Amino acid residue burial

&
Co-evolution in proteins

by

Bhima Auro

A thesis submitted to
The University of Birmingham
for the degree of
Doctor of Philosophy (Sc, PhD)

College of Life and Environmental Sciences
School of Biosciences
The University of Birmingham
May 2013

Abstract

Correlated mutation is probably the most common term used in the literature to refer to the observation that the effects of amino acid substitutions at one part of a protein structure seem to be accompanied by changes elsewhere in the structure. Correlated mutation analysis and other methods for analysing amino acid co-substitution patterns in protein structures have been developed over the last few decades to predict inter-residue contacts within protein structures. The methods have had moderate success and it is clear that there is some observable signal. It is further clear that correlated mutations are not limited merely to contacting residues, although the reasons for this are less clear. This thesis outlines the development of a method for determining the relationship between specific amino acid co-substitution events and spatial distance between them in a protein structure, and presents the results of preliminary analyses of these effects.

To rigorously analyse the relationship between co-substitution events and inter-residue distance it has been necessary to develop a statistical framework for the co-substitution analysis. However, to ensure the analysis was statistically rigorous and precise, it has also been necessary to develop a system for selecting qualitatively similar subsets of protein structures and associated sequence alignments, e.g. cytoplasmic globular proteins in Eukaryota. The bias certain amino acid types have to be either on the protein surface or the solvent excluded volume, results in compositional differences between the two regions, which could confound our interpretation of co-substitution events. These known differences for solvent exposure preference, necessitated that this be investigated as well. In addition to these technically challenging analyses it has also been necessary to develop a statistical framework for the co-substitution analysis itself.

The investigation into the propensity for amino acid solvent exposure and co-substitution distance relationships required very specific selection of the data. This required cross referencing data stored across several on-line databases. To deal with this challenge, a MySQL database was built, which contains data from UniProt/SwissProt, Pfam-A and selected data from the PDB and PiQSi databases. Combined with a regular expression module for Python, I have been able to create an accurate three-way map between Pfam-A, SwissProt and the PDB, including accurate locations of Pfam-domains in PDB structures.

Many different criteria have been used to define a residue in a protein as buried, none of which appear to be founded on a rigorous statistical observation of the solvent exposure of amino acids in protein structures. An investigation of the statistical propensity of amino acid solvent exposure in non-membrane, non-DNA binding cytoplasmic globular proteins, to answer the question "is there a single value of some solvent exposure measure that defines the crossover point from amino acid solvent exposure to burial?" was undertaken. The results suggest half sphere exposure (HSE) is a more reliable measure to determine and define this crossover. Further, the results indicate that a value of 20 HSEu (using a radius of 13 Å) is where the crossover occurs, for the subset of proteins analysed. HSE has a particular advantage over relative solvent accessible surface area (%ASA) in that it provides a measure through the whole range of burial/solvent exposure, being able to measure the depth of burial below the solvent accessible surface, and thus sample more reliably either side of the transition point. Unlike rASA it is not dependent on the definition of some reference state conformation and it is effectively independent of sidechain identity. This latter point means that a measurement of HSE made in a reference structure can be applied to all amino acid types that are found at that position in a multiple sequence alignment. This makes it easier than the accessible surface area (ASA) to work with in more complicated statistical procedures such as bootstrapping calculations. The results produced here by HSE also appear to be more consistent across residue types compared to those produced by ASA.

Finally, a statistical framework for elucidating the propensity for different co-substitution events to occur at different distances has been developed. Initial results indicate that there are

interesting effects to be observed. For the co-substitution type, RD $\leftrightarrow$ KE, the co-substitution behaviour differs between the surface and the interior of proteins. Differences between Eukaryota and Prokaryota are also present in the data, however the statistical significance of these is not certain. Finally, the initial results indicate that the residue types argenine and aspartic acid (RD) preferentially co-substitute to lysine and glutamic acid (KE) through a non-contacting long range interaction. All calculations have been accompanied by extensive bootstrapped calculations to estimate their statistical significance. The observations in the co-substitution calculations require increased data to confirm the statistical significance of these findings, and there is detailed discussion of how to achieve this, and how to further speedup the calculations that we are performing.

Acknowledgements

Firstly I wish to dedicate this thesis to my wife Sudha Auro and recognise the tremendous love, support and encouragement she has given me. She has been with me from the beginning and has given substantially of herself, to help me see this through. Additionally, I want acknowledge my son Siddharth Aditya Auro, who arrived early to be with us when his father defended this document, for being a wonderful addition to our lives.

It has been my great privilege to pursue this project, and I have been humbled by the support that has been afforded me by friends and mentors alike.

David Finch has been a good friend and great source of encouragement and support, even while he suffered through his disability. My family: my sisters Angiras and Aurama, my brother Emmanuel, my parents Judith and Gilles, and my father Pieter; thank you for being there.

Dr. Leon D'Cruz, has been a great champion of my cause for many years and without his support and direction I would likely not have come this far. I wish to recognize and thank him for all he has done. I am also grateful to Prof. ACC Coolen, of King's College London, for his advice, encouragement and active support.

During my time in the Centre for Systems Biology, I have had the great pleasure to get to know a wonderfully eclectic group of people. I want to thank them all for their humour, patience and intellect. Prof. John Heath, Dr. Jan Kreft, Prof. Mark Pallen and his group, and the occasional gaggle of project students who passed through briefly to liven things up. Fellow students and post-docs, Chinmay Kanchi, Sonia Martins, Robert Clegg, Jackie Chan (no relation), Susanne Schmidt and Francis Amrit, have all become good friends. Thank you for making this about more than just research.

I must recognise the invaluable contributions to my work from two colleagues. Firstly, Chinmay Kanchi, who took the time to introduce me to the Python programming language and helped me become a programmer. Secondly, Robert Clegg who was very supportive in checking that the maths was right and his substantial contribution towards the pair-wise sequence weighting method presented in this thesis.

Two undergraduate students were assigned to my joint supervision, while working on specific side projects related to my research. John Le Brun and Matthew Welland. Both of them have been great company, dedicated and hard working, showing initiative and clearly owning their projects. They have both been inspirational in their own way. Their work has enriched my research and working with them has made my experience of this project more memorable.

I would like to recognise the direction given to me by Dr. Klaus Fütterer and Dr. Eva Hyde, in their respective roles as second supervisor and internal assessor. Their insistence on details and clear explanations of my work and their comments have added to the general quality of my work.

I wish to thank and recognise Simon Hubbard, author of the Naccess computer program; for supplying me with the structure files containing the reference state of amino acids, used in the Naccess program.

It would be remiss to not acknowledge the support from members of the University. Anthony Pemberton, our systems administrator. His willingness and availability to sort out all manner of issues, with the computing cluster and individual workstations, has provided a significant contribution to my work. The administrative staff in the School of Biosciences, specifically Anne Begum and Holly Etchell for chasing paperwork on my behalf. And Dr. Neil Hotchin, for his help and support in dealing with the unexpected circumstances.

Lastly, but by no means least, I wish to thank and recognise the contribution of my supervisor Dr. Peter J Winn. Over the course of this project, he has been an invaluable guide through my research. He has always been available to tell me I'm wrong, and encouraged me to prove him otherwise, allowing me to run free with ideas when it was right to do so, and an unbearable micro-manager when required. He has been the supervisor I needed him to be.

CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

The development of an analytical method to asses the relationship between amino acid pairwise substitutions and their inter-amino-acid Euclidean distances, is presented in this thesis. Unlike other methods of investigating co-evolution in protein structures, this work is concerned with elucidating the relationship between inter-residue distance and co-substitution events. The literature is full of studies exploring correlated mutations and co-evolution with the set aim of predicting inter-residue contacts. None of these studies considers the role of non-contact pairwise interactions of amino-acid residues occurring over some distance in the structure. The method presented in this thesis is able to determine the propensity with respect to distance for all combinations of residue types, for both those in pairwise substitutions and for conserved amino acid residue pairs. With additional development the output of the analysis can be incorporated into a Bayesian statistical method to predict protein structures.

The favourability of an amino acid to be involved in a substitution event is affected by its environment. Environment in this context encompasses neighbouring amino acids in the protein structure and also solvent exposure and cellular location. Two additional sub-projects were developed to allow solvent exposure and cellular location to be incorporated into the co-substitution analysis – other environmental effects such as secondary structure were not considered at this time. Firstly, no straightforward method of selecting structural and sequence data based on cellular location was available. To address the need for such a method or tool, led to a bioinformatics project to develop a database consisting of the merged data from Pfam,

the PDB/PiQSi and SwissProt databases. Secondly, there is no consensus value for amino-acid solvent exposure in a protein structure that delimits the crossover from a hydrophilic environment to a hydrophobic one in protein structures. To address this omission in the literature, an investigation into amino acid solvent exposure was undertaken, to see if such a crossover could be determined.

In this introductory chapter, a brief discussion on the context of the main project "the protein structure" is given. This is then followed by two reviews. Firstly a review of co-evolution analysis methods in the literature; secondly a review of solvent expsoure measures. The chapter concludes with a brief discussion of the scope of this thesis.

1.1 Fundamentals of protein structures

1.1.1 Amino acids and protein structures

Protein structures are made up from component parts known as amino acids. There are 20 common proteinogenic amino acids – there are many other naturally occurring ones (e.g. taurine) that are not incorporated into protein structures and the 20 common proteinogenic residues may undergo post-translational modification after incorporation into protein structures (e.g. hydroxyproline or phosphotyrosine). The central dogma of molecular biology stipulates that *DNA is transcribed into RNA and RNA is translated into an amino acid sequence, which folds to form a protein.* During protein synthesis, amino acids are combined into sequences, which then fold to give a functionally active protein. The twenty amino acid types each have specific and unique physico-chemical properties, which when combined together have distinct effects on the shape and the structure formed by the folded sequence.

The number of possible conformations that can be assumed by an unfolded protein sequence is vast, and the amount of time it would take for a protein to "search" all possible conformations to arrive at the native state, could exceed the available time projected to the end of our Universe. Yet, proteins fold spontaneously and reliable into their native state within fractions of a second. This is known as Levinthal's paradox and is a significant consideration in computational methods of protein structure prediction [6]. A statistical description of a protein's potential surface,

known as “energy landscape theory” offers a perspective of the folding process. The theory is based on the assumption that folding is not a series of steps between a set of unique intermediate structural conformations. Rather it assumes that the “folding occurs through organising an ensemble of structures” [7]. A protein’s “energy landscape” can be described as the polypeptide-chain moving from a high energy state, which is flexible, to a stable structure at a lower energy state. This energy state may be a local minima in the “energy-landscape” or it may be the global one, depending on the energy required to overcome any local maxima, in the folding process. Generally the landscape is a rugged funnel-like shape, with the native state of the protein being at the bottom of the funnel [7]. The positions of different amino-acid types at various positions in the sequence is understood to direct the folding pathway in this energy landscape, but how and why this is the case is not fully understood [8].

Though the folding process itself is not completely understood, this has not been a barrier to uncovering fundamentals of protein structures. For example, it is known that the structure and function of proteins are determined by the sequence and order of amino acid residues which form the polypeptide chain. That a protein’s native conformation determines its biological activity, and that the amino acid sequence is what defines the native conformation, was proven by Christian Anfinsen, for which he shared the 1972 Nobel prize in chemistry [9]. His work on the thermodynamic hypothesis – which states that the physiological structure of a protein is the one whose Gibbs free energy for the whole system is the lowest – led him to this discovery. The whole system refers to the physiological environment where the protein would exist in its structural and functional form. This would be defined by specific pH, temperature, ionic strengths, the presence of other components such as metal ions and any other contributing factors. “*That is to say, that the native conformation is determined by the totality of inter-atomic interactions and hence by the amino acid sequence, in a given environment. In terms of natural selection, through the “design” of macromolecules during evolution, this idea emphasised the fact that a protein molecule only makes stable, structural sense when it exists under the conditions similar for which it was selected - the so called physiological state.*” [10].

Protein structures can be classified at three levels, the primary structure, the secondary struc-

ture and the tertiary structure. The primary structure refers to the sequence of amino acid residues, covalently bonded in the polypeptide chain. The secondary structure refers to local structural motifs from which higher structure is created, defined by the inter-residue hydrogen bonding patterns of the hydrogen from the NH of one residue with the oxygen from the CO of another. These secondary structure units include α helices, β strands & sheet, π-helices and 3_{10}-helices; each of which are formed by identifiable patterns of physico-chemical properties of amino-acid residues. These structural units contribute to the stability and function of protein structures. The tertiary structure refers to the complete atomic structure of the protein. The arrangement of amino-acid residues in the primary structure, is what defines the secondary and tertiary structures of proteins. Strictly speaking there is a fourth classification, the quaternary structure; which refers to multiple polypeptide chains combined as a complex. The formation of secondary and higher degrees of protein structure is driven by the hydrophobic effect [11], described below.

1.1.2 The hydrophobic effect and protein stability

By studying the effects of mixing hydrocarbons with water and other solvents, Walter Kauzmann described a phenomenon we know today as the hydrophobic effect. It was this work that proposed the now generally accepted model of the folded protein, a proposal made prior to the first protein structure being resolved by x-ray crystallography. From his experiments he was able to predict that proteins in aqueous environments would fold into complex ribbons; that hydrophobic amino acids would be located away from the solvent because of the hydrophobic effect and hydrophilic ones would be located on the protein's solvent accessible surface [12].

The details of the hydrophobic effect are still not completely understood; although it is clear that it is largely the results of the way water is forced to rearrange itself in response to a non-hydrogen bonding solute. David Chandler has developed a theory describing the balance of forces in the hydrophobic effect. Through molecular dynamic simulations he has shown that the hydrophobic effect is size dependant; such that small hydrophobic particles can be "solvated" if they are small enough not to break the hydrogen bonds between water molecules. Formation of

a small cavity in bulk water will distort the hydrogen bonds between water molecules. Which Chanlder has shown has an average thermodynamic cost which scales with volume. When the cavity size exceeds some maximum it breaks the hydrogen bonds between the water molecules; at which point the thermodynamic cost scales with the surface area of the cavity. Hydrophobic particles which are big enough to break the hydrogen-bonds between water (slightly larger than the size of one methane molecule), will tend to aggregate in order to minimise the surface area of the cavity formed in the water. However, if the hydrophobic particles are too small to break the hydrogen bonds, depending on the concentration of particles in water, the system would tend to an equilibrium between these two regimes. Where on the one hand the cummulative energetic cost of distorting the hydrogen bonds of water is less than breaking the bonds, thus the particles will remain individually solvated in the solute. Or, on the other hand, the cummulative cost of distorting the hydrogen bonding might be more than breaking them and thus aggregation would occur to minimise the surface area of the cavity [13]. How this size dependence of the hydrophobic effect relates to proteins is not well understood, due to their complex amphiphilic surface.

Two similar methods for determining the solvent accessible surface area were proposed in the early 1970s, to investigate the burial of hydrophobic surface area in proteins, the first method was proposed in 1971 by Lee and Richards [14] and this was followed in 1973 by Shrake and Rupley [15], these methods are described in Section 1.3.1. The atoms in a protein molecule can be classed as being either polar or non-polar. Both groups showed that approximately half of the accessible atoms are polar while the remaining are non-polar. Richards in 1977 stated the following on this subject: *"..the grease is by no means all buried. In the folding process there are roughly equivalent decreases in the accessibility of both the polar and non-polar groups"* [16]. However, Rose et al. in 1985 largely repeated the work of Lee and Richards but with a larger data set. They examined the burial of the hydrophobic and hydrophilic surface areas of amino acids in folded proteins; using the same reference states as [14] to measure rASA. They then assessed the mean surface area of each residue which is buried during the folding process. They reported the following in response: *".. we now report findings that*

lead to the opposite conclusion, revealing a strong correlation between hydrophobicity and the surface area residues bury upon folding" [17].

As a result of Kauzmann's work and subsequent investigations, it is now understood that in the folding of the polypeptide chain there will be a loss of protein conformational entropy which must be compensated for, if a protein structure is to be stable. This stability is primarily provided by the hydrophobic effect; although there is an ongoing discussion in the literature regarding the contributions to the stability of a protein fold that arise from the energy of hydrogen bonding and electrostatic interactions. Amino acids in the folded protein interact with each other individually and in groups; whether locally or across some spatial distance. These interactions are considered to be responsible for the specific characteristics of a given protein's structure and its ability to perform its biological function. Specificity of these interactions is delivered by electrostatic interactions, via for example the specific pattern of hydrogen bond donors and acceptors in a protein sequence, as well as the steric constraints imposed by side-chain shape and size.

1.1.3 Amino acid interactions and sequence evolution

Alterations in the amino acid sequence, arising from random mutations in the protein coding DNA, will alter the specific amino acid interactions around the sites of change. As long as the alteration is not detrimental to the function of the protein, and thus the fitness of the organism, then the organism and thus the gene associated with the altered protein will persist. Over time multiple changes in the sequence will occur and this can lead to variations in the amino acid sequence of the same protein found in different species and indeed in individuals of the same species, albeit to a lesser extent. Furthermore gene duplication within a species allows one copy of a duplicated gene to rapidly accumulate changes in its coding region; mutations that are detrimental to protein function will likely not be detrimental to organism survival as long as there is one fully functioning copy of the protein maintained. Such accumulation of substitutions can lead to proteins with novel function. Thus comparison of proteins with a recent common evolutionary ancestor will indicate positions where the amino acid sequence are

different, indicating that amino acid substitutions must have taken place in the history of one or more modern sequences compared with their common ancestral sequence.

Sequence identity [1] is used to classify proteins into families, and deduce common ancestry. Homology is defined as the presence of similar properties or characteristics between two or more species that are a result of common ancestry. There are two types of evolutionary relatedness that apply to homology, orthology and parology. Sequence orthology refers to sequences which are related through a speciation event. While parology refers to sequences related through a gene duplication event. Though it is feasible to differentiate between orthologs and paralogs, it is not necessary in the context of this thesis and the term homology will be used to refer to the evolutionary relatedness of protein sequences.

The definition used by the SCOP database, is that protein sequences with 30% identity with respect to a reference sequence are classified as belonging to that family, with exception made for sequences which score less but are known to have structural and functional similarities [18, 19]. In their 1996 paper on the differences between protein structures as a function of sequence identity, Chothia and Lesk reported that sequences which had a sequence identity of 40% or more would have similar structures and functions [20]. The discrepancy between the two different values of sequence identity has to do with the distinction between structural similarity and functional similarity of proteins. There is a general concept of a "twilight region" between 30%–40% sequence identity where a cut off exists for protein relatedness, which falls between these two reported values.

The variations that can occur between sequences of the same family, in the form of residue substitutions at specific locations in the sequence, will be constrained by the pressures arising from a variety of quarters. An important step in elucidating the way in which protein structures evolve is the identification and characterisation of those pressures and their origins [21]. A study of the substitution behaviour of amino acids in homologous proteins, using hidden Markov models, has shown that the solvation state and the secondary structure environment significantly affect the propensity for substitutions to occur [21]. The solvation state, refers to an amino acid

[1] Sequence identity is mathematically defined in Section 2.5 , equation 2.36.

residue's interaction with the solvent environment surrounding the protein. There are several ways in which this can be determined as discussed later in this chapter, in section 1.3.

The hidden Markov model based study of subsitution behaviour [21], did not consider pairwise substitutions within the protein sequence or structure. The localised replacement or substitution of an amino acid at a given sequence position will alter the physical interactions around the substitution site, this is illustrated in Figure 1.1, shown in the next section. As such a substitution of an amino acid at one position may allow one or more residues at other sites in the structure to undergo a substitution that would otherwise have been deleterious, which may now provide functional or structural benefit or may compensate for minor instabilities arising from the original substitutions. Coordinated changes in amino acid substitution patterns are clearly seen when comparing protein homologues [22] although the exact details of the mechanism of these coordinated changes is not completely clear. For example, if two sites form a Lys-Asp salt-bridge. If Lys is replaced by Asp, then the original Asp will need to be replaced by either Arg or Lys, to maintain the salt bridge. Asp-Asp would be a repulsive interaction and would most likely be disruptive at the very least locally, if not to the entire structure and function of the protein. This correlated substitution behaviour is most commonly referred to in the literature as correlated mutations. Though it will be referred to as co-substitution in this thesis, as this term more accurately describes the process.

The next two sections are reviews. Firstly a review of co-evolution/correlated- mutation/co-substitution analysis methods in the literature is given. This is followed by a review of methods for determining the solvation state of residues. As mentioned earlier, the context of substitutions in amino-acid sequences has an effect on the propensity for the substitution to occur. For this reason amino acid context is explicitly considered in the co-substitution analysis developed in this thesis and requires some introduction.

1.2 Co-Evolution analyses of proteins

The behaviour of amino acid residue co-substitutions between evolutionarily related proteins has been targeted for the purpose of predicting amino acid contact maps for proteins with unre-

solved structures. Further, patterns of amino acid co-substitutions in sequence data may provide great insight into the role of residue interactions in determining protein structure and defining its function. It has been shown by mathematical analysis, that Levinthal's paradox can be resolved when a few amino acid residue interactions are known [23]. As such, using amino-acid co-substitution behaviour as a means to determine which residues in a polypeptide chain are most likely to be in direct contact, is a very attractive intermediate step to resolving the protein structure prediction problem. Valencia and colleagues have done considerable work trying to predict both residue-residue contacts and protein-protein interactions [24–27]. They have had some success with the latter, however their results for the former have been disappointing. There have been other groups who have developed methods to predict residue-residue contacts, yet none have posted results with a success rate greater than around 20% - 35% [28]. In the last 12-18 months, a new method using information theory has been published reporting successful prediction of a membrane protein, using only multiple sequence alignment data [29].

Typically, multiple sequence alignments of homologous sequences have been used for the statistical analyses of co-evolution or correlated mutations within a family or groups of families of proteins. Determining co-evolutionary patterns in the multiple sequence alignments, involves assessing the correlation of substitutions in one column with the substitution pattern of another column in the alignment. This is the subject of correlated mutation analyses and co-evolution analyses in the literature. The aim is to extract underlying trends in the co-substitution patterns of homologous proteins. By analysing multiple families of proteins, generalised trends can be determined.

Figure 1.1 illustrates the common approach to co-evolution analysis found in the literature, described above. In the sequence alignment shown on the right hand side of the image, substitutions are shown in columns 3 and 6 which correspond to the points 3 and 6 of the structure on the left hand side of the image. The same is true for positions 5 and m in the sequence alignment and the structure. For both examples, although the residues are not adjacent in the sequence they are spatially adjacent in the folded structure. This form of co-evolution analysis has been used to predict residue-residue contacts; the success of the applications of the method

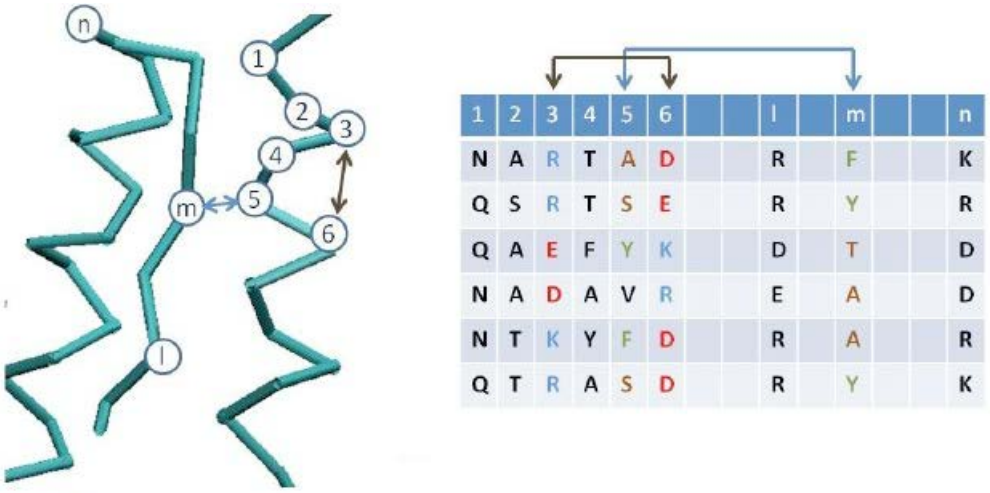

Figure 1.1: Correlated mutations of contacting amino acids: A structural exemplar for the sequences in the alignment on the right, is shown on the left. Interactions between physically proximate residues, which can be distant in the protein sequence, can be determined by correlated mutation analysis. e.g positions 5 and *m* in the structure are close in physical space but distant in the sequence. The different residue types in both columns are suitable substitutions for each other at their respective physical locations, and satisfy any imposed constraints. Such pair-wise substitution behaviour between columns, is what is used to infer contact maps of proteins from sequence data alone.

have then been tested against actual structural data [30, 31]. However, others e.g. Lockless et al [32], have shown that signal clearly arises for reasons other than simple contact.

This type of analysis is predicated on the assumption that positions in the amino acid sequence or regions within the protein with a role in thermodynamic stability, or importance in kinetic stability (e.g. for creating the right breathing motion of the protein), must maintain their specificity otherwise they will lose the ability to perform that role. This is the case for the interface regions of protein surfaces which are involved in protein-protein interactions. This is equally true for residue-residue interactions in a protein structure, for example, where the substitution of a hydrophobic residue with a hydrophilic could alter the folding pathway and result in a different folded conformation. As such an amino acid substitution at a site with imposed constraints, e.g requirements to be hydrophobic or to have a certain size, will need to satisfy those imposed constraints. This also suggests that a greater degree of conservation is likely at sites with strict constraints imposed on them, e.g. a catalytic residue, compared to regions of the protein which are less structurally or functionally important [28].

Proposed co-evolutionary analyses presented in the literature can be catagorised into two

distinct types. Firstly there have been endeavours to use only multiple sequence alignments to predict contacts maps of proteins [30, 31]. These methods have relied exclusively on multiple sequence alignment data from such database sources as HSSP [33] or Pfam [34]. Secondly, efforts have been made to define contact maps by using structural data to parametrise sequence alignments data. The combination of the structural and sequence alignment data has then been used to predict contact maps of proteins. These methods are perhaps better described as co-substitution analyses than co-evolution analyses. The latter has been shown to provide improved accuracy in the prediction of contact maps [35, 36].

A considerable number of different statistical methods have been proposed, for the purpose of investigating co-evolutionary events in multiple sequence alignments. The remainder of this section is divided into two parts. Firstly four co-evolution analysis methods which consider only multiple sequence alignment data are reviewed here: Pearson correlation coefficient [31], statistical coupling analysis (SCA) [32, 37], observed minus expected squared (OMES) [38] and mutual information (MI) [28]. Other methods exist which include perturbation explicit likelihood of subset co-variation (ELSC) [39], two-state maximum likelihood [40], ancestral sequences correlation coefficient [41], however these are not covered here in the interest of brevity. This section then concludes with an overview of the co-evolution analyses developed by [35] and [36], which combines multiple sequence alignment data with structural data from known protein structures.

1.2.1 Co-evolution methods based exclusively on analysing MSA data

Pearson's Correlation Coefficient

The first method applied to the problem of correlated mutations in protein families, was an adaptation of the Pearson Correlation Coefficient, developed by Göbel et al in 1994 [31]. The aim of that work was to predict residue-residue contacts within protein structures. Their results appeared very promising, with a reported accuracy of prediction over 60%. However as the number of known protein families has increased, the applications of this method have reported greatly reduced accuracies at around 20% [28].

In mathematics, the Pearsons correlation coefficient is a measure of the linear dependence (correlation) between two variables X and Y. The coefficient is a real number between -1 and +1. It is defined as the covariance of X and Y, divided by the product of their standard deviations [42].

Here is a generalized description of how the Pearson correlation coefficient has been applied to correlated mutation analysis of multiple sequence alignments.

1. Select or assemble a 20×20 similarity matrix which scores the similarity between each of the 20 amino-acid types. The similarity score between residue types can be derived from either statistical (e.g use of substitution matrices) or physical (e.g. volume and/or hydrophobicity changes) considerations and indicate the degree of change for a mutation from one residue type to the other. The choice of scoring system and the considerations taken into account in the building of the matrix are key to the success of this method. A key assumption here is that residues which are dissimilar will have substitution scores which reflect this. A further fundamental assumption to this method is that a mutation at position i involving two dissimilar residues will correspond to a mutation at position j which will be similarly scored [22, 25, 28].

2. Generate a multiple sequence alignment with N sequences, where each sequence is of length M.

3. For each column i in the sequence alignment build an $N \times N$ matrix, by comparing every amino acid in the column with every other residue in the column. As the matrix will be symmetric only half is needed (minus the diagonal). Every position in the matrix is represented by (u, v).

4. Remove those columns which are perfectly conserved and thus have standard deviations of zero, from the analysis. This is to avoid problems around dividing by zero.

5. Fill each matrix, such that each position (u, v) is the score from the similarity matrix above, for the given residue pair represented by residue u and residue v in the matrix.

6. Use equation 1.1 to calculate the correlation co-efficient.

$$CM_{ij} = r_{ij} = \frac{1}{N^2} \sum_{kl} \frac{W_{kl}(s_{ikl} - \langle s_i \rangle)(s_{jkl} - \langle s_j \rangle)}{\sigma_i \sigma_j} \tag{1.1}$$

Where:

- S_{ikl} and S_{jkl} are the similarity score for residues in column i and j between sequence k and l, respectively.
- $\langle s_i \rangle$ and $\langle s_j \rangle$ are the average similarity score in the $N \times N$ matrix at positions i and j respectively.
- W_{kl} is the fraction of non-identical positions from the multiple sequence alignment in sequences k and l, normalized to sum to 1. This is to down-weight very similar sequences.
- σ_i and σ_j are the standard deviation of the $N \times N$ similarity scores at positions i and j respectively.

There have been variations on this method however in the interest of keeping this brief, these will not be discussed here, refer to Halperin et al [22], Aldrich et al, [28], Pazos and Valencia [43] for further information.

Statistical Coupling Analysis (SCA)

Developed originally by Ranganathan et al. [32, 37] this method is based on determining a pseudo $\Delta\Delta G$ term for a multiple sequence alignment and a sub-alignment. The kT term which was included in the original publication of this method, was dropped in subsequent publications as it was found that it made no difference to the correlation calculation; it was only included to make the correlation coefficient seem like an energetic term. The method is as follows:

1. Build a multiple sequence alignment (MSA) of homologous sequences, referred to below as the parent alignment.
2. Create a sub-alignment derived from the parent MSA in two steps:

i. Choose a column j and determine the most common amino-acid residue-ρ in the column.

ii. Create an alignment with all the sequences in the parent alignment that have residue-ρ present at position j.

3. Determine the pseudo ΔG term for both the parent alignment and the sub-alignment, with equation 1.2, where column i represents any given column in the sub-alignment:

$$\Delta G_i = kT \sqrt{\sum_x \left(ln \frac{P_i^x}{P_{MSA}^x} \right)^2} \tag{1.2}$$

Where:

- P_i^x is the probability of finding residue type x in column i and represents the proportion of residues in column i which are of type x.
- P_{MSA}^x is the probability of finding residue type x in the parent MSA and represents the proportion of all residues in the MSA which are of type x.
- $\frac{P_i^x}{P_{MSA}^x}$ reflects the difference between the two proportions.

4. Use equation 1.3 to calculate the $\Delta\Delta G$ value, which is the co-evolution score. This is the difference between the ΔG term for the parent alignment and the sub-alignment.

$$\Delta\Delta G_{ij} = kT \sqrt{\sum_x \left(ln \frac{P_{i|\delta j}^x}{P_{MSA}^x} - ln \frac{P_i^x}{P_{MSA}^x} \right)^2} \tag{1.3}$$

Where:

- $P_{i|\delta j}^x$ is the probability of finding residue x in column i of the sub-alignment built with residue-ρ fixed in column j.

The kT ([28]) and P_{MSA} ([39]) were found to be unnecessary as reviewed by [22], resulting in equation 1.3 being reformulated as equation 1.4.

$$\Delta\Delta G_{ij} = \sqrt{\sum_{x} \left(ln(P^{x}_{i|\delta j}) - (lnP^{x}_{i}) \right)^2} \tag{1.4}$$

Observed Minus Expected Squared (OMES)

Proposed initially by Kass and Horovitz [38], this statistical method derives from the χ^2 non-parametric test for statistical significance. It is centred on the comparison of an observed distribution of residue pairs in two columns i and j from a multiple sequence alignment, with an expected distribution.

The algorithm follows these steps:

1. Make a list of all distinct residue pairs from column i and j in the sequence alignment; all residues in either column paired with a gap in the other column are excluded.

2. Calculate the expected distribution for each residue with equation 1.5. N_{exp} is an estimation of the number of sequences where residue type ρ is found at position i and residue type τ is at position j, given the frequency with which they both occur individually in their respective columns, and assuming that the distribution of residues in column i is independent of the distribution of residues in column j and vice versa. This is the null hypothesis.

$$N_{exp} = \frac{N_{xi} N_{yj}}{N_{valid}} \tag{1.5}$$

   Where:

   - N_{xi} is the number of times residue x is found in column i.
   - N_{yj} is the number of times residue y is found in column j.
   - N_{valid} is the number of sequences in the alignment that do not have gaps at position i or j.

3. Calculate the correlated mutation score using equation 1.6. This has the property that for column pairs which are perfectly conserved, the correlated mutation score is zero

because $N_{exp} = N_{obs}$ whilst $N_{valid} = N_{exp}$; it does not cause the zero value that arises from $N_{exp} = N_{obs} \Rightarrow r_{ij} = 0$.

$$r_{ij} = \sum_{l=1}^{L} \frac{(N_{obs} - N_{exp})^2}{N_{valid}} \tag{1.6}$$

Where:

- L is the number of distinct residue pairs in column i and j, excluding residues in either column paired with gaps in the other position.
- N_{obs} is the number of times that the distinct pair of residues being considered is observed in the columns i and j.

The considerations of this method are as follows: The statistical significance of the $r_{i,j}$ value (which is a χ^2 statistic), is dependant on the number of unique pairs of residue types in the columns being considered. Evaluation of this value requires that substitution events are observed in both columns being considered, as a conserved column will result in $r_{i,j} = 0$. Finally, there is no need to measure amino acid similarity.

Mutual Information

Mutual information is a measure of the information that two variables contain about each other. It is a measure of how much the uncertainty about one is reduced by knowing the other. Consider two extreme conditions, firstly two independent variables x and y and secondly two variables u and v are always identical. In the case of the two independent variables knowing x tells us nothing about y and therefore the mutual information is zero. In the case where u and v are identical knowing one tells us exactly what the other is.

Mathematically mutual information is defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \frac{p(x,y)}{p_i(x) p_j(y)} \tag{1.7}$$

Where:

- $p(x, y)$ is the joint probability distribution of x and y
- $p_i(x)$ is the marginal distribution of x
- $p_j(y)$ is the marginal distribution of y
- $p_i(x) \times p_j(y)$ is the product distribution of $x \times y$

In effect mutual information is a measure of the distance between the joint distribution $p(x, y)$ and the product distribution $p(x) \times p(y)$ [22]. When applied to the problem of determining correlated mutations in multiple sequence alignments, mutual information can determine how much information one column has about another. Consider two columns in a sequence alignment i and j, mutual information measures how much our knowledge of j is increased if we know i or the other way around. In this set up, there are n and m residue types in i and j respectively. $p_i(x)$ is the probability of finding a residue of type x in column i, while $p_j(y)$ is the probability of finding a residue of type y in column j. Similarly $p_{i,j}(x, y)$ represents the joint probability of finding residues x and y in columns i and j, i.e. the probability that they appear in the same sequence, in their respecitve columns of the alignment. The probability distributions are calculated from the amino acid distribution in each column [22, 28].

This has been applied to co-evolution analysis in multiple sequence alignments by a number of different groups. Early work on this was published by Clarke in his 1995 paper titled 'Co-variation of residues in homeodomain sequence family' [44]. Several years later, Atchley et al, published a couple of papers based on this method [45, 46]. It continues to be popular and in 2007 was included in a procedure to prepare data to train neural networks, by Shackelford and Karplus [47].

EVFold [29, 48, 49] is an adaptation of the mutual information based co-evolution analysis developed by Marks & Sander. By applying a further statistical analysis, called direct coupling analysis (DCA), they try to separate the mutual information scores resulting from direct residue-residue interactions from those arising from transitive interactions and the statistical noise inherent in the set of observed correlations.

Residue pairs found to have high mutual information scores could be in direct contact, yet residue-residue contact maps produced using them often deviate considerably from contact maps produced from actual structure data. This could be interpreted as an indication that non-contacting residues with high mutual information are interacting over some physical distance. Marks & Sander interpreted this as the result of transitive effects; where two residues A and B don't interact directly with each other, but both interact with some residue C. Thus changes in A can affect C which in turn can affect B, resulting in a high mutual information score between A and B that could be misinterpreted as a contact. As a result, mutual information scores can be said to contain both direct and indirect correlation effects. To address this Marks et al. [29] applied DCA to maximise the number of directly interacting residue-pairs and minimise residue-pairs coupled through transitive effects. The product of the DCA is a set of scores for all the observed coupled residue pairs, which is used to build a ranked set of residue-residue interactions, referred to as "evolutionary inferred contacts" (EICs); these are used as constraints in structure prediction.

The authors of the method report having successfully predicted the structures of 11 transmembrane proteins for which no known structure existed [48]. This shows a significant improvement in the performance of co-evolution analysis to predict protein structures from multiple sequence alignment data alone. However, their method has not been entered into CASP [50] as yet and its general performance has yet to be evaluated.

1.2.2 Co-evolution methods parametrised using structure data together with MSA data

The methods parametrised using structural data in combination with multiple sequence alignment data, used by Thoams et al. and Eyal et al. [35, 36] are similar in approach to the method developed for this thesis. However, in this thesis the emphasis is not on the determination or prediction of inter-residue contacts within the protein structure, but the relationship between inter-residue interactions and physical distance. The discussion which follows will only cover the method of Eyal et al. [36], which is very similar to the method used by Thomas et al. [35]

and follows these steps:

1. Select a source of multiple sequence alignments for which known structural representations are available. e.g.Thomas et al. [35] used the HSSP [33] with 'PDB_Select' algorithm [51]. This can also be achieved, as was done for this thesis, by locating structural exemplars for Pfam domains from the PDB. Each pair of multiple sequence alignment and representative structure are analysed independently.

2. Create a contact map from the structure data and map the residue positions from the structure to columns in the multiple sequence alignment. This divides pairs of columns of the alignment into two separate sets: the set of column pairs aligned to contacting residues in the reference structure and the set of column-pairs aligned to non-contacting residues.

3. Solve for the propensity or probability that each possible co-substitution type could occur. The co-substitution type is defined by the residue types at x and y in sequence-k being present at positions i and j respectively, while in an aligned sequence-l residue types u and v are found at the same positions i and j respectively (this is more clearly demonstrated in Chapter 2. The co-substitution event is written as $(x \leftrightarrow u, y \leftrightarrow v)$. Eyal et al. argue that there are $((20\times20)\times(20\times20))$ = 160,000 possible mutations [36]; because there are (20×20) possible pairs of xy and (20×20) uv pairs. This is only correct if one does not account for the symmetry of the matrix. Thomas et al. argue that there are 40,300 [35], to account for symmatry. However a simple combinatorial treatment of two identical alphabets with 20 characters each will reveal that there are only 22,155 unique combinations of residue pairs which includes conservations, considerably fewer than either group published results for. This is becuase it is not straightforward to determine the direction of the mutation, and the symmtery of mutations means many mutations are equivalent.

