perc50_numONE	24.5	24.5	24.6	24.6	24.5	24.5	24.5	24.5	24.5	24.5	24.5	24.5
No Bonferroni Correction												
perc05_numONE	23.8	24.1	23.5	18.8	23.3	24.1	23.9	24.2	24.3	23.7	23.6	23.7
perc95_numONE	25.2	24.9	25.6	29.6	25.7	24.9	25.2	24.9	24.7	25.4	25.5	25.4
tstONE_negBonf	PSIG	PSIG	PSIG	ins	PSIG	PSIG	NSIG	NSIG	PSIG	ins	Ins	PSIG
Bonferroni Correction Applied												
percBL_numONE	23.1	23.8	22.5	14.1	22.1	23.8	23.2	23.9	24.1	22.6	22.4	22.8
percBT_numeONE	25.9	25.2	26.4	33.1	26.4	25.2	25.6	25.1	24.8	26.1	26.1	26.1
tstONE_posBonf	ins	PSIG	PSIG	ins	PSIG	PSIG	NSIG	ins	PSIG	ins	ins	ins

The statistics, as described in table 4.5 below, are calculated by brute force sampling rather than any parametric or non-parametric test. As outlined above, an estimate of the average for each excipient is defined by sampling over a set of integers from 1 to 48 randomly. To estimate the distribution of this average, this calculation is repeated over a set of 1000 iterations. Subsequently, as illustrated in figure 4.9, 5% percentile and 95% percentile values are constructed for each excipient sampled average. These regions define positive (95 % tail) and negative correlations (5% tail) of excipient contact with proteins. This process is done both without and with the Bonferroni correction applied. The average over rankings for all proteins in which an excipient is found

(*average_onelig*) corresponds to the maximal *COSINE* value. For example, the average rank for calcium (table 4.4B, first column) is 26.4, where this value indicates $max(cos\,\theta)$ for calcium in a protein, ranked amongst $max(cos\,\theta)$ for all 47 excipients with that protein. The value 26.4 is above both the non-Bonferroni and Bonferroni ranges and thus is considered non-random. It is worth taking note that given 6723 matched instances of calcium in PDB structures, the 5% and 95% percentile values restrict the insignificance levels to a small range of rankings (24.2 – 24.8 under no Bonferroni, 24 – 25 with Bonferroni applied). The range of random values increases as the number of matches decreases for an excipient, *i.e*. the greater the number of PDBs matched for an excipient, the easier it is for the excipient to be randomly ranked.

The Bonferroni test is used in this case because multiple sampling is being performed, with each excipient sampled as many times as it occurs in PDB structures. As the number of groups being compared in a test increases, so does the probability that they will differ by chance alone. Therefore, the Bonferroni correction for this dataset specifies that for the conventional 5% significance level, rather than the actual excipient average (*average_onelig* in table 4.5) having to be greater than 95%, it must be greater than (100 – 5 / 48) *i.e*. approximately 99.8%, which is considerably more stringent.

The result of the brute sampling approach is a series of four values (listed in table 4.6) for each of the 48 MORPHEUS excipients (tables 4.4A – D), both with and without a Bonferroni correction applied. The value of each of the parameters in table 4.5 is also recorded in tables 4.4A – D summarising the findings for the MORPHEUS dataset. A value of *ins* (statistically insignificant) may result from too few PDB structures matched with that excipient or because genuinely there is no apparent signal in the set of contacts retrieved from PDBeXpress. *PSIG* (positive correlation) indicates that the ranking observed for an excipient is higher than expected by chance. For any PDB, all excipients are ranked by the highest *COSINE* match (table 4.3), denoted by $max(cos\,\theta)$, of their calculated surface patches to the model data for amino acid contacts. Practically, this means that an excipient establishes a statistically preferred contact, *i.e*. the excipient is located in the 95% tail (red tail) of the sampling distribution (figure 4.9). *NSIG* (negative correlation) indicates that an excipient is a statistically unfavourable contact, and hence falls under the 5% (green tail) of the sampling distribution. Such values may appear confusing as they imply a negative correlation between maximum surface patch and model contact vectors, although one reason they are observed may be due to the predicted $max(cos\,\theta)$ vectors for an excipient in all PDBs having an average less than what is defined as random based on the previously described sampling process.

Table 4.5 Statistics for Protein-Excipient Contact

Statistic	Description
Ranked Ligand	Each of the 48 ligands is iterated over and its rank is recorded. Ranks are assigned 1 – 48 (lowest to highest), *i.e*. the median rank is 24.5
average_whole	The average rank for the excipient predicted in all PDB files that are processed
average_nolig	The average rank for the excipient predicated in all PDB files without ligands
average_onelig	The average rank for the excipient in PDB files containing it, *i.e*. the average COSINE value for all instances where an excipient is bound in a PDB file
number_whole	The total number of processed PDB files
number_one	The number of PDBs processed with that excipient (if 0, *average_onelig* is set to *nul*)
perc50numONE	Median rank for a random sampling of excipients using iterative random number generation
perc05numONE	5th percentile for an excipient in sampling statistics
perc95numONE	95th percentile for an excipient in sampling statistics
tstONEnegBonf	Evaluates the significance of the excipient based on random sampling process (see table 4.4)
percBLnumONE percBTnumONE tstONEposBonf	All exactly as the previous three parameters with the Bonferroni correction applied

Table 4.6 Statistics for PDB-MORPHEUS Contacts

Statistic	Description
Nul	No excipients matched
Ins	Excipient is random (not statistically significant)
PSIG	Excipient is statistically preferred (95% percentile)
NSIG	Excipient is statistically not preferred (5% percentile)

4.5 Analysis and Visualisation of PDB-based Protein-Excipient Interactions

An advanced search was performed against the RCSB PDB database for all PDB structures having a molecular weight (MW) in the asymmetric unit (AU) in the range of 0 to 150 kD. The upper threshold was selected because it is close to the average immunoglobulin size, and larger proteins were not desired. This search resulted in 106189 hits (approximately 88% of the entire PDB). This set was downloaded and processed into a file format that could be used as input for the Perl programs. The MORPHEUS excipient contact information compiled from PDBeXpress was tested against this entire set of proteins. The statistical outcome of each excipient association

against PDB structures is recorded tables 4.4A – D above. These results are summarised in table 4.7 below with excipients grouped according to chemical properties. Only positive and negative associations of excipients to PDB structures are recorded (random excipient-PDB association are simply left blank) in order to discover any underlying patterns.

Table 4.7 Summary of MORPHEUS-based Excipient Contacts in PDB Database

Chemical Group	Excipient	Average Rank	PDB matches	non-Bonferroni	Bonferroni
Amino Acids					
neutral side chain	ALA	25.5	58		
neutral side chain	SER	24.7	49		
neutral side chain	GLY	28.2	133	PSIG	PSIG
positive side chain	ARG	22.8	117		
positive side chain	LYS	27.4	68	PSIG	
negative side chain	GLU	24.5	172		
Carboxylic Acids					
	TLA	24	238		
	ACT	25.5	2976	PSIG	PSIG
	FMT	23.2	706	NSIG	
	BCN	31.5	44	PSIG	PSIG
	FLC	27.9	245	PSIG	PSIG
	OXM	29.6	18		
Alcohols					
	PGR	21.1	20		
	PDO	24	14		
	BU1	23.7	44		
	HEZ	28.9	72	PSIG	
	1BO	28.9	11	PSIG	
	IOH	26.4	395	PSIG	
	GOL	24.3	9399		

Monovalent ions					
Cations	NH4	24.7	129		
	IMD	23.4	472	NSIG	
	K	21.6	1242	NSIG	NSIG
	NA	24	4511	NSIG	
Anions	BR	24.7	270		
	CL	24.2	7236	NSIG	
	F	26.1	28		
	IOD	23.3	529	NSIG	
	NO3	27.6	402	PSIG	PSIG
Divalent ions					
Cations	CA	26.4	6723	PSIG	PSIG
	MG	27.1	6978	PSIG	PSIG
Anions	PO4	26.5	12165	PSIG	PSIG
	SO4	26	3437	PSIG	PSIG
Monosaccharides					
	GLA	36.1	119	PSIG	PSIG
	GLC	32.3	430	PSIG	PSIG
	MAN	29.1	1154	PSIG	PSIG
	XYP	40.1	174	PSIG	PSIG
	FUC	31.5	592	PSIG	PSIG
	NAG	33.9	3443	PSIG	PSIG
Ethylene glycols					
	PEG	24.4	1344		
	EDO	24	4528	NSIG	
	1PE	27	331	PSIG	PSIG
	PG4	24.3	660		
	PGE	25.2	514		
Buffers					
	EPE	26.9	539	PSIG	PSIG
	MES	26.1	650	PSIG	PSIG
	MPO	31.1	45	PSIG	PSIG
	MRD	25.4	967	PSIG	
	TRS	25.5	724	PSIG	

As mentioned, only positive association (*PSIG*) and negative association (*NSIG*) under Bonferroni tests are indicated, with blank positions indicating random (Ins) PDB-excipient association. Perhaps the most interesting observation from table 4.7 in terms of contact patterns is that of sugars and divalent ions, which are all positively associated with the PDB subset of protein structures. The only other molecule whose association was non-random under both non-Bonferroni

and Bonferroni conditions was potassium (K^+), which exhibits unfavourable interaction in both cases. The question of interpreting the physical meaning of negative association arises again here, *i.e.* what the 5% percentile region of figure 4.8 is representing. It is possible that *NSIG* is arising from a sparse surface patch vector, which may be simply on account of potassium being a single ion. However, the preferred interaction of all divalent ions with PDB structures presents a counterargument to this, as these molecules are on the same size scale yet have the opposite association patterns.

In order to further understand the excipient-PDB association data, a visualisation of the nature of contacts between proteins and ligands was required. For each of the seven chemical groupings (amino acids, carboxylic acids, alcohols, monovalent/divalent ions, sugars, ethylene glycols, and buffers) two molecules were selected in order to examine how their contacts with interacting proteins occurred on a structural level. Given any one of the 48 MORPHEUS excipients, there is a large number PDB structures whose side chains form contacts with it (obtained from querying PDBeXpress). Hence, for each chemical grouping (table 4.7), one of numerous proteins (PDB structures) forming contacts with a ligand molecule had to be selected in a non-random manner. The output from PDBeXpress provides a summary of contacts grouped by side chain of both natural and synthetic amino acids in the form of a frequency plot (the output of querying arginine against the PDB is illustrated figure 4.6). The output is available for downloading in JSON format, which as outlined provides a framework for data interchanging. The PDBeXpress contact information for ligands is stored as a deeply nested hashmap and can be readily used in a Python environment. A Python script was written to parse all PDB contacts for each amino acid (non-standard side chains were excluded so that only natural amino acids were considered). Subsequently the frequency of each structure match was recorded and the PDB with the highest frequency of contacts throughout all side chains was examined. If more than one occurrence of the highest frequency was recorded (*i.e.* multiple PDB structures), a structure was chosen at random. The results are explored in the figures below.

Figure 4.10A Amino Acid Excipients

Excipient	Highest Matching PDB	Side Chain Contacts
Glycine (GLY)	**4BUO** High-resolution structure of thermostable agonist-bound neurotensin receptor 1 mutant without lysozyme fusion	PHE (1 total)
Lysine (LYS)	**3PUO** Crystal structure of dihydrodipicolinate synthase from Pseudomonas aeruginosa complexed with L-lysine	LEU, SER, GLU, LYS (4 total)

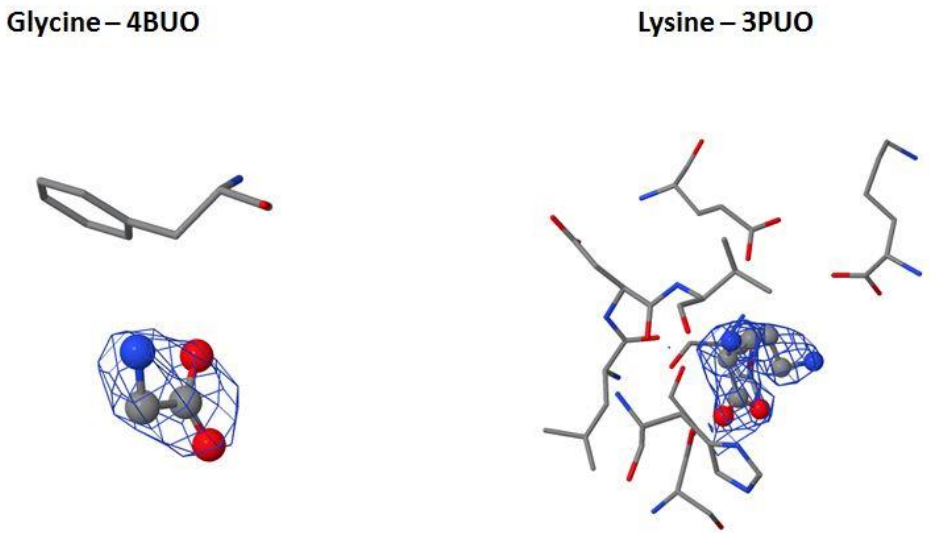


Figure 4.10A. JSMol renderings of glycine – 4BUO (left hand side) and lysine – 3PUO (right hand side) binding pockets. The images were produced using the RCSB PDB 3D viewer.