Eyal et al. [36] refer to the $((20 \times 20) \times (20 \times 20))$ matrix as P2PMAT and calculate a value $M[xy][uv]$ for every element in the matrix, using equation 1.8.

$$M[xy][uv] = ln\frac{f_{obs}^{con}[xy][uv]}{f_{exp}^{con}[xy][uv]} - ln\frac{f_{obs}^{noncon}[xy][uv]}{f_{exp}^{noncon}[xy][uv]} \tag{1.8}$$

Where:

- $f_{obs}^{con}[xy][uv]$ is the observed frequency of contacting pairs in columns of the alignment, see equation 1.9.
- $f_{exp}^{con}[xy][uv]$ is the expected frequency of contacting pairs in the columns of the alignment see equation 1.10.
- $f_{obs}^{noncon}[xy][uv]$ is the observed frequency of non-contacting pairs in columns of the alignment.
- $f_{exp}^{noncon}[xy][uv]$ is the expected frequency of non-contacting pairs in the columns of the alignment.

$$f_{obs}^{con}[xy][uv] = \frac{n_{obs}^{con}[xy][uv]}{\sum_{abcd} n_{obs}^{con}[ab][cd]} \tag{1.9}$$

where:

- $n_{obs}^{con}[xy][uv]$ is a weighted sum for all $x \leftrightarrow u, y \leftrightarrow v$ in columns i and j.

$$f_{exp}^{con}[xy][uv] = \frac{n_{obs}^{con}[x][u]}{\sum_{ab} n_{obs}^{con}[ab]} \cdot \frac{n_{obs}^{con}[y][v]}{\sum_{ab} n_{obs}^{con}[ab]} \tag{1.10}$$

where:

- $n_{obs}^{con}[x][u]$ is a weighted sum for all substitutions $x \leftrightarrow u$ in a column of the alignment.
- $\sum_{ab} n_{obs}^{con}[ab]$ is the sum over all observed substitutions in a column of the alignment.

The application of structural data combined with multiple sequence alignment data, in this fashion, has reportedly improved the accuracy of predicting residue-residue contacts in proteins, by 25-60% [36]. However, by over-estimating the number of possible co-substitution types, they may have under-represented the propensity of each co-substitution type.

1.2.3 Summary

The co-evolution methods in the literature collectively endeavour to predict residue-residue contacts, either using multiple sequence alignment data exclusively, or using sequence data combined with parameters derived from structural data. Though there is evidence that non-contacting residues show correlated mutation behaviour, none of the reported methods have considered investigating this further. The recently published work of Marks et al. [29, 48, 49], has been used to successfully predicted 3D structures for transmembrane proteins, which is an exciting development but needs further generalised testing.

1.3 Solvent Exposure

Solvent exposure, as a measure of solvent accessible surface area (ASA), has been used to investigate the burial of hydrophobic surface area in proteins [14]. ASA has a normalised variant called relative solvent accessible area (rASA). There have been several methods proposed and developed to measure the ASA of proteins since the original method developed by Lee and Richards [14]. ASA and rASA [14, 15] offer a view of the protein surface from the perspective of the solvent or the external environment. Since these methods are applied to static structures they don't give the perspective of water penetrating the protein surface during dynamic motion. In order to distinguish between residues which sit just below the surface and those which are buried in the core of the protein structure, solvent accessibility is not a suitable measure.

Several alternative methods of measuring solvent exposure, which do not consider the surface area of the protein but rather the relative position of individual residues within the protein structure, have been developed, e.g. residue depth(RD) [52], co-ordinate number (CN) and half sphere exposure (HSE) [53]. These methods provide a perspective of the environment from that of amino acids within the structure. The hydrophobic effect causes hydrophilic residues to be preferentially located on the surface and the hydrophobic residues to be buried away from the solvent. With these methods the distribution of residues of the protein interior, which is still defined by the hydrophobic effect, can be investigated.

The rest of Section 1.3 is divided into two parts. The first part a discussion of the methods

used to measure the surface area of proteins, is presented. This is followed by a review of three non-surface area methods of studying solvent exposure, Residue Depth, Co-ordinate number and finally HSE.

1.3.1 Measuring the surface area of a protein

A protein's surface is formed from the side-chains of its amino acids, which during the folding process have been positioned at least partially exposed to the extra-molecular environment. As mentioned earlier in Section 1.1.2, two methods were originally proposed for measuring solvent exposure of amino acids in a protein structure. Both methods measure the same quantity, though they are different in how they achieve the measurement. The first method was proposed by Lee and Richards [14] which was shortly followed by Shrake & Rupley [15]. Lee and Richards termed the quantity they were measuring as the *solvent accessible surface area* of the protein and gave the following definition: *"the area mapped out by the centre of a sphere rolling along the surface of a protein, the sphere represents a solvent molecule - usually having a radius of 1.4 Å, that of water, however different radii can be chosen"* [14]. Their method is sometimes referred to as the "rolling ball" method. The method is still in use today in the *Naccess* computer program [4].

The second method proposed by Shrake and Rupley, uses a different method of calculating the accessible surface area. In the same way as Lee and Richards, they defined a probe radius which was added to the van der Waal's radii of the atoms within the molecule. They then placed 92 points on the resulting sphere and determined which points where accessible to a solvent molecule - and not inside an expanded sphere [15]. Later Connolly [54] developed this further and created a computer software package (called MS) which went through a number of incarnations and is available today as a part of the Chimera package from UCSF. This method of determining the surface area has also been incorporated in the computer program *DSSP* [55].

The question of how to define the surface of a protein is not necessarily straightforward to answer. Figure 1.2 illustrates the rolling ball method of determining solvent accessible surface area, and includes illustrations for several different interpretations of the protein surface. In the

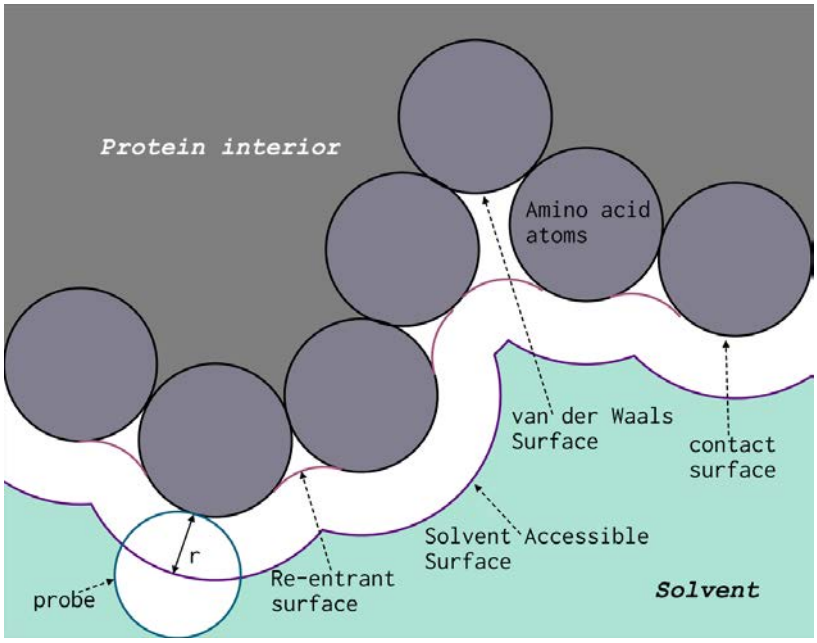

Figure 1.2: The rolling ball method. The probe with radius r is rolled over the external surface of the amino acids' atoms. The Solvent Accessible Area is the area traced out by the centre of the probe. Typically a radius of 1.4 Åis chosen for the probe radius, representing the radius of a single water molecule. The molecular surface is the surface area traced out by the edge of the probe closest to the van der Waals surface. The van der Waals surface of the protein is made up of the non-overlapped van der Waals surfaces of the surface atoms.

strictest sense it could be argued that the surface of a protein is defined by the non-overlapped regions of van der Waals surfaces of all atoms on the exterior of the protein, that are exposed to the environment and not the surface of another atom. However, it could also be argued that the surface of the protein consists only of the regions of the van der Waals surface which can come into contact with a solvent molecule, which is referred to as the contact-surface.

Through the application of the rolling ball method, there are other surfaces which can be defined. For example the contact surface is the traced out trajectory of a probes surface coming into contact with the van der Waals surface; or re-entrant surface, which is the trajectory of the probes surface over regions of the surface, such as crevices which may be too narrow for a solvent molecule to penetrate [16, 56], as shown in Figure 1.2. If we combine the re-entrant surface and the contact surface, we get what is referred to as the "molecular surface" [16, 57]. Alternatively, an extended van der Waals surface can be defined, by offsetting the van der Waals surface of the amino acids by the radius of a solvent molecule, which is equivalent to tracing the trajectory of the centre of a probe rolling over the molecular surface of the protein (the volume

enclosed by this surface is referred to as the solvent excluded volume [54, 58]). This is the solvent accessible surface proposed by Lee and Richards, the area which is determined by the following steps:

1. Assign a van der Waals radius to each atom or group of atoms. Note: Hydrogen atoms are not considered separately but are included in a group radius, e.g. a carbon atom has a radius of *x*-angstroms; but when considering CH_3 as a group a radius of *y*-angstroms is considered (where y > x). This is similarly applied to SH, NH, OH, CH_2 and CH.

2. As the structure is now represented by a set of inter-locking spheres, the continuous structure is sectioned by a set of parallel planes with a predetermined spacing. The resulting cross-sections of the structure shows the inter-locking spheres as circles. The overlapping arcs of which are not eliminated. This is because it helps to distinguish one atom from another. It also helps to easily recognise excessive overlap of symmetry related neighbouring molecules. Their method was proposed before the advent of sophisticated computer graphics systems and so relied on the use of outlines on sheets of plastic.

3. The polar atoms (oxygen and nitrogen) are dotted and labelled. The non-polar atoms (carbon and sulphur) are given solid lines.

4. The sequence number is written at the centre of both the α and β carbons.

5. The skeleton covalent bonds and hydrogen bonds are shown between atom centres to assist in viewing.

Lee and Richards applied their method to investigating the burial of hydrophobic surface area of residues in proteins. To calculate the relevant solvent accessibility, they created two model systems. These were both used to estimate the hydrophilic and the hydrophobic surface area of each amino acid. The models constructed were tri-peptides of Ala-X-Ala and Gly-X-Gly respectively, where X is the residue whose accessibility is being computed. Here a measure of the area accessible to solvent of residue X, in the tri-peptide, is taken as a measure of the residue's accessibility to solvent in an unfolded state. They then measured the accessible

surface area of each residue in a folded protein. This was compared to the unfolded state and the percentage of the surface area that became buried as a result of folding was then known. They called this the "*relative solvent accessibility*" of a residue, it is a percentage area of a residue relative to either of the tri-peptide models discussed earlier. This is distinct from the solvent accessible area which is measured in $Å^2$.

1.3.2 Non-Surface area measures of solvation

Residue Depth

Residue depth (RD) is a different measure of solvent exposure. Here atom depth is defined as the distance between a given atom and the nearest surface water molecule. Residue depth is thus the average atom depth for a given residue [59].

The most accurate method (though also the most computationally intensive) of calculating residue depth was proposed by Chakravarty and Varadarjan [52]. To calculate RD, first the molecular surface is calculated, to have a surface to work from and to know where the crevices and cavities in the structure are. Then the position of the nearest water molecule must be found. This is done by estimating the likely position using a Monte Carlo simulation. Which encapsulates the protein in a box-space and the space is filled with an appropriate density of water molecules; the average distance between each water molecule is 2.8 Å. The protein is then rotated around its centre of mass. Water molecules which fall into two categories are not considered; a) those within 2.6 Å of the atom being considered, and b) those that have less than 2 Neighbouring water molecules in a 4.2 Å radius. Thus water molecules in crevices and cavities are disregarded in these calculations.

The residue depth calculation can be time consuming, which could make it prohibitive to use with large datasets. The method relies on calculating the molecular surface [16] using Connolly's MS program [54]. Having completed the Monte Carlo simulation, it must calculate the average location of each atom in a residue to provide the desired result, the residue depth.

Using the relative accessible surface area, calculated using the Gly-X-Gly reference state, and residue-depth for the amino acids in a single protein having a sequence length of 370

Chakravarty et al. showed a strong correlation between rASA and residue depth. They found that as accessibility decreases, residue depth increases. At approximately 4 Å depth the accessibility decreases sharply and below 6 Å accessibility is 0%. The reported result only considered a single protein and so it cannot be considered statistically significant or generalised.

Residue depth is an involved method of determining solvent exposure, which does not appear to offer any obvious advantage over other methods. There are two online resources available for measuring residue depth, ProDepth [60] and DEPTH [61].

1.3.3 Coordination Number

The coordination number comes from Chemistry and has been adapted to this field. In the context of proteins, the Coordination Number - CN - is defined as the number of C_α atoms in a sphere of chosen radius, centred on the C_α atom of the amino acid residue being considered. To calculate the CN of an amino acid residue, create a sphere of a radius usually between 12 Å and 14 Å on the C_α atom. Then count the number of C_α atoms that are within the sphere.

This is an easy to implement method and is fast to compute. It benefits from not requiring a full atom model of a protein to work. i.e. it only needs the position of C_α atoms. However it provides poor representation of solvent exposure as compared to ASA and relative-solvent-accessibility or even RD [53].

1.3.4 Half Sphere Exposure

Given the shortcomings of the previously discussed measures of solvent exposure, in 2004 Hamelryck proposed a different method, derived from CN. He set about to address two questions: *"...how to construct a measure that combines the best features of the above mentioned solvent-exposure measures and what view of solvent exposure does such a superior measure offer?"* [53]. His solution is called Half Sphere Exposure - HSE.

The calculation is similar to that of CN. A sphere is centred on the C_α atom of the residue being considered, in a protein, see Figure 1.3. The choice of radius is up to the user, however, like CN values are usually chosen in the 12 Åto 14 Årange, where 13 Åis the most common. The next step, is to draw a plane through the C_α atom that is perpendicular to the vector between

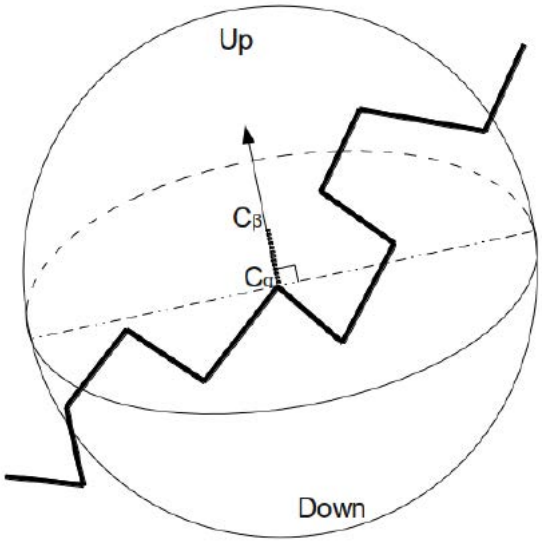

Figure 1.3: HSE: The measure of HSE comes in two forms, HSEu and HSEd. HSEu is the count of C_α atoms in the half sphere in the direction of the side chain. HSEd is the the count of C_α atoms in the other direction.

the C_α and C_β atoms. The plane cuts the sphere into two halves and the number of protein C_α atoms in each half is counted. The value for the half in the direction of the $C_\alpha \rightarrow C_\beta$ is called the HSEu (u for up) and the value for the other half is the HSEd (d for down). HSEu is a measure of the residue's solvent exposure in the direction of the side chain. While HSEd is a measure of the residue's solvent exposure in the other direction, this representing the solvent exposure of the atoms in the main chain that are not shielded by the side-chain. This method of determining HSE is called HSE_β, (because the C_β is used).

There is a slight variation on the method, which is used for cases where no C_β exists in the model. The approach is the same, except that a vector for the $C_\alpha \rightarrow C_\beta$ needs to be generated. This is done, by taking the vector $C_{\alpha-1} \rightarrow C_\alpha$ and the $C_{\alpha+1} \rightarrow C_\alpha$ and extending them. The angle between them is bisected and a vector is drawn at that angle through the C_α atom. The plane used to divide the sphere in half is drawn perpendicular to this vector. This form is called HSE_α,(only the C_α is needed).

With respect to the considerations of sphere radius Hamelryck had this to say: *"The choice of the sphere radius is a compromise between two demands. A radius that is too small misses residue pairs that are obviously shielding each other from the solvent. A radius that is too large includes irrelevant residue pairs. Based on visual inspection of protein structures, 13 Å is a*

good compromise.." [53]

1.3.5 Summary

Of the numerous studies in the literature of solvent exposure, none have tried to address the straightforward question "when is an amino acid residue buried?" The literature is full of papers which define some arbitrary boundary between the surface and the protein interior. This is usually some value of relative solvent accessible surface area and is never justified. Given the relationship between the hydrophobic effect and the distribution of amino acid residue types (hydrophobic/hydrophilic), it is known that hydrophilic amino acids are more likely to be solvent exposed and hydrophobic ones are less likely to be. This suggests that there could be a boundary between two distinct populations, the hydrophobic population and the hydrophilic one. A statistical analysis of the preferred solvent exposure of each residue type, could reveal a crossover point between these two populations.

The implications of the size dependence of the hydrophobic effect, described in Section 1.1.2 , is not well studied with respect to the propensity for amino acid solvent exposure, nor are its implications clear for protein structures. It is possible that a statistical analysis of the preference of amino acid solvent exposure could provide an insight into this.

1.4 Scope of this thesis

The investigations found in the literature on co-evolution and correlated mutations in protein structures have focused on predicting residues in direct contact, because it was assumed that the driver for co-evolution was structural pressure on a local environment. As discussed in Section 1.2, correlated mutation analysis has shown that two columns with changing amino-acid in a sequence alignment, are varying in correlation with each other. However, the relationship between inter-residue distance and correlated mutations has not been explicitly studied because long-range interactions were not considered important factors. We argue that a limitation of correlated mutation analysis is that it does not consider the importance of the residues or residue types involved in the mutations and the lack of physical distance considerations is an oversight.

By analysing the propensity for different pairs of residues to be jointly involved in co-

substitutions, it is possible to determine characteristics of the underlying physics governing protein structures. Consider two positions that appear to be mutating in a correlated fashion, if it is observed that the columns contain exclusively charged residues, and that the mutations appear to conserve the attractive or repulsive nature of the electrostatic potential. Then, if we consider a specific attractive interaction, e.g. a hydrogen bond, to be of primary importance, it could perhaps be replaced by a salt bridge. However, if the attraction is important but the specificity is not, a hydrophobic interaction may be also suitable. In other words, residues may need to be conserved to maintain the folding pathway but don't necessarily need to be close together or interact in the folded structure.

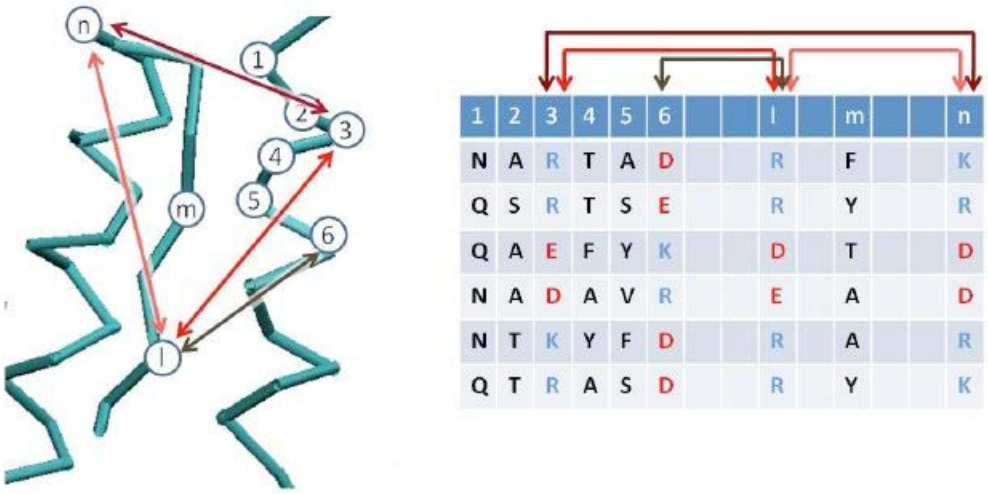

Figure 1.4: Co-substitutions due to long range interactions: The structural exemplar for the sequences in the alignment on the right, is shown on the left. Interactions across some physical distance, e.g. point 3 and *n*, can be determined by the co-substitution behaviour shown in columns 3 and *n* of the sequence alignment.

Here we present an analysis of the propensity for residues to "co-substitute" at different physical distances, to explicitly investigate the role of distance on the propensity to co-substitute. Unlike in a correlated mutation analysis where the emphasis is on determining the correlation of mutation between columns, we investigate individual co-substitutions. Consider Figure 1.4, which is similar to Figure 1.1 and illustrates long-range interactions between positions in the protein structure. In the sequence alignment on the right hand side, in columns 3 and *n*, a conservation of electrostatic repulsion across some distance is shown, while in columns 3 and *l* a conservation electrostatic attraction is shown. These are both examples of electrostatic

interaction being maintained. The objective of the work presented in this thesis has been to develop a method which can elucidate the distance preferences of different types of amino-acid co-substitutions.

Steps have been taken in the analysis, to separate co-substitutions between residues on the surface from those co-substitutions in the protein interior. These steps involved having firstly to define the boundary between the surface and the protein interior. This has been done to investigate if differences in the co-substitution behaviour exists, between the two solvation states. An investigation like this has not been reported in the literature previously.

To allow a rigorous classification of residues as being either surface or buried, a statistical analysis was conducted of the propensity for each of the proteinogenic amino acids to be solvent exposed. This has led to a number of interesting observations regarding the relationship between the solvent exposure measures ASA and HSEu, and some evidence of a correlation between solvent exposure preferences and substitution propensities.

The thesis has 5 chapters beyond the introduction. Chapter 2, introduces the main conceptual ideas of the analytical method. Providing definitions to terms and a derivation of the analytical methods developed and used for this work. Chapter 3 details a bioinformatics project to build a data base which merged the data from Pfam-A, SwissProt and the PDB/PiQSi databases. The database was used to ensure the sequence and structure data used were from the same cellular environment. Chapter 4 presents the analysis of solvent exposure preference for amino-acid types, seeking to determine a crossover point that can be used to define a set of surface residues and a set of buried residues. Chapter 5 presents the co-substitution analysis, indicating evidence for different co-substitution patterns on the surface compared to the buried residues. Chapter 6 closes the thesis with a summary of the main conclusions of this work.

Chapter 2

Development of analytical methods

2.1 Introduction of the statistical functions

The $\frac{O}{E}$ ratio is used in the methods of this thesis to determine the statistical preference for residue co-substitution with respect to distance, and the statistical preference for each of the twenty standard amino acids to be solvent exposed. In this chapter firstly an overview of the $\frac{O}{E}$ statistical method is given. Secondly a discussion of statistical phenomenon known as Simpson's Paradox is given to explain and justify the approach used to conduct the statistical analysis. Thirdly an explanation of how the $\frac{O}{E}$ statistical method can be applied to co-substitution and solvent exposure analyses respectively. Finally the weighting of protein sequences is explained with a discussion of Henikoff weighting and a new method of weighting pairs of sequences developed for this work, is presented.

2.2 Determining Bias, $\frac{O}{E}$

We wish to determine the natural bias of co-substitution events in the data with respect to distance, compared to what we might see by chance. *The observed* data O, the distribution of events with respect to distance, needs to be compared to the distribution of co-substitution events expected if there were no such bias; i.e what we would expect if the data followed a distribution unbiased by the consideration of distance, which will be referred to as *the expected* data E. A deviation of O from E indicates a bias in the co-substitution data due to distance effects.

In a similar fashion the range of preference for a residue to be solvent exposed can be determined by comparing the distribution of amino acid residues with respect to some measure of solvent exposure (*the observed*), with the unbiased distribution (*the expected*). Physico-chemical properties of amino acids mean that types of amino acids are not partitioned randomly between the surface and the interior, but rather are biased as discussed in Section 1.1.2. The existence of this bias has implications on the distribution for co-substitution events; since we expect substitutions in the core biased for hydrophobic residues and substitutions on the surface to be biased towards hydrophilic residues. To calculate an appropriate expected (E) value, for the co-substitutions analysis, we need to account for this bias in distributions.

2.2.1 The $\frac{O}{E}$ Ratio

Consider the generalisation of this concept: If one were trying to determine if a dependence exists between some constraint and an observable phenomenon in the data, one could determine the frequency of the phenomenon while subjected to a constraint (*the observed frequency, O*) and this can be compared with the frequency of the phenomenon when not subject to the constraint (*the expected frequence, E*). A deviation of O from E is an indication that the phenomenon is dependant on the constraint.

In the limit of large-numbers, i.e. given a sufficient representative sampling of the complete data space, O is a conditional probability and E is the independent probability distribution of the event. The ratio $\frac{O}{E}$ checks the null-hypothesis that the phenomenon is independent of the condition or constraint, by determining if they are equal or not. Both can be represented as probability distributions: $O = P(event|condition)$, while $E = P(event)$, which will be abbreviated to: $O = P(e|c)$ and $E = P(e)$, which gives:

$$\frac{O}{E} = \frac{P(e|c)}{P(e)} \tag{2.1}$$

From probability theory it is known that [62]:

$$P(e|c)P(c) = P(c|e)P(e) = P(e,c) \tag{2.2}$$

Which is an expression of Bayes Theorem.

Using these relationships with a little algebraic manipulation we arrive at:

$$\frac{P(e|c)}{P(e)} = \frac{P(c|e)}{P(c)} = \frac{P(e,c)}{P(e)P(c)} = \frac{O}{E} \tag{2.3}$$

Which shows there are three mathematically equivalent approaches to framing the problem to determine the $\frac{Observed}{Expected}$ ratio:

- Method 1:

$$\frac{O}{E} = \frac{P(e|c)}{P(e)} \tag{2.4}$$

- Method 2:

$$\frac{O}{E} = \frac{P(c|e)}{P(c)} \tag{2.5}$$

- Method 3:

$$\frac{O}{E} = \frac{P(e,c)}{P(e)P(c)} \tag{2.6}$$

Equations 2.4 and 2.5 represent the ratio between conditional probabilities and their unconditioned counterparts. The former is the ratio between the conditional probability of some event e, given a condition c, and the probability of the event e unconstrained by the condition c. While in the latter, the ratio between the conditional probability of some condition c given an event e, and the probability of the condition c. For these two cases the constraint is clearly visible, in the conditional probability of the Observed. In equation 2.6, the lack of a conditional probability in the observed makes it less obvious as to where the constraint is. For this case the observed is a joint probability distribution; probability theory tells us that if e and c are independent the joint probability distribution is equal to the product of the independent probability distributions, i.e. $P(e,c) = P(e) \times P(c)$. However, if they are not independent then the joint distribution will not be equal. Thus the constraint is implicit in the joint distribution of the Observed.

The $\frac{O}{E}$ ratio has the following characteristics:

- $\frac{O}{E} = 1 \Rightarrow O = E$: The observed is equivalent to the unbiased distribution. This suggests

that the constraint has no effect on the observed phenomenon. This represents support of the null-hypothesis, that the event and the constraint are independent of each other.

- $\frac{O}{E} > 1 \Rightarrow O > E$: A dependence exists between the constraint and the event. Such that the constraint makes the event favourable, i.e. O occurs more than we would expect were there no constraint.

- $\frac{O}{E} < 1 \Rightarrow O < E$: A dependence exists between the constraint and the event. Such that the constraints makes the event unfavourable, i.e. O occurs less than we would expect were there no constraint.

Further testing is required to see if the deviation of O from E is statistically significant. In this thesis, bootstrapping of the data to test whether $\frac{O}{E}$ values can be achieved by chance is used.

In this form $\frac{O}{E}$ is a lower bound function, with no upper bound. Since $0 \leq O \leq 1$ and $0 \leq E \leq 1$ this means that $\frac{O}{E} \in [0, \infty]$. This causes a problem when attempting to interpret the results, because values in the range $[0 - 1]$ show the dependent observed distribution of the phenomenon in the data to be lower than the unbiased distribution - indicating a preference for the phenomenon not to occur as frequently when the constraint is applied. Conversely results in the range $[1 - \infty]$ represents a preference for the phenomenon to occur more frequently when subject to the constraint. This suggests that if the ratio where equal to 1×10^{-3} or 1×10^{3} they would represent the same "strength" of propensity, negative or positive respectively. Taking the logarithm of the $\frac{O}{E}$ ratio, will make $\frac{O}{E}$ symmetric about 0. Taking $\log_2$ would make it simpler to interpret the results because an increase by a single log_2-unit is representative of a doubling of the effect on the ratio. In this case:

- $\log_2 \frac{O}{E} = 0 \Rightarrow O = E$

- $\log_2 \frac{O}{E} > 0 \Rightarrow O > E$

- $\log_2 \frac{O}{E} < 0 \Rightarrow O < E$

2.2.2 Using $\frac{O}{E}$ to make predictions

Equation 2.2, is an expression representing Bayes Theorem, which is used extensively in Bayesian methods to make predictions using observations and "prior" knowledge of the data. If we consider the equality of equations 2.4 and 2.5, shown in equation 2.3:

$$\frac{P(e|c)}{P(e)} = \frac{P(c|e)}{P(c)} \tag{2.7}$$

Now consider two manipulations of this equality where:

1. $P(c|e)$ can be determined by:

$$\frac{P(e|c) \times P(c)}{P(e)} = P(c|e) \tag{2.8}$$

$$\frac{O}{E} \times P(c) = P(c|e) \tag{2.9}$$

2. $P(e|c)$ can be determined by:

$$\frac{P(c|e) \times P(e)}{P(c)} = P(e|c) \tag{2.10}$$

$$\frac{O}{E} \times P(e) = P(e|c) \tag{2.11}$$

In equations 2.8 and 2.9, $P(c|e)$ is the probability of getting the condition given the event. The term $P(c)$ represents the intrinsic probability distribution of the condition, in the data. In Bayesian statistics this is referred to as "the prior", because it represents some prior knowledge of the data. Using the $\frac{O}{E}$ value and the prior, it is possible to make estimates of $P(c|e)$, or predictions for some condition, given an event. This similarly holds for equations 2.10 and 2.11, where the difference is that our prior is now $P(e)$ and it would be used for making predictions on $P(e|c)$ instead. The use of Bayesian statistics for the purpose of predictions is extensive, and

is a reasonable extension of this work. Though discussed in the context of co-substitutions later in this chapter, it was not applied in this thesis due to time considerations.

2.3 Simpson's Paradox

Simpson's Paradox is the change of the trends observed in a data set when an analysis considers the entire data compared with an analysis of segregated sub-groups or categories in the data. A very public example of this occurred in the early 1970s when the University of California, Berkeley, had to defend itself against a legal challenge to its admissions policy. They were accused of being biased against women who had applied to join their graduate school. The admission figures for the academic year starting in the latter part of 1973 are shown in Table 2.1. The data appears to show a very clear cut case against the University, until the data is broken down into more detail; in Table 2.2 [1] the applications and admissions data is shown broken down into individual departments.

Table 2.1: Simpson's paradox example, summary applicants to UC Berkeley: The total number of men and women who applied to the UC Berkeley graduate school for the fall of 1973.

	Men	Women
Number of applicants	8442	4321
% Of applicants admitted	44	35

[1]Both table 2.1 and table 2.2 are reproduced from the Wikipedia article on Simpson's Paradox on January 31 2013

Table 2.2: Simpson's paradox example, summary applicants to UC Berkeley by department: The number of men and women who applied to the UC Berkeley graduate school for the fall of 1973, divided into departments.

Dept.	Number of male applicants	Number of female applicants	% Male applicants admitted	% Female applicants admitted
A	825	108	62	82
B	560	25	63	68
C	325	593	37	34
D	417	375	33	35
E	191	393	28	24
F	272	341	6	7

The data shows that in most departments the percentage of female applicants being admitted to a course exceeds the percentage of male applicants. O'Connel et al. [63], concluded that women were applying to departments with generally lower admission rates than men were. Further they concluded that the bias in admissions was slightly in favour of women over men. This conclusion being the opposite of what can be concluded from the data in Table 2.1.

Considerations in the context of this work

The analyses presented in this thesis, were performed on a selection of different protein structures. Historically, analyses similar to this work would calculate their $\frac{O}{E}$ or other statistical measure based on the total data. In doing so, it is possible that their results did not reflect the true trend in their data as a result of Simpson's Paradox.

In calculating the $\frac{O}{E}$ ratio discussed in this chapter and applied in Chapters 4 and 5, the historic approach would be to perform an analysis using the data harvested from the entire dataset without considering the potential categories and sub-groupings possible within the data. In

the context of this work, it would be the equivalent of measuring either the solvent exposure or the distance between co-substitution events and calculating the $\frac{O}{E}$ using every value measured in all proteins. However this ignores the nuances in individual proteins, introduced by variations in shape, size and amino acid composition, which will affect the expected distributions for each protein. To ignore this will result in the loss of the context of information and force some potentially misleading global average on all proteins being considered. To help conceptualise this, consider a study of global freight traffic through sea ports. Assume that all countries are being considered, including landlocked ones, the final result will be misleading as no landlocked country will have freight passing through sea ports in their territory; therefore we clearly need to segregate countries into those with sea ports and those without.

Our solution to this issue, is to perform $\frac{O}{E}$ ratio calculations for each protein individually for the solvent exposure data for the solvent exposure analysis, and for every pair of sequences in which a co-substitution is observed for the co-substitution analysis. Furthermore, to deal with the issue of evolutionary relatedness, an average value is calculated for each residue-type or co-substitution type per Pfam family. The value for each Pfam family is an average over all sequence in that family with each sequence or sequence pair being weighted according to their identity to all other sequences in that Pfam alignment, as will be discussed later. Finally an average of averages is calculated using the number of contributing Pfam families for each residue-type or co-substitution type. The process of setting up the data into reasonable sub-categories is the subject of the next chapter, in which a discussion is presented on the method of data selection and the supporting arguments for how the data was segregated.

2.4 Application of O:E Ratio

2.4.1 $\frac{O}{E}$ Analysis of Co-Substitutions

Defining Co-Substitutions

In the literature a common term to discuss correlated substitutions of amino acids, is *correlated mutations*. The introduction presented a discussion of correlated mutations analysis. The work

presented here is a *co-substitution* analysis, and here we give the precise definition of a *co-substitution* event, in the context of this thesis. This serves to draw a strict distinction between the two types of analysis.

Consider two homologous sequences, which have been aligned with each other, in a sequence alignment with two-sequences. The columns of the alignment in which neither sequences has a gap, represent locations where it can be assumed that the structures of both sequences overlay each other in space. Often those positions have conserved residues, i.e. the residues in those positions in both sequences are the same. However it is also the case that variation exists between the two sequences, with some residue-type x at some given position i, in the first sequence being replaced by some residue-type u in the second sequence, where $x \neq u$. This we refer to as a *substitution* event: $(x \rightarrow u)$, which is illustrated in Table 2.3. It should be noted that $(x \rightarrow u) \equiv (u \rightarrow x)$.

Table 2.3: A Substitution: in column i of sequence k, residue x is present, while in sequence l residue u is present. Through the course of evolution, the residue at position i has been substituted from x to u or vice-versa as it is difficult to determine temporal events from a sequence alignment.

	Columns					
	1	2	...	...i...	...	N
Sequence k	T	R	...	x	...	L
				$\downarrow$		
Sequence l	E	R	...	u		R

The subject of investigation in this work is the co-substitution of amino acids between homologous sequences. Co-substitutions can be defined as two substitutions taking place simultaneously at two different positions in the sequence pair, as shown in Table 2.4.

Table 2.4: A Co-Substitution: in sequence k at positions i residue-type x is present and in sequence l residue-type y is present. Simultaneously at position j residue-type y is observed in sequence-k with residue-type v in sequence-l. The investigation is concerned with determining the statistical propensity of these events to occur at different euclidean distances within the protein structure.

	Columns					
	1	2	...i...	...	...j...	N
Sequence k	T	R	...x...	...	...y...	L
			↓		↓	
Sequence l	E	R	...u...	...	...v...	R

Consider the alignment of sequence-k and sequence-l shown in Table 2.4. In column i it is observed that $(x \rightarrow u)$, and in column-j it is observed that $(y \rightarrow v)$. In the implementation we impose the condition $i > j$ to avoid double counting. It should be noted that though $(x \rightarrow u) \equiv (u \rightarrow x)$ for a single substitution, when considering a co-substitution it is necessary that the pairing of residues in each sequence is maintained, such that $(x \rightarrow u, y \rightarrow v) \equiv (u \rightarrow x, v \rightarrow y)$.

The notation of a co-substitution event can be difficult to agree on. Firstly, consider $(x \rightarrow u)$, this represents the substitution of x with u. Now consider, $(x \rightarrow u, y \rightarrow v)$, this shows clearly that x is substituted by u and y is substituted by v. However, the notation does not necessarily imply a co-substitution, it could also imply two independent substitutions. As such the following can be used to more explicitly state the co-substitution event: $(xy \rightarrow uv)$, i.e. that the residues-types xy are being substituted by residue-types uv. In this form the residue pair from each sequence are shown together, which can reduce any confusion surrounding which sequence a residue is in. Yet, it would be more truthful to use $(xy \leftrightarrow uv)$, because it is difficult to be certain of the direction of the co-substitution in evolutionary time. Throughout this thesis the notation that will be used to represent a co-substitution event is of the form $(xy \leftrightarrow uv)$ as it was felt that it was the most descriptive notation.

Finally, when we consider the positions i and j in a pair of aligned sequences, such that (x, y) in sequence-k are aligned with (u, v) in sequence-l, we are investigating the propensity of the event that $x \rightarrow u$ and $y \rightarrow v$ (or $u \rightarrow x$ and $v \rightarrow y$) occur together at different euclidean

distances. Since we are dealing with what are effectively different pairs of aligned-residue-types, the mathematics remains the same for the following conditions: ($x \neq u, y \neq v$), ($x \neq u, y = v$), ($x = u, y \neq v$), ($x = u, y = v$), ($x = y, u = v$) and ($x = y = u = v$). The important step in our analysis is to define the combination of residue-types (xy, uv) that we want to investigate. A useful consequence of this property could make it possible to determine the range of an interaction. For example, consider ($VI \rightarrow SL$), if ($VI \rightarrow VL$) is also analysed, it may be the case that we observe a distance dependent decrease, or cut off, in the propensity of the former to occur which coincides with an increase in the propensity of the latter. This would be an indication that VI behaves in a similar fashion to SL in the structure up to a certain distance separation. A further useful feature of this method is the fact that the conservation ($xy \leftrightarrow xy$) can be analysed with no modification to the method, due to the mathematical equivalence described earlier.

What follows is a development of a statistical analysis method to determine the propensity for co-substitution events to occur when separated by different physical distances.

Concepts and Notation

The single, but crucial, difference between the co-substitution analysis in this work and the co-evolution analyses and correlated mutation analysis found in the literature, is the investigation into the explicit relationship between distance and co-substitution events. In Figure 2.1, a macromolecule is shown in conjunction with a sequence alignment and a distance matrix. The sequence alignment is the search space, where co-substitutions can be observed (and their frequency recorded) but there is no physical distance information obviously available. The distance information is gathered from a protein whose sequence is in the sequence alignment and for which an experimentally defined structure exists. The inter-residue distances measured between each pair of residues in the structure is stored in the disntance matrix, and provides inter-column distance information for the alignment. In this way, it is possible to determine the frequency of a given co-substitution-type (e.g. AG $\leftrightarrow$ GS) at different inter-residue distances in the structure.

The mathematical notation used in the following application of $\frac{O}{E}$ to the co-substitution

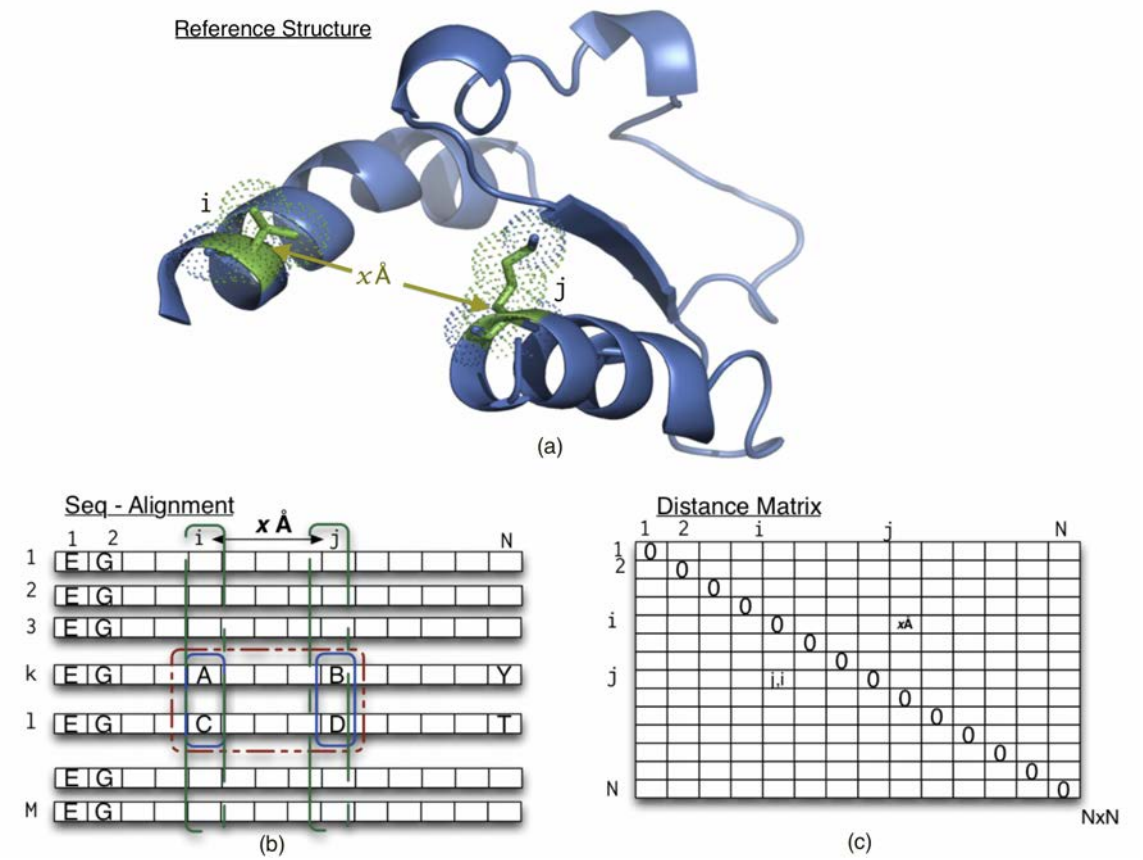

Figure 2.1: Capturing the distance information for co-substitution event AB ↔ CD: (a) is a segment of tertiary structure with the physical spearation of x Å between two residues i and j highlighted. (b) is a sequence alignment of homologuos sequences, for which the structure segmenant in (a) is a representative structure. The columns i and j are aligned to the positions i and j in the strucutre. (c) is a distance matrix, which is used to store all inter-residue distance from the structure shown in (a). The inter-residue distnaces in the distance matrix are used as the physical distances between columns in the sequence alignment shown in (b).

analysis is as follows:

$$d \in D;\ D = \{\text{all inter-residue distances}\}$$

d is an inter-residue distance and D is the set of all inter-residue distances, retrieved from the structure.

$$c \in C;\ C = \{\text{all pairs of aligned columns in the sequence-pair}\}$$

c is any co-substitution in the two sequences being considered, and C is the set of all possible

such pairs of aligned residues from the two sequence. As discussed in Section 2.4.1, no special distinction is needed between conservations and substitution, for a more detailed mathematical treatment of this problem please see Appendix A.

$\sum_d$ is the sum with respect to a specified distance, d.

$\sum_D$ is the sum with respect to all distances, D.

The three forms of $\frac{O}{E}$ applied to co-substitution analysis

In the context of the co-substitution analysis, the Observed, O, is a measure of dependency of some co-substitution c on some distance d in a given protein structure. The Expected, E, is to control for the biases that arise from the shape and composition of that protein structure. It addresses the question, does what we see in the Observed derive from an intrinsic bias in the distributions of c and d with respect to each other? or do these simply arise as a result of the shape and amino acid composition of the protein? A divergence between the Observed and Expected would be indicative that an intrinsic dependence exists between co-substitution events and physical distance.

In section 2.2.1 it was shown that there are three methods which are mathematically equivalent, by which an Observed distribution and the Expected distribution can be calculated. Here follows an application of each method to the analysis of co-substitutions, to determine the propensity with which co-substitution event types might occur at different distances apart.

$\frac{O}{E}$ **Method 1:** The first method presented here, considers the distribution of distance with respect to co-substitution events.

$$\text{Observed}_1\text{: } O_1 = \frac{\sum_d c}{\sum_D c} = P(d|c) \text{ for a given pair of proteins} \tag{2.12}$$

This is the proportion of all co-substitutions of type c with inter-residue distance d, in the sequence-pair being considered. In the limit of many observations it is the conditional probability of a distance given that we observe a co-substitution c.

Compositional bias in the given sequence is implicitly corrected for, in the Observed O, since it looks at the proportion of individual co-substitutions (c), occurring at a given distance. Thus, if the absolute quantity of c increases for that co-substitution-type, then this does not affect the observed.

$$\text{Expected}_1\text{: } E_1 = \frac{\sum_d C}{\sum_D C} = P(d) \text{ for the reference protein} \tag{2.13}$$

This represents the proportion of all amino acid-pairs in the protein that are separated by a given distance. Which is the intrinsic bias for any two residues to be some distance-d apart in the structure under consideration. It is ignorant of the amino acid composition of the protein and is only concerned with the probability of some distance d to exist between any two points in the structure. To measure this only requires a structural example representing the aligned positions in the multiple sequence alignment being considered.

The null hypothesis for this method states: 'co-substitution does not tell us anything about distance.'

$$\frac{O_1}{E_1} = \frac{P(d|c)}{P(d)} \tag{2.14}$$

$\frac{O}{E}$ **Method 2:** The second method presented here considers the distribution of co-substitution events with respect to distance.

$$\text{Observed}_2 \text{ :} O_2 = \frac{\sum_d c}{\sum_d C} = P(c|d) \text{ for a given pair of proteins} \tag{2.15}$$

This is the proportion of all co-substitutions of type c separated by distance d, in the sequence pair being considered. For that protein pair it is the conditional probability, of a co-substitution event given the physical distance between the two positions.

$$\text{Expected}_2\text{: } E_2 = \frac{\sum_D c}{\sum_D C} = P(c) \text{ for a given pair of proteins} \tag{2.16}$$

This is the proportion of all co-substitution types C which are of type c, in the sequence-

pair being considered. This form of the Expected, represents the intrinsic bias in the data for a specific co-substitution event to occur. It is ignorant of distance, but is concerned with the total number of possible pairings of aligned residue positions in the alignment of the two sequences being considered. i.e. If there were no distance bias in the distribution of co-substitution events then c should occur in each distance bin proportional to its existence in the set C. Further it represents the bias in the data arising from the amino acid composition of both sequences.

The null hypothesis for this method states: 'inter-residue distance does not bias the types of co-substitutions that can be observed.'

$$\frac{O_2}{E_2} = \frac{P(c|d)}{P(c)} \tag{2.17}$$

$\frac{O}{E}$ Method 3, for completeness:

$$\text{Observed}_3 : O_3 = \frac{\sum_d c}{\sum_D C} = P(c,d) \tag{2.18}$$

This is the joint probability distribution of c and d.

$$\text{Expected}_3 : E_3 = \frac{\sum_d C}{\sum_D C} \times \frac{\sum_D c}{\sum_D C} = P(c)P(d) \tag{2.19}$$

$$\frac{O_3}{E_3} = \frac{P(c,d)}{P(c)P(d)} \tag{2.20}$$

Deriving equivalence of the three methods

In Section 2.2.1 an equivalence between each form of the $\frac{O}{E}$ was provided in Equation 2.3. The same equivalence applies here in the form:

$$\frac{P(d|c)}{P(d)} = \frac{P(c|d)}{P(c)} = \frac{P(c,d)}{P(c)P(d)} \tag{2.21}$$

$$\Rightarrow \frac{O_1}{E_1} = \frac{O_2}{E_2} = \frac{O_3}{E_3}$$

Discussion of $\frac{O}{E}$ -ratio for Co-substitution and usefulness in predictions

The great historical interest in correlated mutations (represented by the 20,300 or so results from a PubMed search for "correlated mutations"[2]) has been driven by the prospect of being able to perform useful protein structure predictions. Ultimately it would be a great step forward in protein-structure prediction if this work could be used to predict the probability of different inter-residue distances in unsolved protein structures.

Section 2.2.2 describes the relationship between $\frac{O}{E}$ and Bayes Theorem. Here is an application of that relationship applied to co-evolution, to show how it could be used to make predictions of inter-residue distances, between amino acids in protein structures.

From equations 2.14 and A.43:

$$\frac{O}{E} = \frac{P(d|c)}{P(d)} = \frac{P(c|d)}{P(c)} \tag{2.22}$$

Let us consider method 1 and method 2 from section 2.4.1:

$$\frac{P(d|c)}{P(d)} = \frac{P(c|d)}{P(c)} \tag{2.23}$$

As shown in section 2.2.2, there are two possible probability distributions which can be predicted with the available data:

1. $P(d|c)$, the probability of some distance, given a co-substitution:

$$P(d|c) = \frac{P(c|d) \bullet P(d)}{P(c)} \equiv \frac{P(c|d)}{P(c)} \bullet P(d) \Rightarrow \frac{O}{E} \bullet P(d) \tag{2.24}$$

2. $P(c|d)$, the probability of some co-substation, given a distance:

$$P(c|d) = \frac{P(d|c) \bullet P(c)}{P(d)} \equiv \frac{P(d|c)}{P(d)} \bullet P(c) \Rightarrow \frac{O}{E} \bullet P(c) \tag{2.25}$$

In the context of being able to predict inter residue distances, equation 2.24 would be the

[2]Search performed on the 31^{st} January 2103, the number was rounded to the nearest hundred.

one to use. This is discussed further in Chapter 5. However, it is useful here to show that the combined data gathered from the distance-matrix and the sequence alignment, could be used to derive predictions using a Bayesian approach. I have not done this; it is one of the progressions that could follow from this work.

2.4.2 Application of $\frac{O}{E}$ to the Analysis of Solvent Exposure

In Section 1.3, different methods of measuring solvent exposure were introduced and an explanation of some of these was given. Here follows an explanation of how the statistical preference or propensity for each amino-acid type to have a given measured value of solvent exposure can be calculated using the $\frac{O}{E}$ analysis. Solvent exposure can be either a measure in Å^2 or in the case of HSE a dimensionless count. However because $\frac{O}{E}$ is based on the frequency of occurrence of an event, the units are not needed in the calculations. This allows for a single generalised application of the method which can be applied to either measure.

Notation and concepts

We consider a protein molecule and measure the solvent exposure of each amino acid. Thus we have a value of solvent exposure assigned to each amino acid residue in the protein sequence. We are interested in determining if solvent exposure is dependent on residue type r.

The set of possible amino acid types, considered in this work was the 20 naturally occurring residue types:

$$\{\text{A,R,N,D,C,Q,E,G,H,I,L,K,M,F,P,S,T,W,Y,V}\}$$

For our analysis:

$$r \in R; R = \{\text{all amino acid types in the protein being considered}\}$$

This is an important definition, which states that we only consider the residues present in the protein under consideration. This has implications in support of our approach to deal with Simpson's Paradox, discussed in Section 2.3. Consider residues such as cysteine or histidine,

both of which are relatively rare. By calculating the $\frac{O}{E}$ for each residue-type present in a protein on a per protein basis and determining an average based on the number of proteins with a residue present, we avoid the assumption that all proteins have all residue types.

Having measured the solvent exposure of each amino acid in the protein, we have a set of solvent exposure measures, for the protein, we denote a single measured value as a. Formally:

$$a = \text{A measured value of solvent exposure}$$

$$a \in A; A = \{\text{all measured values of solvent exposure in the protein being considered}\}$$

Here a can be either a value for the solvent accessible surface area of a residue which is solvent exposed, or the Half Sphere Exposure of the residue.

$\sum_a$ is the sum with respect to a specific value of solvent exposure

$\sum_A$ is the the sum with respect to all values of solvent exposure

The three forms of $\frac{O}{E}$ applied to solvent exposure analysis

In the context of the solvent-exposure analysis, the Observed O, is a measure of dependency of some measured value of solvent exposure on residue type. While the Expected E, is again a control for the biases that arise from the shape and composition of the protein structure we are considering. It addresses the question, does what we see in the Observed derive from a bias in the distributions of a and r with respect to each other? or do these simply arise as a result of the shape and amino acid composition of the protein? A divergence between the Observed and Expected would be indicative that a dependence exists between solvent exposure and residue type. It is known that such a relationship exists and so our objective is to determine what the preferential values of solvent exposure are for each residue type. This analysis is presented in Chapter 4.

As shown in Section 2.2.1, there are three methods for calculating $\frac{O}{E}$ that are mathematically equivalent.

$\frac{O}{E}$ **Method 1:** This first method considers the distribution of surface area with respect to residue-type r:

$$\text{Observed}_1{:}O_1 = \frac{\sum_a r}{\sum_A r} = P(a|r) \text{ for that protein} \tag{2.26}$$

This is the proportion of all residues of type r which have a measured solvent exposure value a. In the limit of many observations the Observed represents the conditional probability for a value of solvent exposure a given a specific residue-type r.

$$\text{Expected}_1{:}\ E_1 = \frac{\sum_a R}{\sum_A R} = P(a) \text{ for that protein} \tag{2.27}$$

This is the proportion of all solvent exposure measures A which are some specific value a. It represents the intrinsic bias for a given value of solvent exposure to occur and as such reflects the shape and size of the protein. It is ignorant of the amino-acid composition of the protein and is only concerned with the probability of the a to occur, although composition is to some degree implicitly accounted for in O.

$$\frac{O_1}{E_1} = \frac{P(a|r)}{P(a)} \tag{2.28}$$

The null hypothesis for this method states: 'a residue's type does not determine its solvent exposure.'

$\frac{O}{E}$ **Method 2:** This second measure of the observed considers the distribution of a residue-r with respect to its solvent exposure.

$$\text{Observed}_2{:}\ O_2 = \frac{\sum_a r}{\sum_a R} = P(r|a) \text{ for that protein} \tag{2.29}$$

This is the proportion of all residues having a solvent exposure value of a that are of residue-type r, for the protein being considered.

$$\text{Expected}_2{:}\ E_2 = \frac{\sum_A r}{\sum_A R} = P(r) \text{ for that protein} \tag{2.30}$$

This is the proportion of all residues in the protein which are r, and represents the bias introduced by the amino-acid composition of the protein. It is ignorant of the solvent exposure,

but is concerned entirely with population variety of amino acids in the data. Since O measures the proportion of residues with exposure a that are of type r, it implicitly accounts for some degree of protein shape.

The null hypothesis for this method states: 'solvent exposure does not affect the distribution of residue-type r in the protein.'

$$\frac{O_2}{E_2} = \frac{P(r|a)}{P(r)} \tag{2.31}$$

$\frac{O}{E}$ Method 3:

$$\text{Observed}_3 : O_3 = \frac{\sum_a r}{\sum_A R} = P(r,a) \text{ for that protein} \tag{2.32}$$

This represents the joint probability of r and a.

$$\text{Expected}_3 : E_3 = \frac{\sum_a r}{\sum_a R} \times \frac{\sum_a R}{\sum_a R} = P(r) \times P(a) \text{ for that protein} \tag{2.33}$$

Equivalence of the three methods

In Section 2.2.1 an equivalence between each form of the $\frac{O}{E}$ was provided in Equation 2.3. The same equivalence applies here in the form:

$$\frac{P(a|r)}{P(a)} = \frac{P(r|a)}{P(r)} = \frac{P(r,a)}{P(r)P(a)} \tag{2.34}$$

$$\Rightarrow \frac{O_1}{E_1} = \frac{O_2}{E_2} = \frac{O_3}{E_3} \tag{2.35}$$

2.5 Sequence weighting

The search-space for co-substitution events is all pairs of homologous sequences, taken from multiple sequence alignments. A consideration that must be addressed is the unequal distribution of data that can arise from sets of homologous sequence pairs with differing degrees of sequence identity, i.e. we would like to sample the structure/sequence space evenly. Homology in this context refers to those proteins or genes which share a common evolutionary ancestor

have the same fold and the same, or very similar, function. Protein sequence pairs which have a high sequence identity do not offer much new information and each Pfam family is given equal weights in the final average, if all sequences were counted equally this would lead to a biasing of the results. To sample all pairs of sequences equally a weighting of the sequences was introduced. The Henikoff & Henikoff method of weighting sequences based on similarity [64] is used extensively in the literature to achieve this type of weighting. This approach gives similar sequences a reduced score, that reflects their collective influence on the data.

Henikoff weighting for our purposes, has some limitations. Firstly, if two identical sequences are aligned with a set of similar but non-identical sequences, the weight of the duplicate sequences is not half that of a single copy of the same sequence aligned with the others, as would be ideally expected, an example of this is shown in Table 2.6. Secondly, the method was not designed for weighting pairs of sequences compared with other pairs of sequences.

To ensure that the $\frac{O}{E}$ calculated for each co-substitution type for each pair of sequences was appropriately weighted, it was necessary to develop a method of weighting pairs of sequences. Our method of weighting, which is described in Section 2.5.2 weights a pair of sequences based on the similarity of the pair, with the set of all other pairs of sequences from an alignment of sequences. The method is based on the idea that sequences with high identity will have many columns with the same type of amino acids as each other; so each residue type at each position is given a score based on its abundance in the column, and each sequence has a score determined from the residue type it has in each column. The aim of this method is to return a weighting which is analogous to the Henikoff & Henikoff weighting, but for pairs of sequences. The weakness of this new method arises from its reliance on the Henikoff method and the inherent weakness therein.

Before discussing our method, I discuss Henikoff & Henikoff weighting first.

2.5.1 Henikoff & Henikoff weighting

Table 2.5 is an example of calculating the sequence score for an alignment of 4 sequences with a total of 5 columns. Table 2.6 shows the same example as shown in Table 2.5, with a duplicated

sequence to illustrate the effect of this on the weights overall.

Table 2.5: Example of the Henikoff weighting method: In column 1 there are 3 types of residue "TES", for the first letter of the sequence TRIAL, the letter T would have a score of $\frac{1}{3\times2}$ because there are 2 letter Ts in the column. The sum of all the scores for the sequence letters is $\frac{13}{12}$ and the number of columns is 5. Thus the weight for the first sequence is $\frac{13}{12} \div 5 = 0.2167$

Residues positions	1	2	3	4	5	**Sum**	Normalised
TRIAL	$\frac{1}{(3\times2)}$	$\frac{1}{(2\times3)}$	$\frac{1}{(3\times1)}$	$\frac{1}{(4\times1)}$	$\frac{1}{(3\times2)}$	$\frac{13}{12}$	$\frac{13}{12} \div 5 = 0.2167$
TRAIL	$\frac{1}{(3\times2)}$	$\frac{1}{(2\times3)}$	$\frac{1}{(3\times2)}$	$\frac{1}{(4\times1)}$	$\frac{1}{(3\times2)}$	$\frac{11}{12}$	$\frac{11}{12} \div 5 = 0.1833$
ERRQR	$\frac{1}{(3\times1)}$	$\frac{1}{(2\times3)}$	$\frac{1}{(3\times1)}$	$\frac{1}{(4\times1)}$	$\frac{1}{(3\times1)}$	$\frac{17}{12}$	$\frac{17}{12} \div 5 = 0.2833$
STAND	$\frac{1}{(3\times1)}$	$\frac{1}{(2\times1)}$	$\frac{1}{(3\times2)}$	$\frac{1}{(4\times1)}$	$\frac{1}{(3\times1)}$	$\frac{19}{12}$	$\frac{19}{12} \div 5 = 0.317$
Sum	1	1	1	1	1	5	1

Table 2.6: An Example of the Henikoff weighting method, with two identical sequences: This contains the same set of sequences as shown in Table 2.5, with the sequence 'TRAIL' duplicated.

Residues positions	1	2	3	4	5	**Sum**	Normalised
TRIAL	$\frac{1}{(3\times3)}$	$\frac{1}{(2\times4)}$	$\frac{1}{(3\times2)}$	$\frac{1}{(4\times2)}$	$\frac{1}{(3\times3)}$	$\frac{23}{36}$	$\frac{23}{36} \div 5 = 0.128$
TRIAL	$\frac{1}{(3\times3)}$	$\frac{1}{(2\times4)}$	$\frac{1}{(3\times2)}$	$\frac{1}{(4\times2)}$	$\frac{1}{(3\times3)}$	$\frac{23}{36}$	$\frac{23}{36} \div 5 = 0.128$
TRAIL	$\frac{1}{(3\times3)}$	$\frac{1}{(2\times4)}$	$\frac{1}{(3\times2)}$	$\frac{1}{(4\times1)}$	$\frac{1}{(3\times3)}$	$\frac{169}{216}$	$\frac{169}{216} \div 5 = 0.156$
ERRQR	$\frac{1}{(3\times1)}$	$\frac{1}{(2\times4)}$	$\frac{1}{(3\times1)}$	$\frac{1}{(4\times1)}$	$\frac{1}{(3\times1)}$	$\frac{7}{4}$	$\frac{11}{8} \div 5 = 0.275$
STAND	$\frac{1}{(3\times1)}$	$\frac{1}{(2\times1)}$	$\frac{1}{(3\times2)}$	$\frac{1}{(4\times1)}$	$\frac{1}{(3\times1)}$	$\frac{19}{12}$	$\frac{19}{12} \div 5 = 0.317$
Sum	1	1	1	1	1	5	1

The steps for calculating the sequence score are:

1. Calculate the score for each amino acid in a column by dividing 1 by the product of the number of occurrences of the amino acid type at a given position, multiplied by the total number of amino acid types in the column.

2. Calculate the sum of all the scores assigned to each residue in a sequence which gives an unnormalised score for each sequence.

3. Calculate the sum of the unnormalised scores for all the sequences in the alignment, this must be equal to the total number of columns in the alignment.

4. Divide each sequence score by the total number of columns in the alignment. The sum of the normalised scores must be equal to 1.

Inspection of the normalised weights in Table 2.6 , reveals that the combined weight of the two'TRAIL' sequences is not quite half of the single sequence on its own shown in Table 2.5. Further the weights of 'TRIAL' and 'ERRQR' are slightly down-weighted. This arises because weights are assigned to every residue in a column and both TRIAL & ERROR have residues in common with 'TRAIL'. The result is that the combined weight of the two 'TRAIL' sequences is slightly greater than the weight of the individual sequence alone. The weight for the sequence 'STAND' is unaffected, because the additional sequence shares no common letters with it. This shows that the weighting change is distributed amongst the sequences which have common letters with the duplicate sequence. This is not ideal for our purposes and is a compromise in light of the lack of an appropriate alternative.

2.5.2 Weighting Sequence Pairs

As stated earlier, our aim is to generate a weight for every pair of sequences from an alignment. To demonstrate the procedure developed here, I shall use the alignment in Table 2.5. Table 2.7 shows all possible pairs of sequences in Table2.5. A further mathematical proof of this method can be found in Appendix B. The procedure is as follows:

Table 2.7: Sequence pairs: The unique sequence pairs that can be made from the sequences in Table 2.5.

					Weight
T	R	I	A	L	??
E	R	R	Q	R	
T	R	I	A	L	??
S	T	A	N	D	
T	R	I	A	L	??
T	R	A	I	L	
E	R	R	Q	R	??
S	T	A	N	D	
E	R	R	Q	R	??
T	R	A	I	L	
S	T	A	N	D	??
T	R	A	I	L	

To calculate the sequence-pair weights, the following method was developd:

1. Calculate the Henikoff & Henikoff weighting for all sequences, such that $h(x_k)$ = weighting for sequence k, shown in Table 2.5. This could be replaced by an appropriate method for sequence weighting but the Henikoff method is the method that was used in this thesis.

2. Assemble all pairs of sequences and measure their sequence-identity with respect to each other, where:

$$\text{sequence identity} = \frac{\sum \text{matched positions}}{\text{length of the aligned sequences}} \tag{2.36}$$

For the sequences shown in 2.5 we have the following results:

	TRIAL	ERRQR	STAND	TRAIL
TRIAL	1	0.2	0	0.6
ERRQR		1	0	0.2
STAND			1	0.2
TRAIL				1

3. Calculate the sequence difference score:

$$d(x_k, x_l) = 1 - (\text{sequence identity}) \tag{2.37}$$

for our data this is:

	TRIAL	ERRQR	STAND	TRAIL
TRIAL	0	0.8	1	0.4
ERRQR		0	1	0.8
STAND			0	0.8
TRAIL				0

4. Calculate the un-normalised sequence-pair weighting using:

$$W(x_k, x_l) = h(x_k) \times h(x_l) \times d(x_k, x_l) \tag{2.38}$$

which gives:

	TRIAL	ERRQR	STAND	TRAIL
TRIAL	0	0.049111111	0.068611108	0.015888884
ERRQR		0	0.089722222	0.041555544
STAND			0	0.046444430
TRAIL				0

5. Calculate normalisation factor:

$$\sum \frac{1}{\text{all sequence-pair-weights}} = \sum_{x<l} \frac{1}{W(x_k, x_l)} \tag{2.39}$$

$$= \frac{1}{0.0491 + 0.0686 + 0.0158 + 0.0897 + 0.0415 + 0.0464}$$

$$\boxed{= 3.212}$$

6. Multiply all pair-weights by the normalisation factor, to give our final weighting:

	TRIAL	ERRQR	STAND	TRAIL
TRIAL	0	0.15774449	0.22037832	0.051034965
ERRQR		0	0.28818704	0.13347607
STAND			0	0.14917913
TRAIL				0

Sum of Normalised weights = **1**

The case where two identical sequences are paired is not explored here. This is because two identical sequences do not affect the result. However, a provision was made in the analysis to remove sequences which had very high sequence identity with others, because of the weakness of the Henikoff weighting method.

Additional, it can be noted from the table above, that no weighting can be assigned to a pair of identical sequences, simply because the product of Henikoff weightings for both sequences and the sequence difference will always be zero, because the sequence difference will always be zero. If a sequence S is paired with a each of a pair of duplicated sequences, A and A', then the weight W of each SA and SA' are equal to half the weight of SA in the case where A is not duplicated. i.e. the weighting reduces the weight of duplicate sequences pairs by exactly $\frac{1}{N}$ where N is the number of times a sequence is duplicated.

2.6 Acknowledgement

The pair-wise weighting method developed for this thesis, presented in section 2.5.2 and Appendix B, which is implemented in Chapter 5, was developed in conjuction with Mr. Robert Clegg. The formal proof presented in Appendix B was prepared by Mr. Clegg.

Chapter 3

Development of data selection

3.1 Introduction

The primary aim of this thesis is to investigate the relationship between inter-residue distance in 3D structures and the propensity for different types of amino acid co-substitutions. For this an appropriate set of data must be collated for statistical analysis. The first, requirement of this data is that it must consist of both structural data and sequence data. There are different ways in which sequence data and structure data can be combined in order to provide input for the statistical analysis. However they all require the assembly of protein sequence data mapped to protein structure data, which can be sourced from several on-line data-banks, such as the PDB [65] and UniProt/TREMBLE [66]. Additionally on-line resources such as the Blocks database, HSSP [33] and Pfam exist, which are curated data-banks of protein families. The data requirements for the analyses in this thesis, could be acquired by marrying up structural data from the PDB with protein families from any of the protein family databases. The decision was made to use data from the Pfam-A database as it is based on a manually curated seed alignment.

The statistical methods developed in the previous chapter endeavour to ensure a uniform statistical weighting of the data. This would be in vain if the data on which the method is to be applied did not account for the inherent variety that is known to exist in protein data. Amino acid composition of proteins varies between locations, e.g. extra-cellular vs. intra-cellular, mitochondria vs. cytoplasm [67]. Differences in pH between organelles could affect the pro-

tonation state of any acidic or basic residues in a protein structure, most notably histidine. Similarly changes in redox potential will affect cysteine's propensity to form disulphide bonds. Further membrane proteins have hydrophilic regions exposed to solvent with hydrophobic regions buried into the membrane, with a suitable shape to accommodate their function. These constraints have implications for the appropriate structural positioning of residues and for the variance of protein sequence composition. Efforts have been made to exploit these differences for the purpose of predicting cellular locations of proteins [68].

From the perspective of the developed statistical method, it must be noted that the variation of amino acid composition arising from different organelle locations, will alter the Expected distribution – described in the previous chapter – and therefore an imperative to segregate the data into appropriate subsets exists. Therefore, when performing a rigorous analysis of amino acid propensity to be solvent accessible or to be involved in a correlated substitution with other amino acids, we should not ignore these considerations. The effect of these differences has, to the best of our knowledge, not previously been considered in the context of statistical analysis of studying protein attributes. Accounting for these considerations, using just the PDB and the Pfam databases would be extremely difficult as neither provides easy access to information regarding such things as taxonomic classification, or cellular location. This additional information exists in the UniProt databases: TREMBLE and SwissProt. TREMBLE is an automatically annotated database while SwissProt is manually curated and as such can be considered more reliable.

There is currently no straightforward method to simultaneously query the PDB, SwissProt and Pfam data-banks. Thus, selection of protein structures or domain structures based on cellular location, taxonomy or other biological context is not simple, restricting our ability to perform context dependent analyses. Presented in this chapter is a description of a relational database developed to cross references the sequence, functional and contextual data from the SwissProt with the domain and sequence alignment information in Pfam-A, and the structural information in the PDB and the PiQSi database of quaternary structures [69] (the PiQsi database was chosen to provide the biological unit of proteins, to make the results biologically relevant). The result-

ing database made it easier and faster to search and cross reference these data sources with SQL, thus allowing for subsets of the available data to be easily selected. This provides a versatile means for novel selection of data from either Pfam or the PDB (or both) based on information contained in the UniProt/SwissProt.

This chapter covers an overview and discussion of the data selection method developed and used for this thesis. A more technical coverage of this work can be found in Appendix C.

3.2 The Merging of SwissProt, Pfam and the PDB

The analyses presented in the subsequent chapters of this thesis are studies of the biological context of amino acids and amino acid substitutions. To address the need to quickly and reliably select subsets of protein structural data filtered for cellular location and taxonomy, I created a relational database that cross references the UniProt/SwissProt, Pfam-A, the PDB and the PiQSi database of the quaternary structures. Thus, using contextual information found in the UniProt/SwissProt data it is possible to quickly and easily select subsets of PDB, PiQSi or Pfam data. Additionally, I have also developed a tool that allows the selection of atomic co-ordinates of Pfam domains from larger structural ensembles. Thus selections of data can be made, for example, based on: taxonomy, host taxonomy, cellular location, Pfam domain, key words in the SwissProt and regular expression searches of the comments section of a SwissProt entry.

A MySQL database with the necessary tables to store data from Pfam, UniProt and from the DBRef section of the PDB headers was created. A python program was written to parse data from the freely available text files for all three online databases (Pfam in FASTA format) and populate each table in the database (shown in Tablele 3.1). With some slight modification the code could be extended to run on alternative versions of the UniProt and Pfam data sets, such as UniProtKB and Pfam-B.

Table 3.1: List of database tables and description of their contents: These tables were created to store the data from each of the on-line data-banks.

Table Name	Description
PDB UniProt cross reference data	The DBRef data from the PDB header that matches UNP & SWS codes.
Pfam data	The Pfam FASTA data, including Sequence and UniProtID
Pfam cross reference data	The pdbmap data from Pfam.
UniProt to PDB cross reference data	The PDB cross reference data for SwissProt entries.
UniProt to Pfam cross reference data	The Pfam cross reference data for SwissProt entries.
UniProt PDB Pfam cross reference data	The Combined PDB, Pfam and SwissProt cross reference data from SwissProt.

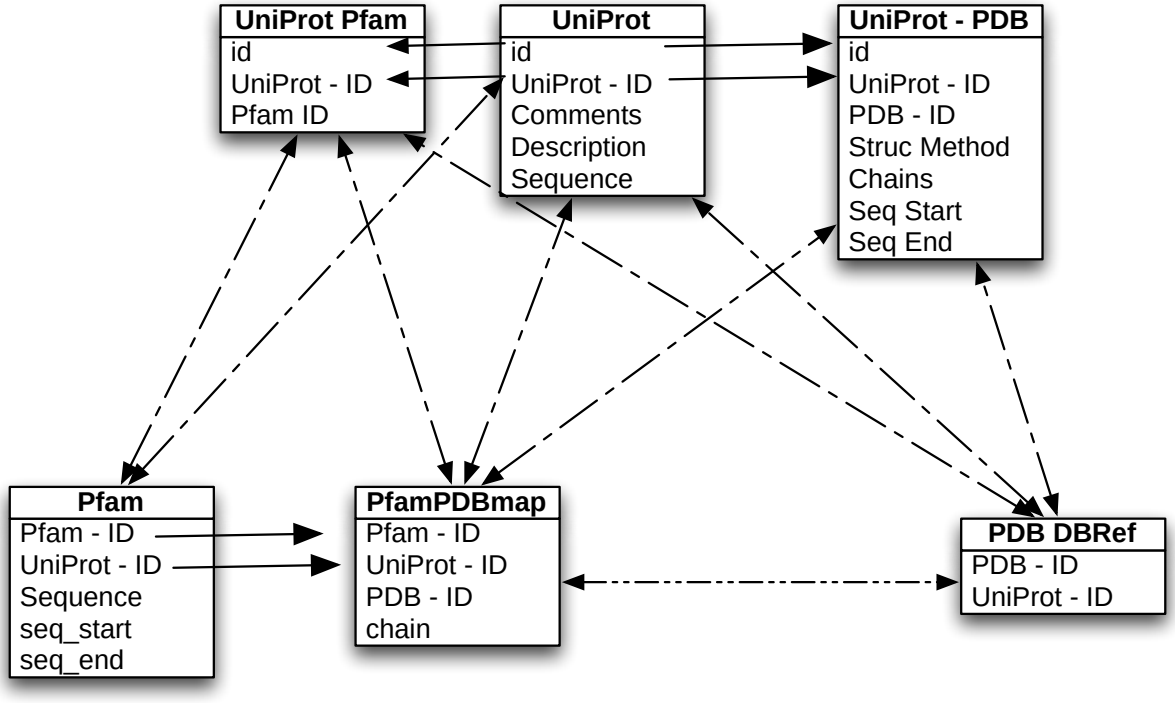

Figure 3.1: Available cross-referencing: The three selected on-line data-banks reference each other. The single-headed arrows with unbroken lines indicate relationships between data from the same on-line database. The double-headed rows with broken lines, represent cross-referencing from one on-line data source to another.

The data available from each of the on-line resources contain cross-references to each other, as shown in Figure 3.1. The data was downloaded from the respective repositories in plain text format. This was parsed using a Python script and stored in a set of linked tables in a MySQL database, for each of the online databases. The cross reference data from each was checked against the cross references from the other two using a purpose written Python program. This was achieved by retrieving a list of all PDB structures that are referenced in either SwissProt or Pfam. UniProt/SwissProt and Pfam each have their own cross reference record to a Pfam, and a UniProt ID. For each entry in this list of PDB IDs, the UniProt and Pfam IDs stored in the Pfam table are compared to those in the UniProt table. If these agree then the corresponding sequences are retrieved for Pfam, UniProt and PDB and the location of the Pfam domain is found in the UniProt and PDB sequences using the Tre[1] module for Python. Tre returns the start and end positions of the Pfam domain in the UniProt sequence and the residue numbers for the start and end points in the PDB structure. A tolerance of 10% mismatch is allowed to account for missing residues in the PDB structure and potential disagreements between databases. Note that the sequence for the PDB structure is retrieved from the atomic data and not the header data, so that the residue number can be returned. For a given PDB ID in the list of PDB structures, should the Pfam and UniProt IDs stored in the Pfam table not match those in the UniProt table then it is still possible that these entries do match but that the databases are not synchronised in their cross referencing. In this case, corresponding amino acid sequences for each entry are retrieved and Tre is used to try to locate the Pfam domain in the UniProt and PDB sequences as well. In the event that either a sequence cannot be retrieved or Tre can not find a match with a maximum of 10% error, the cross reference is discarded. All correctly matched data was recorded in its own table in the MySQL database. The result of this procedure is to produce a relationship between the databases that is illustrated in Figure 3.2. This step ensures that every entry in the table has been checked and found to be consistent and correct, thus minimising any errors, known or unknown, in the source data. The program developed for the above was used to create a separate cross reference table, for the PiQSi database of quaternary structures, for

[1]Tre is an open-source regular expression matcher, which measures the Debye distance between to strings. It is available from: http://hackerboss.com/approximate-regex-matching-in-python/ correct on 7 May 2013

the list of all PDB IDs present in the PiQSi, i.e this is the same procedure but on a smaller list of PDB ids.

The UniProt/SwissProt database offers a wealth of information about each protein entry, including taxonomic classification, known functional information and cellular location amongst others. To expand the versatility of this database, as much of the data included in each entry as was available has been put in the UniProt data table, including the comments, keywords, organism, and host organism. It is possible to modify the source code to include those which are relevant to a given use. This makes possible novel selection of data from either Pfam or the PDB (or both) based on parameters set against data in UniProt/SwissProt.

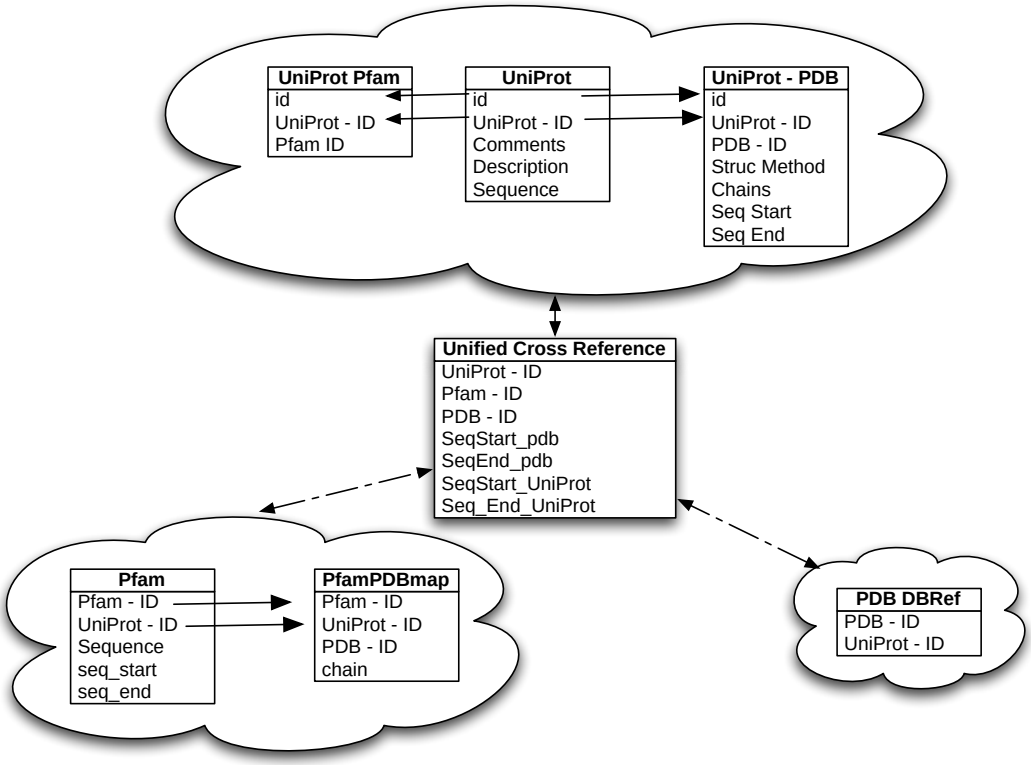

Figure 3.2: The Unified Cross-Reference produced here: The merging of the data from the three data-banks and the use of the regular expression matcher Tre has made it possible to find the set of proteins and protein families for which data exists in all three databases. This facilitates rapid data selection. The single headed arrows with an unbroken line represent relationships between data from the same on-line data-bank. The double headed arrows represent relationships between data from different on-line data-banks.