Figure 4.10A shows glycine (*PSIG*, statistically preferred under both Bonferroni and non-Bonferroni criteria) having only 1 contact and lysine (*PSIG* only under non-Bonferroni) having 4 contacts. The signal from amino acid excipients is concentrated between glycine and lysine, as all other side chains all yielded random association propensities. It is interesting that lysine, which has a positive association propensity under only the less stringent statistical test (non-Bonferroni) has more contacts than glycine, which is a preferred interaction under both tests.

Figure 4.10B Carboxylic Acid Excipients

Excipient	Highest Matching PDB	Side Chain Contacts
Acetate (ACT)	**4GKH** Aminoglycoside phosphotransferase APH(3')-Ia	SER, GLY, VAL (3 total)
Formic Acid (FMT)	**3F98** Human plasma platelet activating factor acetylhydrolase covalently inhibited by tabun	GLY, ILE (2 total)

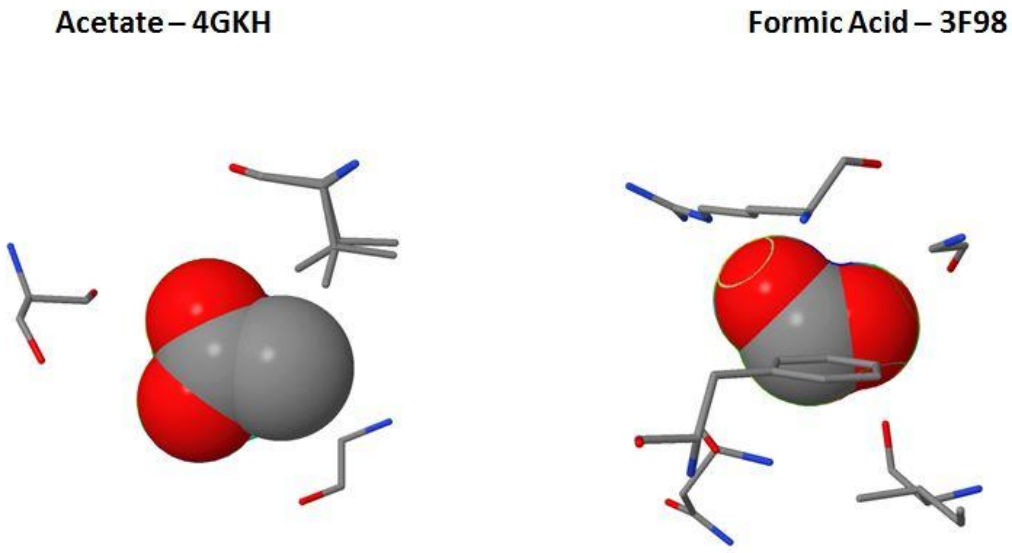


Figure 4.10B. JSMol renderings of acetate – 4GKH (left hand side) and formic acid – 3F98 (right hand side) binding pockets. The images were produced using the RCSB PDB 3D viewer.

Figure 4.10B shows acetate (*PSIG*, statistically preferred under both Bonferroni and non-Bonferroni criteria) having 3 contacts and formic acid (*PSIG* only under non-Bonferroni) having 2 contacts. The signal from carboxylic acids is generally positive for association, and in this case the negatively associated excipient (formic acid) has less contacts than one of the positively associated ones (acetate).

Figure 4.10C Alcohol Excipients

Excipient	Highest Matching PDB	Side Chain Contacts
Isopropyl (IPA)	**4MLA**	ASP, THR, VAL, ILE (4 total)
Glycerol (GOL)	**2OKX**	GLU, GLN (2 total)

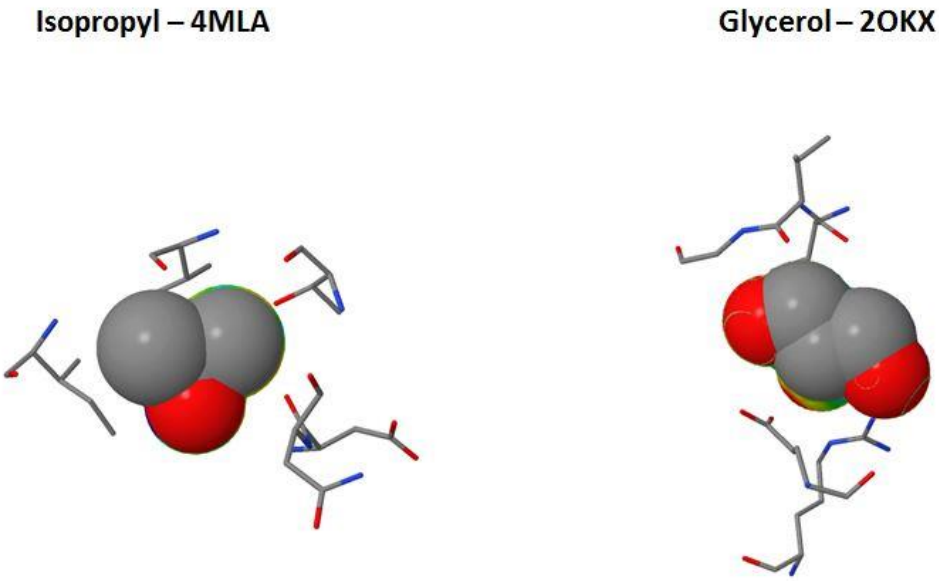


Figure 4.10C. JSMol renderings of isopropyl – 4MLA (left hand side) and glycerol – 2OKX (right hand side) binding pockets. The images were produced using the RCSB PDB 3D viewer.

Figure 4.10C shows isopropyl (*PSIG*, statistically preferred under non-Bonferroni criteria) having 4 contacts and glycerol (statistically random under both Bonferroni and non-Bonferroni criteria) having 2 contacts. Alcohols give weak signal, with most of the compounds considered being statistically insignificant for contacts.

Figure 4.10D Ionic Excipients (Monovalent Ions)

Monovalent Ions		
Excipient	**Highest Matching PDB**	**Side Chain Contacts**
Potassium cation (K^+) K	**2HZV** NikR-operator DNA complex	VAL, GLU, ASP, ILE (4 total)
Nitrate anion (NO_3^-) NO3	**2FBB** Lysozyme (hexagonal)	LYS, GLN, SER, SER, LEU (5 total)
Divalent Ions		
Calcium cation (CA^{2+}) CA	**2L51** Cacium-bound S100A16	SER, SER, TYR, LEU, LYS, VAL (6 total)
Phosphate anion (PO_4^{2-}) PO4	**2NT1** Acid-beta-glucosidase at neutral pH	TRP, TYR, ASP, ARG, SER, SER (6 total)

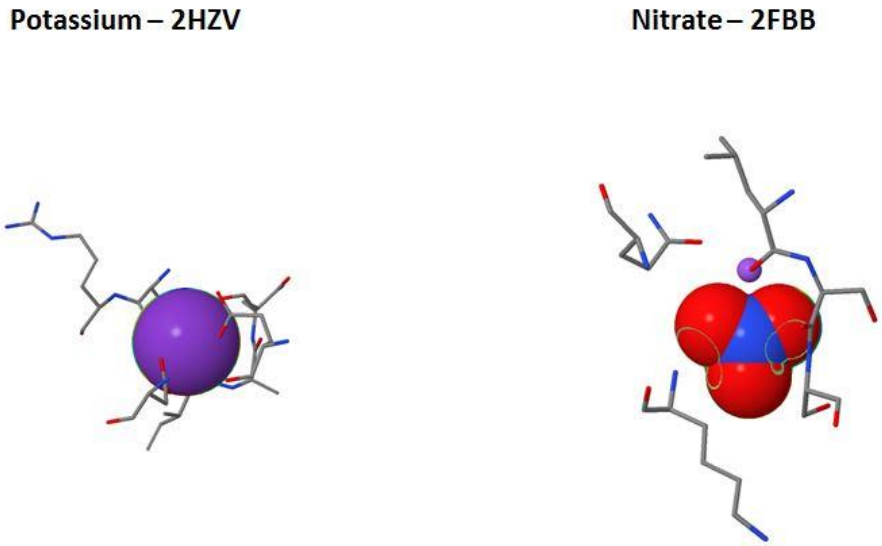


Figure 4.10D(i). JSMol renderings of potassium – 2HZV (left hand side) and nitrate – 2FBB (right hand side) binding pockets. The images were produced using the RCSB PDB 3D viewer.

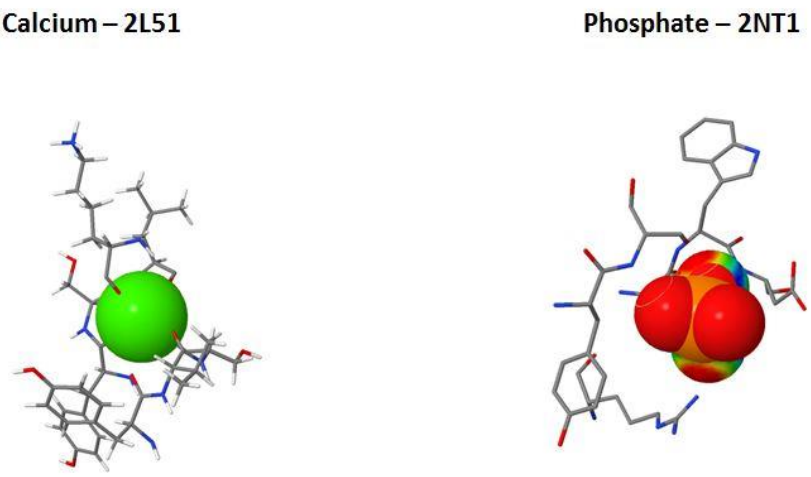


Figure 4.10D(ii). JSMol renderings of calcium – 2L51(left hand side) and phosphate – 2NT1 (right hand side) binding pockets. The images were produced using the RCSB PDB 3D viewer.

Figure 4.10D is divided into two subfigures illustrating monovalent and divalent ionic excipients separately. Potassium (*NSIG*, statistically unfavourable under both Bonferroni and non-Bonferroni criteria) has 4 contacts and nitrate (*PSIG* under both Bonferroni and non-Bonferroni criteria) also has 4 contacts. This is perhaps surprising as the statistical relationship between these two excipients and their representative PDB structures is opposite (unfavourable for potassium and preferential for nitrate). Monovalent cations show the strongest *NSIG* (negative association) signal of all excipient classes studied, while anions are either *NSIG* or *Ins* (nitrate being the sole exception). As discussed above, this may be due to small test patches not giving the extent of amino acid background that is observed in the average model vector. Interestingly, the potassium cation and nitrate anion differ by only a single contact in the PDB structures they were examined in, despite K^+ being the only of the 48 excipients that has a statistically negative association propensity under both statistical criteria examined.

The divalent ions calcium and phosphate also have the same number of contacts with their selected PDB structure (6 each), although in both cases they are statistically preferred interactions under both Bonferroni and non-Bonferroni criteria. This trend extends to all four divalent cation excipients.