To ensure that the correct data has been included in the database, the total number of UniPro-

t/SwissProt entries were checked against the on-line database, similarly for the Pfam. For the PDB there are a number of entries which do not have a cross reference to the UniProt database and these were excluded. Roughly ten random checks were made of the data to ensure that the entries were correct.

3.3 Discussion

Though the intention was to develop a method to facilitate rapid selection of Pfam and PDB data based on information stored in the UniProt/SwissProt database; the merging of the data in the three databases, into a single relational database, resulted in tool of some considerable capability. For example, it could be used to address such questions as "Is some Pfam domain only observed in a specific organelle?", or "what are all the known enzymes involved in digestion in Eukaryota and what structural data is available for them?"

One of the applications to which I applied it was a comparison of the consistency of annotation between the different databases. Since the databases are updated and released at different times it is more or less impossible for them to be completely synchronised. This relational database makes it possible to identify such inconsistencies in the cross-reference data supplied by each of the databases. The current implementation of the database allows cross referencing between the different data-sets and provides a fully merged cross reference. In Table 3.2 is presented the number of complete cross-reference matches for each data-bank, available from each.

Table 3.2: Summary of the different cross reference data content in each of the on-line data-banks: Each column shows the total number of entries for which a cross reference exists. Consider the Pfam-A map, the first entry shows 5,580 Pfam domains, 53,748 associated PDB structures and 19,451 associated UniProt entries, this should be read as follows: 'There are 5,580 Pfam-domains present in 53,748 PDB structures which corresponds to 19,451 UniProtKB entries.' This is because the Pfam-A cross references UniProtKB and not just UniProt/SwissProt. This is a summary of the cross reference data available from the three on-line databases as found in our database.

	Pfam-A	SwissProt	PDB
References in data source to Pfam	5,580	4,493	0
References in data source to PDB	53,748	50,006	43,848
References in data source to UniProt	19,451	15,077	12,572

3.3.1 Selecting data for subsequent analyses

In the context of the co-substitution analysis and solvent-exposure analysis, a suitable subset of the available data needed to be chosen for statistical analysis. Much of the literature concerns itself with globular proteins and so it was decided to try to study a set of globular proteins which were exclusively cytoplasmic, non-membrane and non-DNA binding.

Proteins themselves are composed of functional units known as domains. Protein domains are the molecular building blocks used by evolution in different combinations and arrangements to build proteins with different functions. A domain has its own tertiary structure, and many proteins are composed of multiple domains. The Pfam data-bank is a store of extracted domain sequences from protein sequences sourced from translated genomic sequence data. The domains are grouped into families, with the intention of creating a periodic table of protein domains [34]. The Pfam curators provide multiple sequence alignments built from the domain sequences in individual proteins. The multiple sequence alignment are derived from a manually curated seed alignment using Hidden Markov Models.

The data selection was first performed to prepare data for the solvent exposure analysis presented in Chapter 4. The decision to study exclusively cytoplasmic, non-membrane non-DNA binding globular proteins was refined, to consider only protein domains that are found exclusively in proteins matching that criteria. For the co-substitution analysis this was further refined to separate homo-oligomers and hereto-oligomers for reasons described in Chapter 5. The MySQL database was used to select the Pfam families which met these conditions and a list of all protein structures in the PiQSi which contained at least one of the domains were retrieved. This provided the structural data for the solvent exposure analysis, and the sequence and structural data for the co-substitution analysis.

3.4 Conclusion

A tool, in the form of a MySQL database, was developed to make selections of protein domains stored in Pfam and protein structures stored in the PDB and PiQSi, based on contextual information stored in the UniProt database. To the best of our knowledge no such tool existed before

this one. It has been used to highlight the existence of discrepancies between UniProt/SwissProt, Pfam-A and the PDB. The objective of developing this tool was to select Pfam domains based on classifications of taxonomy and cellular location. The database is capable of this and due to its versatility, significantly more.

Chapter 4

Determining amino acid solvent exposure preferences

4.1 Introduction

The likelihood of a given type of amino acid substitution event is dependent on its context [21] [70]. Context here, refers to the solvent exposure of the residue position, the secondary structure and the general environment of the protein (e.g. inter-cellular or intra-cellular, organelle etc.). To satisfy these considerations for the co-substitution analysis, presented in the next chapter, steps were taken to address the issues of amino acid context. In the previous chapter, the method for determining cellular location of a protein was discussed. Presented in this chapter is an investigation into accurately determining a meaningful measure and value that defines the boundary between residue burial and solvent exposure. Due to the already complex nature of our investigation, secondary structure context was ignored potentially to be investigated at a later date.

Hydrophilic residues will preferentially partition towards the solvent exposed surface, with the hydrophobic residues mostly being packed into the protein interior. This will affect the type of co-substitutions we might expect in the core of the protein compared to the surface. The implication of this on the co-substitution analysis, is that the potential for residue-types to be involved in co-substitutions on the surface will be different to the core. This means that the

Expected E (described in Section 2.2.1), the unbiased distribution of co-substitution events, would be biased by the difference in the types of amino acids in the population of the two environments. The question being addressed in this chapter is, "What is a statistically meaningful measurement to distinguish this compositional change?" Defining this boundary, could have implications to our understanding of protein-protein interactions, protein function, protein evolution and drug design.

A small consideration of nomenclature is required. The reader will be aware that the author jumps between the use of solvent exposure and solvent accessibility. The author has chosen to use the term "solvent exposure" as a generic term covering any measure of an amino acid's proximity to solvent, which includes residue depth, HSEu, ASA, rASA. The term " solvent accessibility" will be used to refer exclusively to either the solvent accessible surface area or the relative solvent accessible surface area of a residue.

In Section 1.3 several different methods of measuring amino acid proximity to solvent were presented. The two most common approaches to determining amino acid residue burial are Solvent Accessible Surface Area (ASA) and Relative Solvent Accessible Surface Area(rASA). The former was originally proposed by Lee and Richards in 1971 and is referred to as the 'rolling ball method' [14]. The latter, is a pseudo normalisation of the former, achieved by dividing the ASA by some approximate maximum ASA of residue X, which is measured using an extended conformation of a tri-peptide of either Ala-X-Ala or Gly-X-Gly. The choices of rASA to delimit residue burial vary from 6% to about 20% [71–73], however these choices are quite literally arbitrary with no analysis of appropriateness presented to support them. The inherent limitation of ASA and rASA is its inability to provide information concerning residues which are below the solvent accessible surface.

Of the alternative measures of solvent exposure, two potential candidate methods are Residue Depth [52, 59] and Half Sphere Exposure [53]. The Residue Depth method requires the determination of the ASA of each residue in a protein and requires solvent molecules to be placed in the vicinity of the protein to provide reference points from which to measure, making it computationally expensive. Half Sphere Exposure is a much simpler method. There are two flavours

of HSE, these are HSEu and HSEd. HSEu is the count of the number of C_α atoms in the direction of the side-chain, contained in a half-sphere of a chosen radius (usually 13 Å), whose plane is perpendicular to the C_α – C_β vector. HSEd is the number of C_α atoms in the other half-sphere. HSEd, does not offer any information regarding side chain environment and is not considered here. There are two distinct advantages of using HSEu to measure solvent exposure. Firstly by its very design it is not limited to residues that can come in contact with solvent - i.e. it can penetrate the solvent excluded volume. Secondly because HSEu is a count of the number of residues in the direction of the side-chain, it provides a component of directionality to the measure. Thus indicating if a residue is pointing inward toward the protein interior or outward toward the solvent. Thirdly it does not require a complete side chain to be present in the crystal structure, as it only takes into account the C_α atoms present in the half sphere being considered. Even a lack of side-chain can be accommodated, as the method can infer the C_α - C_β vector by bisecting the angle between the extended vectors $C_{\alpha-1}$ – C_α and $C_{\alpha+1}$ and C_α- as is necessary for Glycine.

To address the question posed above "What is a statistically meaningful measurement to distinguish the amino acid compositional change between the surface and the protein core?" the propensity ($\frac{O}{E}$) for amino acid residues to have a given measure of HSEu was analysed. HSEu appeared to be the better choice but to compare it with other studies in the literature was necessary to contrast the results with ASA. Because HSEu is dependent on the direction of the side chain, it is only comparable with the ASA of amino acid side chains; ignoring backbone atoms. Side chain ASA was measured from the C_β atom and above.

4.2 Methods

To determine the preferred solvent exposure of each residue type, a set of Pfam domains was selected. Available structural information for each was compiled. The solvent exposure of each residue was measured while the domain was in the biological unit as defined by the PiQSi database [69]. The domain was then removed from the biological unit and stored separately. The frequency of each residue having a given measure of solvent exposure within specified

range-bins was recorded. The frequency was used to determine the $\frac{O}{E}$ for each residue in each range-bin of solvent exposure. The analysis was performed on each structural example of each Pfam family independently. The weighted average across all examples of each domain family was calculated. Finally, the average of all families was determined, weighting each family equally. Here follows the procedure used to generate the presented results.

4.2.1 Procedure for analysis

The method below applies to two separate analyses, one performed on HSEu data and the other on side-chain ASA data. The procedure for both are the same except for the determination of the Expected in the ASA analysis, which needed to allow for the variation of side chain size between the different amino acids, as described later.

Preparing the data

The representative structures for Pfam domains were selected using the database described in Chapter 3. The selection criteria were to have two sets of domains, one exclusively eukaryotic and the other exclusively prokaryotic. Further, in each set, the domains were selected to be exclusively cytoplasmic, non-membrane, non-DNA and non-RNA binding. i.e. domains exclusively from cytosolic globular proteins. The set of quaternary structures that were selected from the PiQSi database contained at least one of the chosen Pfam domains. Initial results included some unusual points inconsistent with the trend indicated by the rest of the data. Investigation into this revealed that there were errors in some structure files taken from PiQSi. Thus a "cleaned" set of PiQSi structures was assembled, with the problem structures removed, leaving a total of 12,234 unique structures from which to select representatives for our Pfam families.

A script calling the program Whatif [74] was used to check the quality of the selected structures. Three operations were performed on the structures: i) Adding missing side-chain atoms into the structure. This was especially important for the ASA analysis as missing atoms would reduce the measure of ASA and introduce an error into the results. ii) Checking and correcting bond-lengths and bond angles. iii) The numbering of amino acids in the structure was checked and residues were renumbered to correct for duplications. This was done to resolve problems

that had arisen with the BioPython PDBParser module that could not handle inconsistencies in residue numbering.

For each of the 12,234 structures the solvent exposure of every residue in the structure was measured. Solvent exposure in this case refers to both HSEu and the side chain ASA in $Å^2$. The hsexpo.py script included in the BioPython module [75] for Python, was used to develop a script that worked within the work-flow to calculate HSEu. The Naccess [4] program was used to determine the ASA for the side chains. A copy of each structure file was made with the B-factor column used to store the solvent exposure measure. Thus, the solvent exposure was measured for each residue, in it's crystallised biological unit, as defined by PiQSi. Quaternary structures were used for the calculation of solvent accessibility values so that our analysis had the greatest biological and physical relevance possible.

The selected Pfam domains were located and extracted from the the crystal structures provided by PiQSi. Locating the domain in the structure required cross referencing the chosen Pfam families with the UniProtID associated with each protein structure. Although Pfam does provide start and end points for domains in protein structures, these are not recorded in a consistent fashion and therefore could not be used. A purpose built Python script was used to find the start and end points of the domains in the structure, by using the regular expression matcher Tre – mentioned in Chapter 3. The Python script located the sequence provided by Pfam in the structure sequence from PiQSi, within a 10% margin of error. The PDBParser module in BioPython provided a method for extracting the structure of the domain from the structure file from PiQSi in PDB format. Representative structures for each of the Pfam domains were each individually stored in a labelled ASCII file in PDB format. The set of representative structures stored with HSEu and ASA values were consistent with each other.

Differences in Calculating the Expected for ASA Compared with HSEu Data

Differences in the size of amino acid side chains required a different method of determining the Expected E for the ASA analysis compared with HSE. HSEu has the property that it is completely independent of residue size, it considers the amino acid population in the half sphere of a specified radius in the direction of the C_α - C_β vector for a given residue. To determine the

Expected for HSEu it was sufficient to use the method described in Chapter 2 Section 2.4.2, detailing the application of $\frac{O}{E}$ to solvent exposure analysis.

While ASA is explicitly size dependant and restricted to some specific maximum value that varies between residue types. The nature of the Expected in the analysis is to offer some unbiased or 'unconditioned' distribution with which to compare the Observed distribution. However residues such as alanine and lysine, which are of very different sizes, could never be 'expected' to share the same Expected value of absolute ASA. To estimate the Expected for each residue type in the ASA analysis 100 randomised data sets were created yielding 100 bootstrapped Observed values, based on the data from the selected Pfam domain structures. The average of these bootstrapped Observed values was taken as an approximation of the Expected ASA given no distributional bias of residues. This was used as the Expected value for each residue-type, for the analysis using normal, unrandomised, data to calculate the $\frac{O}{E}$ for each residue type. It was also used as the Expected for all the bootstrap analyses to calculate the $\frac{O}{E}$ for each residue type in the analysis of randomised bootstrap data. This was used for determining the chance variation possible in the data analysis, and thus the likely statistical significance of the true data, as described later in Section 4.2.2

The Analysis

For each Pfam family the frequency of each residue type having a value within a given range (representing a range-bin) of solvent exposure, was determined. For HSEu the range-bin size was set to 4 HSEu counts and the highest range-bin was 56-60 counts, while for ASA it was set at 10 $Å^2$ with the highest range bin set at 240-250 $Å^2$. The maximum values for each analysis were chosen by searching the data set for the highest single value of solvent exposure for each residue type. The true range of each bin was from the lower value to strictly less than the upper value i.e. the range (0 to 10) is in fact set as: $i \in [0, < 10]$, this means that 10 goes into the next bin (10 to 20), such that: $j \in [>= 10, < 20]$. The frequency of each residue in a range-bin and the frequency of each residue in the entire structure were then used to determine the $\frac{O}{E}$ for each residue-type in each range-bin. For the HSEu analysis the $\frac{O}{E}$ was calculated using the equations given in Section 2.4.2:

$$\text{Observed:} O_1 = P(a|r) = \frac{\sum_a r}{\sum_A r} \tag{4.1}$$

$$\text{Expected: } E_1 = P(a) = \frac{\sum_a R}{\sum_A R} \tag{4.2}$$

$$\frac{O_1}{E_1} = \frac{P(a|r)}{P(a)} \tag{4.3}$$

For the ASA analysis, the Observed was determined as per equation 4.1 above, the Expected was calculated as described in Section 4.2.1 had already been calculated earlier. N.B. The Expected for the ASA analysis is only calculated once, and is used for both the analysis of the normal data and the randomised bootstrap data. This was done to accommodate time considerations, it would be prohibitively expensive in terms of computational resources (both CPU cycles and data storage) to do this once for Observed value, i.e. for the normal data and once for each bootstrap analysis.

Our objective was to calculate a value of $\frac{O}{E}$ which was generic and unbiased by differences in the number of representative structures available for each Pfam domain. Thus the $\frac{O}{E}$ for each residue type in each range-bin for a given structure example, needed to be weighted to adjust for the similarity of the sequence to others for that family. To achieve this the Henikoff & Henikoff weighting [64] described in Section 2.5.2 was calculated for the sequences of all the representative structures for each Pfam family being considered. This weighting was then applied to each $\frac{O}{E}$ for each residue type in each range-bin for each structural example of the Pfam family. The sum of all weighted $\frac{O}{E}$ for a given range-bin for all structures was calculated, giving a weighted $\frac{O}{E}$ for the Pfam family. This provided a representative $\frac{O}{E}$ for each residue type in each range-bin for each Pfam family. Finally an average value of all $\frac{O}{E}$ from all Pfam families was calculated to give the values reported in the results section. This final average of averages was calculated by generating the sum of all $\frac{O}{E}$ in a given range bin and dividing by the number of families that had contributed to the value in that range-bin for that residue type.

4.2.2 Bootstrapping

The purpose of the bootstrapping is to provide a comparison of the actual data with an estimate of what could have been seen by chance. In the case of the $\frac{O}{E}$ analysis this required the randomisation of the solvent exposure value assigned to each residue within each structure for each Pfam family. The solvent exposure measures within each family were randomised, using a process designed to conserve the alignment of possible solvent exposure of residues in columns of the Pfam alignment. The bootstrapping was performed with the following steps:

1. Sequences were found representing the subset of the Pfam family alignment for which structural data was available in the PiQSi database.

2. Two matrices were built, the first containing the sequences from the alignment; the second containing the solvent exposure values for the corresponding residue positions in the sequence matrix.

3. Columns were filtered to remove those which contained 33% or more gaps in the alignment of selected sequences. The threshold of gaps was an arbitrary choice, based on the assumption that columns which were a third or more gaps would not contain sufficient information to provide a reliable value towards the average solvent exposure for that location in the domain.

4. For the remaining columns that contained gaps ($< 33\%$) the average solvent exposure for a given column was calculated and assigned to the gapped positions. This was to ensure that no residue was assigned a null value of solvent exposure during the subsequent randomisation process.

5. Using the inbuilt random function in Python, columns of solvent exposure values were randomly assigned to new columns in the alignment, with replacement, i.e. the a column of solvent exposure could be assigned to more than one location.

The process of randomly assigning columns of solvent exposure to columns of residues, in the alignment was repeated 100 times to produce 100 bootstrap datasets, for each Pfam family

in the data. The $\frac{O}{E}$ analysis was then performed on each of the randomised datasets.

The random assignment of columns of solvent exposure to columns of residues in the alignment, ensured that the resulting dataset, though randomised preserved the alignment of positions between the individual structures for the domain family. Since the real data has correlations in the columns and the bootstrapping is used to investigate the sort of $\frac{O}{E}$ line that might arise by chance in the real data, then the correlations needed to be conserved in bootstrap.

Generating bootstrap data for the ASA analysis necessitated a consideration of the size of each residue type. The maximum permissible surface area of a residue-type depends on the size of the side-chain. Assigning a surface area to a residue that is in excess of its physical maximum would result in a bootstrap which does not reflect a randomisation of the physically possible surface area of each residue. Consider alanine, which has one methyl in it side chain, as an example, there is a range of physically possible values of surface area that it can have with the maximum being the area of one methyl group, whereas lysine has a 4 carbon chain with an amino group at the end and would thus have a very different range of possible values of surface area. Assigning a value of ASA that exceeds alanine's absolute maximum surface area would result in a non-zero expected value, in a range of values where alanine cannot exist. Ignoring this would result in the Expected value being incorrect and thus our $\frac{O}{E}$ being inaccurate. A method of assigning a random value of ASA to each residue type that conforms to the physically permissible range of values for that type and is still randomly assigned, was needed.

In Section 1.3 a variant of ASA called rASA (relative Accessible Surface Area) was discussed. rASA is reduction of ASA to a percentage, using some maximum ASA (to represent 100%) obtained by a theoretical reference state of an extended conformation of the tri-peptide Gly-X-Gly or Ala-X-Ala. Where X is the residue being considered. This reduction of ASA into a percentage, offers a method to address the size consideration, which was applied as follows. The ASA values in the solvent exposure matrix were first converted to rASA. The random assignment of columns of solvent exposure, with replacement, to columns of the sequence matrix was performed. Then the rASA values were converted back to ASA using the appropriate reference size for that amino acid type, as discussed below.

4.2.3 Amino acid reference states for rASA

The reference state areas derived from the tri-peptide Ala-X-Ala or Gly-X-Gly will vary depending on the conformation of the tri-peptide and the side-chain conformation in the reference state. When Lee and Richards [14] introduced the idea of a reference state peptide it was intended as an extended conformation representing the residue type in an unfolded state. The rASA measure was proposed as an estimate of the change in solvent accessible surface area a residue would experience when going from the unfolded to the conformational state and local environment in which the measure was taken. Naccess was used to calculate ASA and rAS; it uses the Ala-X-Ala tri-peptide as a reference to calculate rASA . The original journal article about the program offers little explanation as to the choice of tri-peptide conformation [4]. However the Naccess author Simon Hubbard, kindly supplied the reference structures for each of the tripeptides. The torsion angle data for each of the tri-peptides is shown in Appendix D. The ASA of the residues side chain was measured from the C_β atom onwards, for each residue-type in the reference state. These measured values were used for the conversion from ASA to rASA and back again for each side chain ASA in our data.

4.2.4 Interpolation

The question that is being addressed by this analysis is: "Is there a a value of HSEu or ASA that defines a compositional change between the surface and core of a protein?"

The simplest choice of crossover point is the point at which there is a change of value of $log_2\langle\frac{O}{E}\rangle$ from greater than 0 to less than 0, or vice versa, representing a change in preference for that amino acid type to have that value of solvent accessibility. To determine where the crossover takes place for each residue type, the $log_2\langle\frac{O}{E}\rangle$ was plotted against solvent exposure. The data points for each residue type, were used as input to a spline interpolation algorithm, built into the Numpy package for Python. The interpolation provided the root points (when $log_2\langle\frac{O}{E}\rangle = 0$) for each residue-type, providing the crossover points. Where the interpolation gave multiple crossover points, the results of the interpolation were compared with the actual graphs by eye, to determine which points were most likely to indicate a true transition from a statistical preference

state to another, as opposed to local fluctuations in the line due to unconverged sampling.

4.3 Results

Results are presented for three $\frac{O}{E}$ analyses, which were undertaken to address three questions:

(i) What is an appropriate radius to be used for HSEu?

(ii) Which measurement of solvent exposure best delimits the crossover from residue burial to solvent exposed, ASA or HSEu?

(iii) Is there a difference in the solvent exposure behaviour of amino acids in proteins between Eukaryota and Prokaryota?

Five data sets were compiled for analysis. Firstly two sets of Pfam domains were compiled using the database described in Chapter 3, one was exclusively eukaryotic and cytoplasmic, non-DNA/RNA binding and non-membrane binding, the other was exclusively prokaryotic and cytoplasmic, non-DNA/RNA binding and non-membrane binding. The eukaryotic set was used to compile a set of representative structures with four solvent exposure measurements, one for each of three radii of HSEu (10 Å, 13 Å and 16 Å) and one for ASA – to denote the radius of an HSE measurement, it will be denoted HSEu_{radius}, e.g. HSEu radius 13 will be written HSEu_{13}. The prokaryotic set was used to create a single data set of structures with HSEu_{13} measurements. As described in Section 4.2.1, the solvent exposure was measured for each residue in the biological unit taken from the PiQSi database; the domains were extracted from the quaternary structures after the solvent exposure had been recorded in the B-factor column of the PDB file. The total number of Pfam families and the structures that were used to produce the following results are shown in Table 4.1.

Table 4.1: The data selected for analysis: The total number of Pfam families and representative structures used for each analysis, for which results are presented in this chapter. The difference between the number of families and structures for the HSEu data and ASA data is a result of Naccess not being able to correctly parse some structure files, this was not a solvable problem in the available time.

Analysis	Subset of data	Pfam Families	Structures
$HSEu_{10}$	Eukaryota	143	3742
$HSEu_{16}$	Eukaryota	143	3742
$HSEu_{13}$	Eukaryota	143	3742
$HSEu_{13}$	Prokaryota	142	4306
ASA side chain	Eukaryota	139	3683

To address question (i) above, the propensity for each residue type for solvent exposure was compared between each of the three HSEu radii; this is covered in Section 4.3.1. Question (ii) was approached similarly, using a comparison between the propensity for residue types to have a given ASA and the propensity to have a given $HSEu_{13}$ count. Further analysis was done to evaluate the behaviour of residue types with respect to each other, by determining correlation coefficients between residue types, which is explained further in Section 4.3.2. Finally, question (iii) is covered in Section 4.3.3, where the propensity for solvent exposure, measured in $HSEu_{13}$, between Eukaryota and Prokaryota is considered.

Deciding if a residue is buried or exposed to solvent, is dependent on the value of the given solvent exposure measure. A low value of HSEu implies fewer neighbours and therefore closer proximity to the protein surface; while for ASA the larger the number the greater the surface area and hence the greater the solvent exposure. The boundary between the surface and the protein interior, is likely to be the crossover from $log_2\langle\frac{O}{E}\rangle > 0$ to $log_2\langle\frac{O}{E}\rangle < 0$ (and vice versa), i.e. when $log_2\langle\frac{O}{E}\rangle = 0$. These points were determined using a spline interpolation function built into the Numpy package for Python. The $log_2\langle\frac{O}{E}\rangle$ data used to produce the plots of $log_2\langle\frac{O}{E}\rangle$ vs. solvent exposure, presented in this section, was used as input for the spline function, providing the crossover points for each analysis. For ASA the range-bins were chosen to be of width 10 $Å^2$, and for HSEu the range-bins were chosen to be of width 4 HSEu counts.

The figures shown in the following subsections include comparisons between actual data and bootstrap data. A convention for the figures containing bootstrap data was adopted; the trend line for the actual data is shown as a red line with points and error bars; the average of 100 bootstrap analyses is shown as a purple line. Plots showing all of the 100 bootstrap lines, have the trend line for the actual data drawn over by the bootstrap lines. This was done to highlight where the actual data deviated significantly from random. The plots comparing the actual data with the bootstrap data, show plots of the average bootstrap data compared to the actual data, in the top part of the figures; the comparison between the actual data and the individual bootstrap data are shown in the lower part of the figures. The reasoning for showing both sets of plots was to demonstrate bootstrap lines average approximately around $log_2\langle\frac{O}{E}\rangle = 0$. But the average bootstrap lines don't show clearly when the sample size in a range bin is too small to be useful. Therefore to show the variance of the random data, which is a reflection of the amount of data for each residue type in each range-bin, the complete set of bootstrap lines is shown. Additionally, showing all the bootstrap lines makes it possible to check if any single bootstrap line closely resembles the line for the actual data. The significance of the entire trend line for the data can be interpreted on the basis of the number of equivalent random bootstrap lines. However the significance of each point in the trend line can be questioned based on the amount of variance in the bootstrap data. i.e. if there is a significant amount of "fanning" in the bootstrap data near a point, this is an indication that the data in that region is sparse and prone to sampling errors.

4.3.1 Comparison of $\frac{O}{E}$ analysis of HSEu using different sphere radii, 10 Å, 13 Å and 16 Å

The objective of this analysis was to determine which was an appropriate radius for HSEu. Only four representative figures are shown here, the plots for the other residue types can be found in Appendix E. Figures 4.1 – 4.4, show $log_2\langle\frac{O}{E}\rangle$ vs. range-bins for a selection of four residue types (Arg, Cys, Ile, Trp). Each plot contains 6 subplots; (a) – (c) show $log_2\langle\frac{O}{E}\rangle$ vs range-bins with the average trend line for 100 bootstrap analysis, for $HSEu_{10}$, $HSEu_{13}$ and $HSEu_{16}$ respectively; (d)

– (f) show $log_2\langle\frac{O}{E}\rangle$ vs range-bins with the individual trend line and 100 bootstrap analysis lines for $HSEu_{10}$, $HSEu_{13}$ and $HSEu_{16}$ respectively. The crossover in $log_2\langle\frac{O}{E}\rangle$ propensity between buried and exposed, for all residue-types, is summarised for each of the HSEu radii in Table 4.2.

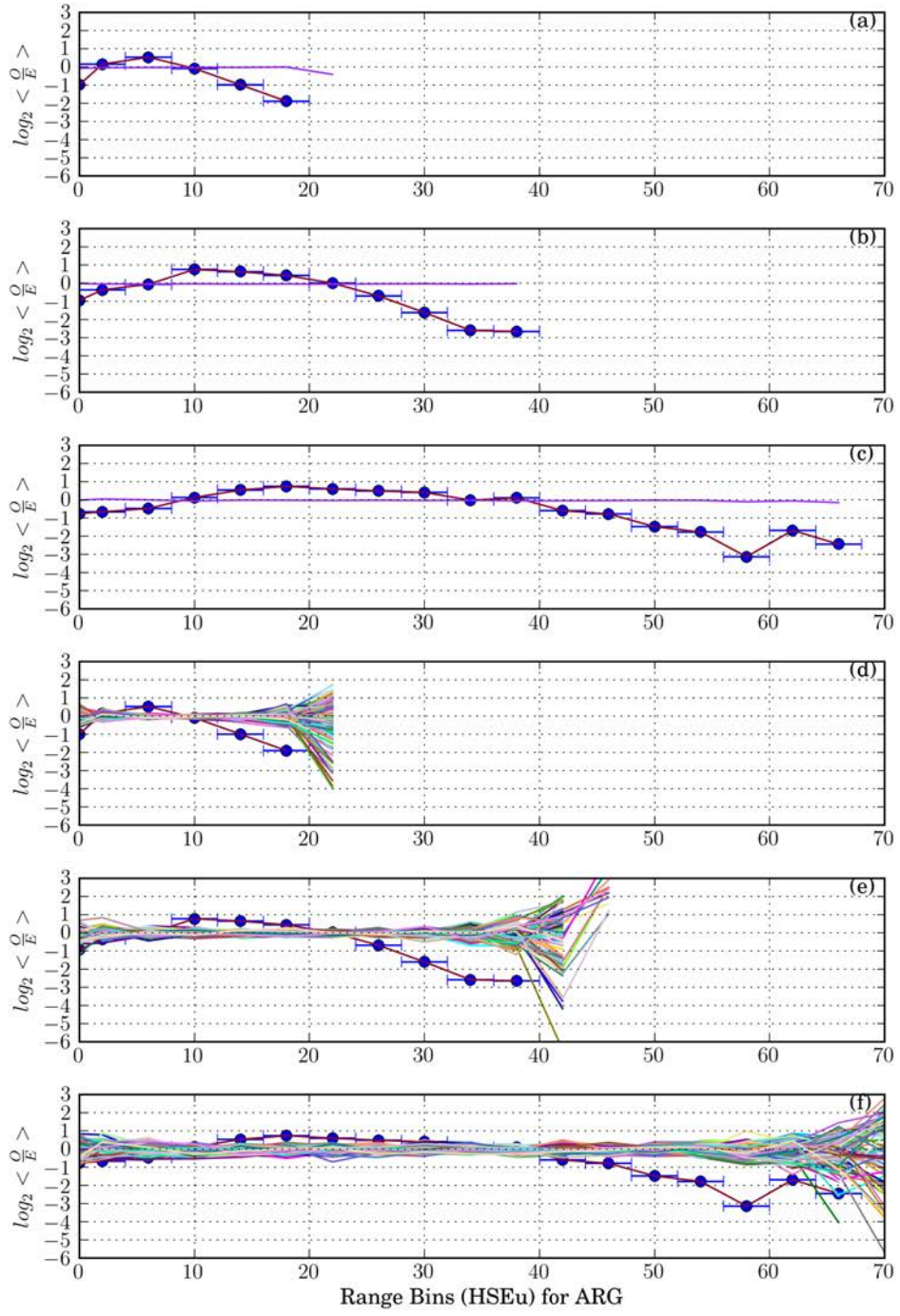

Figure 4.1: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Arg: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with $HSEu_{10}$, (b) with $HSEu_{13}$, (c) with $HSEu_{16}$. The individual bootstrap lines are shown in (d) with $HSEu_{10}$, (e) with $HSEu_{13}$ and (f) with $HSEu_{16}$.

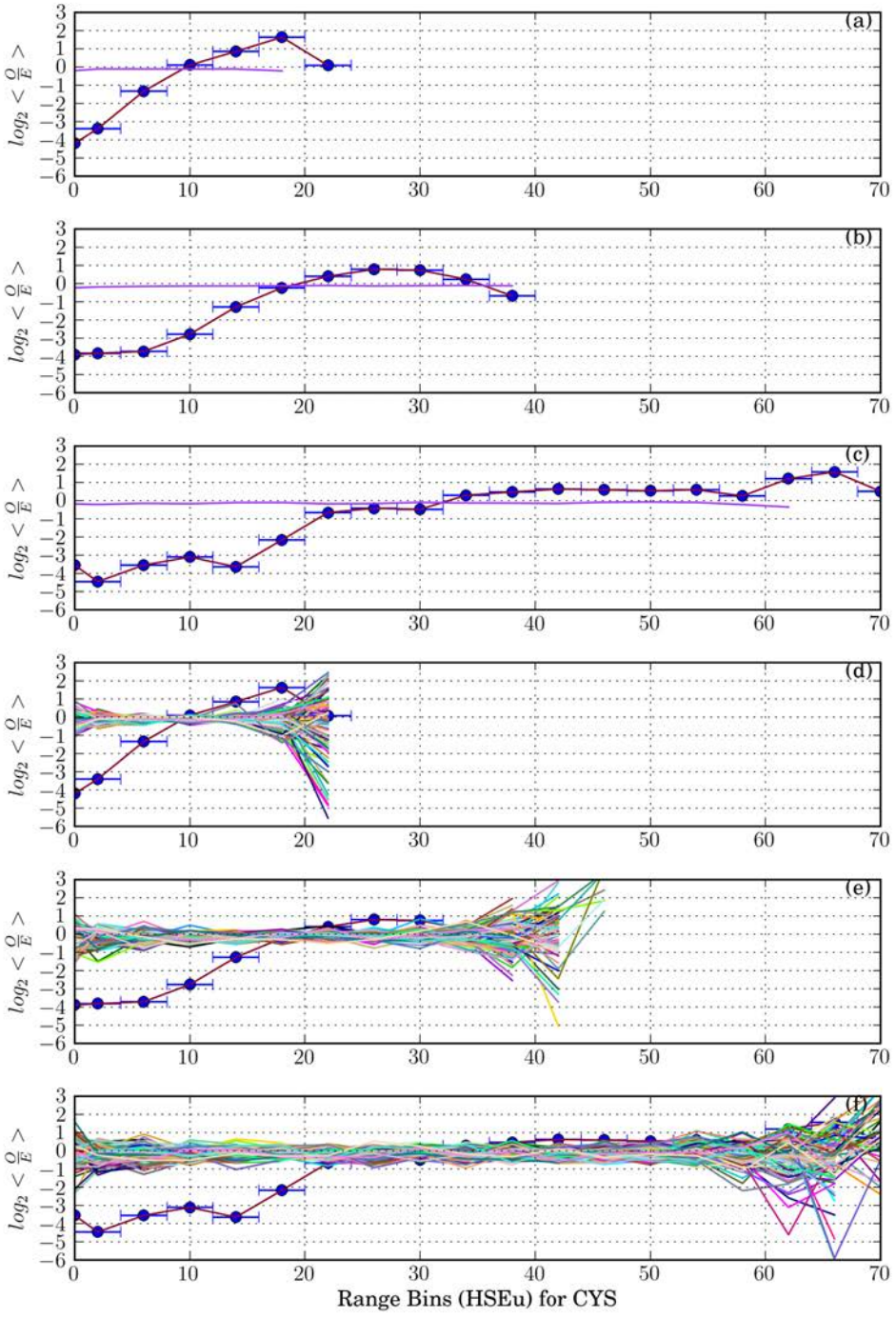

Figure 4.2: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Cys: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with $HSEu_{10}$, (b) with $HSEu_{13}$, (c) with $HSEu_{16}$. The individual bootstrap lines are shown in (d) with $HSEu_{10}$, (e) with $HSEu_{13}$ and (f) with $HSEu_{16}$.

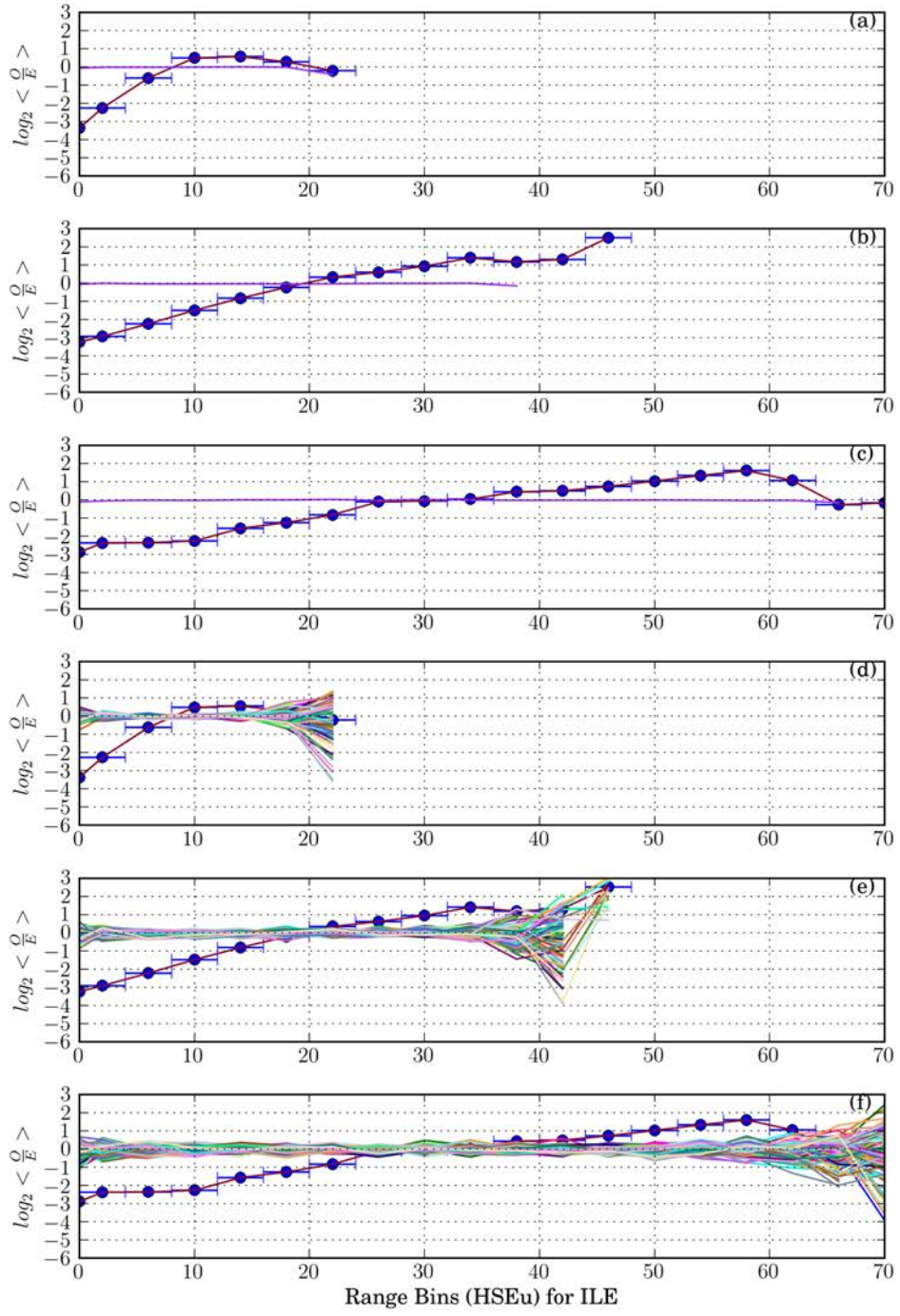

Figure 4.3: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Ile: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with $HSEu_{10}$, (b) with $HSEu_{13}$, (c) with $HSEu_{16}$. The individual bootstrap lines are shown in (d) with $HSEu_{10}$, (e) with $HSEu_{13}$ and (f) with $HSEu_{16}$.

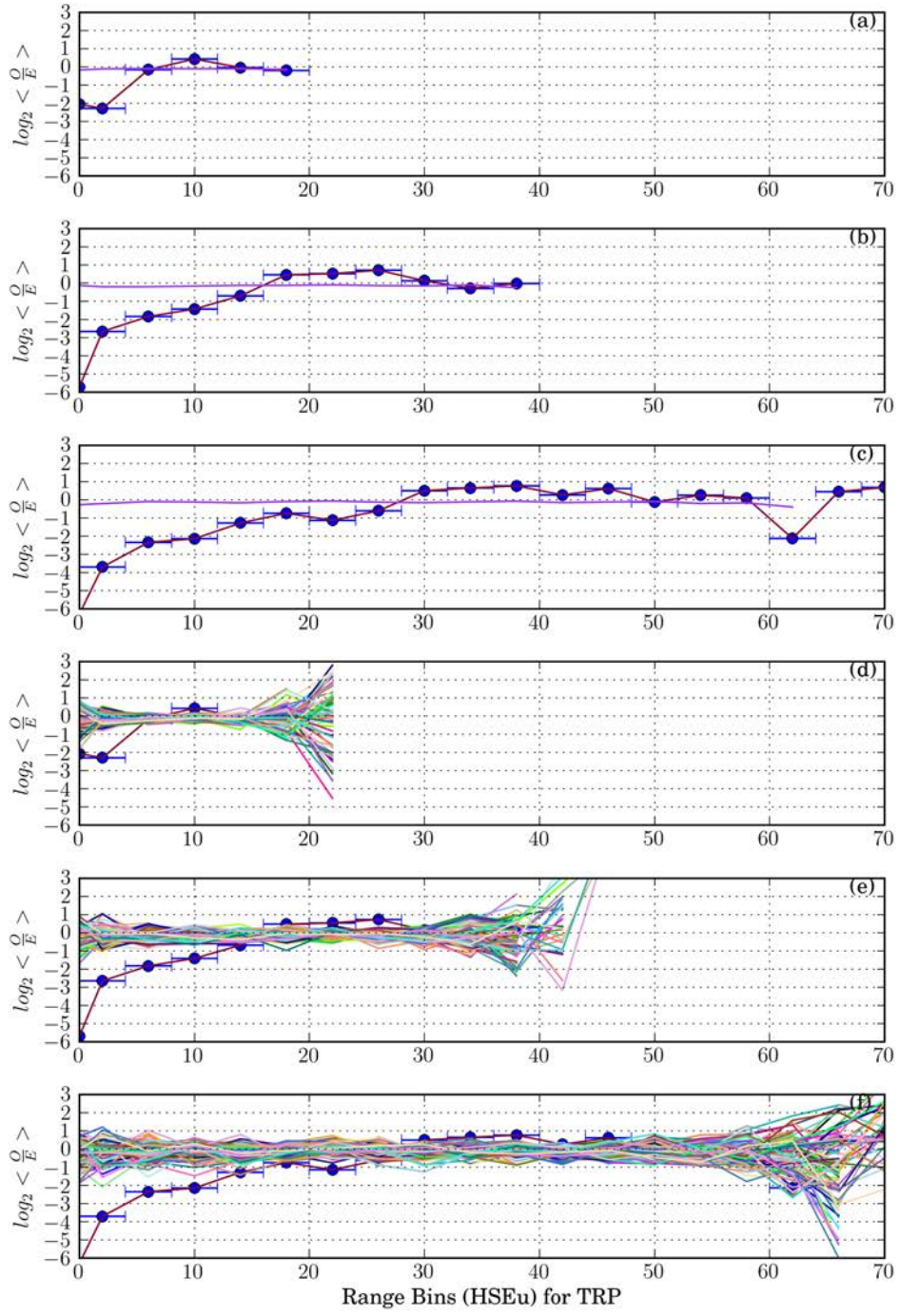

Figure 4.4: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Trp: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with $HSEu_{10}$, (b) with $HSEu_{13}$, (c) with $HSEu_{16}$. The individual bootstrap lines are shown in (d) with $HSEu_{10}$, (e) with $HSEu_{13}$ and (f) with $HSEu_{16}$.

Table 4.2: Crossover points for each amino acid type for three different HSEu radius: $HSEu_{10}$, $HSEu_{13}$, $HSEu_{16}$ shown. The points were determined using a spline interpolation function in the Numpy package for Python. The roots of the interpolated line, when $log_2\langle\frac{O}{E}\rangle = 0$, are crossover points. They are indicative of a change between +/- values which is a change from over represented by chance to under represented by chance. These values are used to infer the transition from solvent exposed to buried. The range-bin width is the same in all three analyses, 4 HSEu.

Residue	$HSEu_{10}$	$HSEu_{13}$	$HSEu_{16}$
ALA	10.6	17.8	21.6
ARG	9.6	22.1	33.4
ASN	8.4	16.5	26.1
ASP	8.1	17.1	28.2
CYS	9.6	19.2	32.5
GLN	8.7	19.2	32.6
GLU	8.3	17.9	29.9
GLY	4.3	10.2	18.0
HIS	5.8	13.7	25.4
ILE	7.8	19.5	33.6
LEU	8.0	19.9	32.7
LYS	8.7	18.7	32.0
MET	8.8	20.7	34.3
PHE	7.7	18.6	28.4
PRO	5.4	16.9	25.7
SER	6.0	13.8	23.5
THR	8.2	16.0	27.0
TRP	6.4	16.1	28.0
TYR	7.4	15.4	27.6
VAL	8.3	18.1	30.7
Mean	7.8	17.4	28.6
Stnd Dev	1.6	2.73	4.3

4.3.2 Comparison of $HSEu_{13}$ and Side-Chain ASA $\frac{O}{E}$

The objective of this analysis was to compare the performance of $HSEu_{13}$ and ASA for delimiting residue burial. These two analyses were conducted on a set of Pfam domain structures from Eukaryota only.

To investigate the correlation behaviour between ASA and $HSEu_{13}$, scatter plots of $HSEu_{13}$ vs. ASA were generated and a linear regression analysis was done on the data. Only the side-chain ASA was considered, measured from the C_β atom and above, as HSEu is only concerned with the direction of the side-chain. A selection of these scatter plots of $HSEu_{13}$ vs. ASA for four residue-types are shown in Figures 4.5 – 4.8; the scatter plots for the remaining residue types are in Appendix F. The linear regression method built into the statistics package for Scipy (the scientific programming module for Python) was used to perform the analysis. The scatter plots were generated by taking the two sets of representative structure files for each Pfam family, those with $HSEu_{13}$ data and those with side chain ASA data stored in the B-factor column. The $HSEu_{13}$ and ASA data was harvested from every pair of copies for each domain structure, providing the data points for all residue types. The scatter plots represent the raw data used in the analysis, with no weighting. The output from the linear regression analysis for each residue-type is shown in Table 4.3 , the table is shown after the figures for this subsection. The linear regression analysis was repeated using all structures from the PiQSi, no plots are shown of that analysis, but the results are included in Table 4.3. This PiQSi data included some NMR structures which resulted in unusually high $HSEu_{13}$ values, as a result of the entire ensemble being treated as a single structure; however this did not seem to have a significant effect on the results.

The scale of the scatter plots is consistent for all twenty residue types. The x-axis was restricted to a maximum of 190, which truncated the plots for a few of the larger residue-types. This was done to keep the scale consistent across all plots so a comparison could easily be made. The smaller residue types, such as alanine and serine have no points above 90 $Å^2$, while argenine and glutamic acid go well over 200 $Å^2$. The higher data points are not of direct interest in the comparison and so the scale was truncated. The red line through the data points represent

the best fit line determined using a linear regression analysis.

To provide a comparison of the behaviour of the bootstrap data for $HSEu_{13}$ with side chain ASA, Figures 4.9 – 4.12 are shown. The plots for the remaining residue types are shown in Appendix G. The average bootstrap lines for both the $HSEu_{13}$ and the ASA $\frac{O}{E}$ analysis are mostly consistent around $log_2\langle\frac{O}{E}\rangle$ =0. However, the variation of the bootstrap lines is less in $HSEu_{13}$ plots compared with ASA plots. In the scatter plots the y-intercept of the linear regression analysis was 26 $HSEu_{13}$ (+/- 2) for all residue types. Therefore ASA values greater than 0 $Å^2$ will mostly correspond to $HSEu_{13} > 28$.

Figures 4.13, 4.14 and 4.15 compares $log_2\langle\frac{O}{E}\rangle$ vs. $HSEu_{13}$ with $log_2\langle\frac{O}{E}\rangle$ vs. side chain ASA. In this set of plots, the residue types have been grouped together on the basis of their physico-chemical properties. The groupings are: aliphatic, charged negative, charged positive, polar uncharged and special cases. The comparative results from the analysis, including the crossover points for each residue type, is summarised in Table 4.4.

The results from the linear regression analysis of the scatter plots were then combined with the points where $HSEu_{13}$ crossover $log_2\langle\frac{O}{E}\rangle = 0$, to try to predict the ASA crossover points from the $HSEu_{13}$ crossover points. The predicted ASA cross over points, using the results from the linear regression analysis of both the selected data and the full PiQSi are shown in Table 4.5.

To determine how the preference for solvent exposure varies between residues of different physico-chemical properties, a Pearson's correlation coefficient was calculated for the $log_2\langle\frac{O}{E}\rangle$ of all residue pairs. To assess how the preference for solvent exposure compared with the substitution behaviour of amino acids, a Pearson's correlation coefficient was calculated between the results of the comparison between residue solvent exposure preference and several popular substitution matrices, e.g. PAM30 and Blosum60. The results of the second analysis are summarised in Table 4.6.

N.B. Comparison of HSEu with ASA data can be confusing. With the HSEu scale, low values are indicative of a sparsely populated neighbourhood in the direction of the side-chain and as such indicate proximity to the surface. High values of HSEu are indicative of a densely populated neighbourhood and as such represent burial of the side chain. The ASA scale reads

in the opposite sense. A high value of ASA represents a large surface area and suggests a large solvent exposure. A low value represents a small (or no) surface area and indicates limited solvent exposure. It is important to be mindful of this difference.

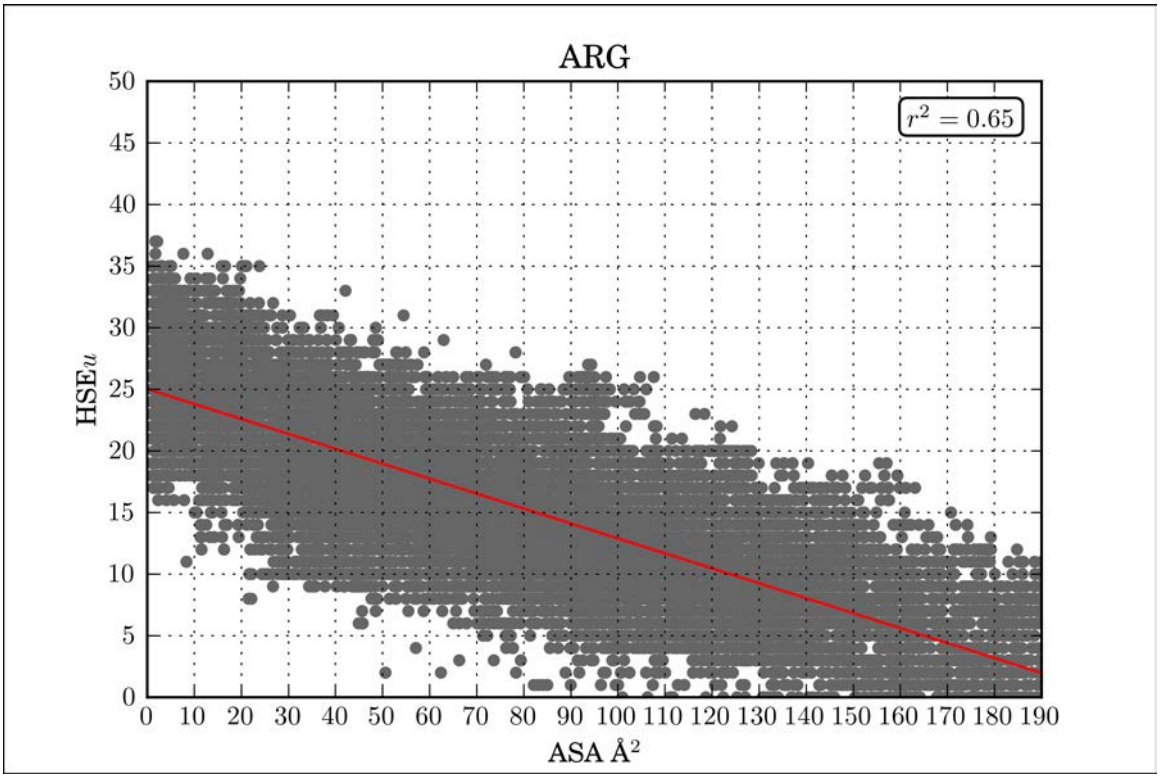

Figure 4.5: Scatter plot of $HSEu_{13}$ vs Side chain ASA, for Argenine: Each point represents a single instance of the residue-type, for which both $HSEu_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between $HSEu_{13}$ and ASA for this residue type.

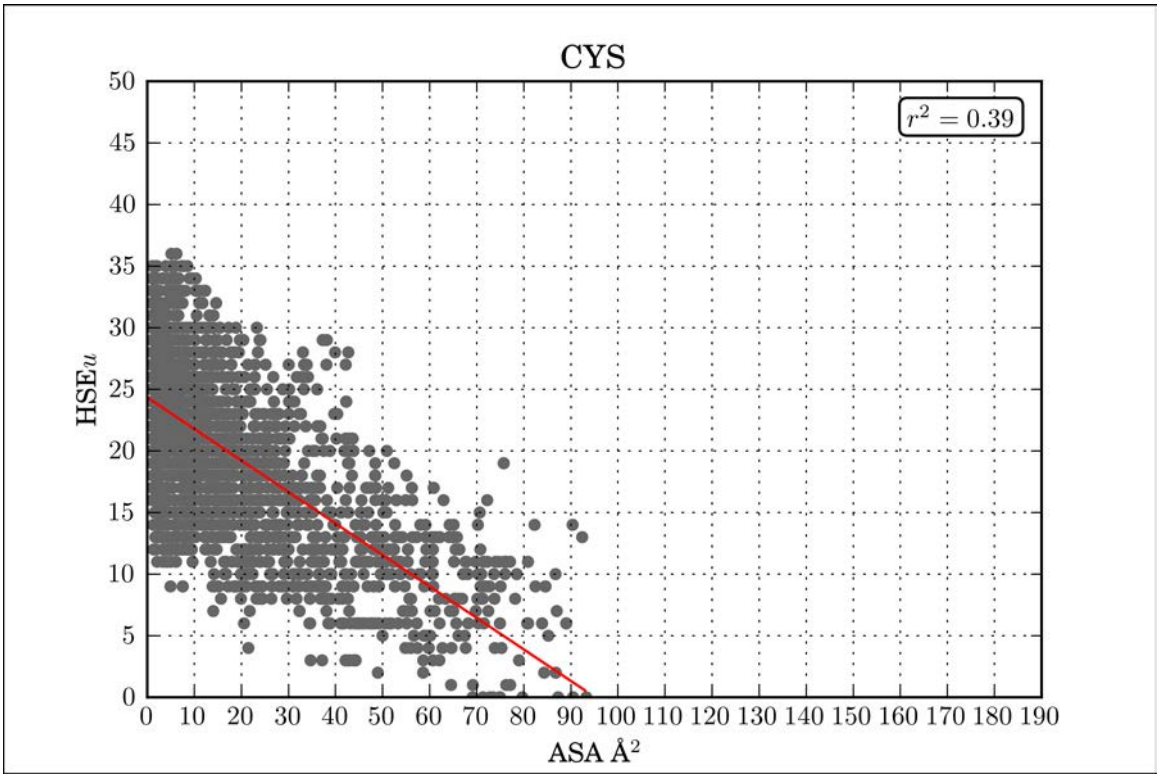

Figure 4.6: Scatter plot of $HSEu_{13}$ vs Side chain ASA, Cysteine: Each point represents a single instance of the residue-type, for which both $HSEu_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between $HSEu_{13}$ and ASA for this residue type.

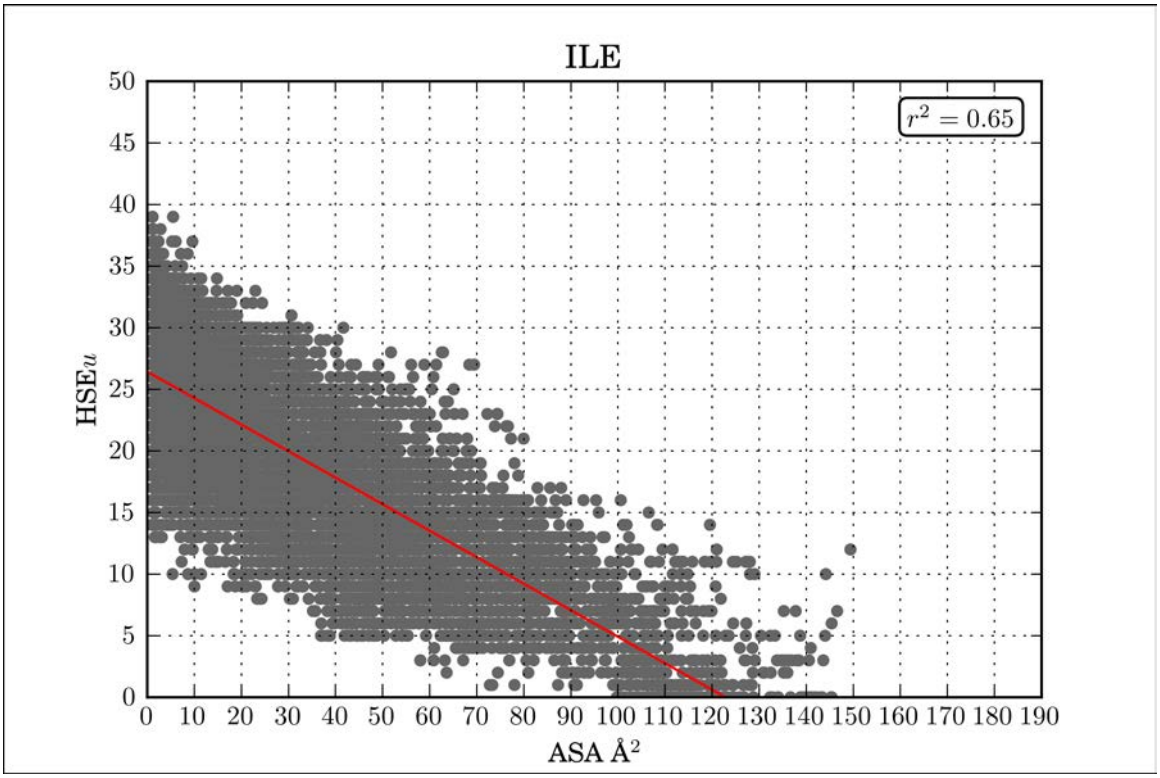

Figure 4.7: Scatter plot of $HSEu_{13}$ vs Side chain ASA, Isoleucine: Each point has represents a single instance of the residue-type, in the same protein structure for which both $HSEu_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between $HSEu_{13}$ and ASA for this residue type.

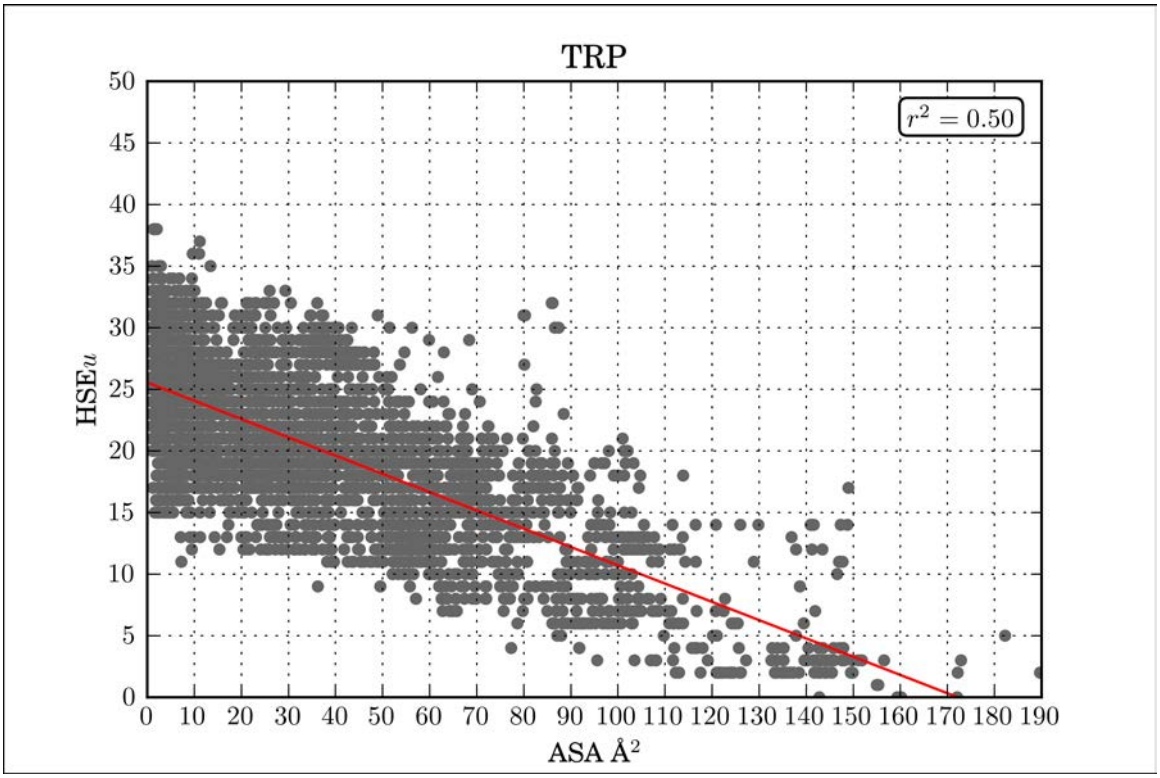

Figure 4.8: Scatter plot of HSEu$_{13}$ vs Side chain ASA, Tryptophan: Each point represents a single instance of the residue-type, for which both HSEu$_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between HSEu$_{13}$ and ASA for this residue type.

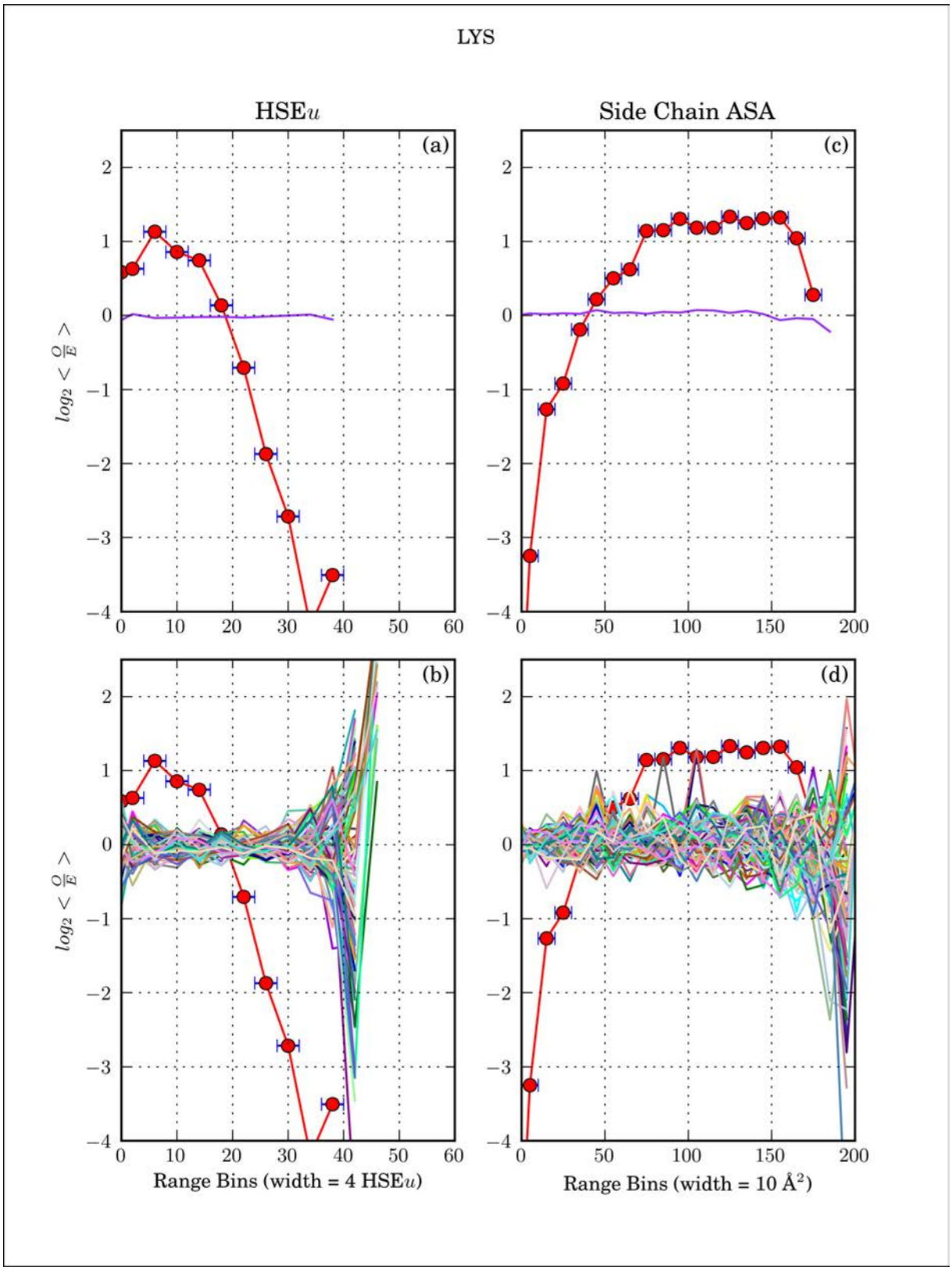

Figure 4.9: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for Lys: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

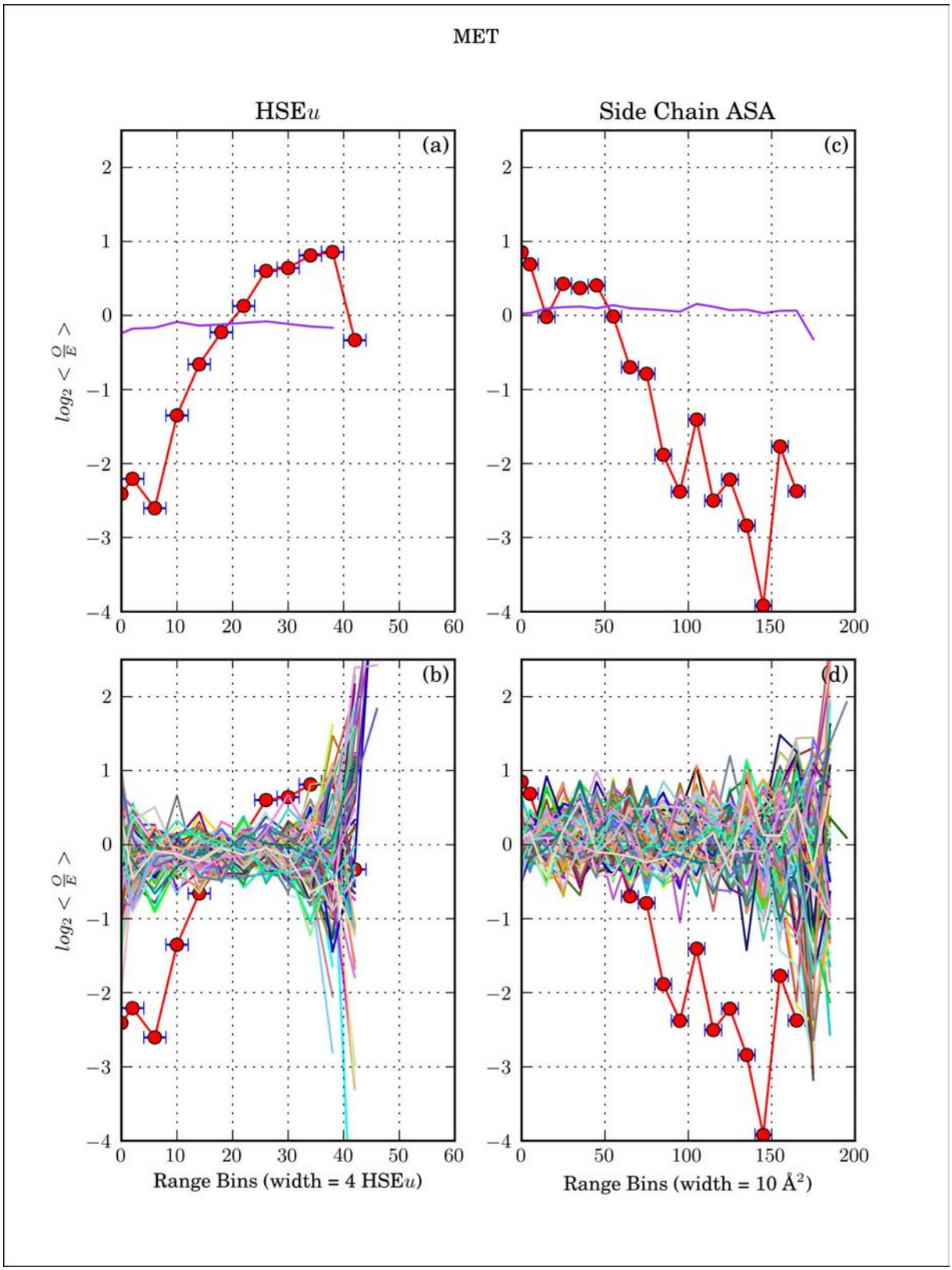

Figure 4.10: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for Met: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

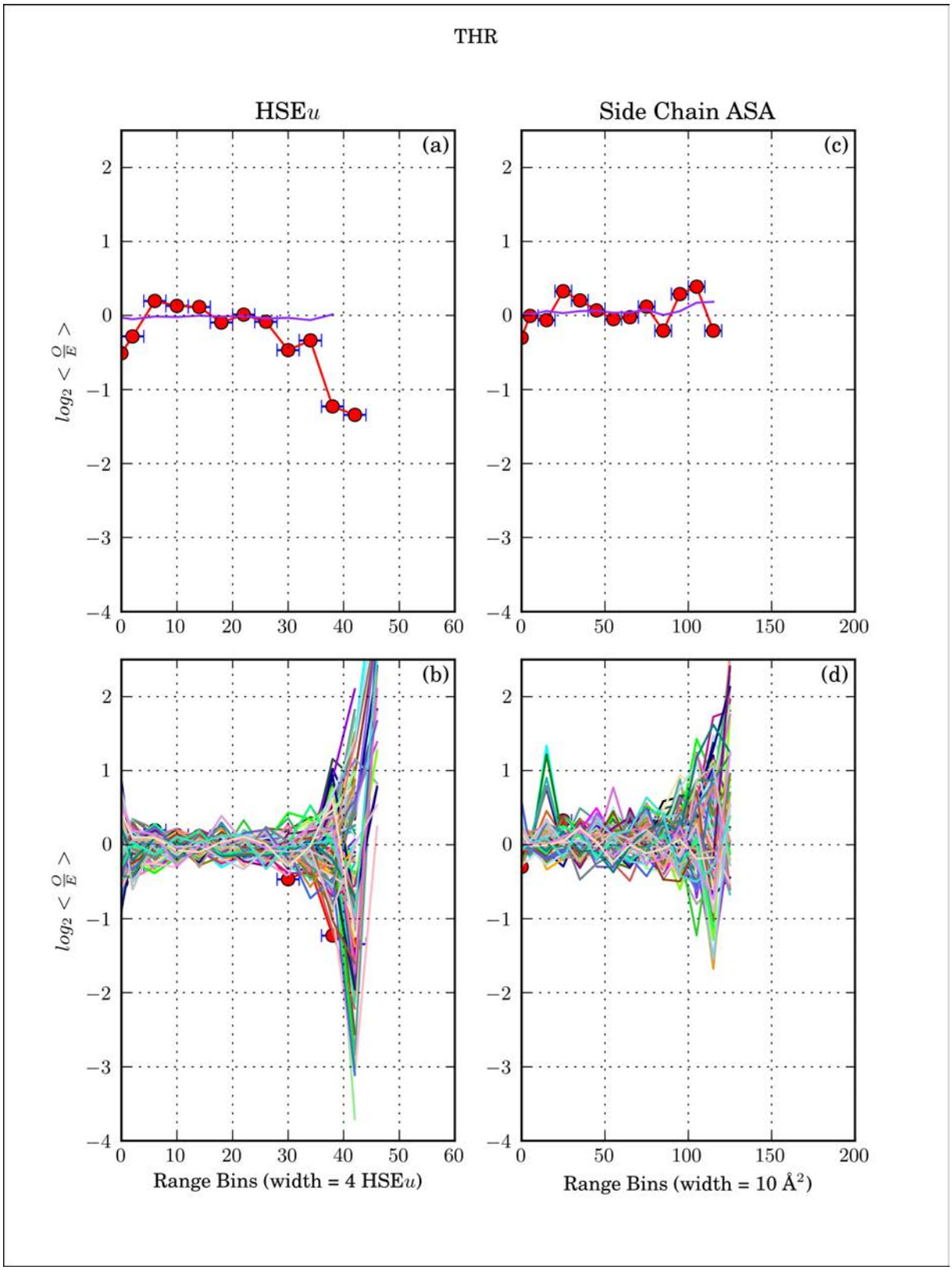

Figure 4.11: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for Thr: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

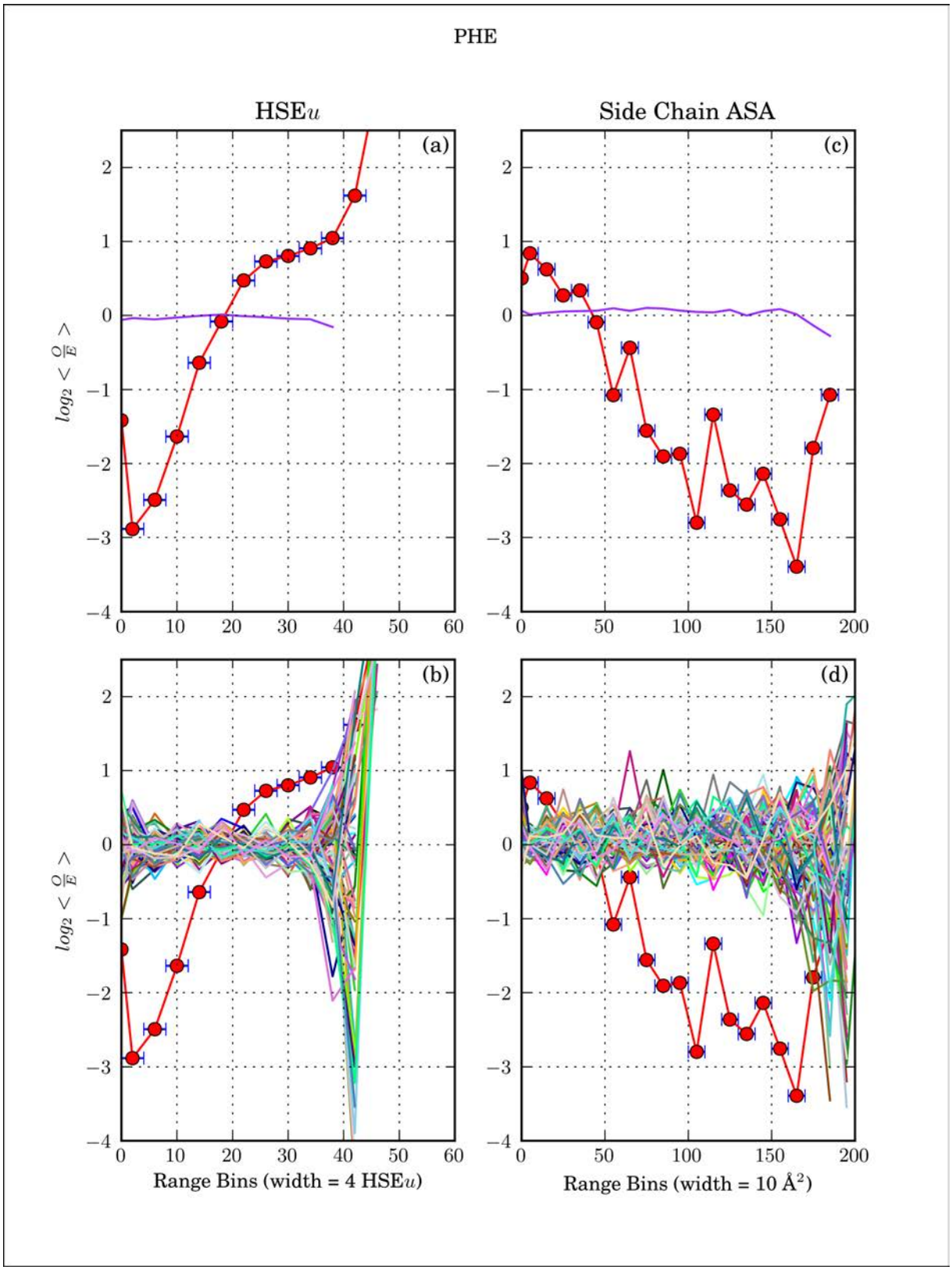

Figure 4.12: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for Phe: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

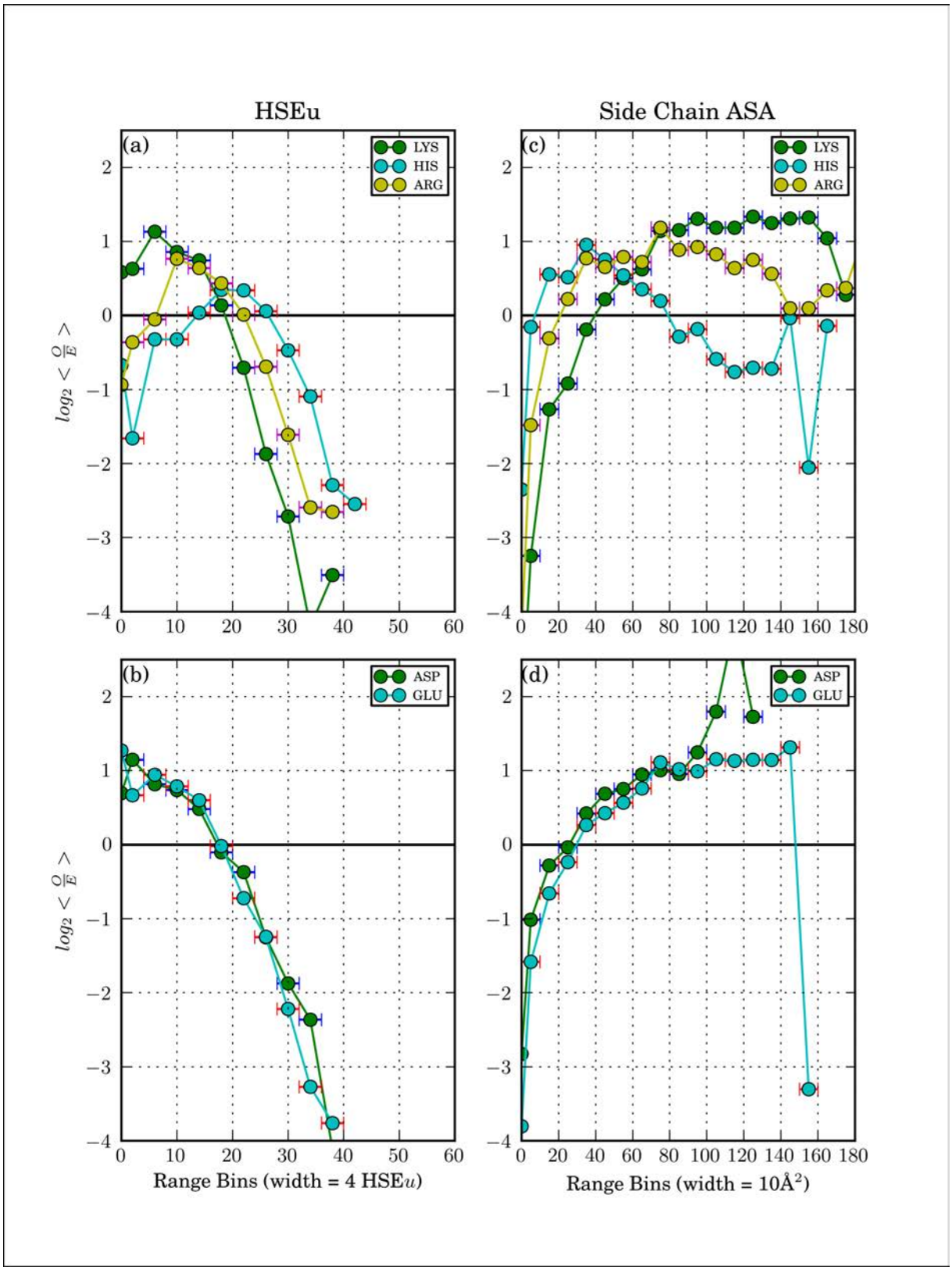

Figure 4.13: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu$_{13}$ and $log_2\langle\frac{O}{E}\rangle$ vs. ASA for eukaryotic charged residues: (a) HSEu$_{13}$ positively charged residues, (b) HSEu$_{13}$ negatively charged residues, (c) Side Chain ASA positively charged residues, (d) Side Chain ASA negatively charged residues