Figure 4.10E Monosaccharide Excipients

Excipient	Highest Matching PDB	Side Chain Contacts
Mannose (MAN)	**1GAH** Glucoamylase-471 complexed with acarbose	GLN, SER, THR (3 total)
Fucose (FUC)	**4AHA** Antibody VRC01 complexed with HIV-1 gp120)	GLU, TRP, ARG, LEU (4 total)

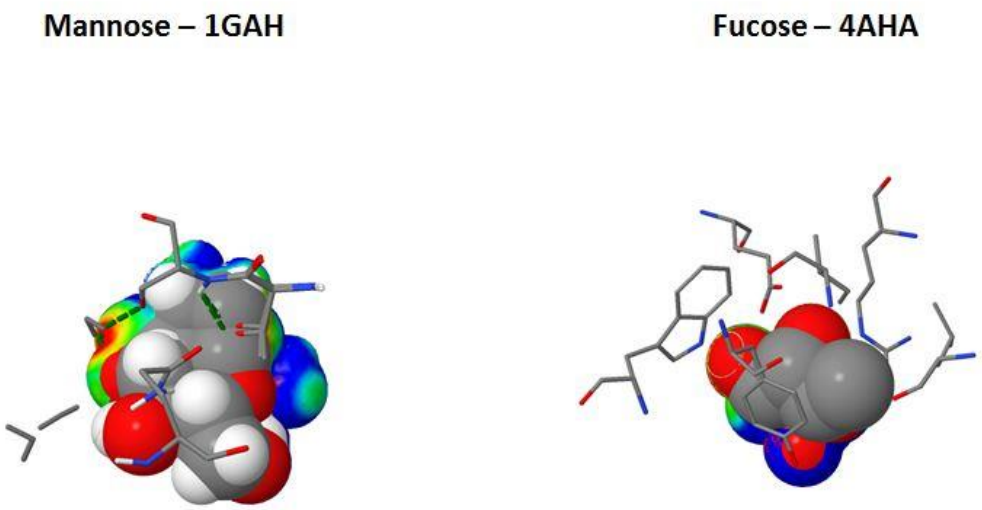


Figure 4.10E. JSMol renderings of mannose –1GAH (left hand side) and fucose – 4AHA (right hand side) binding pockets. The images were produced using the RCSB PDB 3D viewer.

Figure 4.10E shows mannose (*PSIG*, statistically preferred under both Bonferroni and non-Bonferroni criteria) having 3 contacts and fucose (also having preferential interactions under both statistical tests) having 4 contacts. The sugars comprise the only excipient class to give statistically positive association for all compounds considered. The reason for sugars having the strongest signal of all excipient classes in the MORPHEUS dataset is open to speculation, and could perhaps be related to their prominence in glycosylation reactions. Specifically, N-acetyl-D-glucosamine has a strong presence in the PDB as determined from PDBeXpres, although there is only one contact with asparagine in the PDB structure where it is most commonly found (3NGB – data not shown).

Figure 4.10F Ethylene Glycol Excipients

Excipient	Highest Matching PDB	Side Chain Contacts
Polyethylene glycol PEG-400 (1PE)	**3KQZ** Protease 2	TYR, TYR, ASN (3 total)
Ethylene glycol (EDO)	**1X0R** Thioredoxin peroxidase from Aeropyrum pernix K1	GLU, GLU, ARG (3 total)

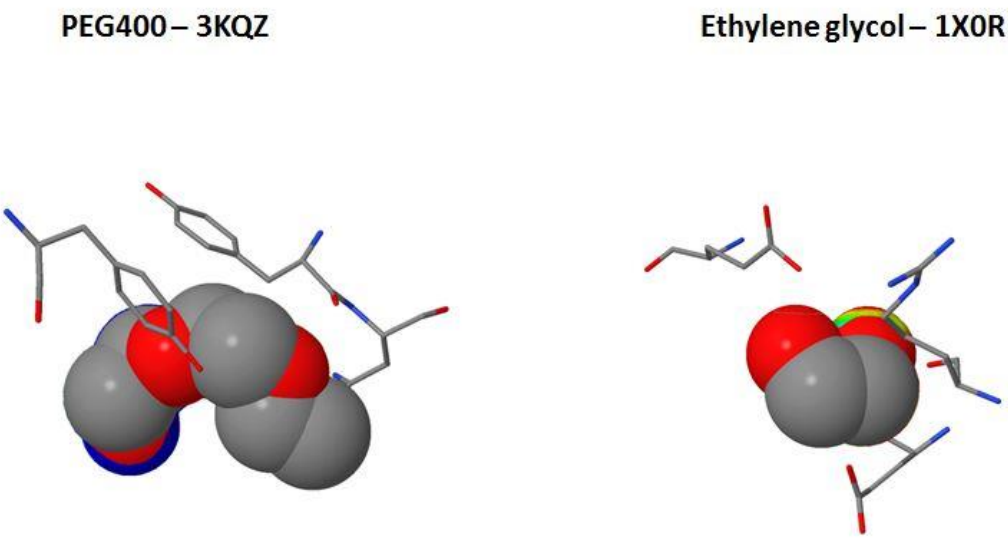


Figure 4.10F. JSMol renderings of PEG400 – 3KQZ (left hand side) and ethylene glycol –1X0R (right hand side) binding pockets. The images were produced using the RCSB PDB 3D viewer.

Figure 4.10F shows polyethylene glycol (PEG400) (*Ins*, statistically insignificant under both Bonferroni and non-Bonferroni criteria) having 3 contacts and ethylene glycol (statistically unfavourable under non-Bonferroni criteria) also having 3 contacts.

Figure 4.10G Buffer Excipients

Excipient	Highest Matching PDB	Side Chain Contacts
Ethanesulfonic Acid - HEPES (EPE)	**1EWK** Metabotropic Glutamate receptor subtype 1 complexed with glutamate	ASP, GLN, THR, HIS, VAL, TYR, ASN (7 total)
Tris (TRS)	**1PMO** *E. coli* GadB at neutral pH	ASN, TRP, LEU, THR, PRO, ASP, VAL, ALA, ILE (9 total)

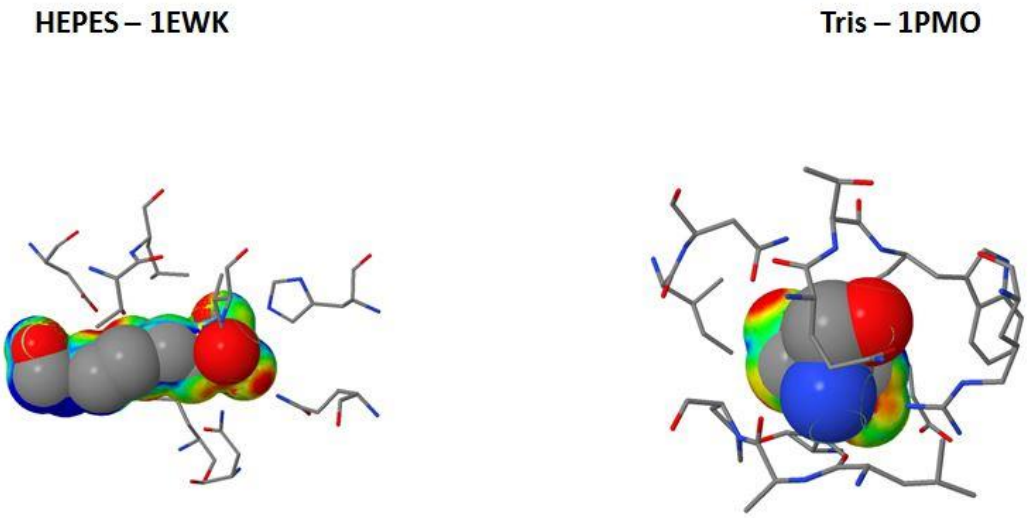


Figure 4.10G. JSMol renderings of EPE –1EWK (left hand side) and Tris –1PMO (right hand side) binding pockets. The images were produced using the RCSB PDB 3D viewer.

Figure 4.10F shows the buffer excipients, which had the greatest number of interactions of all excipient classes investigated. HEPES was statistically favourable under both statistical tests and had 7 interacting side chains in its representative PDB structure (1EWK). Tris was statistically favourable only under non-Bonferroni criteria but had 9 interacting side chains in tis representative PDB structure (1PMO). Buffers have a high positive association signal, and replicate this in their contact environments, with HEPES and Tris having more contacts in their selected PDB structures than any other of the sampled excipients.

4.5 Conclusions

The investigation undertaken in this chapter comprises a low-resolution analysis of protein-excipient contact environments based on structural annotations of a set of crystallisaiton excipients found in the PDB database. The analysis has not yielded any conclusive results relevant to charactersing patterns in protein-excipient interactions. The approach used in this chapter involved a cosine metric to measure the co-linearity of vectors encoding excipient contacts and amino acid composition within surface regions of structurally annotated proteins. In this sense, the findings presented are little less than raw indicators of which compounds could be preferred interacting partners for proteins with solved crystal or NMR structures.

Although there is some weak signal (positive association of sugars and buffers in PDB structures, negative association of monovalent cationic species) in the data, the underlying reasons are not understood well enough to use this type of analysis in a predictive capacity. Negative associations could be due to systematic failures in the modelling process related to the size of patches and numbers of contacts for small ions. For example, it is puzzling that divalent ions show positive association while monovalent ions show almost expressly either negative or zero association. More sophisticated computational techniques such as molecular docking and scoring function-based methods are required for robust structural predictions. Furthermore, it is almost certainly the case that electrostatics (charge-charge interactions) have to be taken into consideration when modelling protein-ion dynamics in solution for meaningful interpretation of results. However, the work described in this chapter could provide the foundation of more precise structural studies to characterise protein-excipient interaction prediction.

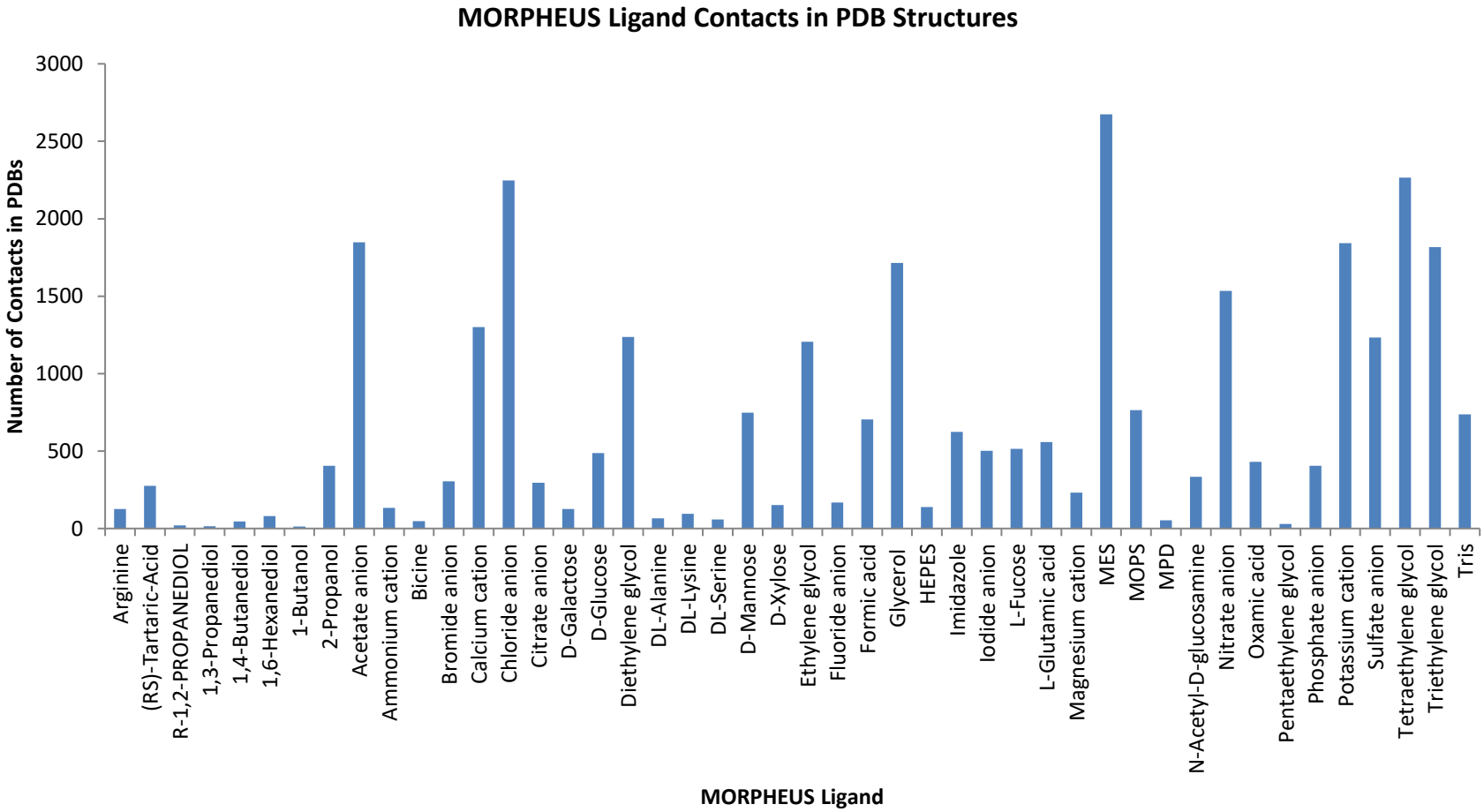
Perhaps the most important caveat in using the cosine metric-based co-linearity between amino acid vectors as a proxy for interaction propensity is the lack of validation of PDBeXpress results. The uncertainty in what constitutes a contact (*i.e.* a distance criterion) renders it difficult to validate the contact vector co-linearity approach, although it is the case that all matches returned from a query have the ligand bound to the protein, with varying number of side chain contacts. Furthermore, even with statistically significant of sugars and buffers in PDB structures (table 4.7), it is very difficult to interpret these finding within a protein stabilisation framework, which would be of interest in therapeutic design applications. Future work in this direction perhaps could focus on a larger set of excipients/ligands, and on closer investigation of the crystallisation conditions of PDB structures in which they are found to be in complex with.