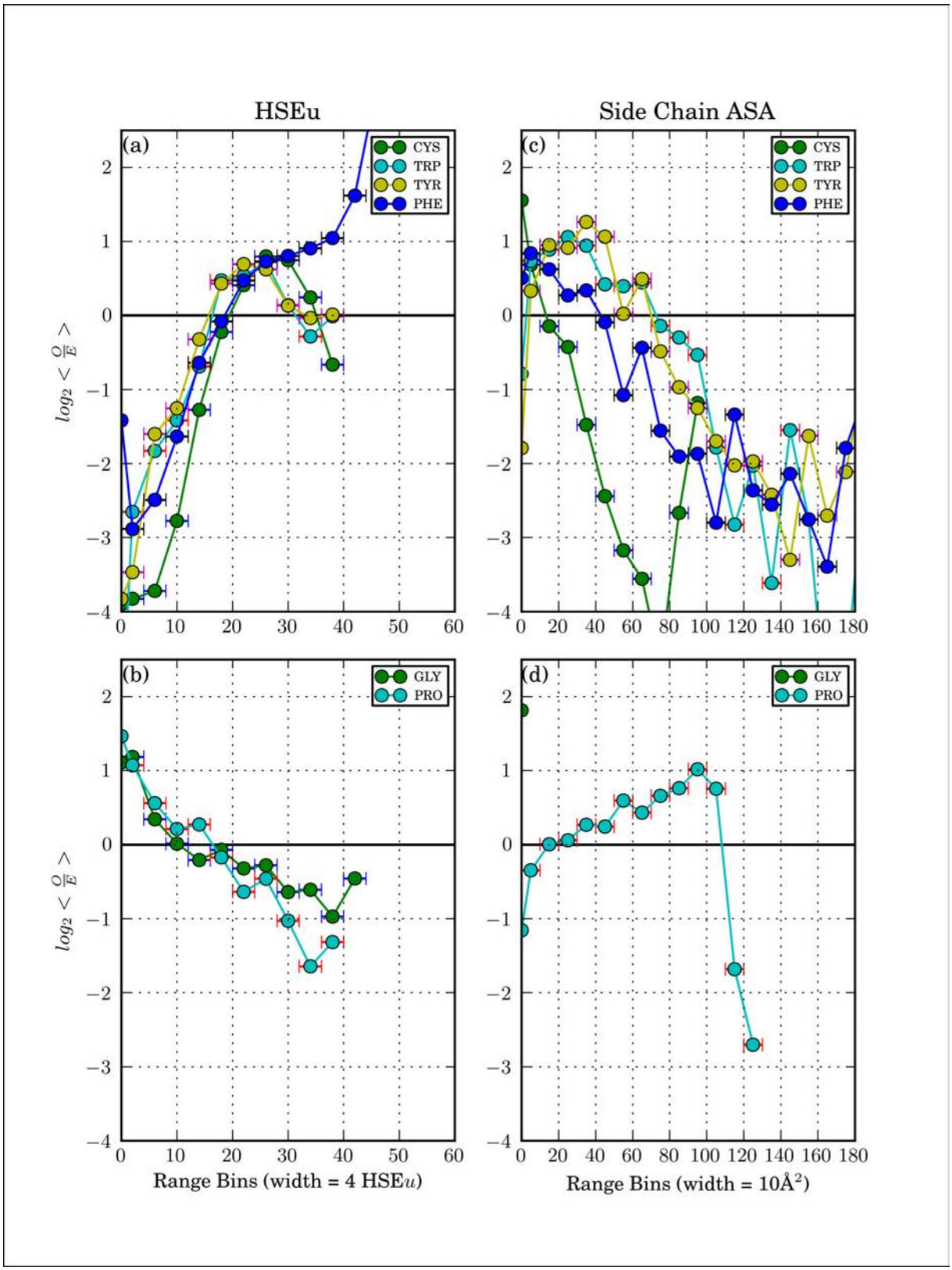

Figure 4.14: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu$_{13}$ and $log_2\langle\frac{O}{E}\rangle$ vs. ASA for aromatic residues and "special cases:" (a) HSEu$_{13}$ aromatic residues, with CYS, (b)HSEu$_{13}$ special case residues, (c) Side Chain ASA aromatic residues, with CYS, (d) Side Chain ASA, special case residues, GLY has no side chain.

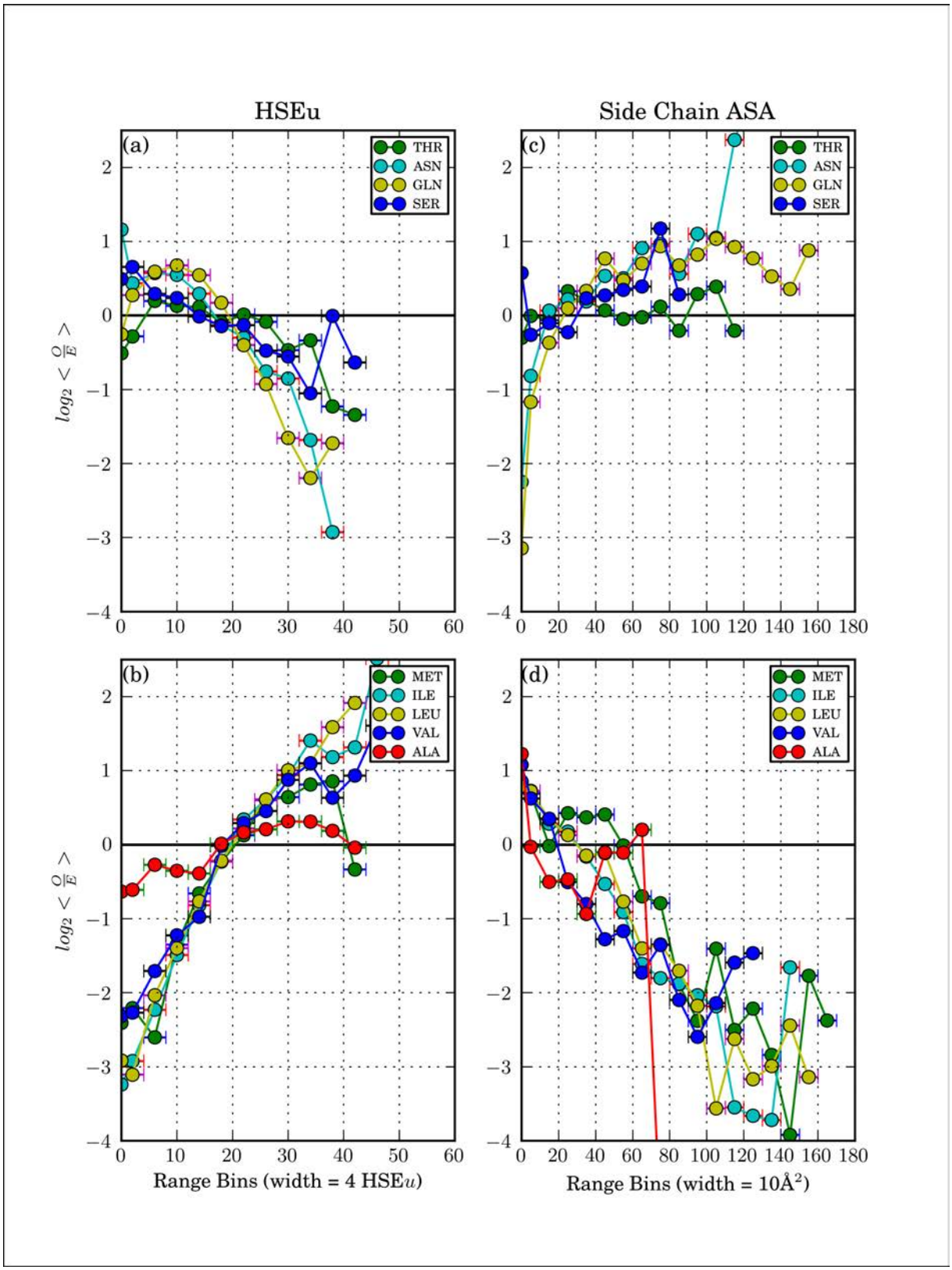

Figure 4.15: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu$_{13}$ and $log_2\langle\frac{O}{E}\rangle$ vs. ASA for uncharged and hydrophobic residues: (a)HSEu$_{13}$ polar uncharged residues, (b) HSEu$_{13}$ aliphatic residues, (c) Side Chain ASA polar uncharged residues, (d) Side Chain ASA aliphatic residues.

Table 4.3: Comparison of linear regression analysis of scatter plots of $HSEu_{13}$ vs. ASA : Shown here are the results of linear regression analyses performed on two sets of data. The "Selected Data" columns contain the results from the analysis performed on the structural data which was selected for the $\frac{O}{E}$ analysis, Figures 4.5 – 4.8. The "All Data" columns contain the result for the analysis performed on the entire PiQSi database, figures not shown. There is an issue with this latter data set, as it contains NMR data, where the whole ensemble is being analysed as one structure leading to unrealistically high $HSEu_{13}$values. However please note that the difference between the two results is very slight.

Residue	Selected Data: Slope	Selected Data: y-intercept	Selected Data: r-value	All Data: Slope	All Data: y-intercept	All Data: r-value
ALA	-0.41	26.38	-0.82	-0.42	26.36	-0.84
ARG	-0.12	25.03	-0.81	-0.13	25.35	-0.81
ASN	-0.21	24.24	-0.81	-0.22	24.7	-0.84
ASP	-0.23	23.86	-0.82	-0.23	24.12	-0.83
CYS	-0.32	26.51	-0.63	-0.31	26.32	-0.65
GLU	-0.18	23.93	-0.84	-0.18	24.3	-0.84
GLN	-0.18	25.16	-0.84	-0.18	25.59	-0.84
HIS	-0.19	26.14	-0.83	-0.19	26.52	-0.82
ILE	-0.24	27.97	-0.80	-0.24	27.69	-0.77
LEU	-0.22	27.50	-0.78	-0.22	27.49	-0.76
LYS	-0.14	24.89	-0.79	-0.15	24.95	-0.8
MET	-0.19	27.21	-0.78	-0.2	27.31	-0.79
PHE	-0.19	26.83	-0.71	-0.18	26.68	-0.72
PRO	-0.25	24.71	-0.89	-0.26	25.1	-0.87
SER	-0.36	25.18	-0.84	-0.35	25.16	-0.85
THR	-0.28	25.59	-0.82	-0.27	25.71	-0.83
TRP	-0.16	26.25	-0.73	-0.14	25.83	-0.71
TYR	-0.16	26.62	-0.74	-0.16	26.32	-0.75
VAL	-0.28	27.17	-0.77	-0.27	27.17	-0.79

Table 4.4: Comparison of the crossover points for $HSEu_{13}$ with ASA crossover and crossover point converted to rASA equivalent of the ASA crossover point: Some residue-types have more than one cross-over point in the $HSEu_{13}$ and ASA plots. The ASA points have been converted to rASA, to show the variation in crossover points for rASA. In the last column on the right are the maximum observed side-chain ASA for residue-type in the entire data set. In some cases 2 crossover points were seen in the solvent exposure data, which was present in one measure but not in the other, these points have been highlighted using "–" in the table. A value in brackets for Arg, was taken from visual inspection of the graph, where the trend line appears to be very close to crossing over, similarly for Thr.

	$HSEu_{13}$ Crossover		Side Chain ASA Crossover in Å^2		rASA Crossover in %		Maximum Observed ASA in Å^2
ALA	–	17.84	59.85	7.73	84.87	10.96	76.98
ARG	6.33	22.07	24.02	(143.0)	11.95		268.52
ASN		16.55	16.71		15.86		132.03
ASP		17.11	28.97		28.27		132.03
CYS		19.22	15.35		15.89		109.06
GLN	0.79	19.25	25.32	–	18.00	–	159.77
GLU		17.91	32.41		24.08		232.79
HIS	13.72	26.49	81.72	9.03	55.91	6.18	225.94
ILE		19.46	34.02		24.72		159.36
LEU		19.88	31.91		22.61		167.05
LYS		18.74	41.5		25.30		180.18
MET	–	20.74	57.82	19.72	36.93	12.59	171.87
PHE		18.6	47.07		28.76		190.91
PRO		16.88	17.47		14.50		134.55
SER	–	13.82	33.3	3.49	42.76	4.48	86.26
THR	3.97	16.04	20.07	(42.0)	19.71	–	120.67
TRP	–	16.15	75.54	2.84	35.95	1.35	229.7
TYR	–	15.37	73.97	6.09	42.08	3.46	202.75
VAL		18.08	22.16		19.56		133.41
GLY	10.2						

Table 4.5: Predicted Solvent Accessible Surface Area Crossover Points: The y-intercept and slope from the linear regression analyses of the scatter plot data, were used to estimate the ASA crossover point for each residue type, using the equation $x = \frac{(y-c)}{m}$. The linear regression analysis was performed on the $HSEu_{13}$ and ASA data for only the structures in the data set selected for the $\frac{O}{E}$ analysis, and on all the structures in the full PiQSi database. These two analyses produced slightly different results. The prediction of the ASA crossover point was done using both sets of results.

	Predicted ASA in Å^2 *Selected Data*		Predicted rASA in % *Selected Data*		Predicted ASA in Å^2 *Full PiQSi*		Predicted rASA in % *Full PiQSi*	
ALA	–	20.83	–	29.54	–	20.29	–	28.77
ARG	155.84	24.65	77.50	12.26	146.31	25.22	72.76	12.54
ASN	–	36.64	–	34.77	–	37.06	–	35.16
ASP	–	29.36	–	28.65	–	30.49	–	29.75
CYS	–	22.79	–	23.58	–	22.92	–	23.72
GLN	128.56	26	91.39	18.48	130.61	28.06	92.85	19.95
GLU	–	40.27	–	29.92	–	42.66	–	31.69
HIS	65.39	-1.86	44.74	-1.27	67.39	0.14	46.10	0.10
ILE	–	35.46	–	25.77	–	34.29	–	24.92
LEU	–	34.64	–	24.54	–	34.59	–	24.51
LYS	–	43.91	–	26.77	–	41.38	–	25.23
MET	–	34.06	–	21.75	–	32.86	–	20.99
PHE	–	43.34	–	26.48	–	44.91	–	27.44
PRO	–	31.32	–	25.99	–	31.62	–	26.24
SER	–	31.57	–	40.54	–	32.41	–	41.62
THR	77.21	34.1	75.82	33.49	80.52	35.8	79.07	35.16
TRP	–	63.14	–	30.05	–	69.16	–	32.92
TYR	–	70.32	–	40.00	–	68.44	–	38.93
VAL	–	32.48	–	28.67	–	33.68	–	29.73

Table 4.6: Solvent exposure inter-residue correlation, correlation with substitution matrices: The $log_2\langle\frac{O}{E}\rangle$ data for each residue type was compared with every other residue type, using a Pearson's correlation analysis. These correlation coefficients were compared in a second Pearson's correlation analysis, against a set of popular substitution matrices [1–3], the results of the second analysis are shown here.

Substitution Matrix	$HSEu_{13}$ r-value	$HSEu_{13}$ r^2	ASA r-value	ASA-r^2
Blosum30	0.33	0.11	0.37	0.13
Blosum35	0.41	0.17	0.47	0.22
Blosum40	0.48	0.23	0.55	0.30
Blosum45	0.56	0.31	0.60	0.36
Blosum50	0.58	0.34	0.60	0.36
Blosum55	0.58	0.34	0.60	0.36
Blosum60	0.60	0.35	0.59	0.35
Blosum62	0.60	0.36	0.58	0.34
Blosum65	0.61	0.37	0.60	0.36
Blosum70	0.59	0.35	0.60	0.37
Blosum75	0.59	0.34	0.60	0.36
Blosum80	0.58	0.34	0.58	0.33
Blosum85	0.57	0.32	0.58	0.33
Blosum90	0.55	0.30	0.56	0.32
Blosum95	0.56	0.32	0.56	0.32
Blosum100	0.55	0.30	0.56	0.32
Gonnet	0.65	0.42	0.57	0.33
PAM30	0.34	0.11	0.34	0.11
PAM60	0.36	0.13	0.35	0.12
PAM90	0.38	0.14	0.38	0.14
PAM120	0.37	0.13	0.38	0.15
PAM180	0.40	0.16	0.40	0.16
PAM250	0.43	0.19	0.42	0.18
PAM300	0.42	0.18	0.42	0.18

4.3.3 Comparison of Eukaryota and Prokaryota HSEu $\frac{O}{E}$

Finally I present a comparison of $HSEu_{13}$ results for Eukaryota with those of Prokaryota, with the objective of determining if the $HSEu_{13}$ distribution for amino acids types differed between the two taxonomies. A complete set of plots comparing $log_2\langle\frac{O}{E}\rangle$ vs $HSEu_{13}$Eukaryota with Prokaryota are shown in Figures 4.16 4.17 4.18. A summary of the crossover points between both data sets is shown in Table 4.7, which has been presented as two histograms in Figures 4.19 and 4.20.

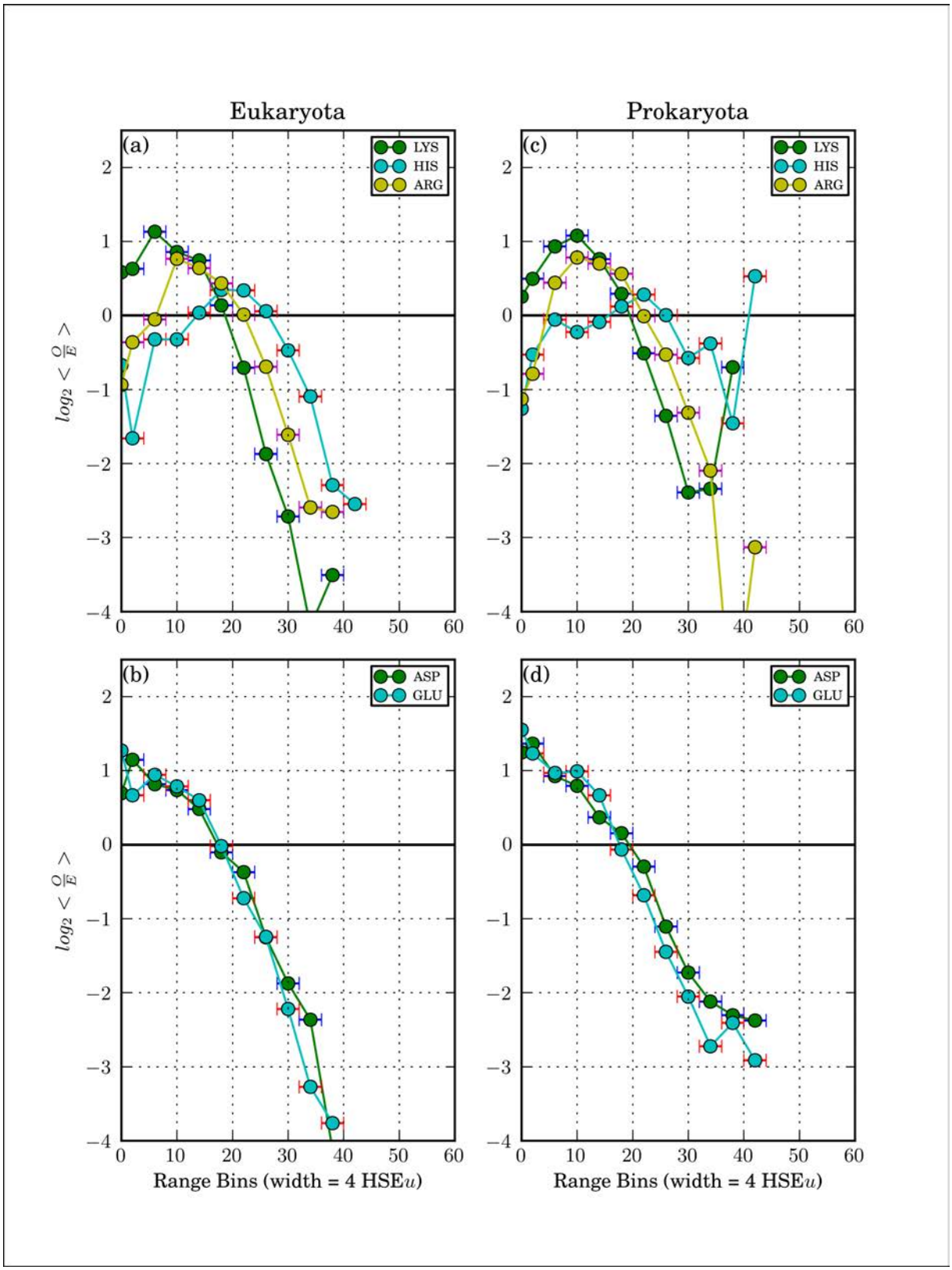

Figure 4.16: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. $HSEu_{13}$ for eukaryotic and prokaryotic charged residues:(a) Eukaryota, positively charged residues, (b) Eukaryota, negatively charged residues, (c) Prokaryota, positively charged residues, (d) Prokaryota negatively charged residues.

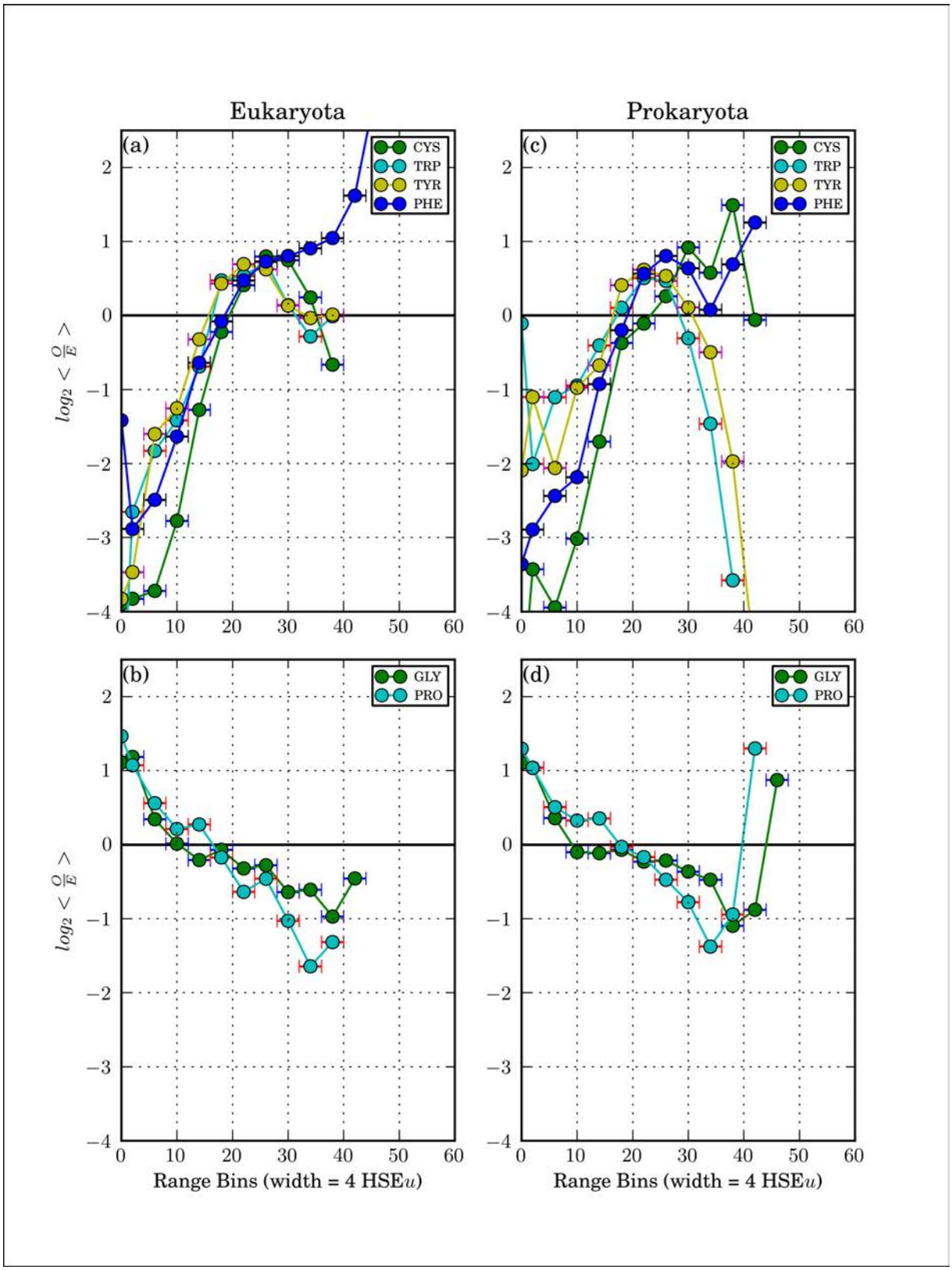

Figure 4.17: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu$_{13}$ for eukaryotic and prokaryotic aromatic residues with "special cases:" (a) Eukaryota, aromatic residues, with cysteine, (b) Eukaryota, special case residues, (c) Prokaryota, aromatic residues, with cysteine, (d) Prokaryota, special cases.

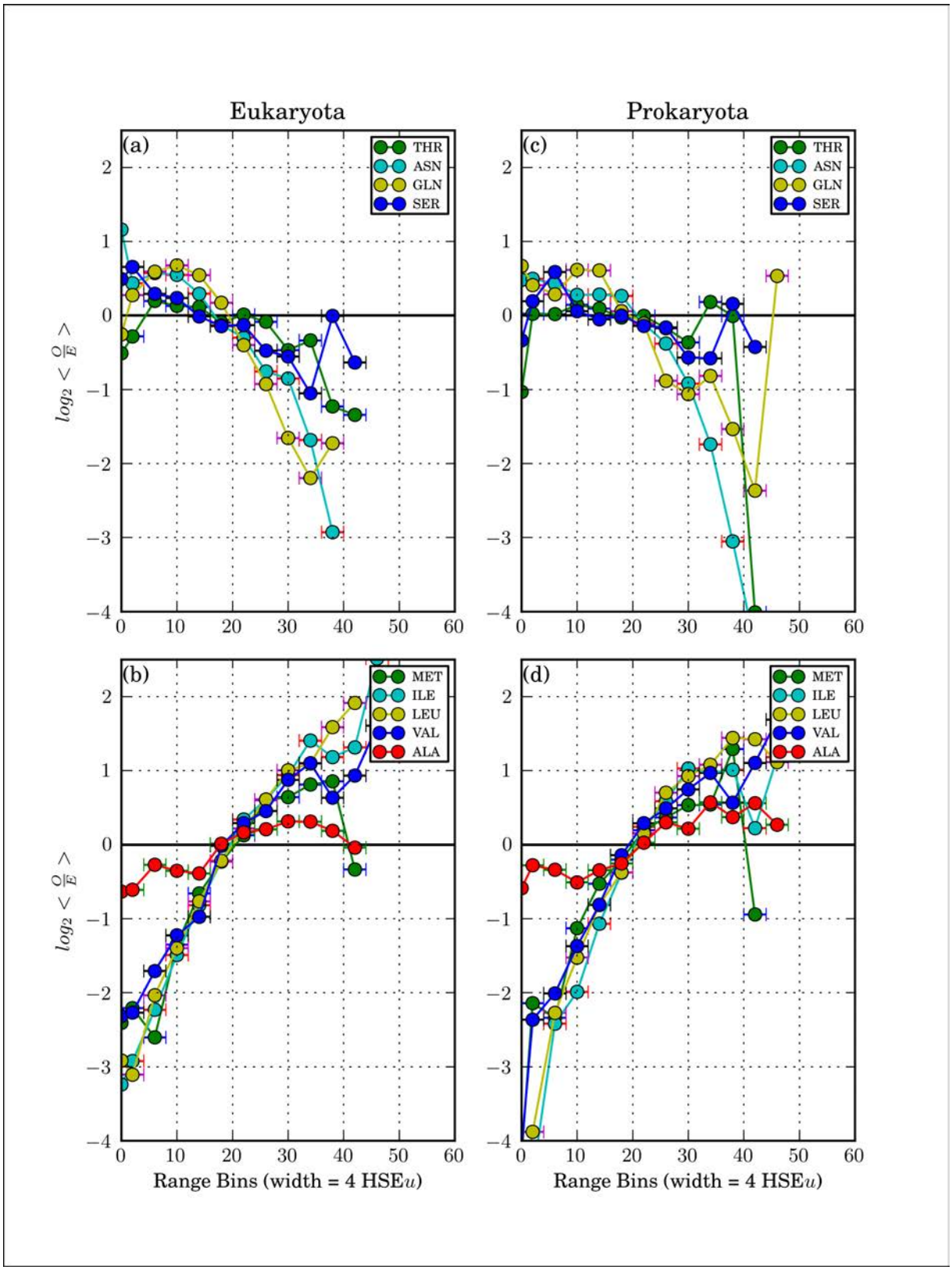

Figure 4.18: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. $HSEu_{13}$ for eukaryotic and prokaryotic uncharged and hydrophobic residues: (a) Eukaryota, uncharged polar residues, (b) Eukaryota, aliphatic residues, (c) Prokaryota, uncharged polar residues, (d) Prokaryota, aliphatic residues.

Table 4.7: Crossover points for $HSEu_{13}$, Eukaryota and Prokaryota.

Residue	Eukaryota	Prokaryota
ALA	17.8	21.7
ARG	22.1	21.9
ASN	16.5	20.9
ASP	17.1	19.9
CYS	19.2	24.2
GLN	19.2	19.4
GLU	17.9	17.6
GLY	10.2	8.6
HIS	13.7	15.6
ILE	19.5	20.2
LEU	19.9	20.9
LYS	18.7	19.6
MET	20.7	19.8
PHE	18.6	19.0
PRO	16.9	17.6
SER	13.8	10.6
THR	16.0	16.5
TRP	16.1	17.1
TYR	15.4	16.5
VAL	18.1	19.0

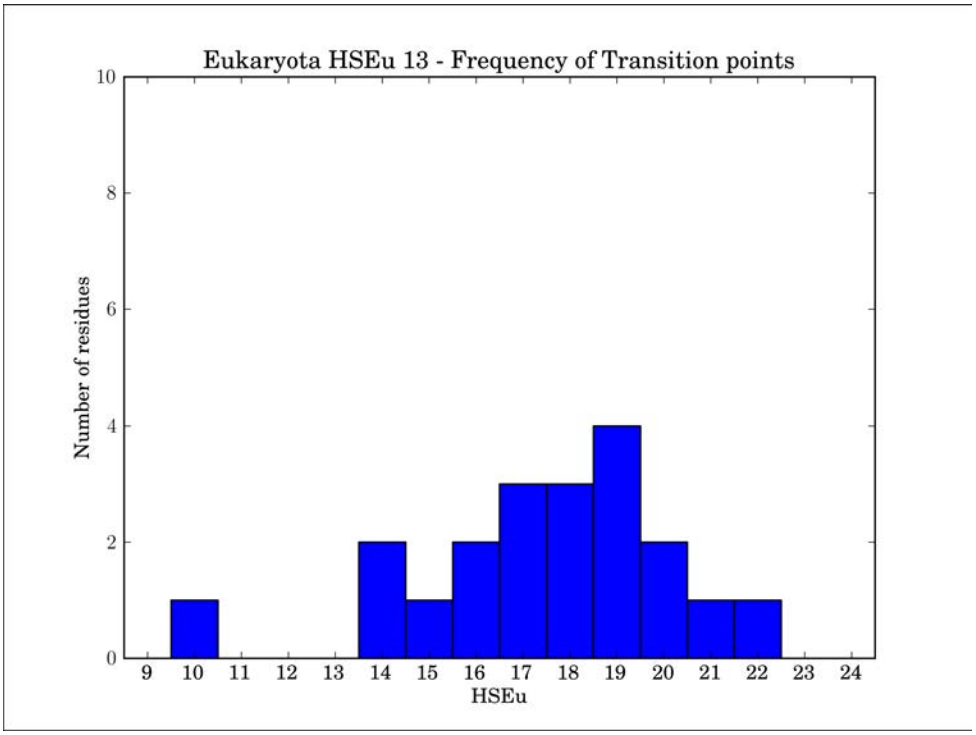

Figure 4.19: Histogram of $HSEu_{13}$ crossover points in Eukaryota

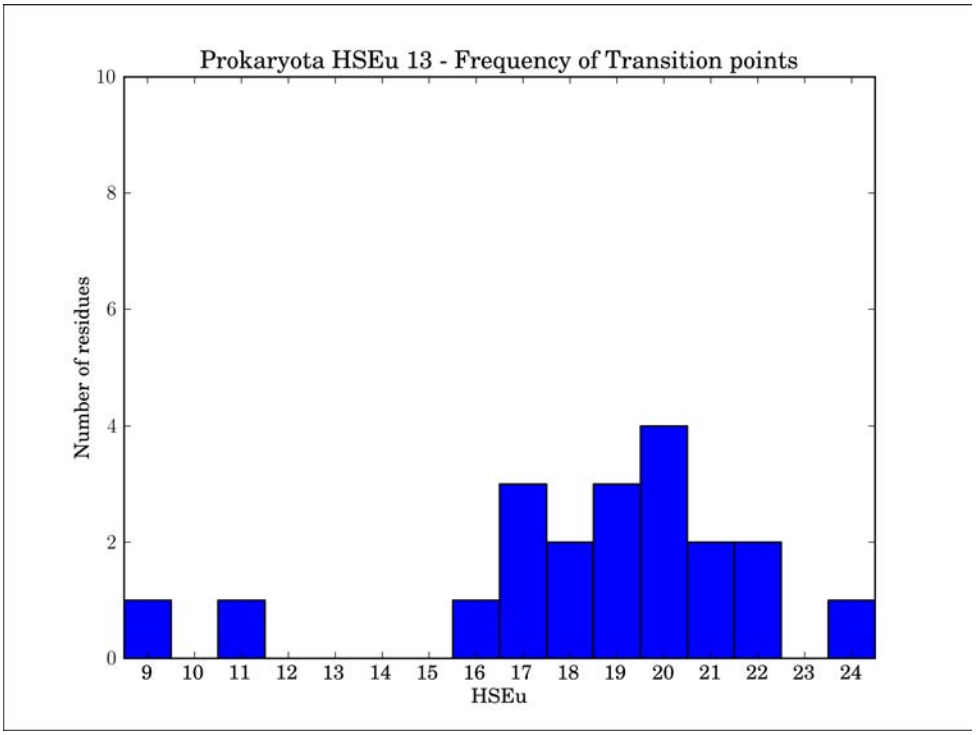

Figure 4.20: Histogram of $HSEu_{13}$ crossover points in Prokaryota

4.4 Discussion

The primary objective of these analyses was to address the following question: *"What is a statistically defined measure to distinguish the amino acid compositional change between the surface and the protein core?"* The result which would offer the strongest answer would be a single value of solvent exposure representing a uniform crossover from preferred solvent exposure values ($log_2\langle\frac{O}{E}\rangle > 0$) to not-preferred ($log_2\langle\frac{O}{E}\rangle < 0$), consistent across all residue types; the hydrophilic and hydrophobic residues sharing a crossover point between a preference and dis-preference of either HSEu or ASA, or both. Such a universal crossover point could be interpreted as the change in statistical preference between the surface and residue burial. Additionally the solvent exposure measure in which the universal crossover is observed, would be the correct choice to delimit residue burial.

As a result of the hydrophobic effect, we expect to see an increased preference for hydrophobic residues to be buried and hydrophilic residues to be at the surface, consistent with the physico-chemical properties of each residue-type. Therefore, it is reasonable to expect that there is a crossover point between measures of solvent exposure where different residue types become more or less preferred, thus separating the residue types into distinct populations, based on their preference to be solvent exposed. Expressed in terms of the $\frac{O}{E}$ ratio, when comparing the distribution of a given residue type with respect to solvent exposure with an unbiased distribution, we expect to see an over-representation of hydrophobic residues with high values of $HSEu_{13}$ (indicating burial) and low values of ASA (also indicating burial). Similarly we expect to see hydrophilic residues to be over-represented for low values of $HSEu_{13}$ and high values of ASA.

The three analyses for which results have been presented, individually attempted to address their own specific question. However, each contributed towards addressing the primary question of this chapter. This section, presents a discussion of the three analyses separately. The conclusions of the analyses are summarised in the conclusions to the chapter.

4.4.1 Comparison of HSEu using different sphere radii $\frac{O}{E}$

The purpose of completing the $\frac{O}{E}$ analysis for three different radii of HSEu was to assess the difference between each and to confirm or dismiss use of a radius of 13 Å, as recommended by Hamelryck [53].

The HSEu measure represents the total number of residues – determined by counting the C_α atoms – in the direction of the side-chain within a hemisphere of a chosen radius. A noticeable feature of all the bootstrap plots of the HSEu data shown in Figures 4.1 – 4.4, is the behaviour of the individual bootstrap lines, where they tend to be fairly close together in one range-bin and diverge considerable in the next. This "fanning" of the individual bootstrap data lines is an indication of sparse data; it is present in the plots for all three radii, occurring at both the lower values and higher values of HSEu, with the most pronounced divergence in the upper range-bins for each radius. For the $log_2\langle\frac{O}{E}\rangle$ vs. $HSEu_{10}$ plots the range-bins 0-4 and 16-20 show indications of sparse data. For the $log_2\langle\frac{O}{E}\rangle$ vs. $HSEu_{13}$ indications of sparse data are in the range-bins 0-4, with some residues this includes the range-bin 4-8; the upper region of sparse data appear in the bins 32-36 and 36-40. For the $log_2\langle\frac{O}{E}\rangle$ vs. $HSEu_{16}$, regions of sparse data cover the range-bins 0-12 and anything above 56. The upper regions of sparse data are probably because there must exist a theoretical maximum number of residues that can be "packed" into the volume of the hemisphere. These regions of extreme fanning likely represent the region bordering the maximum optimum packing, where more dense packing becomes less frequent and less optimal.