Appendix 4A. MORPHEUS Dataset

Ligand
Arginine*
(RS)-Tartaric-Acid
R-1,2-PROPANEDIOL
1,3-Propanediol
1,4-Butanediol
1,6-Hexanediol
1-Butanol
2-Propanol
Acetate anion
Ammonium cation
Bicine
Bromide anion
Calcium cation
Chloride anion
Citrate anion
D-Galactose
D-Glucose
Diethylene glycol
DL-Alanine
DL-Lysine
DL-Serine
D-Mannose
D-Xylose
Ethylene glycol
Fluoride anion
Formic acid
Glycerol
HEPES
Imidazole
Iodide anion
L-Fucose
L-Glutamic acid
Magnesium cation
MES
MOPS
MPD
N-Acetyl-D-glucosamine
Nitrate anion
Oxamic acid
Pentaethylene glycol
Phosphate anion
Potassium cation
Sulfate anion
Tetraethylene glycol
Triethylene glycol
Tris

*Arginine is not part of the MORPHEUS screen, but was included in the dataset due to its well established role in protein stabilisation.

Appendix 4B. PDBeXpress Ligand Contacts



Chapter 5. Conclusions

The research undertaken in this thesis has applied computational methods to address sequence-based and structure-based protein solubility prediction. The problems that were considered included separation of soluble and insoluble therapeutic protein subsets using both sequence-based and structural features (Chapter 2), and subsequently expanding the range of features to form the basis for a predictive model for solubility (Chapter 3). Protein-excipient interactions were studied in a low-resolution analysis of the co-linearity of contact environments between proteins and ligands (Chapter 4). The aim was to improve our understanding of the roles of protein sequence and structure in aggregation-related phenomena and to identify features that could be useful in a predictive capacity.

In Chapter 2, building on work from cell-free *E. coli*-based data (Niwa *et al*., 2009) reporting structural features that perform well in discriminating soluble and insoluble proteins (Chan *et al*., 2013), datasets of protein-based therapeutics (both antibody-derived and non-antibody) were construced. Surface charge, polarity, and KR-ratio were subsequently used to separate the datasets into predicted soluble and insoluble subsets. Surface non-polarity (patch-based) produced the most uniform discrimination across therapeutic datasets (over 85% of all therapeutics fell on the soluble side of the threshold), while for positive surface charge (patch-based) and KR-ratio, none of the datasets showed a clear preference for either the soluble or insoluble side. The antibody-derived Fab fragments fell on the soluble side of the threshold for all three descriptors used, although this there was a notable bias for patch-based surface non-polarity (99%) and KR-ratio (96%) as opposed to patch-based positive surface charge (64%). This is consistent with the role of non-polar patches in protein insolubility, since Fab fragments have presumably evolved to avoid non-polar patches, and indicates that positive surface charge has a less clear role in determining solubility. At the sequence-level, the bias of Fab fragments towards a KR-ratio above the *E. coli*-based threshold suggests that antibodies may have adapted this sequence trait from evolutionary pressures, possibly reflecting their functional requirement to exist at high concentrations in plasma. Although engineered fragments (scFv's) and biologics also show a bias toward solubility for the three descriptors considered, it was relatively weak for positive charge and KR-ratio.

Because KR-ratio is a sequence-based property it was possible to use all quantified sequences from the cell-free *E. coli* solubility dataset instead of being restricted to those having structural annotations. Solubility and KR-ratio correlate with $R = 0.22$ ($p > 1 \times 10^{-8}$), indicating a positive correlation. Further work showed that eukaryotic protein families existing at high concentrations *in vivo*, notably serum albumin and myoglobin have significantly higher KR-ratios than paralogues existing at lower concentration. The findings from this work suggest the emergence of KR-ratio as a hitherto unknown sequence-based feature that has a role in determining protein solubility. Although the mechanistic basis of this effect is not yet clear, it may be related to the more interactive arginine side chain, which may potentially promote non-specific protein-protein interactions such as aggregation.

A sequence-based and structural framework for building a predictive model was investigated in Chapter 3. Building on the relatively novel findings with respect to the role of KR-ratio in solubility, as well as the more established correlations regarding surface charge and non-polarity, a more comprehensive approach was formulated. In our predictive model, heatmap analysis was used to compare sequence and structural features enriched in soluble and insoluble proteins. For sequence-based features, this was based on computing the difference of *z*-scores for amino acid proportions, K-R, D-E (difference of lysine/arginine and aspartate/glutamate percentage compositions), composition of aromatic residues, as well as charge-based (net charge, total charge and absolute charge) and fold-based properties.

Quantitative proteomics datasets from various organisms (*E. coli*, *S. cerevisiae* and *A. niger*) both intracellular and secreted proteomes were used to compare disparities in sequence properties. This was subsequently extended to repositories of protein abundance and concentration data (PaxDb and PPD) in order to obtain adequate amounts of proteomic data for solubility or related properties (mRNA abundance, protein abundance/concentration). Datasets were either separated into soluble and insoluble (or a related property) as part of the study or were separated into *ad hoc* subsets of high and low solubility, abundance, or concentration based on cumulative frequency distributions of the measured property (section 3.5.3).

Although there is significant variation across individual high-throughput datasets and the PaxDb and the plasma subproteome repositories, certain trends do appear to exist. Charge-based features and properties related to ionisable residues (*e.g.* K-R) are elevated in soluble/high abundance proteins, while sequence entropy and aromatic residues are enriched in insoluble/low

abundance proteins. Among proteomics datasets, an interesting divergence between intracellular and secreted proteins was observed, with several properties having opposite enrichments. Perhaps most distinctly, asparagine (N) and aromatic residues (F+W+Y) occur more frequently in highly produced secreted *A. niger* proteins whereas they are enriched in low solubility/abundance intracellular proteins. This divergence was also observed when comparing sequence repositories, as charge-based and aromatic properties are enriched in low concentration proteins of the plasma. Data from the PaxDb database shows these properties elevated in highly abundant proteins of the human, *E. coli*, *S. cerevisiae* and mouse proteomes. This may suggest an evolutionary adaptation based on the functional environment of a protein (*e.g.* intracellular vs. extracellular), although this trend presents an interesting case for further investigation (Tartaglia *et al*., 2007). Sequence length is also well preserved across all datasets, with longer sequences being less soluble when in cellular environments and more soluble when secreted. This is consistent with the macromolecular crowding dogma, in which cells have to accommodate very large numbers of proteins whilst maintaining their functionality. Importantly, the consistency of K-R enrichment in soluble and high abundance proteins across the datasets supports the original hypothesis concerning the role that KR-ratio may have in determining solubility. Lysine to arginine content hence appears to be a feature that separates soluble and insoluble proteins, at least to a certain degree. As discussed in previous chapters, this would comprise a minimally invasive way of engineering a protein of interest for increased solubility.

Structural calculations were performed mostly for charge-based properties, using both whole protein and patch-based approaches. Features not directly related to charge include contact order (used as a measure of packing), a solvation parameter (measuring the contribution to solvation per amino acid) and protein size (sequence length). The statistical comparison between high and low solubility/abundance proteins in this case employed the Pearson correlation coefficient. Positive correlations were observed consistently in all datasets only for negative charge. Inverse correlations dominate on the heatmap, mainly for uncharged and positively charged regions. These findings are not surprising, as structural calculations were based largely on cell-free *E. coli* proteins, where positive surface charge has been shown to contribute to insolubility.

Our predictive model reiterates previous findings related to protein length, surface charge and non-polarity as features discriminating soluble and insoluble proteins, with the novel aspect

of a simple, sequence-based descriptor in the lysine/arginine composition ratio. While it is true that validation of the model will inevitably require larger-scale solubility data, this will remain a constraint until further high-throughput studies measuring pure protein solubility emerge. The use of protein abundance and concentration as proxies for solubility has important caveats, as discussed in Chapter 3, but nonetheless provides useful insight into the sequence-level and structure-level dependencies of these quantities. The sequence-based findings from this chapter have been made publically available via the Protein-sol webserver tool (www.protein-sol.manchester.ac.uk), which uses cell-free *E.coli* solubility data as a benchmark to predict solubility characteristics (as shown in figure 3.12) for a queried protein, which can be derived from any organism (both prokaryotes and eukaryotes).

Future work in this area should focus on augmenting the model in terms of benchmark data, with the possible inclusion of extremophile proteomes to probe the sequence and structural adaptations that such organisms have made to their proteomes. The previously described web-based tool incorporating this predictive model is currently available and under expansion, and even in its basic form does calculate solubility predictions based on single FASTA sequence inputs. Investigation of 3D structure-based analysis will also have to be expanded as high-throughput "omics" studies of eukaryotic protein solubilities and native aggregation rates are developed. Until such data exist at the large-scale level (several thousands of proteins with PDB-annotated structures having quantified solubilities), it may be reasonable to use structural homology modelling to generate artificial soluble/insoluble or high-/low-abundance "datasets" and study how well the features described in chapter 3 perform in separating subsets. However, until solubility-related data such as those for cell-free *E. coli* become largely available for eukaryotic proteomes, it is sensible to employ protein abundance and concentration-based quantities as proxies. This conclusion is largeley based on the fact that the encrichment patterns of charge- and polarity-based features were observed to replicate from bacterial systems (*E. coli*) to larger proteomes (PaxDb) as is shown in figure 3.13. It is hoped that the tool will continue to expand to accommodate more advanced inputs, such as multiple sequence alignments, and will become a useful resource for prediction of protein solubility and aggregation, thereby increasing its userbase.

The contact environment between a set of excipients used in a crystallisation screen and PDB-annotated proteins was investigated in the final chapter usinga a vector co-linearity

approach. A dot product metric was used to measure similarity of excipient contacts from the PDBeXpress tool and amino acids on protein surface patches. This co-linearity of the specified vectors was used in a very raw manner to measure the association between excipients and proteins. Small molecules (sugars and buffers) were enriched (statistically preferred) throughout the PDB, although it remains unclear why other molecules such as monovalent cations, exhibit non-enrichment (statistically not preferred). Hence there is currently limited scope for this approach to be used in a predictive capacity or to be integrated into the Protein-sol prediction model without a better understanding of the findings.

Clearly, this type of 3D structure-based analysis of preferred protein-excipient interactions is at the early stages, with observations of some siganal being established through a very rough analysis of PDB structures. Further work in this direction will require more sophisticated methods, *e.g*. molecular docking, and a thorough consideration of charge-charge interactions when measuring association between ionic excipients and proteins is necessary. Building on the work presented in chapter 4, a broader range of molecular properties will have to be considered and techniques such as molecular dynamics exploited, so that atomic-level details about the movement of molecules in the context of protein-excipient interactions can be ascertained.