Comparison of the plots for the measured data, shows that the line of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu are consistent for each radius. But the line for $HSEu_{10}$ is a compressed version of $HSEu_{13}$, and $HSEu_{16}$ an extended version $HSEu_{13}$. Table 4.2 presents a comparison of the crossover points for each radius, for each residue type; these crossover points should be considered +/- 2 given the range-bin width used in the analyses. Inspection of the data in the table, would suggest that $HSEu_{10}$, with an almost universal crossover point around 8, would be the appropriate measure to use. If we assign the interpolated crossover points for $HSEu_{10}$ to the range-bins used in the analysis, all crossover points fall within the range-bins 4-8 and 8-12.

$HSEu_{13}$ has crossover points ranging from 10 through to 22. Assigned to the range-bins used in the analysis, the crossover points would be placed in bins from 8-12 through to 20-24 covering a total of four range-bins. This is double the number for $HSEu_{10}$. This is consistent with there being twice the number of HSEu range-bins for $HSEu_{13}$ compared to $HSEu_{10}$. $HSEu_{16}$ spreads the crossover points across a wider range of bins than the other two, which likely reflects a sparse spread of data across the range-bins, suggesting larger range-bins might be more appropriate - which would reduce the accuracy of the results.

Based on the results of the three comparative analyses, it is argued that $HSEu_{13}$ for HSEu is an appropriate value to use. This is consistent with the findings of [76], who investigate the distribution of HSEu and HSEd for 5 different radii (8, 10, 1 2, 13 and 14 Å). They showed that HSEu 12, 13, and 14 had distributions approaching normal distributions. Further they argue that analysis of Coordinate Number by Karchin et al. and Yuan et al. [77, 78] has shown a radius of 12-14 Å is better for fold recognition and structure prediction. This is consistent with the radius recommended by Hamelryck (the author of HSE) "because it looks right" [53].

4.4.2 Comparison of HSEu and Side Chain ASA $\frac{O}{E}$

The use of ASA to investigate amino acid solvent exposure is long standing and well documented in the literature. As such it could be considered a benchmark against which to test any alternative measure and compare relative performance. The objective of performing the $\frac{O}{E}$ analysis using the side-chain ASA data and comparing it with the $HSEu_{13}$ $\frac{O}{E}$ data, is to investigate which of the two measures is most appropriate to delimit residue burial. It also afforded the opportunity to investigate the propensity of each residue type to be solvent exposed from two distinct perspectives, one from that of the solvent and the other from that of the amino acid residue.

HSEu has a linear correlation with side chain ASA

The scatter plots in Figures 4.5 – 4.8 show clearly that a linear relationship between $HSEu_{13}$ and ASA exists for each residue-type. Further, the method used to perform the linear regression analysis returned the correlation coefficient (r-value) between ASA and $HSEu_{13}$ for each residue

type (except Glycine, which could not be measured as it has no C_β atom), these are summarised in Table 4.3. These range from -0.65 for cysteine, to -0.87 for proline, with most residue types having a correlation coefficient between -0.75 and -0.87. Interestingly Song et al. [76], who were investigating methods of predicting residue HSE values, reported a correlation coefficient between HSEu and full-residue ASA of -0.76. However their method did not investigate the behaviour of individual residue-types as ours did. Additionally they showed a non-linear relationship between ASA and HSEu, again by considering all residue-types together and not individually.

The $HSEu_{13}$ measure where ASA = 0 for all residue types, summarised in Table 4.3, are all approximately the same, 26 $HSEu_{13}$ +/-2. If all numbers are rounded to the nearest integer (as HSEu strictly speaking can only be represented by a positive integer), then the y-intercepts of the linear regression lines for all residue-types are in the range [24–28], which is the width of one range-bin used in the $HSEu_{13}$ $\frac{O}{E}$ analysis, and by coincidence is also one of the actual range-bins used in the analysis. With the exception of histidine, the hydrophilic residues all have y-intercepts below 26 and the hydrophobic residues all have y-intercepts at or above 26. A given HSEu value effectively measures the depth of the C_α position, assuming an even protein C_α density. Thus, for a given HSEu value, for some HSE sphere radius, there will be a maximum and minimum possible ASA for each residue type, which depend on the maximum and minimum extensions that each side chain type can have. Although some exceptions may exist due to strange geometric locations e.g. semi-buried active sites. This seems to have some relevance for explaining the linear relationship between HSEu and ASA and also for explaining why the choice of data set does not have a great affect on the linear regression line, see Table 4.3 The fact that the y-intercept between residue types are so close together, to be effectively one value within a small margin of error, is undoubtedly a significant results, unfortunately interpreting it has proved to be challenging.

A useful observation that can be made from the data in the Table 4.3, is that there is only a slight variation in the numbers between the data selected for the $\frac{O}{E}$ analysis and the entire PiQSi. The selected data contained data from exclusively cytoplasmic globular proteins. The

PiQSi data included structures from multiple contexts (there were a handful of NMR ensembles which produced very high HSEu values as well) and yet the linear relationship between ASA and HSEu was hardly affected.

$HSEu_{13}$ and ASA crossover points are in close agreement

Hydrophobic theory predicts that hydrophobic residues will be preferentially distributed in the protein interior, while hydrophilic residues will be preferentially distributed on the surface of the proteins. The $log_2\langle\frac{O}{E}\rangle$ vs. solvent exposure for both $HSEu_{13}$ and ASA are in agreement with the expectations of individual residue types based on their physcio-chemical properties. For example, Figures 4.13, 4.14 and 4.15, show that charged residues have a greater propensity to be solvent exposed than buried, while aliphatic residues have a greater propensity to be buried.

The preference for solvent exposure for each residue type, is described by both measures without contradiction. The $\frac{O}{E}$ data for both $HSEu_{13}$ and ASA are generally in agreement across all residue types (see Figures 4.13, 4.14 and 4.15). For all residue types, when $log_2\langle\frac{O}{E}\rangle$ is positive for a low value of $HSEu_{13}$, there is a region of positive $log_2\langle\frac{O}{E}\rangle$ for high value of ASA, and vice versa; similarly for negative values. Consider, for example aspartic acid and glutamic acid, in Figure 4.13, subplots (b) and (d); for both residue types there is a region of $HSEu_{13}$ $O > E$ from [0 – 18] in (b), complemented by a region of ASA with $O > E$ from 30 $Å^2$ and above in (d).

The crossover points from $O > E$ to $O < E$, shown in Table 4.4, can be interpreted as the point where the two measures indicate a change from solvent exposed to buried and vice versa. Given the linear relationship between HSEu and side chain ASA one might expect a correspondence between the HSEu crossover point and the side chain ASA crossover point. The ASA crossover points predicted from HSEu crossover values, shown in Table 4.5, show that 9 residues (Arg, Asp, Gln, Ile, Leu, Lys, Phe, Ser, Tyr) have a converted/transformed $HSEu_{13}$ crossover point in ASA within +/- 5 $Å^2$ of the ASA crossover point derived by the $\frac{O}{E}$ analysis, see Table 4.4, 3 are within +/- 10 (Cys, Glu, Val), 5 are within +/- 15 (Trp, Thr, Pro, Met, Ala) and 2 of 19 are within +/- 20 $Å^2$, (Asn, His), i.e. all residues have predicted ASA crossover point within +/- 20 $Å^2$, with almost half being within a margin of error of +/-5

Å^2. The discrepancy between the actual crossover values from the ASA $\frac{O}{E}$ analysis and the converted HSEu$_{13}$ crossover point, may have arisen due to compromises which had to be made in the $\frac{O}{E}$ analysis method used for the ASA data, to account for the variation in size between residue types. This raises the question as to whether the converted HSEu$_{13}$ crossover points are a better indication of ASA crossover points, than those produced by the $\frac{O}{E}$ analysis of ASA data; this would require further investigation but was not directly relevant to the aims of this thesis so was not pursued further.

The relative solvent exposure area included in Table 4.4 was generated by converting the ASA crossover to rASA using the Ala-X-Ala reference state. The predicted values of rASA crossover points compared with ASA crossover points shown in Table 4.5 don't show a significant benefit to using rASA over ASA.

HSEu allows for investigation of glycine and reveals close correlation with proline

Using side-chain ASA analysis does not allow for the behaviour of glycine to be determined, which is possible using HSEu. This is clearly shown in the bottom plots, (b) and (d) in Figure 4.14. The two plots provide a comparison of glycine and proline between the two measures. However the ASA data shows a single point for glycine and the HSEu$_{13}$ data shows a complete distribution.

Pearson's correlation analysis of the HSEu$_{13}$ $log_2\langle\frac{O}{E}\rangle$ data revealed Proline and glycine have an r-value of 0.9. The two residue-types behave almost identically; showing the same over-representation in the lower range-bins of HSEu and under-representation in the higher-range bins, indicating a preference for solvent exposure. The slope of the two lines appear to be quite similar descending at similar rates, though their crossover points, shown in Table 4.4, are different.

Proline is strictly speaking an aliphatic residue; as such this would suggest that it should be statistically more present in the protein interior than on the surface. However proline shows a generally negative correlation to most hydrophobic residues, r = -0.9 with respect to alanine, and a generally positive correlation with the hydrophilic residues, e.g. with respect to asparagine r = 0.8. Proline is often found in loops and turns between secondary structure units,

providing quick turns in the backbone allowing it to minimise its penetration into the solvent. Moreover it is rarely present in α-helices and when present it kinks it. Proline is also involved in protein-protein interactions e.g. polyproline interacts with WW & SH2 domains. The interesting thing to note here is that proline presents an exception to the rule that hydrophobic residues are preferentially buried. This exception may be a result of a biological need, which over-rides our physical expectations.

Glycine is just backbone with a hydrogen for a side-chain. It is a weakly hydrophobic residue and so it would not necessarily be expected on the surface, based on its physico-chemical properties. It is known that glycine's lack of a hydrocarbon side chain makes it more flexible around its ϕ, ψ bonds. As such it is often seen in loops and turns between secondary structure units, just like proline. It is an interesting observation that two residue-types of such distinct physico-chemical properties have such similar propensities for solvent exposure as has been shown here. The preference for both glycine and proline to be solvent exposed was observed and reported by Rose et al. [17].

Propensity for solvent exposure reflects amino acid physico-chemical properties

The hydrophobic effect has been shown to be size dependant. The smallest hydrophobic particle that can be "solvated" without disrupting the hydrogen bonding pattern of water is a single molecule of methane. There is a gradual increase in the strength of the hydrophobic effect with the increase of the particle size, until the hydrogen bonding pattern of water is so disrupted that it forms a pseudo gas layer around the particle [13]. The relationship between the size dependence of the hydrophobic effect and amino acid solvent exposure is not well understood. However, if we consider the size of any hydrophobic amino acid side chain other than alanine, complete solvent exposure of the side chain from the protein surface, would likely affect a larger volume than a single methane molecule. To account for this in greater detail hydrophobicity measurements would be useful. In the absence of these, residues are considered here in terms of both their size and other properties.

The grouping of residue types in Figures 4.16, 4.17 and 4.18 was based on a combination of the known physico-chemical properties of residues and visual comparison of the trend lines

to decide which closest resembled each other. Here the solvent exposure preferences of residue types with respect to each other and their physcio-chemical properties is discussed.

Firstly, consider the residue-types which could be classified as small polar and uncharged: serine and threonine, shown in figure 4.15, with the larger residues asparagine and glutamine. Both serine and threonine show a slight preference for solvent exposure over being buried, which is seen in plot (a), showing the HSEu data, and is confirmed in the ASA data shown in plot (c). The side-chain ASA data indicates that the proportion of residues for threonine and serine, below the solvent accessible surface, is not significantly more or less than we would expect by chance. This is an interesting observation, given that threonine has an additional methyl group in the side chain, thus having a C_γ compared to serine which only has a C_β in the side-chain. Making serine the smaller of the two. Yet it is threonine which appears to be the closest to random in our data. This is either indicative of a genuine result or a result of insufficient data in the analysis, though the bootstrap data suggest the former.

Now consider, the other two residue-types in the two plots, asparagine and glutamine. These two residue-types are bigger than threonine and serine. Asparagine is slightly bigger than threonine because it has an amine group and an oxygen bonded to C_γ, while the C_γ in Thr is part of a methyl group. Glutamine has a longer side-chain with a C_δ and an amide group. Notice that asparagine has a higher propensity for low values of HSEu compared with glutamine, i.e. it is more likely to be solvent exposed. Similarly the ASA data shows that at low values of ASA, asparagine is less under-represented compared to glutamine. This is possibly a function of the length of the side-chain. Glutamine has a longer side chain which is largely aliphatic (i.e. hydrophobic), except at the end. Therefore it would be energetically favourable to be less exposed to the surface, whereas asparagine with a smaller side-chain is less affected. This is a trend that can be seen with argenine and lysine as well, shown in figure 4.13.

Focusing now on alanine, shown in Figure 4.15 (b) and (d) with the aliphatic residue types, it appears to have the opposite behaviour of serine. Showing a slight preference to be over-represented in the higher range-bins of HSEu and lower range-bins of ASA, the opposite of serine. This is illustrative of their behaviour as a result of their differing physico-chemical

properties, with respect to the hydrophobic effect. The difference between serine and alanine being the OH group in serine replacing a hydrogen in the methyl group of alanine.

The behaviour of all residue-types shown in the plots can largely be attributed to their physico-chemical properties. Yet, there are a few unexpected results, e.g. proline as described in Section 4.4.2, and cystine. The aromatic residues shown in 4.14 (a) and (c), reveal an expected result. In the $log_2\langle\frac{O}{E}\rangle$ data for $HSEu_{13}$ cystine behaves very similar to the aromatic residues, while in the $log_2\langle\frac{O}{E}\rangle$ data for ASA the similarity is less apparent.

The behaviour of cysteine, in the $HSEu_{13}$ analysis, closely resembles the behaviour of the aromatic residues Trp, Tyr and Phe. Examination of Figure 4.14 (a), reveals that Phe behaves somewhat differently to the other two aromatic residues, while cysteine more closely resembles them. Though, it is known that cysteine has a propensity to form cys-cys disulphide bonds, these are normally found in extra-cellular proteins. The data set chosen for this analysis, was exclusively cytosolic and thus attributing this result to Cys-Cys bonding would be inappropriate. Cysteine is itself a relatively rare amino-acid, which is known to occur in active sites. Active sites are often obscured from the solvent accessible surface, thus affecting cysteine's propensity to be located away from the surface of the protein, which might offer a possible explanation of this observed behaviour. Further study would be required to verify this explanation.

$HSEu_{13}$ better describes the distribution of residues within the protein structure

The comparative analysis of side-chain ASA and HSEu, has shown that the two measures offer results which are in general agreement. Both methods have yielded results that indicate the same behaviour of residue-types based on physico-chemical properties, with respect to hydropathy.

A comparison of the residue pair-wise correlation coefficient for propensity for solvent exposure was compared with popular substitution matrices, the results of which are shown in Table 4.6. The highest correlation coefficient is for the Gonnet substitution matrix [2] and $HSEu_{13}$ with r = 0.65, while for ASA r = 0.57. Though both solvent exposure measures score similarly with several of the other substitution matrices. The true relationship between the substitution matrices and the residue-type propensity for solvent exposure requires further investigation.

However, the objective of the analysis was to determine which of the two measures is more

suited to determining a statistically defined measure to distinguish the amino acid compositional change between the surface and the protein interior. HSEu is capable of providing a distribution for each residue-type ranging from the surface to the protein core. While side-chain ASA cannot, it is restricted to residues with side-chains that come into contact with solvent; this excludes the capability of analysing glycine. Further, as was shown from the bootstrap plots, the ASA analysis effectively uses a subset of all residues in a protein structure compared to the HSEu analysis. In the context of complicated statistical analysis where, for example there is a need to account for difference in side chain size, as was the case for the bootstrapping of the data in this chapter, HSEu is simply easier to work with; because HSEu is not obviously dependent on residue size.

4.4.3 Comparison of Eukaryota & Prokaryota HSEu $\frac{O}{E}$

The third and final comparative analysis, was undertaken to determine if there was a statistically significant difference in behaviour of each amino acid type, between the cytosolic proteins from eukaryota, and those from prokaryota. The results of the prokaryota analysis are shown alongside the eukaryota results, for an $\frac{O}{E}$ analysis using $HSEu_{13}$, in Figures 4.16, 4.17 and 4.18. Comparison of the $\frac{O}{E}$ crossover points for HSEu in Eukaryota and Prokaryota are shown in Table 4.7 and further summarised in Figures 4.19 and 4.20,

The results of the analyses shows a slight difference between the two data sets. Though the mean value for both fall in the range 18-22, it would appear that in Prokaryota the crossover is slightly deeper (having a mean value of 18.3 $HSEu_{13}$ which would be rounded to 18) than in Eukaryota (with a crossover of 17.3 $HSEu_{13}$ which would be rounded to 17). We cannot determine if this is a genuine result or an artefact of the method, because the difference is so slight, covering two range-bins.

The conclusion of this analysis is, that there appears to be a unique and largely uniform crossover point consistent between the strictly hydrophobic residues I(euk: 20, prok: 20), L(euk: 20, prok:21), V(euk: 18, prok:19), M(euk: 21, prok: 20) and the strictly hydrophilic residues, D(euk: 17, prok: 20), E(euk: 18, prok: 18), K(euk: 19, prok: 20) , R(euk: 22, prok:

22). (The mean cross over point for these residues (both for Eukaryota and Prokaryota) is 19.7.) These residues having physico-chemical properties which make them most susceptible to the hydrophobic effect, are the ones which will be most drawn to or repelled by the solvent. With the exception of Aspartic Acid in the Eukaryota, all these residues crossover in the range 18 - 22, The majority of the residues in both sets, crossover in this range. Given that the analytical set up placed all the residues in to range-bins of width 4, the crossovers extrapolated from the graphs, cross the higher region of the range-bin 16-20 and the lower range of the range-bin 20-24. For this reason, the proposed cut-off to delimit residue burial is 20 $HSEu_{13}$. Though from the scatter plot data it could be argued that 24 might be more suitable, but given the lack of adequate weighting of the points in the scatter plots, it is difficult to justify relying heavily on those results other than as a general indication of the correlation between HSEu and side-chain ASA. It could be argued that the range 20-24 is some "twilight" region between absolute exposure and absolute burial.

However, for the purposes of the co-substitution analysis presented in the next chapter, a single figure was required to define the limit and this was set at 20 $HSEu_{13}$. This value was chosen for two reasons. Firstly, it is the boundary of the range-bin in which the two mean crossover points for the eukaryota and prokaryota data were present. Secondly, the mean cross over value for the combined set of strictly hydrophilic and hydrophobic residues, described above, has a mean cross-over point of 19.7. HSE is an integer count, so rounding up to 20 seemed appropriate. It could be argued that using the midpoint of the range-bin (i.e. 18 $HSEu_{13}$) might be more appropriate. Future analyses may need to determine which is the better value to use.

4.5 Conclusions

The objective was to address the question: "What is a statistically meaningful measurement to distinguish the amino acid compositional change between the surface and the protein core?" to address this, three comparative analyses were undertaken. Firstly, it was necessary to determine an appropriate radius to use with HSEu. Secondly, it was necessary to determine which was

better suited to addressing the question posed, side-chain ASA or HSEu. Finally, it was necessary to compare the crossover points in the eukaryota data with the crossover points in the prokaryota.

First, it was determined that a radius of 13 Å was appropriate for the HSEu measure. Second, it was determined that ASA suffered limitations compared with HSEu. Finally it was determined that the eukaryota and prokaryota data have small differences in their crossover points, but it is not possible to determine if these are statistically significant or not.

However, the conclusion of this work, does answer the question posed. Yes, there is a meaningful measurement, which can be used to distinguish the amino acid compositional change between the surface and the protein interior. It is a measure of 20 $HSEu_{13}$, for cytosolic globular proteins. Analysis of other data sets composed of proteins belonging to different contexts, will need to be analysed to determine the same limit in those environments. It is likely that there will be differences, notably with the charged residues K, R, D and E, given the effect that changes in pH can have on the protonation state of different residues.

4.6 Acknowledgement

Computer programs were written in the Python programming language, making considerable use of the BioPython packages [75]. The methodology described here was developed and designed by Mr. Bhima Auro, however some of the work was completed with the help of a final year undergraduate student, Mr. John Le Brun. A break down of our relative contributions is given in Appendix I.

Chapter 5

Determining the propensity for co-substitutions with respect to distance

5.1 Introduction

Elucidating the rules governing amino acid interactions, both inter- and intra-protein, is a challenging problem. Presented here is the development of a method that can provide at the very least a partial solution to this problem. The method is presented and a small set of co-substitution graphs provide a proof of principle. This will provide the basis and framework to investigate further the principles of co-substitution in intra- and inter-protein interactions. With additional development it will quantify the relationship between co-substitution events and amino acid euclidean separation in protein structures, and possibly provide a method for protein-structure prediction.

The substitution of contacting amino acids in a protein structure should be correlated [79]. Consider, for a thought experiment, a buried salt-bridge between K and E e.g as discussed in the introduction to this thesis. If K were replaced by E, maintaining the salt-bridge would require that the original E be replaced by either K or R. Having E and E in place of a K-E salt-bridge would result in a localised repulsive interaction. This could be disruptive to the local structure of the protein at very least, if not to the entire structure and function of the protein. Further, if either K or E were substituted with an apolar residue, such as I, the remaining charged residue

would now be in a hydrophobic environment. This too would be unfavourable and could be disruptive to the protein structure.

Many hypotheses as to how to analyse protein co-evolution, are in the literature [22]. These primarily assume the correlation of physico-chemical properties [22] or attempt to average residue interaction propensities [80]. Protein co-evolution has been considered in two contexts, inter-protein co-evolution and intra-protein co-evolution. Efforts by Valencia et al. Haussler et al. [30, 81] and others have attempted to use inter-protein co-evolution to predict protein interactions, while Valencia et al. Halperin et al. and Marks et al. [22, 29, 30, 48, 49] have attempted to predict intra-protein contacts.

Co-evolution, in the form of "correlated-mutations" or co-substitutions –as described in Chapter 2– is likely to occur within protein families to maintain the integrity of the folding pathway and maintain the protein's functional and structural properties. Historically, investigations into correlated mutations have concentrated efforts on predicting inter-residue contacts, while assuming non-contact interactions are not relevant. Non-contacting interactions can arise for different reasons, such as electrostatic attractions or repulsions; or to accommodate adjustment in the packing of the protein transmitted through the structure. However developing a method to explicitly single out a specific causal source of a co-substitution would be far from a trivial task. If correlated mutations can occur for many reasons [70], then we aim to develop a method to quantify distance effects. This will provide evidence of which co-substitutions give information about inter-residue distance and what that information about distance is.

As yet, there has been no attempt to determine which co-substitutions are most likely to occur and if certain co-substitution types correlate with inter-residue distance in any way. The following question has not been addressed in the literature: *"are there specific co-substitutions which occur preferentially when the residue positions are in direct contact, and others which occur at some preferred distance from each other?"* The method presented here, has been developed to address these questions. A consideration that has been applied to this method, as explained in the previous chapter, is the solvation state of the residue positions involved in the co-substitution, as well as the cellular location and whether a protein is from either a eukaryotic

or prokaryotic organism. This investigation will attempt to determine how the propensity for each co-substitution type differs between the surface and the protein interior. This, to the best of our knowledge, has not been attempted before.

5.2 Methods

5.2.1 Development of methodology

Determining the propensity for a given type of co-substitution in a given context requires calculating $\frac{O}{E}$. Context refers to the cellular environment/location and solvation state of the residues under consideration. The emphasis on context arises because the distribution/composition of amino acid residues in the protein structure is dependent on environmental factors [68]; for example pH of the solvent will affect the protonation state of some residue-types causing them to be either more exposed or more buried. Additionally due to the hydrophobic effect there will be a preference for hydrophilic residues to be on the surface and the hydrophobic residues to be buried. These variations will affect the likelihood of seeing a given type of co-substitution in a given environment and thus the Expected. To account for this variance of distribution of residue types, proteins were selected on the basis of cellular location and residue types on the basis of solvent exposure. The data set used, was taken from the data used in the solvent exposure analysis discussed in chapter 4, which was selected using the database tool discussed in chapter 3.

As a first approximation to evenly sampling the phylogenetic tree, sequence weighting was applied. The question being considered is "what co-substitutions are acceptable to the evolutionary process?" Consequentially, it is necessary to ensure that the same evolutionary data is not duplicated. To address this, weighting is applied to pairs of sequences, as described in Chapter 2. This is different from methods of co-evolution analysis in the literature, where weighting is applied on the basis of the similarity of the data in the set of e.g. columns exhibiting correlated mutation behaviour, rather than weighting sequence pairs between which mutations are observed.

Locating co-substitution events in a sequence alignment requires a comparison of every

possible unique pair of sequences in the alignment. A fundamental assumption being made here considers residues in the same columns in multiple sequence alignment as representing the equivalent position in the structures of these homologues. By comparing two column positions in a pair of sequences, a co-substitution event is identified when at least one residue differs at either position. However, in our analysis we are not restricted to determining the relationship between co-substitution events and distance; we can equally study conservations in the same way, i.e. we can address the question *"do certain conservations occur preferably at certain inter-residue distances?"*

For this work a structural example representing the homologous sequences in an alignment was required to provide the distances between each pair of columns in the alignment. As described in Chapter 3, the PiQSi database was used to provide biologically relevant structural data and the Pfam-A database was used to provide sequence alignments of protein domains. Several steps were taken to filter the data and prepare it prior to identifying co-substitutions, as described below. All computer programs were written in Python.

5.2.2 Preparation of sequence alignment data

Preparing the sequence alignment

For the co-evolution analysis, the Pfam families and structures used for the solvent exposure analysis discussed in Chapter 4 were used; Pfam families that have examples known to exist as homo-oligomers were identified and removed using the database discussed in Chapter 3. This step is taken as it would be impossible to determine if a co-substitution detectable by this analysis was seen as a result of long range intra-domain interactions, or short-range inter-domain interactions between neighbouring sites in the biological unit between the two domains, which is explained more clearly in Figure 5.1. The analysis was performed independently on 46 Pfam families with a sequence and structural reference taken from Eukaryota and 51 Pfam families with a sequence and structural reference taken from Prokaryota.

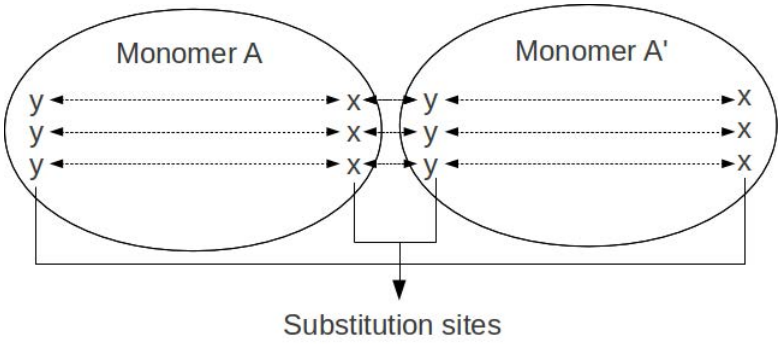

Figure 5.1: Co-substitution events in a homo-oligomer: Residues x could influence the residue type at sites y in A and A', i.e. residue-type x in Monomer A might influence y in A and A'

Each Pfam family has a sequence alignment consisting of known examples of the domain; these are segments of the protein sequences stored in the UniProt data-bank. The alignment for each selected family was processed as follows. The sequences in the alignment were compared with the sequences from the structural example of the domain. The structural example with the most residues and the highest sequence identity to a sequence in the alignment was selected. Since crystal structure data is sometimes from a mutated variant of the wild type, this means that it is not always possible to have a perfect match between the structure sequence and the protein sequence derived from genomic data. The two mentioned sequences formed a pair, the structure sequence and the alignment-reference-sequence. As in the solvent exposure analysis, the separation into taxonomic groups (Eukaryota/Prokaryota) was maintained. This selection step was applied to the Pfam alignment, such that all sequences not from the taxonomic group were removed and a new alignment was compiled, consisting of only those sequences belonging to the taxonomic group of the selected reference structure. A threshold sequence identity of at least a 35% with the reference sequence was also applied. It has been shown that structural similarity is maintained between homologous sequences of from c.35% and higher [20].

Two further filtration steps were applied to the alignment. Firstly the percentage of gapped entries in each column was determined. Columns consisting of 45% gaps or more were excluded from the analysis, as it was deemed that these columns probably represented poor quality regions in the alignment and could most likely reduce the reliability of our results.

Secondly, steps were taken to remove duplicate sequences from the alignment. This was

done, because of the weakness of the Henikoff weighting described in Section 2.5. The discussion of the weighting showed that the Henikoff weighting method does not reduce the weight of two equal sequences by one half, three equal sequences by one third and so on. Due to this limitation we decided to remove sequences which were close to identical, with the threshold being set at 95% sequence identity; the sequence identity for every possible pair of sequences in the alignment was calculated. If a sequence pair was found to have identity at or above the threshold the first of the pair was added to a list of excluded sequences. This is not the ideal solution, which suggest that the optimal possible data for analysis was not necessarily achieved. However, as there were time constraints on developing an optimal solution, this was the simplest solution permissible. Importantly the 95% threshold will actually remove little co-substitution data; but the reduction will speed up the algorithm to locate co-substitutions. The above completed the preparation of the sequence alignment data for the search of given co-substitution events in the data.

Mapping the structure to the alignment

The structure sequence and the alignment-reference-sequence described above, were used to create a map between positions in the reference structure and columns in the sequence alignment. This was done by recording the position of all the gaps from the alignment sequence. The gaps were then removed and ClustalW [82] was used to align the structure sequence with the alignment sequence. Positions from the structure sequence which had to be gapped in order to align properly with the alignment sequence were recorded. The positions in the alignment sequence matching positions in the structure sequence were recorded as these are the only column positions in the alignment for which any distance information will be available in the distance matrix, and thus the only ones relevant to our analysis.

Generating the Distance Matrix

The inter-residue distance between pairs of residues in the domain structure was calculated, using the PDB module for BioPython [75]. All inter-residue distances were measured between C_β atoms. Glycine, having no C_β atom, posed a problem in this regard. To address this, firstly,

all the glycine positions in the structure sequence were relabelled to alanine using a Bash-shell script. Secondly the CORALL command of the program WHATIF [74] was used to correct the atomic positions of the structure to incorporate a C_β atom for those glycine changed to alanine. This inserted the missing C_β atom for the glycines which had been changed to alanine; but the structure was not optimised to accommodate the additional atoms; thus the inter-residue distances remained unchanged for all residue pairs, although CORALL did add any atoms missing from residues in the structure. The inter-residue distances were then determined. The glycine positions were recorded to ensure no confusion between the actual alanine residues in the structure and those representing the replaced glycines. As such the correct residue is recorded for each position in the distance matrix. However the inter-residue distances recorded between any residue and glycine is measured with respect to the pseudo C_β position.

Permissible co-substitution events in a given environment may also be constrained by the requirements for packing chains in the tertiary structure, and they may be constrained by the requirements for the formation of secondary structures. Given that for secondary structure predictions there are reports of above 90% accuracy [83]; it makes sense to ignore interactions that may be intra-secondary structure and focus on tertiary structure associated packing interactions. To avoid intra-secondary structure interactions, substitutions between residue pairs 15 residues or fewer apart from each other in the sequence, were not considered. The choice of this separation is supported by Brunak et al. [84], who studied inter-atomic distances between C_α atoms. They reported that the distribution of inter-atomic distances in the tertiary structure reflects the secondary structure units within the protein structure. They saw the strongest indication of secondary structure effects between residues separated by fewer than 10 residues. However they claim to have seen some effect up to a separation of 20 residues. This is further supported by the work of Mr. Welland, who calculated the $P(d|s)$ – the probability of a distances given sequence separation, between C_β atoms. Figures 5.2 and 5.3 [1], show $P(d|s)$ against the distance of separation in the structure. The first figure illustrates the distribution in the short ranges of sequence separation up to 16 residues apart and shows that secondary structure effects

[1] Figures supplied by, and included with the permission of, Mr. Welland

disappear at about a separation of about 15 residues. The second figure is shown to illustrate that the secondary structure effects don't appear again at any separation greater than 15 and that at a separation of 15 residues and above, the distribution of $P(d|s)$ very strongly resembles a Poisson distribution. For these reasons, co-substitutions were considered for analysis only for pairs of residues separated by 15 residues or more, in the sequence alignment.

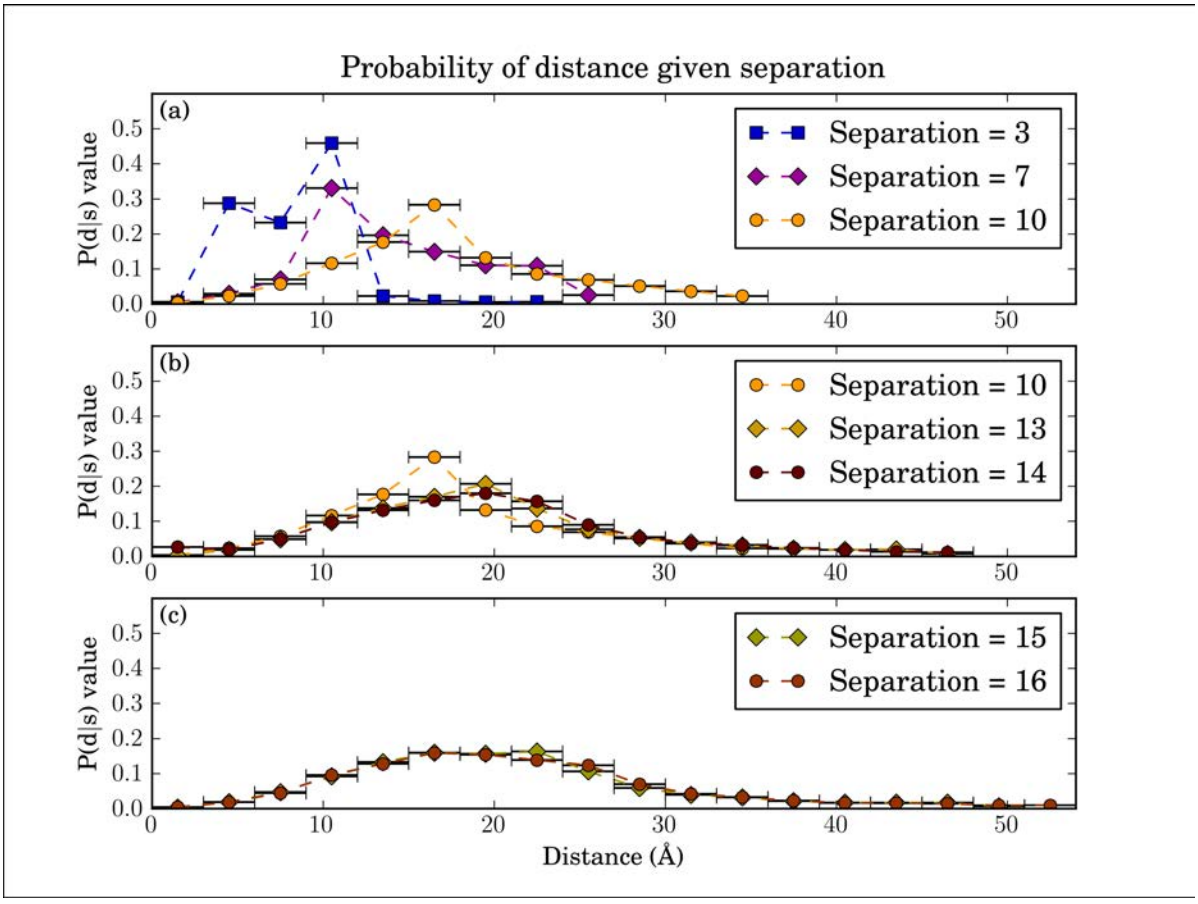

Figure 5.2: $P(d|s)$ **vs. inter-residue separation.**

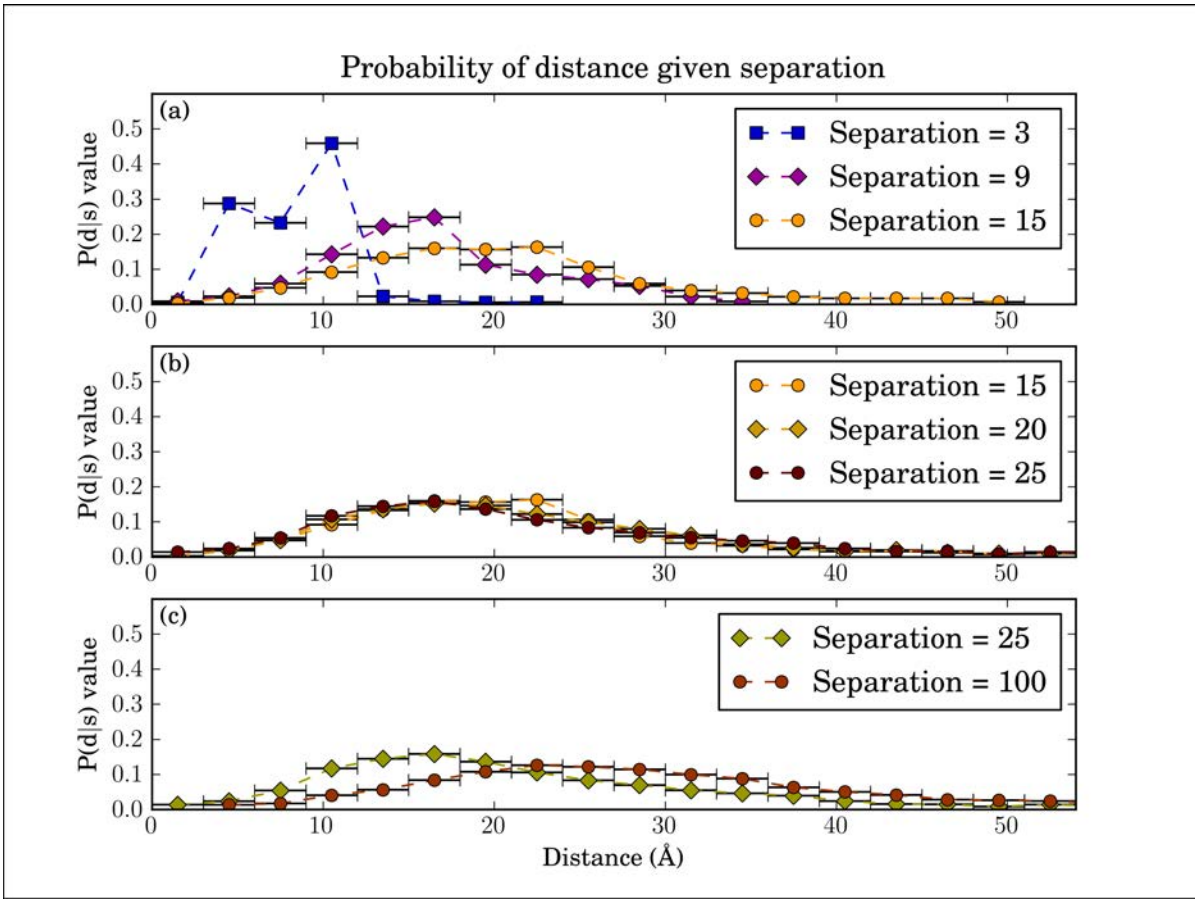

Figure 5.3: $P(d|s)$ **vs. inter-residue separation.**

Building the distance matrix was done while taking into account solvation states of residue pairs. The inter-residue distance data, was measured between residues in the reference structure and sorted according to the solvent exposure of each residue in the residue-pair. Three sets of residue pairs were generated: surface pairs, where both residues were on the surface; buried-pairs, where both pairs were in the protein interior; mixed pairs, where one residue was buried and the other was on the surface. The solvent exposure used to delimit the boundary between the surface and the protein interior was a measure of 20 $HSEu_{13}$, as determined by the solvent exposure analysis discussed in Chapter 4. The reference structure had been selected from the structures which had been used for the analysis of solvent-exposure, with the solvent-exposure, values stored in the B-factor column of the PDB structure file. The inter-residue positions in the reference structures (which represents inter-column distances in the alignment) were assigned to distance range-bins of width 3 Å . Thus, the final distance matrix was represented, firstly into

three sets for each solvation state of the residue pairs, and secondly with the range-bin for each residue-pair assigned.