In summary, the efforts made in this thesis to contribute to the field of protein solubility prediction can be divided in three general directions: (i) sequence-based features that opitmise solubility, (iii) structure-based features that optimise solubility and abundance, and (iii) structure-based protein-excipient interactions. The most novel finding of the work undertaken arises in the sequence-based category. In addition to replicating the findings relevant to surface charge and non-polarity, the hitherto uncharacterised KR-ratio and DE-ratio features were explored and established as being correlated to high solubility and concentration levels of proteins (Warwicker *et al*., 2014; Charonis *et al.*, in writing). Experimental validation using mutant-based proteins where arginine and lysine residues have been swapped is currently underway. Structural analyses established that the importance of charge- and polarity-related features in determining solubility carries over to abundance-related quantities. This allows such proxies to be used until large-scale solubility data becomes more widely available, although further work in this direction should take greater advantage of sequence repositories (*e.g*. PaxDb and PPD) to expand analyses and include more organisms (*e.g*. extremophiles) so that “outlier”

proteomes can be investigated and compared against. The findigns of structure-based PDB anlaysis presented in chapter 4 is are too raw to be considered for anything other than a foundation upon which to build future work that will ideally employ atomic-scale simulations of ligand molecules found to be statistically "enriched" throughout the PDB. Expansion of the current approach in all three directions will be necessary in order for this feature-based model to be someday integrated into a functional predictive tool that will have industrial applications.

Finally, building upon existing industrial collaborations with pharmaceutical companies will be pivotal extending the predictive model so that it can at some point be useful as a therapeutic developability screening tool. Perhaps the most important advantage of establishing and strengthening collaborations at the industrial level is the access to proprietary data relevant to therapeutic formulations that are otherwise largely non-public. Specifically, if the model presented in this thesis is to be someday useful as an *in silico* pre-screening tool to assess the developability of protein-based therapeutics, it will have to be tested on actual novel products in the upstream processing phase. The model presented here can realise this potential only if the sequence- and structure-based feature approach carries over to proteins that are being relevant as novel therapeutics. Given the breadth of protein sequences that were analysed in chapter 3, as well as the well-established antibody- and non-antibody-based therapeutics analysed in chapter 2, it is strongly argued that this is well within the means of a further refined and expanded version of our predictive model. In any case, this wil remain an active field of study as the market for biopharmaceuticals continues to grow and rapid, inexpensive means of assessing solubility and aggregation propensity of novel products become ever more indispensible.

References

Agostini, F., Vendruscolo, M., Tartaglia, G.G. (2012). Sequence-based prediction of protein solubilty. *J Mol Biol*. **421**(2-3), 237-241

Agrawal, N.J., Kumar, S., Wang, X., Helk, B., Singh, S.K., Trout, B.L. (2011). Aggregation in protein-based biotherapeutics: computational studies and tools to identify aggregation-prone regions. *J Pharm Sci*. **100**(12), 5081-5095

Alexov, E., Mehler, E.L., Baker, N., Baptista, A.M., Huang, Y., Miller, F., Nielsen, J.E., Farrell, D., Carstensen, T., Olsson, M.H., Shen, J.K., Warwicker, J., Williams, S., Word, J.M. (2011). Progress in the prediction of pKa values in proteins. *Proteins.* **79**(12), 3260-3275

Anderson, N.L. and Anderson, N.G. (2002). The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics*. **1**(11), 845-867

Andersson, M.M. and Hatti-Kaul, R. (1999). Protein stabilising effect of polythyleneimine. *J Biotechnol*. **72**(1), 21-31

Antonini E. and Brunori, M. (1971). Hemoglobin and Myoglobin in Their Reactions With Ligands. North-Holland, Amsterdam. 161 – 167

Arakawa, T., Ejima, D., Tsumoto, K., Obeyama, N., Tanaka, Y., Kita, Y., Timasheff, S.N. (2007). Suppression of protein interactions by arginine: a proposed mechanism of the arginine effects. Biophys Chem. **127**(1-2), 1-8

Arakawa, T. and Timasheff, S.N. (1985). Theory of protein solubility. *Methods Enzymol.* **114,** 49-77

Arakawa, T. and Timasheff, S.N. (1985). The stabilisation of proteins by osmolytes. *Biophys J.* **47**, 150-153

Arakawa, T. and Tsumoto, K. (2003). The effects of arginine on refolding of aggregated proteins: not facilitate refolding, but suppress aggregation. *Biochem Biophys Res Commun*. **304**(1), 148-152

Arakawa, T., Tsumoto, K., Nagase, K., Ejima, D. (2007). The effects of arginine on protein binding and elution in hydrophobic interaction and ion-exchange chromatography. *Protein Expr Purif*. **54**(1), 110-116

Arbabi-Gharoudi, M., To, R., Gaudette, N., Hirama, T., Ding, W., MacKenzie, R., Tanha, J. (2009). Aggregation-resistant VHs selected by in vitro evolution tend to have disulfide-bonded loops and acidic isoelectric points. *Protein Eng Des Sel*. **22**(2), 59-66

Astbury, W.T., Dickinson, S., Bailey, K. (1935). The x-ray interpretation of the denaturation and the structure of the seed globulins. *Biochem J.* **29**(10)**,** 2351-2360

Back, J.F., Oakenfull, D., Smith, M.B. (1979). Increased thermal stabilityof proteins in the presence of sugars and polyols. *Biochemistry*. **18**, 5191 – 5196

Bahar, I., Atilgan, A.R., Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des*. **2**(3), 173-181

Bahadur, R.P., Chakrabarti, P., Rodier, F., Janin, J. (2004). A dissection of specific and non-specific protein-protein interfaces. **336**(4), 943 – 955

Baierlein, R. (2000). The elusive chemical potential. *Am J Phys*. **69**(4), 423 – 434

Balbirnie, M., Grothe, R., Eisenberg, D. (2001). An amyloid-forming peptide from yeast prion Sup35 reveals a dehydrated beta-sheet structure for amyloid. *Proc Natl Acad Sci USA.* **98**(5)**,** 2375-2380

Banks, D.D., Latypov, R.F., Ketchem, R.R., Woodard, J., Scavezze, J.L., Siska, C.C., Razinkov, V.I. (2012). Native-state solubility and transfer of free energy as predictive tools for electing excipients to include in protein formulation development studies. *J Pharm Sci*. 101(8), 2720 – 2732

Basak, A., Bateman, O., Slingsby, C., Pande, A., Asherie, N., Ogun, O, Benedek, G.B., Pande, J. (2003). High-resolution X-ray crystal structures of human gammD Crystallin (1.25 A) and the R58H mutant (1.15 A) associated with aculeiform cataract. *J Mol Biol*. **328**(5), 1137-1147.

Bashford, D. and Karplus, M. (1990). pKa's of ionizable groups in proteins: atomic detail from a continuum electrostatic model. *Biochemistry*. **29**(44), 10219-10225

Berman, H.M., Westbrook, J., Feng, Z, Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E. (2000). The Protein Data Bank. *Nucleic Acids Res*. **28**(1), 235-242

Bianchi, E., Venturini, S., Pessi, A., Tramontano, A., Sollazzo, M. (1994). High level expression and rational mutagenesis of a designed protein, the minibody. From an insoluble to a soluble molecule. *J Mol Biol*. **236**(2), 649-659

Bloemendal, H., de Jong, W., Jaenicke, R., Lubsen, N.H., Slingsby, C., Tardieu, A. (2004). Ageing and vision: structure, stability and function of lens crystallins. *Prog Biophys Mol Biol*. **86**(3), 407-485

Bourne, P.E. and Weissig, H. (2003). *Structural Bioinformatics*. Hoboken, NJ.

Bratko, D., Cellmer, T., Prausnitz, J.M., Blanch, H.W. (2006). Molecular Simulation of Protein Aggregation. *Biotechnol Bioeng.* **96**(1)**,** 1-8

Bryan, A.W. Jr., Menke, M., Cowen, L.J., Lindquist, S.L., Berger, B. (2009). BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis. *PLoS Comput Biol.* **5**(3)**,** e1000333

Burnsteiner, M., Flock, M., Nidetzky, B. (2013). Structure based descriptors for the estimation of colloidal interactions and protein aggregation propensities. *PLoS One*. **8**(4), e59797

Campbell. *Biology*. http://www.studyblue.com

Caflisch, A. (2006). Computational models for the prediction of polypeptide aggregation propensity. *Curr Opin Chem Biol*. **10**(5), 437 – 444

Castillo, V., Graὰa-Montes, R., Ventura, S. (2011). The aggregation properties of Escherichia coli proteins associated with their cellular abundance. *Biotechnol J*. **6**(6), 752 – 760

Cellmer, T., Bratko, D., Prausnitz, J.M., Blanch, H.W. (2007). Protein aggregation *in silico*. *Trends Biotechnol*. **25**(6), 254 – 261

Chan, P., Curtis, R.A., Warwicker, J. (2013). Soluble expression of proteins correlates with a lack of positively charged surface. *Sci Rep*. **26**(3), 1-6

Chang, C.C., Song, J., Tey, B.T., Ramaman, R.N. (2013). Bioinformatics approaches for improved recombinant protein production in Escherichia coli: protein solubility prediction. *Brief Bioinform*, **15**(3)

Chapman, D.L. (1913). A contribution to the theory of electrocapillarity. *Phil Mag*. **25**, 475- 481

Chavan, A.J., Haley, B.E., Volkin, D.B., Marfia, K.E., Verticelli, A.M., Bruner, M.W., Draper, J.P., Burke, C.J., Middaugh, C.R. (1994). Interaction of nucleotides with acidic fibroblast growth factor (FGF-1). *Biochemistry*. **33**(23), 7193-7202

Chaudhuri, R., Cheng, Y., Middaugh, C.R., Volkin, D.B. (2014). High-throughput biophysical analysis of protein therapeutics to examine interrelationships between aggregate formation and conformational stability. *AAPS*. **16**(1), 48-64

Chen, L., Oughtred, R., Berman, H.M, Westbrook, J. (2004). TargetDB: a target registration database for structural genomics projects. *Bioinformatics*. **20**(16), 2860-2862

Chennamsetty, N., Voynov, V., Kayser, V., Helk, B., Trout, B.L. (2009). Design of therapeutic proteins with enhanced stability. *Proc Natl Acad Sci USA*. **106**(29), 353-358

Chennamsetty, N., Voynov, V., Kayser, V., Helk, B., Trout, B.L. (2010). Prediction of aggregation prone regions of therapeutic proteins. *J Phys Chem B*. **114**(19), 6614-6624

Chi, E.Y., Krishnan, S., Kendrick, B.S., Chang, B.S., Carpenter, J.F., Randolph, T.W. (2003) Roles of conformational stability and colloidal stability in the aggregation of recombinant human granulocyte colony-stimulating factor. *Protein Sci*. **12**(5), 903-913

Chiti, F. and Dobson, C.M. (2006). Protein misfolding, functional amyloid, and human disease. *Annual Rev Biochem.* **75,** 333-366

Chiti, F., Stefani, M., Taddei, N., Ramponi, G., Dobson, C.M. (2003). Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature.* **424**(6950), 805-808

Cioci, F. and Lavecchia, R. (1998). Sorbitol-mediated stabilisation of human IgG against thermal inactivation. *Biotechnol Technol*. **11**, 855 – 858

Clark, E.D. (2001). Protein refolding for industrial processes. *Curr Opin Biotechnol.* 12(2), 202-207

Cole, C. and Warwicker, J. (2002). Side-chain conformational entropy at protein-protein interfaces. *Protein Sci.* **11**(12)**,** 2860-2870

Conchillo-Sollé, O., de Groot, N.S., Avilés, F.X., Vendrell, J., Daura, X., Ventura, S. (2007). AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics.* **8**(65)

Costatini, S., Colonna, G., Facchiano, A.M. (2006). Amino acid propensities for secondary structures are influenced by the protein structural class. *Biochem Biophy Res Commun*. **342**(2), 441-451

Costatino, H.R., Schwendeman, S.P., Langer, R., Klibanov, A.M. (1998). Deterioration of lyophilized pharmaceutical proteins. *Biochemistry*. **63**(3). 357-363

Cromwell, M.E.M., Hilario, E., Jacobson, F., (2006). Protein aggregation and bioprocessing. *AAPS J*. **8**(3), 572-579

Czepas, J., Devedjiev, Y., Krowarsch, D. Derewenda, U., Otlewski, J., Derewenda, Z.S. (2004). The impact of Lys->Arg surfact muations on the crystallization of the globular domain of RhoGDI. *Acta Crystallogr D Biol Crystallogr*. **60**(Pt 2), 275-280

De Felice, F.G., Vieira, M.N., Meirelles, M.N., Morozova-Roche, L.A., Dobson, C.M., Ferreira, S.T. (2004). Formation of amyloid aggregates from human lysozyme and its disease-associated variants using hydrostatic pressure. *FASEB J*. **18**(10), 1099-1101

De Simone, A., Kitchen, C., Kwan, A.H., Sunde, M., Dobson, C.M., Frenkel, D. (2012). Intrinsic disorder modulates protein self-assembly and aggregation. *Proc Natl Acad Sci USA.* **109**(18)**,** 6951-6956

Debye, P. and Hückel, E. 1923. Zur Theorie der Elektrolyte. *Phys Zeitschr*. **24**, 185-206

Demarest, S.J. and Glaser, S.M. (2008). Antibody therapeutics, antibody engineering, and merits of protein stability. *Curr Opin Drug Discovery Dev*. **11**(5), 675-687

Der, B.S., Kluwe, C., Miklos, A.E., Jacak, R. Lyskov, S., Gray, J.J., Georgiou, G. Ellington, A.D., Kuhlman, B. (2013). Alternative computational protocols for supercharging protein surfaces for reversible unfolding and retention of stability. *PLoS One*. 8, e64363

Dimitrov, D.S. (2012). Therapeutic Proteins. *Methods Mol Biol*. **899**, 1-26

Dobson, C.M. (2003). Protein folding and misfolding. *Nature.* **426**(6968)**,** 884-890

Dobson, C.M. (2004). Principles of protein folding, misfolding, and aggregation. *Semin Cell Dev Biol*. **15**(1), 3-16

Domenico, R.D. and Lavecchia, R. (2000). Thermal stability of human haemoglobin in the presence of sarcosine and sorbitol. *Biotechnol Lett*. **22**, 335 – 339

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., Arnold, F.H. (2005). Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA*. **102**(40), 14338-14334

Dudgeon, K., Rouet, R., Kokmeijer, I., Schofield, P., Stolp, J., Langley, D., Stock, D., Christ, D. (2012). General strategy for the generation of of human antibody variable domains with increased aggregation resistance. *Proc Natl Acad Sci USA*. **109**(27), 10879-10884

Duncancel, F. and Muller, B.H. (2012). Molecular engineering of antibodies for therapeutic and diagnostic purposes. *MAbs*. **4**(4), 445-457

Eisenberg, D. and McLachlan, A.D. (1986). Solvation energy in protein folding and binding. *Nature*. **319**(6050), 199-203

Ellis, R.J. (2001). Macromolecular crowing: obvious but underappreciated. *Trends Biochem Sci*. 26(10), 597-604

Fernandez-Escamilla, A.M., Rousseau F., Schymkowitz, J., Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol*. **22**(10), 1302-1306

Ferrao-Gonzales, A.D., Souto, S.O., Silva, J.L., Foguel, D. (2000). The preaggregated state of an amloidogenic protein: hydrostatic pressure converts native transthyretin into the amyloidogenic state. *Proc Natl Acad Sci USA*. **97**(12), 6445-6450

Frousios, K.K, Iconomidou, V.A., Karletidi, C.M., Hamodrakas, S.J. (2009). Amyloidogenic determinants are usually not buried. *BMC Struct Biol*. **9**(1), 44

Gallivan, J.P. and Dougherty, D.A. (1999). Cation-pi interactions in structural biology. *Proc Natl Acad Sci USA*. **96**, 9459-9464

Galzitskaya, O.V., Garbuzynskiy, S.O., Lobanov, M.Y. (2006). Prediction of amyloidogenic and disordered regions in protein chains. *PLoS Comput Biol*. 2(12), e177

Garcia-Moreno, B. (2009). Adaptations of proteins to cellular and subcellular pH. *J Biol*. **8**(11), 98

Gaynor, P.M., Bonnette, R., Garcia, E., Kahl, L., Valerio, L. (2006). FDA's approach to the GRAS provision: a history of processes. *FDA Science Forum*

Gekko, K. (1982). Calorimteric study on thermal denaturation of lysozyme in polyol-water mixtures. *J Biochem*. **91**, 1197-1204

Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., Weismman, J.S. (2003). Global analysis of protein expression in yeast. *Nature*. **425**(6959), 737-741

Gilson, M.K. (2006). Introduction to continuum electrostatics, with molecular applications.

Goh, C.S., Lan, N., Douglas, S.M., Wu, B., Echols, N., Smith, A. Milburn, D., Montelione, G.T., Zhao, H., Gerstein, M. (2004). Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. *J Mol Biol*. **336**(1), 115-130

Golavonov, A.P., Hautbergue, G.M., Wilson, S.A., Lian, L.Y. (2004). A simple method for improving protein solubility and long-term stability. *J Am Chem Soc*. **126**(29), 8933-8939

Goncearenco, A., Ma., B.G., Berezovsky, I.N. (2014). Molecular mechanisms of adaptation emerging from the physics and evolution of nucleic acids and proteins. *Nucleic Acids Res*. **42**(5), 2879-2892

Gonzalez, M., Murature, D.A., Fidelio, G.D. (1995). Thermal stability of human immunoglobulins with sorbitol. *Vox Sang*. **68**(1), 1 – 4

Gorrec, F. (2009). The MORPHEUS protein crystallization screen. *J Appl Crystallogr*. 42(6), 1035-1042

Goldschmidt, L., Teng, P.K., Riek, R., Eisenberg, D. (2010). Identifying the amylome, proteins capable of forming amyloid-like fibrils. *Proc Natl Acad Sci USA*. **107**(8), 3487-3492

Gouy, M. (1910). Sur la constitution de la charge électrique a la suface d'un électrolyte. *J Phys*. **9**, 457-468

Greaves, R. and Warwicker, J. (2005). Active site identification through geometry-based and sequence-profile based calculations: burial of catalytic clefts. *J Mol Biol*. **349**(3), 547-557

Habibi, N., Mohd Hashim, S.Z., Norouzi, A., Samian, M.R. (2014). A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in Escherichia coli. *BMC Bioinformatics*. 15:134

Hamodrakas, S.J. (1988). A protein secondary structure prediction scheme for the IBM PC and compatibles. *Comput Appl Biosci*. **4**(4), 473-477

Hartl, F.U., Bracher, A., Hayer-Hartl, M. (2011). Molecular chaperones in protein folding and proteostasis. *Naure*. **475**(7356), 324-332

Hellstrand, E., Boland, B., Walsh, D.M., Linse, S. (2010). Amyloid β-protein aggregation produces highly reproducible kinetic data and occurs by a two-phase process. *ACS Chem Neurosci*. **1**(1), 13-18

Hileman, R.E., Fromm, J.R., Weiler, J.M., Linhardt, R.J. (1998). Glycosaminoglycan-protein interactions: definition of consensus sites in glycosaminoglycan binding proteins. *Bioessays*. **20**(2), 156-167

Hirano, A., Shiraki, K., Arakawa, T. (2012). Polyethylene glycol behaves like weak organic solvent. Biopolymers. **97**(2), 117 – 122

Ho, J.G. and Middelberg, A.P. (2004). Estimating the potential refolding yield of recombinant proteins expressed as inclusion bodies. *Biotechnol Bioeng*. **87**(5), 584-592

Hofmeister, F. (1888). *Arch Exp Pathol Pharmacol*. 24, 247 – 260

Holliger, P. and Hudson, P.J. (2005). Engineering antibody fragments and the rise of single domains. *Nat Biotechnol*. **23**(9), 1126-1136

Homouz, D., Stagg, L., Wittung-Stafshede, P., Cheung, M.S. (2009). Macromolcular crowding modulates folding mechanism of alpha/beta protein apoflavodoxin. *Biophys J*. **92**(2), 671-680

Honig, B. and Nicholls, A. (1995). Classical Electrostatics in Biology in Chemistry. *Science*. **268**(5214), 1144-1149

Howe, P.W. (2004). A straight-forward method of optimising protein solubility for NMR. *J Biomol NMR*. **30**(3), 283-286

Idicula-Thomas, S. and Balaji, P.V. (2005). Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in *Escherichia coli*. *Protein Sci*. **14**(3), 582-592

Idicula-Thomas, S. Kulkarni, A.J., Kulkarni, B.D. (2006). A support vector machine-based method for predicting the propensity of a protein to be solube or to form inclusion body on overexpression in *Echerichia coli*. *Bioinformatics*. **22**(3), 278-284

Invernizzi, G., Papaleo, E., Sabate, R., Ventura, S. (2012). Protein aggregation: mechanisms and functional consequences. *Int J Biochem Cell Biol*. **44**(9), 1541-1554

Ishibashi, M., Tsumot, K., Tokunaga, M., Ejima, D., Kita, Y., Arakawa, T. (2005). Is arginine a protein-denaturant? *Protein Expr Purif*. **42**, 1-6

Ivankov, D.N., Garbuzynskiy, S.O., Alm, E., Plaxco, K.W., Baker, D., Finkelstein, A.V. (2003). Contact order revisited: influence of protein size on the folding rate. *Protein Sci*. **12**(9), 2057-2062

Jarrett, J.T. and Lansbury, P.T. (1993). Seeding “one-dimensonal crystallization” of amyloid: a pathogenic mechanism in Alzheimer’s disease and scrapie? *Cell*. **73**(6), 1055-1058

Jones, S., Marin, A., Thornton, J.M. (2000). Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng*. **13**(2), 77-82

Jorgensen, W.L. and Duffy, E.M. (2000). Prediction of drug solubility from Monte Carlo simulations. *Bioorg Med Chem Lett*. **10**(11), 1155-1158

Kamerzell, T.J., Esfandiary, R., Joshi, S.B., Middaugh, C.R., Volkin, D.B. (2011). Protein-excipient interactions: mechanisms and biophysical characterization applied to protein formulation development. *Adv Drug Deliv Rev*. **63**(13), 1118-1159

Karlin, S. (1995). Statistical significance of sequence patterns in proteins. Curr Opin Struct Biol. **5**(3), 360-371

Kato, A., Maki, K., Ebina, T., Kuwaijima, K., Soda, K., Kuroda, Y. (2007). Mutational analysis of protein solubility enhancement using short peptide tags. *Biopolymers*. **85**(1), 12-18

Kauzmann, W. (1959). Some factors in the interpretation of protein denaturation. Adv Protein Chem. **14**, 1-63

Kayser, V., Chennamsetty, N., Voynov, V., Helk, B., Trout, B.L. (2011). Conformational stability and aggregation of therapeutic monoclonal antibodies studies with ANS and thioflavin T binding. *MAbs*. **3**(4), 408-411

Kheddo, P., Tracka, M., Armer, J., Dearman, R.J., Uddin, S., van der Walle, C.F., Golovanonv, A.P. (2014). The effect of arginine glutamate on the stability of monoclonal antibodies in solution. *Int J Pharm*. **473**(1-2), 126-133

Klapper, I., Hagstrom, R., Fine, R., Sharp, K., Honig, B. (*1986)*. Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: effects of ionic strength and amino-acid modification. *Proteins*. **1**(1), 47-59

Konno, T. (2001). Amyloid-induced aggregation and precipitation of soluble proteins: an electrostatic contribution of the Alzheimer's beta(25-35) amyloid fibril. *Biochemistry*. **40**(7), 2148-2154

Kramer, R.M., Shende, V.R., Motl, N., Pace, C.N., Scholtz, J.M. (2012). Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility. *Biophys J*. **102**(8), 1907-1915

Kudou, M., Shiraki, K., Fujiwara, S., Imanaka, T., Takagi, M. (2003). Prevention of thermal inactivation and aggregation of lysozyme by polyamines. *Eur J Biochem*. **270**(22), 4547-4554

Kuhlman, B. and Baker, D. (2000). Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci USA*. **97**(19), 10383-10388

Kusumoto, Y., Lomakin, A., Teplow, D.B., Benedek, G.B. (1998). Temperature dependence of amyloid beta-protein fibrillization. *Proc Natl Acad Sci USA*. **95**(21), 12277-12282

Kuznetsova, I.M., Zaslavsky, B.Y., Breydo, L., Turoverov, K.K., Uversky, V.N. (2015). Beyond the Excluded Volume Effects: Mechanistic Complexity of the Crowded Milieu. *Molecules*. **20**(1), 1377 – 1409

Kyte, J. and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*. **157**(1), 105-132

Lauer, T.M., Agrawal, N.J., Chennamsetty, N., Helk, B., Trout, B.L. (2012). Developability index: a rapid in silico tool for the screening of antibody aggregation propensity. *J Pharm Sci*. **101**(1), 102-115

Laurent, T.C. (1963). The interaction between polysaccharides and other macromolecules. 5. The solubility of proteins in the presence of dextran. *Biochem*. 89, 253-257

Lawrence, M.S., Phillips, K.J., Liu, D.R. (2007). Supercharging proteins can impart unusual resilience. *J Am Chem Soc*. **129**(33), 10110-10112

Lee, D., Smallbone, K., Dunn, W.B., Murabito, E., Winder, C.L., Kell, D.B., Mendes, P., Swainston, N. (2012). Improving metabolic flux predictions using absolute gene expression. *BMC Syst Biol*. 6:73

Lee, J.C. and Timasheff, S.N. (1981). The stabilisation of proteins by sucrose. *J Biol Chem*. **256**(14), 7193 – 7201

Li, Y. and Roberts, C.J. (2009). Lumry-Eyring nucleated-polymerization model of protein aggregation kinetics. 2. Competing growth via condensation and chain polymerization. *J Phys Chem B*. **113**(19), 7020 – 7032

Linding, R., Russell, R.B., Neduva, V., Gibson, T.J. (2003). GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res*. **31**(13), 3701-3708

Linse, B. and Linse, S. (2011). Monte Carlo simulations of protein amyloid formation reveal origin of sigmoidal aggregation kinetics. *Mol Biosyst*. **7**(7), 2296-2303

Lovell, S.C., Davis, I.W., Arendall, W.B. III, de Bakker, P.I., Word, J.M., Prisant, M.G., Richardson, J.S., Richardson, D.C. (2003). Structure validation by C-alpha geometry: phi, psi and C-beta deviation. *Proteins*. **50**(3), 437-450

Lu, P., Vogel, C., Wang, R., Yao, X., Marcotte, E. (2007). Absolute protein expression profiling estimates the relative contributions of transcriptional and translation regulation. *Nat Biotechnol*. **25**(1), 117-124

Ma, J.C. and Dougherty, D.A. (1997). The Cation- π interaction. *Chem Rev*. **97**, 1303-1324

Machida, S., Yu, Y., Singh, S.P., Kim, J.D., Hayashi, K., Kawata, Y. (1998). Overproduction of beta-galactosidase in active form by an *Escherichia coli* system coexpressing the chaperonin GroEL/ES. *FEMS Microbiol Lett*. **159**(1), 41-46

Magee J. and Warwicker J. (2005). Simulation of non-specific protein-mRNA interactions. *Nucleic Acids Res.* 33(**21**), 6694-6699

Magnan, C.N., Randall, A., Baldi, P. (2009). SOLPro: accurate sequence-based prediction of protein solubility. *Bioinformatics*. **25**(17), 2200-2207

Maji, S.K., Perrin, M.H., Sawaya, M.R., Jessberger, S., Vadodaria, K., Rissman, R.A., Singru, P.S., Nilsson, K.P., Simon, R., Schubert, D., Eisenberg, D., Rivier, J., Sawchenko, P., Vale, W., Riek, R. (2009). Functional amyloids as natural storage of peptide hormones in pituitary secretory granules. *Science*. **325**(5938), 328-332

Manning, M.C., Patel, K., Borchardt, R.T. (1989). Stability of protein pharmaceuticals. *Pharm Res*. **6**(11), 903-9118

Martis, R.L., Singh, S.K., Gromiha, M.M, Santhosh, C. (2008). Role of cation-pi interactions in single chain 'all-alpha' proteins. *J Theor Biol*. **250**(4), 655-662

Mendez, C.M., McClain, C.J., Marsano, L.S. (2005). Albumin therapy in clinical practice. *Nutr Clin Pract*. **20**(3), 314 – 320

Minton, A.P. (2005). Models for excluded volume interaction between an unfolded protein and rigid macromolecular cosolutes: macromolecular crowding and protein stability revisited. *Biophys J*. **88**(2), 971-985

Mirceta, S., Signore, A.V., Burns, J.M., Cossins, A.R., Campbell, K.L., Berenbrink, M. (2013). Evolution of mammalian diving capacity traced by myoglobin net surface charge. *Science*. **340**(6138): 1234192

Molinari, M. (2007). N-glycan structure dictates extension of protein folding or onset of disposal. *Nat Chem Biol*. **3**(6), 313-320

Mosavi, L.K. and Peng, Z.Y. (2003). Structure-based substitutions for increased solubility of a designed protein. *Protein Eng*. **16**(10), 739-745

Nanjappa, V., Thomas, J.K., Marimuthu, A., Muthusamy, B., Radharkrishnan, A., Sharma, R., Ahmad, Khan A., Balakrishnan, L., Sahasrabuddhe N.A., Kumar S., Jhaveri B.N., Sheth, K.V., Kumar, Khatana, R., Shaw, P.G., Srikanth, S.M., Mathur, P.P., Shankar, S., Nagaraja, D., Christopher, R., Mathivanan, S., Raju, R., Sirdeshmukh, R., Chatterjee, A., Simpson, R.J., Harsha, H.C., Pandey, A., Prasad, T.S. (2014). Plasma Proteome Database as a resource for proteomics research: 2014 update. *Nucleic Acids Res*. **42**, 959-965

Neves-petersen, M.T. and Petersen, S.B. (*2003).* Protein electrostatics: A review of the equations and methods used to model electrostatic equations in biomolecules – Applications in biotechnology. *Biotechnol Annu Rev*. **9**, 315-395

Nielsen, J.E. (2007). Analysing the pH-dependent properties of proteins using pKa calculations. *J Mol Graph Model*. **25**(5), 691-699

Nielsen, L., Khurana, R., Coats, A., Frokjaer, S., Brange, J., Vyas, S., Uversky, V.N., Fink, Al.L. (2001). Effect of environmental factors on the kinetics of insulin fibril formation: elucidation of the molecular mechanism. *Biochemistry*. **40**(20), 6036-6046

Nimmo, I.A., Atkins, G.L., Strange, R.C., Percy-Robb, I.W. (1977). An evaluation of ways of using equilibrium dialysis to quantify the binding of ligand to macromolecule. *Biochem J*. **165**(1), 107-110

Niu, X., Li, N., Chen, D., Wang, Z. (2013). Interconnection between the protein solubility and amino acid and dipeptide compositions. *Protein Pept Lett*. **20**(1), 88-95

Niwa, T., Ying, B., Saito, K., Jin, W., Takada, S. Ueda, T. Taguchi, H. (2009). Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins. *Proc Natl Acad Sci USA*. **106**(11), 4201-4206

Nozaki Y. and Tanford, C. (1963). The solubility of amino acids and related compounds in aqueous urea solutions. *J Mol Biol*. **238**, 4074 – 4081

Obrezanova, O., Arnell, A., de la Cuesta, R.G., Berthelot, M.E., Gallagher, T.R., Zurdo, J., Stallwood, Y. (2015). Aggregation risk prediction for antibodies and its application to biotherapeutic development. *MAbs*. **7**(2), 352-363

Ohtake, S., Kita, Y., Arakawa, T. (2011). Interactions of formulation excipients with proteins in solution and in the dried state. *Adv Drug Deliv Rev*. **63**(13), 1053-1073

Olsen, S.N., Andersen, K.B., Randolph, T.W., Carpenter, J.F., Westh, P. (2009). Role of electrostatic repulsion on colloidal stability of Bacillus halmapalus alpha-amylase. *Biochim Biophys Acta*. **1794**(7), 1058-1065

Otzen, D.E., Kristensen, O., Oliveberg, M. (2000). Designed protein tetramer zipped together with a hydrophobic Alzheimer homology: a structural clue to amyloid assembly. *Proc Natl Acad Sci USA*. **97**(18), 9907-9912

Pelegrine, D.H.G. and Gasparetto, C.A. (2005). Whey proteins solubility as function of temperature and pH. *LWT*. **38**(1), 77-80.

Pennacchio, L.A., Olivier, M., Hubacek, J.A., Cohen, J.C., Cox, D.R., Fruchart, J.C., Krauss, R.M., Rubin, E.M. (2001). An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science*. **294**(5), 169-173.

Perchiacca, J.M, Bhattacharya, M., Tessier, P.M. (2011). Mutational analysis of domain antibodies reveals aggregation hotspots within and near the complementarity determingin regions. *Proteins*. **79**(9), 2637 – 2647

Pikal, M.J. (1990). Freeze-drying of proteins. I Process design. *BioPharm* 3, 18-27

Price, W.N. II, Handelman, S.K., Everett, J.K., Tong, S.N., Bracic A., Luff, J.D., Naumov, V., Acton, T., Manor, P. Xiao, R. Rost, B., Montelione, G.T., Hunt, J.F. (2011). Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility *in vivo* in *E. coli*. *Microb Inform Exp*. **1**(1):6

Pushker, R., Mira A., Rodriguez-Valera, F. (2004) Comparative genomics of gene-family size in closely related bacteria. *Genome Biol*. **5**(4), R27

Ralston, G.B. (1990). Effects of "crowding" in protein solutions. *J Chem Educ*. **67**(10), 857 – 860

Ramachadran, G.N. and Sasisekharan, V. (1968). Conformation of polypeptides and proteins. *Adv Protein Chem*. 23, 283-438

Richardson, J.S. and Richardson, D.C. (2002). Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci USA*. **99**(5), 2754-2759

Roberts, C.J. (2007). Non-native protein aggregation kinetics. *Biotechnol Bioeng*. **98**(5), 927 – 938

Roberts, C.J. (2014). Therapeutic protein aggregation: mechanisms, design, and control. *Trends Biotechnol*. **32**(7), 372 – 380.

Roberts, D., Warwicker, J., Curtis, R.A. (2015). Molecular Modeling for Protein Aggregation and Formulation. In: Ouyang, D. and Smith, S.C. eds. *Computational Pharmaceutics: Application of Molecular Modeling in Drug Delivery*. Wiley, pp. 123-147

Rosano, G.L. and Ceccarelli, E.A. (2014). Recombinant protein expression in Escherichia coli: advances and challenges. *Front Microbiol*. **5**: 172

Rosenberg, A.S. (2006). Effects of protein aggregates: an immunologic perspective. *AAPS J*. **8**(3), E501-507

Rousseau, F., Serrano, L., Schymkowitz, J.W. (2006). How evolutionary pressure against protein aggregation shaped chaperone specificity. *J Mol Biol*. **355(**5), 1037-1047

Saluja, A., Badkar, A.V., Zeng, D.L., Kalonia, D.S. (2007). Ultrasonic rheology of a monoclonal antibody (IgG2) solution: implications for physical stability of proteins in high concentration formulations. *J Pharm Sci*. **96**(12), 3181-3195

Sawaya, M.R., Sambashivan, S., Nelson, R., Ivanova, M.I., Sievers, S.A., Apostol, M.I., Thompson, M.J., Balbirnie, M., Wiltzius, J.J., McFarlane, H.T., Madsen, A.O., Riekel, C., Eisenberg, D. (2007). Atomic structures of amyloid cross-beta spines reveal varied steric zippers. *Nature* **447**(7143), 453-457

Schneider, C.P. and Trout, B.L. (2009). Investigation of cosolute-protein preferential interaction coefficients: new insight into the mechanism by which arginine inhibits aggregation. *J Phys Chem B*. **113**(7), 2050 – 2058

Schreiber, G. and Keating, A.E. (2011). Protein binding specificity versus promiscuity. *Curr Opin Struct Biol*. **21**(1), 50-61

Schwartz, R., Istrail, S., King, J. (2001). Frequencies of amino acid strings in globular protein sequenes indicate suppression of blocks of consecutive hydrophobic residues. *Protein Sci*. **10(**5), 1023-1031

Sek, D.C. (2008). Protein Formulations Containing Sorbitol. US 12/032, 478

Shah, D., Li, J., Shaikh, A.R., Rajagopalan, R. (2012). Arginine-aromatic interactions and their effects on arginine-induced solubilization of aromatic solutes and suppression of protein aggregation. *Biotechnol Prog*. **28**(1), 223-231

Shiraki, K., Kudou, M., Fujiwara, S., Imanka, T., Takagi, M. (2002). Biophysical effect of amino acids on the prevention of protein aggregation. *J Biochem*. **132**(4), 591-595

Shukla, D. and Trout, B.L. (2010). Interaction of arginine with proteins and the mechanism by which it inhibits aggregation. *J Phys Chem B*. **114**(42), 13426-13438

Shukla, D. and Trout, B.L. (2011). Understanding the synergistic effect of arginine and glutamic acid mixtures on protein solubility. *J Phys Chem B*. **115**(41), 11831-11839

Shukla, D. and Trout, B.L. (2011). Preferential interaction coefficients of proteins in aqueous arginine solutions and their molecular origins. *J Phys Chem B*. **115**(5), 1243-1253

Sinha, N. and Smith-Gill, S.J. (2002). Electrostatics in protein binding and function. *Curr Protein Pept Sci*. **3**(6), 601-614

Smialowski, P., Doose, G., Torkler, P. (2012). PROSO II – a new method for protein solubility prediction. *FEBS J*. **279**(12), 2192-2200

Smialowski, P., Martin-Galiano, A.J., Mikolajka, A. (2007). Protein solubility: a sequence based prediction and experimental verification. *Bioinformatics*. **23**(19), 2536-2542

Sokalingam, S., Raghunathan, G., Soundrarajan, N., Lee, S.G. (2012). A study on the effect of surface lysine to arginine mutagenesis on protein stability and structure using green fluorescent protein. *PLoS One*. **7**(7), e40410

Su, Y. and Chang, P.T. (2001). Acidic pH promotes the formation oftoxic fibrils from beta-amyloid peptide. *Brain Res*. **893**(1-2), 287-291

Tanford, C. (1979). Interfacial free energy and the hydrophobic effect. *Proc Natl Acad Sci USA*. 76(9), 4175-4176

Tartaglia, G.G., Cavelli, A., Pellarin, R., Caflisch, A. (2005). Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Sci*. **14**(10), 2723-2734

Tartaglia, G.G., Pawar, A.P., Campioni, S., Dobson, C.M., Chiti, F., Vendruscolo, M. (2008). Prediction of aggregation-prone regions in structured proteins. *J Mol Biol.* **380**(2), 425-436

Tartaglia, G.G., Pechmann, S., Dobson, C.M., Vendruscolo, M. (2007). Life on the edge: a link between gene expression levels and aggregation rates of human proteins. *Trends Biochem Sci.* **32**(5), 204-206

Tartaglia, G.G., Pechmann, S., Dobson, C.M., Vendruscolo, M. (2009). A relationship between mRNA expression levels and protein solubility in *E. coli*. *J Mol Biol.* **388**(2), 381-389

Tartaglia, G.G and Vendruscolo, M. (2009). Correlation between mRNA expression levels and protein aggregation propensities in subcellular localisations. *Mol Biosyst*. **5**(12), 1873-1876

The UniProt Consortium (2015). UniProt: a hub for protein information. *Nucleic Acids Res*. **43**: D204-D212

Thompson, M.J., Sievers, S.A., Karanicolas, J., Ivanova, M.I., Baker, D., Eisenberg, D. (2006). The 3D profile method for identifying fibril-forming segments of proteins. *Proc Natl Acad Sci USA.* **103**(11), 4074-4078

Timasheff, S.N. (1998). Control of protein stability and reactions by weakly interacting cosolvents: the simplicity of the complicated. *Adv Protein Chem*. **51**, 355 – 432

Timasheff, S.N. (2002). Protein-solvent preferential interactions, protein hydration, and the modulation of biochemical reactions by solvent components. *Proc Natl Acad Sci USA*. **99**, 9721 – 9726

Tirumalai, R.S., Chan, K.C., Prieto, D.A., Issag, H.J., Conrads, T.P., Veenstra, T.D. (2003). Characterization of the low molecular weight human serum proteome. Mol Cell Proteomics. **2**(10), 1096-1103

Tiwari, A. and Bhat, R. (2006). Stabilisation of yeast hexokinase A by polyol osmolytes: correlation with the physico-chemical properties of aqueous solutions. *Biophys Chem*. **124**, 90-99

Tjernberg, L., Hoisa, W., Bark, N., Thyberg, J., Johansson, J. (2002). Charge attraction and beta propensity are necessary for amyloid fibril formation from tetrapeptides. *J. Biol Chem.* **277**(45), 43243-43246

Tjong, H. and Zhou, H.X. (2008). Prediction of protein solubility from calculation of transfer free energy. *Biophys J*. **95**(6), 2601-2609

Tompa, P. (2002). Intrinsically unstructured proteins. *Trends Biochem Sci*. 27(10), 527-53

Traube, J. (1910). *J Phys Chem*. **14**, 452 – 470

Trevino, S.R., Scholtz, J.M., Pace, C.N. (2007). Amino acid contribution to protein solubility: Asp, Glu, and Ser contribute more favorably than the other hydrophilic amino acids in RNase Sa. *J Mol Biol*. **366**(2), 449-460

Trinquier G. and Sanejouand, Y.H. (1998). Which effective property of amino acids is best preserved by the genetic code? *Protein Eng*. **11**(3), 153 – 169

Trovato, A., Chiti, F., Maritan, A., Seno, F. (2006). Insight into the structure of amyloid fibrils from the analysis of globular proteins. *PLoS Comput Biol.* **2**(12), e170

Tsai, A.M., van Zanten, J.H., Betenbaugh, M.J. II (1998). Electrostatic effect in the aggregation of heat-denatured RNase A and implications for protein additive design. *Biotechnol Bioeng*. **59**, 281-285

Tsai, P.K., Vokin, D.B., Dabora, J.M, Thompson, K.C., Bruner, M.W., Gress, J.O., Matuszewska, B., Keogan, M., Bondi, J.V. (1993). Formulation of acidic fibroblast growth factor. *Pharm Res*. **10**(5), 649-659

Uversky, V.N., Gillespie, J.R., Fink, A.L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*. **41**(3), 415-427

van den Berg, B., Reinders, M., M.J., Hulsman, M., Wu, L., Pel, H.J., Roubos, J.A., de Ridder, D. (2012). Exploring Sequence Characteristics Related to High-Level Production of Secreted Proteins in Aspergillus niger. *PLoS One*. **7**(10).

van den Berg, B., Reinders., M, Roubos, J.A., de Ridder, D. (2014). SPiCE: a web-based tool for sequence-based protein classification and exploration. *BMC Bioinformatics.* **31**, 15:93

Vázquez-Rey, M. and Lang, D.A. (2011). Aggregates in monocolonal antibody manfucaturing processes. *Biotechnol Bioeng*. **108**(7), 1495-1508

Velankar, S., Dana, J.M., Jacobsen, J., van Ginkel, G., Gane, P.J., Luo, J., Oldfield, T.J., O'Donovoan, C., Martin, M.J., Kleywegt, G.J. (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequence. *Nucleic Acids Res*. **41**, 483-489

Ventura, S. and Villaverde, A. (2006). Protein quality in bacterial inclusion bodies. *Trends Biotechnol*. **24**(4), 179-185

Vogel, C. Bashton, M., Kerrison, N.D., Chothia, C. Teichmann, S.A. (2004). Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol*. **14**(2), 208-216

Voynov, V., Chennamsetty, N., Kayser, V., Helk, B, Trout, B.L. (2009). Predictive tools for stabilization of therapeutic proteins. **1**(6), 580-582

Wang, G. and Dunbrack, R.L. Jr. (2005). PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acid Res*. **33**, W94-98

Wang , W., Singh, S., Zeng, D.L, King, K., Nema, S. (2007). Antibody structure, instability, and formulation. *J Pharm Sci*. **96**(1), 1 – 26

Wang, M., Weiss, M. Simonovic, M., Haertinger, G., Schrimpf, S.P., Hengartner, M.O., von Mering, C. (2012). PaxDb, a database of protein abundance averages across all three domains of life. *Mol Cell Proteomics*. **11**(8), 492-500

Warwicker, J. (1986). Continuum dielectric modeling of the protein-solvent system, and calculation of the long-range electrostatic field of the enzyme phosphoglycerate mutase. *J Theor Biol*. **121**, 199-210

Warwicker, J. (1999). Simplified methods for pKa and acid pH-dependent stability estimation in proteins: removing dielectric and counterion boundaries. *Protein Sci*. **8**(2), 418-425

Warwicker, J. (2011). pKa predictions with a coupled finite difference Poisson-Boltzmann and Debye-Hückel method. *Proteins*. **79**(12), 3374-3380

Warwicker, J., Charonis, S., Curtis, R.A. (2014). Lysine and Arginine Content of Proteins: Computational Analysis Suggests a New Tool for Solubility Design. *Mol Pharm*. **11**(1), 294-303

Waugh, D.S. (2005). Making the most of affinity tags. *Trends Biotechnol*. **23**(6), 316-320

West, M.W., Wang, W.X., Patterson, J., Mancias, J.D., Beasley, J.R., Hecht, M.H. (1999). *De novo* amyloid proteins from designed combinatorial libraries. *Proc Natl Acad Sci USA.* **96**(20), 11211-11216

White, A.D., Keefe, A.J., Ella-Menye, J.R., Nowinski, A.K., Shao, Q., Pfaendtner, J., Jiang, S. (2013). Free energy of solvated salt bridges: a simulation and experimental study. *J Phys Chem B*. **117**(24), 7254-7259

Wilkinson, D.L. and Harrison, R.G. (1991). Predicting the solubility of recombinant proteins in *Escherichia coli*. *Nat Biotechnol*. **9**(5), 443-448

Won, C.M, Molnar, T.E., McKean, R.E., Spenlehauer, G.A. (1998). Stabilisers against heat-induced aggregation of RPR 114849, an acidic fibroblast growth factor (aFGF). *Int J Pharm*. **167**, 25-36

Wu, H., Kroe-Barrett, R., Singh, S., Robinson, A.S., Roberts, C.J. (2014). Competing aggregation pathways for monoclonal antibodies. *FEBS Lett*. **588**(6), 936 – 941

Xie, G. and Timasheff, S.N. (1997). Mechanism of the stabilisation of ribonuclease A by sorbitol: preferential hydration is greater is greater for the denatured then for the native protein. *Protein Sci*. **6**, 211 – 221

Yancey, P.H., Clark, M.E., Hand, S.C., Bowlus, R.D., Somero, G.N. (1982). Living with water stress: evolution of osmolyte systems. *Science*. **217**(4566): 1214 – 1222

Yadav, S., Laue, T.M., Kalonia, D.S., Singh, S.N., Shire, S.J. (2012). The influence of charge distribution on self-association and viscosity behavior of monoclonal antibody solutions. *Mol Pharmaceutics*. **9**(4), 791-802

Zhang, H, Saiani, A., Guenet, J., Curtis, R.A. (2007). Effect of stereoregular polyelectrolyte on protein thermal stability. *Macromolec Symp*. **251**(1), 25-32

Zhou, H.X. (2005). Interactions of macromolecules with salt ions: an electrostatic theory for the Hofmeister effect. *Proteins*. **61**(1), 69-78

Zurdo, J., Guijarro, J.I., Jimenez, J.L., Saibil, H.R., Dobson, C.M. (2001). Dependence on solution conditions of aggregation and amyloid formation by an SH3 domain. *J Mol Biol*. **311**(2), 325-340

Zwanzig, R., Szabo, A., Bagchi, B. (1992). Levinthal's paradox. *Proc Natl Acad Sci USA*. **89**(1), 20-22

Supplementary Files (Available on the web)

Supplementary File 2.1: pdb2fasta_scfv.sh (Chapter 2)
Description: Unix shell script for parsing sequence from a directory of scFv structures (PDB files)

Supplementary File 2.2: pdb2fasta_fab.sh (Chapter 2)
Description: Unix shell script for parsing sequence from a directory of Fab fragment structures (PDB files)

Supplementary File 2.3: get_biol_chains.sh (Chapter 2)
Description: Unix shell script for parsing biological chains from a directory of non-antibody biologics structures (PDB files)

Supplementary File 2.4: pdb2csv.py (Chapter 2)
Description: Python program for reading a PDB file and extracting sequence-based information (four additional Unix scripts required for running are included)

Supplementary File 3.1: fasta2csv.py (Chapter 3)
Description: Python program for reading a FASTA file with multiple sequences in order to compute and plot the mean residue composition

Supplementary File 4.1: extract_pdbexpress_stats.py (Chapter 4)
Description: Python program for extracting data fields from the PDBeXpress output of a queried ligand

Supplementary File 4.2: extract_pdbexpress_pdbs.py (Chapter 4)
Description: Python program for extracting PDB annotations from the PDBeXpress output of a queried ligand

Supplementary File 4.3: lig_stats.py (Chapter 4)
Description: Python program for plotting the distribution generated from brute force sampling of excipients

Supplementary File 4.4: search_pdb_contacts.py (Chapter 4)
Description: Python program for parsing PDB contacts from the PDBeXpress output for each amino acid of a given ligand