Bootstrap generation

The distance data was used to generate randomised distance-matrices for the purpose of performing a bootstrap analysis. The residue-pair and measured inter-residue distance data for a given solvation state was taken. From this data a list of all residue pairs and a list of all inter-residue distances was compiled. An inter-residue distance was then randomly selected from the list, with replacement, and assigned to a residue-pair. This was repeated for all residue pairs. One hundred randomised distance matrices were generated for each set of position pairs in each Pfam family selected for the analysis. Due to time considerations, the residue-position-pairs which were in the "mixed" set (i.e. one residue on the surface and one buried), were excluded, this will need to be analysed at a later date, to determine the co-substitution behaviour between the surface and the protein interior.

5.2.3 Searching and the Co-Substitution Analysis

The next step is to determine the frequency of each co-substitution type in each inter-residue distance bin, in every pair of sequences associated with each filtered Pfam alignment. There are different ways in which this can be done. However, each co-substitution type must be searched for and analysed independently of all others. Two different approaches were tried, both are described in this subsection. Firstly, a method which attemtped to search for co-substitution events and calcate the $\frac{O}{E}$ value for that specific co-substitution type, within the same workflow. Secondly, a method which attempted to perform the search step first and store the location data of each co-substitution type being analysed, prior to performing the $\frac{O}{E}$ calculations.

Search combined with the statistical analysis

In the first attempt of this, a program was written which would make use of all the data amassed in the preparation of multiple sequence alignment data and the distance data, described earlier. Every pair of sequences from the alignment was searched for each co-substitution being investigated. The search was performed by considering every position in the sequences. As both

sequences had to be of equal length, this searched both sequences simultaneously. Every set of four residues present in each pair of positions was used to check against a list of mutations/substitutions. If the four residues were a mutation of interest then a counter was incremented for that co-substitution type for the distance range-bin for that inter-residue separation. A mutation in this context includes all symmetry related equivalents in the same event, i.e. (Gly, Ala) → (Pro, Ser) ≡ (Pro, Ser) → (Gly, Ala) ≡ (Ala, Gly) → (Ser, Pro) ≡ (Ser, Pro) → (Ala, Gly), as described in Chapter 2. The frequency of all co-substitution events was recorded for every range-bin in the distance-matrix. This was done by combining the residue position data with the structure-to-alignment-map generated earlier. For every pair of sequences this count data was used to calculate the $\frac{O}{E}$ for each co-substitution event in each distance range-bin, using the approach to the statistical analysis described in Chapter 2.

On examination of the steps involved, it became obvious that a major bottleneck in the process was the comparison of every pair of positions in every pair of sequences. For a sequence of length M this amounts to approximately $\sum_{x=1}^{M-1} x = \frac{M^2}{2}$ comparisons for every pair of positions in the sequence and similarly for every pair of sequences in the alignment with N sequences, $\sum_{x=1}^{N-1} x = \frac{N^2}{2}$. Thus the number of comparisons increases at a rate of $\frac{M^2}{2} \times \frac{N^2}{2}$. i.e. longer and many sequences make the search step incredibly slow. This is regarded in software design to be one of the most inefficient search algorithms known [85]. Taking nearly a week to complete an analysis of 2,000 pairs of sequences having a length of over 300 residues long, on a single computer-cluster node with 8GB RAM using a single core from a 2.4GHz AMD Opteron CPU. Considering the presence of Pfam families providing up to c.20,000,000 sequence pairs to the search space, the above approach is too slow to be of any real use.

Separating the search and statistical analysis

In the time available it was not possible to develop an intelligent algorithm which would be able to optimally search out the locations of co-substitutions. However, an attempt was made to reduce at very least one part of the bottleneck. By still relying on a comparison of all pairs of sequences, but removing the need for a comparison of all pairs of positions in the sequence-pairs, the search speed could be increased significantly.

The search method:

The revised method separated the searching for and locating of mutations from the analysis. The search of a pair of sequences for a mutation made use of a very useful function built into the Python programming language, for dealing with strings of characters. This is the index function for Python strings, which returns the index of a character in a string, from some optimised look up table. This was applied in this context as follows:

Consider a co-substitution (AB $\leftrightarrow$ CD), determining if the event occurs between sequences k and l requires locating the event by comparing residue i in sequences k and l and residues j in both sequences as well. The search is optimised, by requesting the location of the first A in sequence-k, this also returns the position we need to look for C in sequence-l. By checking if C is also present in sequence-l at the same location informs the algorithm whether to search for the location of B in sequence-k or not. In this way the location of each co-substitution-type can be located.

To account for the set of equivalent co-substitutions, the sequence-pair must be searched two or four times depending on the whether A=D and B=C or not. The set of equivalent co-substitution types includes all events which maintain (AB $\leftrightarrow$ CD). This covers all cases where residues A and B are in the same sequence, while C and D are in the other sequence and A $\leftrightarrow$ C can be at either position i or j as long as B $\leftrightarrow$ D are at the other position. To locate each equivalent mutation requires searching sequences-k and l forwards and then in reverse; then swapping the order of the sequences and again searching both ways through the sequences. However when trying to locate a symmetric mutation like AB $\leftrightarrow$ BA, this only requires searching the two sequences forwards and reverse once for a fixed order of the sequences.

The location data, i.e. position of each co-substitution, was stored in a MySQL database, which made it straightforward to later request data for a given substitution in a given family, returning the location of each position. The significant advantage afforded by storing the locations of a co-substitution type in a database was that it allowed for the analysis to be performed for the actual data and on the bootstrap data as well, without having to repeat the search of the data for each bootstrap analysis. Since, during the generation of the bootstrap data, it is only

the distances that are reassigned, not the location of the co-substitutions in the sequence data.

The statistical analysis:

The co-substitution location data stored during the search step, above, can now be analysed to calculate $\frac{O}{E}$ for each mutation, for each pair of sequences. The final $<\frac{O}{E}>$ is a weighted average across all pairs of sequences in a family and then across all families. The structure-to-alignment map described earlier, contains the distance range-bins for each pair of columns, combining it with the co-substitution location data in the MySQL databases, the frequency of each co-substitution type in a given distance range-bin for a pair of sequences was determined. The calculation of $\frac{O}{E}$ was done per-mutation for each of the surface and buried sets.

Sequence weighting is needed to ensure the contribution of those sequences between which a given co-substitution-type was observed is not under or over emphasised. If a co-substitution type is not observed in a sequence, then that sequence contributes nothing towards the determination of $< \frac{O}{E} >$ for that co-substitution type. The weighting was calculated for the full sequence, not merely the buried or surface regions of the sequence. This is because we are interested in accounting for the evolutionary relatedness of the whole domain, not merely the variability of e.g. the surface. Weighting of the full sequence is done because we want to sample equally from the phylogenetic tree.

For each co-substitution type, the set of all contributing sequences is needed in order to calculate the sequence-pair-weighting, described in Section 2.5. The set of contributing sequences, per co-substitution-type, solvation state and for a given Pfam family, could only be determined once the sequence-pairs where a mutation was observed were known. The calculated $\frac{O}{E}$ value for a mutation-type in a distance range-bin was multiplied by the appropriate sequence-pair weight. The weighted $\frac{O}{E}$ for each co-substitution in each distance range-bin was stored in a table in the MySQL database. For each Pfam family, the sum of all weighted $\frac{O}{E}$ for a given co-substitution-type was calculated for every distance range-bin, providing the average $\frac{O}{E}$. The final values, shown in the results section is an average of the average value calculated for each Pfam family. The average of averages is arrived at by calculating the sum of all $< \frac{O}{E} >$ in a range bin for each family and dividing by the total number of families in which the co-substitution

was observed in that distance range-bin.

The analysis for each co-substitution type in each of the selected Pfam families, was firstly performed using the actual distance data from the reference structures. The randomised distance-matrices were then used to calculate the bootstrap data. During the generation of the randomised data, the solvent accessibility does not change, but the occupancy of a given distance range-bin does. Since the sequence pair weights are calculated per sequence, not per distance range-bin, the weightings only needed to be calculated once for each co-substitution type per Pfam family.

5.3 Software Testing

The development of the software for the analysis required extensive testing to ensure confidence that the presented results are correct. Testing took several forms at different stages of the development. Firstly each individual method/function was tested as it was written to ensure the output was correct. Secondly, each class-file was tested with test data to ensure that all methods/functions were working correctly with each other and the class-file was returning the correct output and performing as expected. Finally all the class files and scripts had to be tested as per the work-flow to ensure that they were working correctly together. This was done by compiling a small test set of data and passing it through the work-flow. The input and outputs had been worked out manually and the program output was checked against these results. The tests comprised the use of the sequences in the example given in Section 2.5.2 and verifying that the program's output with manually calculated results.

5.4 Results

There are a total of 22,155 pairs of residue-types that could be involved in a co-substitution event, including conservations. Evaluating all types will take a a considerable amount of time. Due to time constraints only a single co-substitution event was calculated for: RD $\leftrightarrow$ KE (incomplete results for one other co-substitution type is shown in Appendix H). This co-substitution type is a conservation of charge at both sites, but with both substitutions resulting in local volume changes. The co-substitution does maintain an electrostatic attraction.

The results show firstly, in Figures 5.4 and 5.5, two plots showing a $log_2\langle\frac{O}{E}\rangle$ line for each Pfam family from the eukaryota and prokaryota data respectively. This is followed by Figures 5.6 and 5.7 the results for average over multiple families, with a $log_2\langle\frac{O}{E}\rangle$ line for 45 eukaryotic families and one for 50 prokaryotic families respectively; both include a line of $log_2\langle\frac{O}{E}\rangle$ for the average of 100 bootstrap analyses. A technical fault meant that the 46^{th} eukaryotic family and the 51^{st} prokaryotic family had to be excluded from the result. Figures 5.8 and 5.9 are the plots with an individual $log_2\langle\frac{O}{E}\rangle$ lines for each bootstrap data set performed on the eukaryota and prokaryota data respectively. Finally two plots showing a merged data-set composed of all eukaryotic and prokaryotic data averaged together are presented. Firstly the $log_2\langle\frac{O}{E}\rangle$ data from a total of 78 Pfam families taken with the average of 99 bootstrap analyses, is shown in Figure 5.10; in Figure 5.11 is shown the data for the same 78 Pfam families, with the individual $log_2\langle\frac{O}{E}\rangle$ plots for 99 bootstrap analyses.

In the plots shown in Figures 5.6, 5.7 and 5.10, which show the plot of $log_2\langle\frac{O}{E}\rangle$ with the average $log_2\langle\frac{O}{E}\rangle$ for the 99 bootstrap analyses, the red line with points represents the plot of actual data; the magenta line represents the average of the bootstrap data. In Figures 5.8, 5.9 and 5.11 the plots showing a single line for each bootstrap analysis, the red-line with points shown in other figures is also included, but has been drawn over to highlight where it deviates greatest from the bootstrap data.

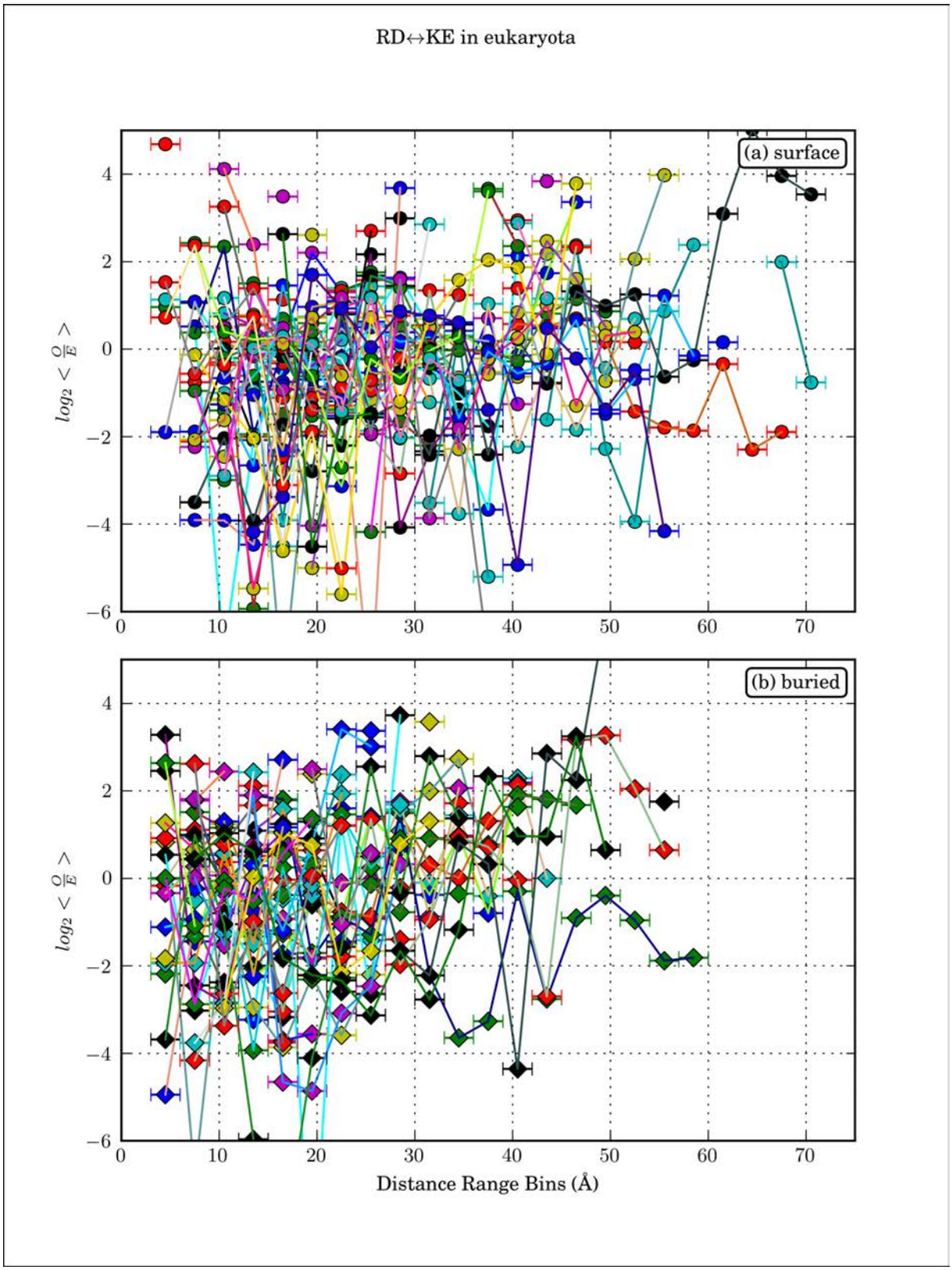

Figure 5.4: Co-substitution propensities of RD ↔ KE in individual Pfam families, derived from eukaryotic sequences: A single line is shown for each of the 45 Pfam families included in the analysis. The points show the $\frac{O}{E}$ value for an individual Pfam family in a given distance range-bin, with width 3 Åindicated by the error bar.

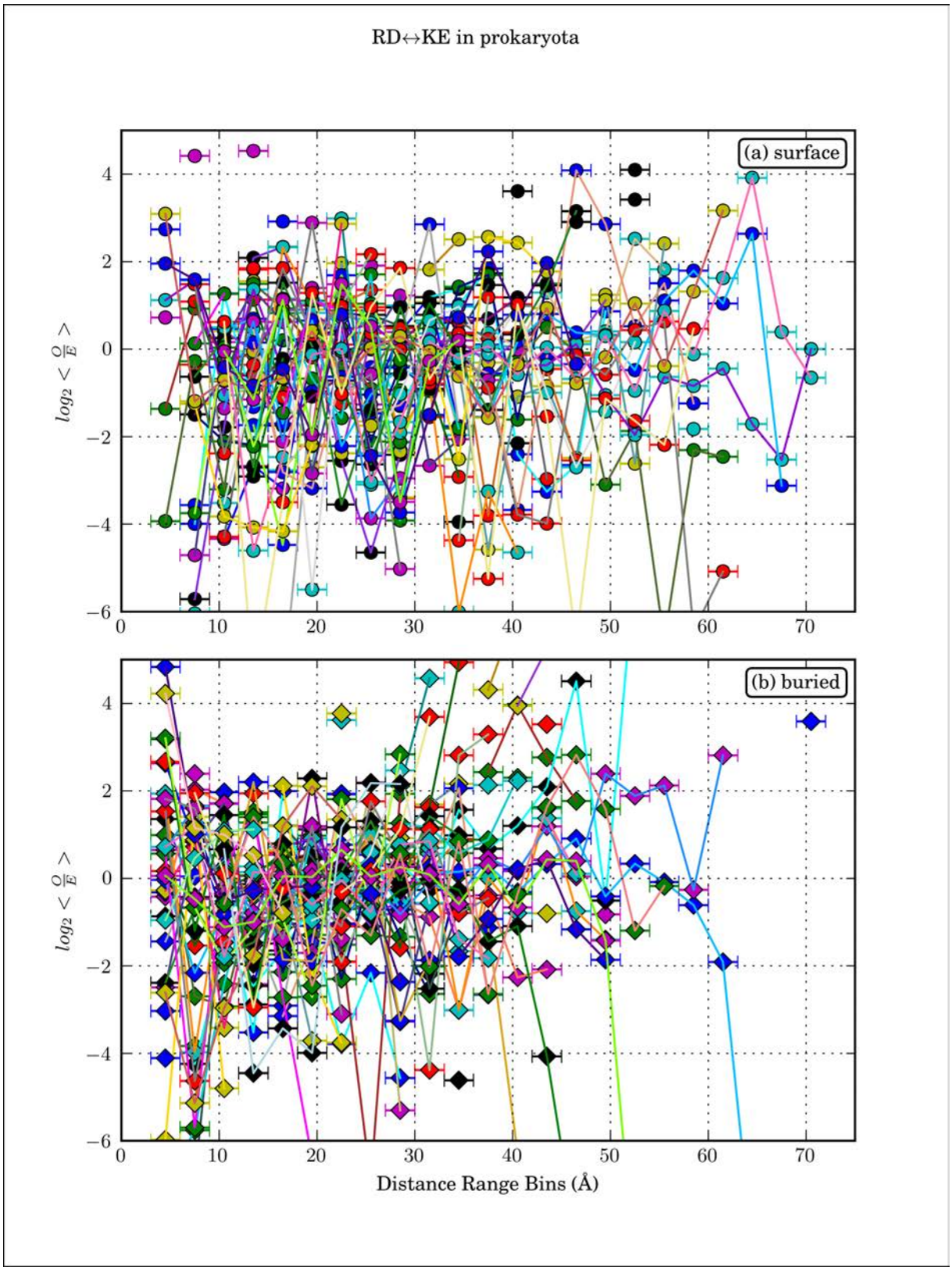

Figure 5.5: Co-substitution propensities of RD ↔ KE in individual Pfam families derived from prokaryotic sequences: A single line is shown for each of the 50 Pfam families included in the analysis. The points show the $\frac{O}{E}$ value for an individual Pfam family in a given distance range-bin, with width 3 Åindicated by the error bar.

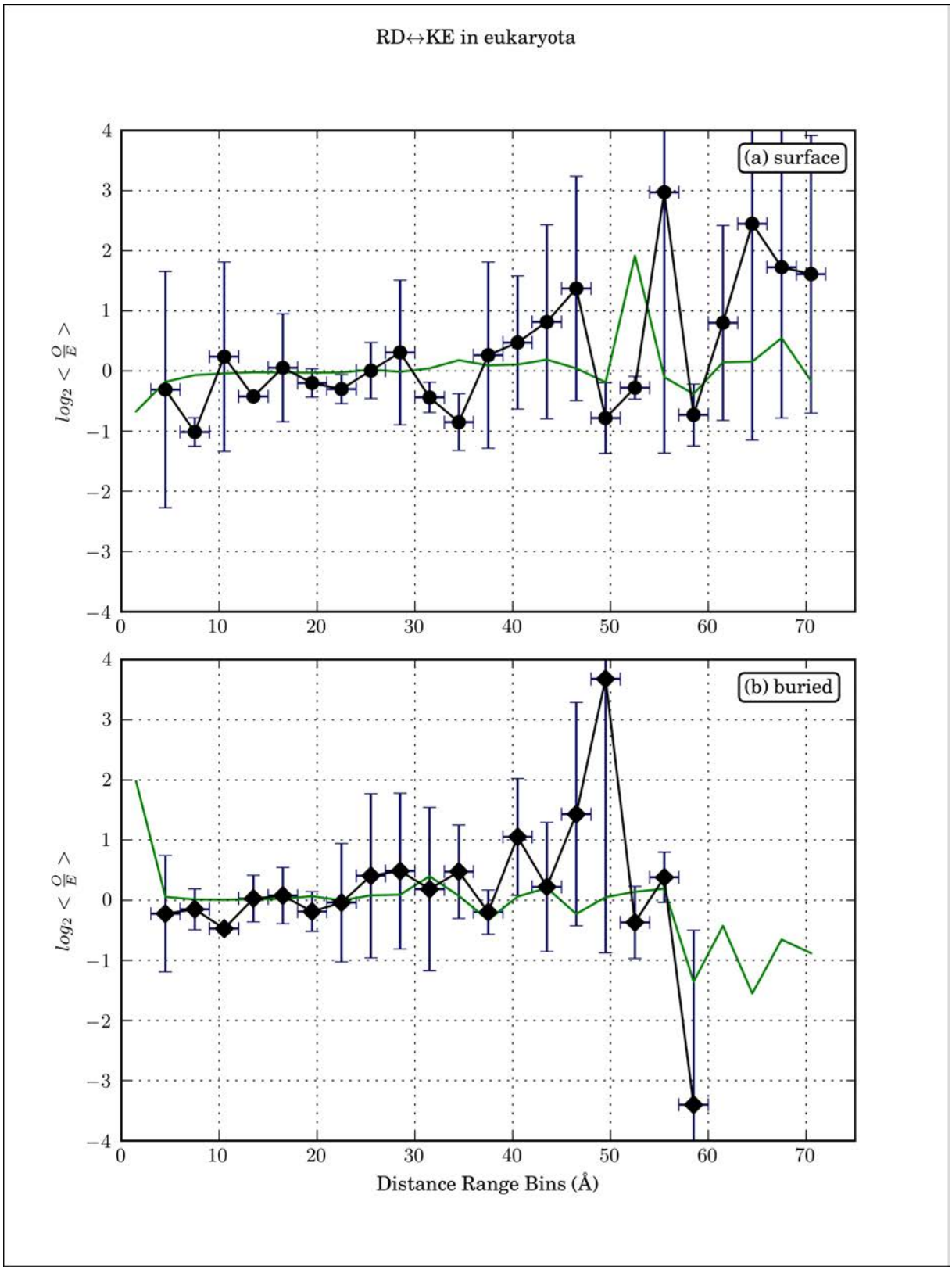

Figure 5.6: The average co-substitution propensity RD ↔ KE derived from eukaryotic sequences: The black line with points show the average of 45 Pfam families and represents the independence of a 3 Årange-bin, indicated by the horizontal error bar. The vertical error bars are the log_2 of the standard deviation of $\frac{O}{E}$ for each Pfam family (shown in Figure 5.4). The average $\frac{O}{E}$ of 99 bootstrap analyses is shown in green.

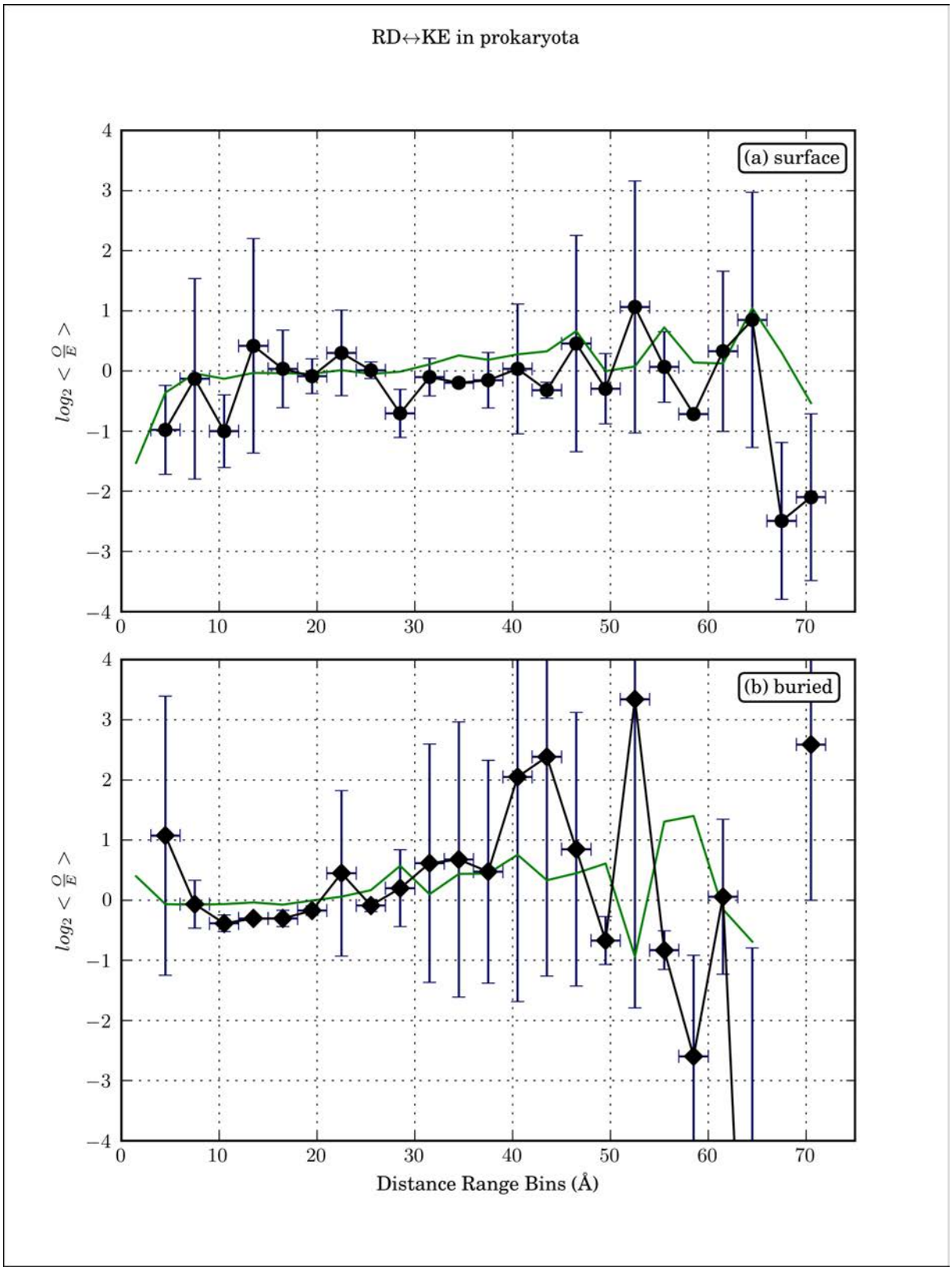

Figure 5.7: The average co-substitution propensity RD ↔ KE derived from prokaryotic sequences: The black line with points points show the average of 50 Pfam families and represent the independence of a 3 Årange-bin, indicated by the horizontal error bar. The vertical error bars are the log_2 of the standard deviation of $\frac{O}{E}$ for each Pfam family (shown in Figure 5.5). The average $\frac{O}{E}$ of 99 bootstrap analyses is shown in green.

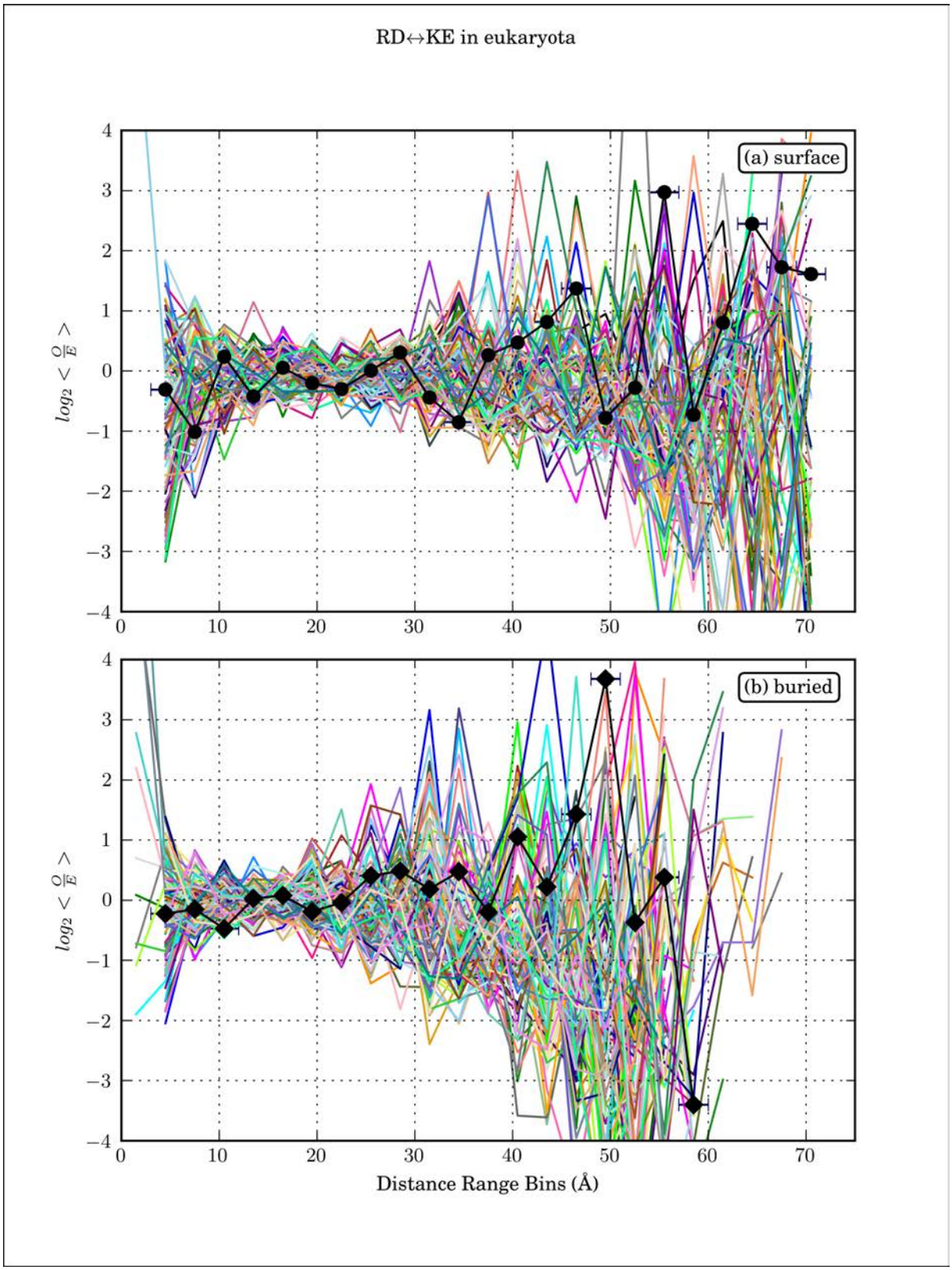

Figure 5.8: Co-substitution propensity RD ↔ KE for 99 bootstrap analyses, derived from eukaryotic sequences: The individual lines shown represent the average $\frac{O}{E}$ values calculated from a randomised distance matrix for each Pfam family. The Bootstrap data is incomplete for 4 families, however it is included here to show the behaviour of the bootstrap data.

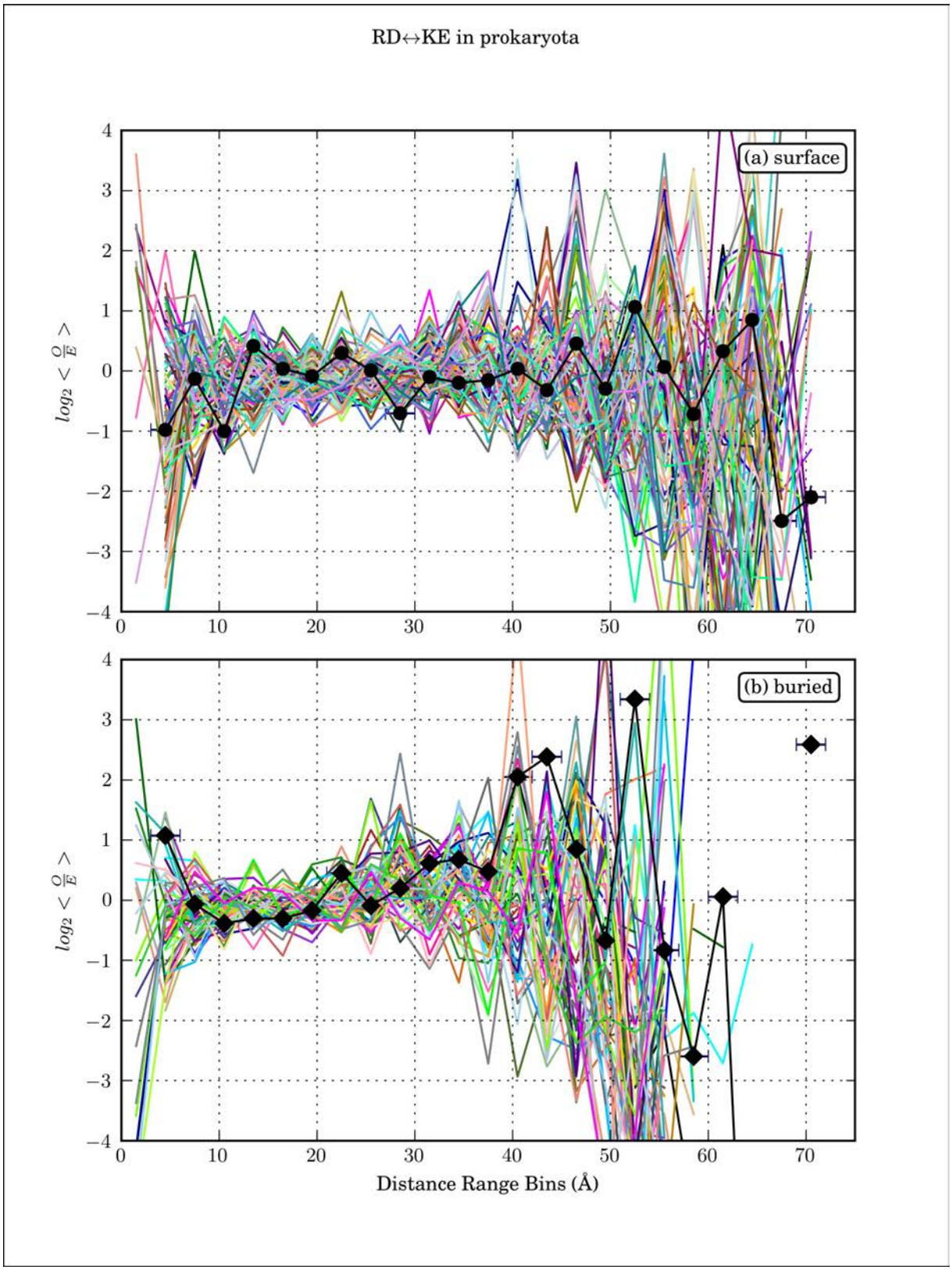

Figure 5.9: Co-Substitution propensity RD ↔ KE for 99 bootstrap analyses, derived from prokaryotic sequences: The individual lines shown represent the average $\frac{O}{E}$ values calculated from a randomised distance matrix for each Pfam family. The bootstrap data analyses was not completed for 13 families, however it is included here to show the behaviour of the bootstrap line.

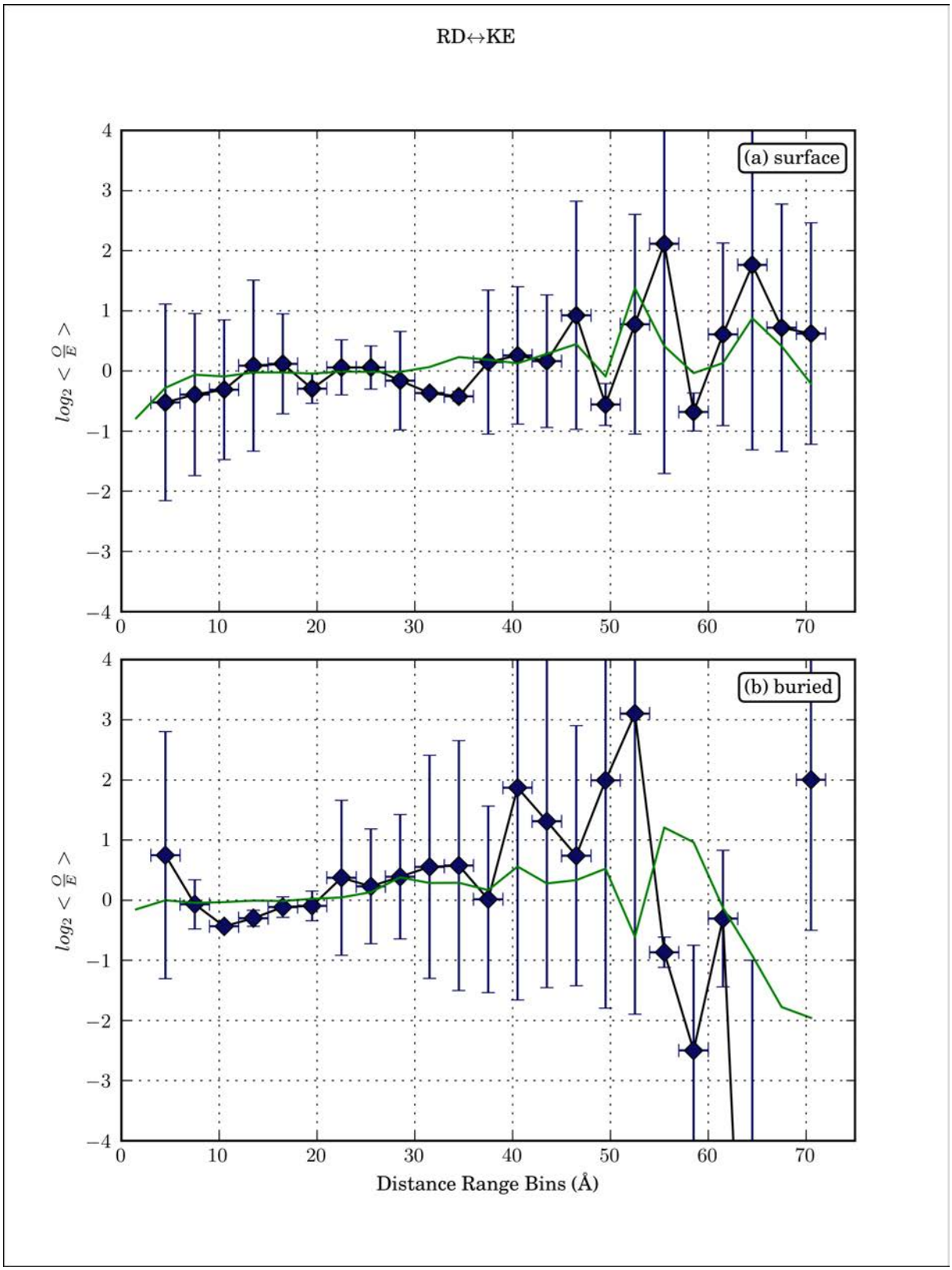

Figure 5.10: Co-substitution propensity RD ↔ KE derived from the merger of eukaryotic and prokaryotic data at distance increments of 3 Å.

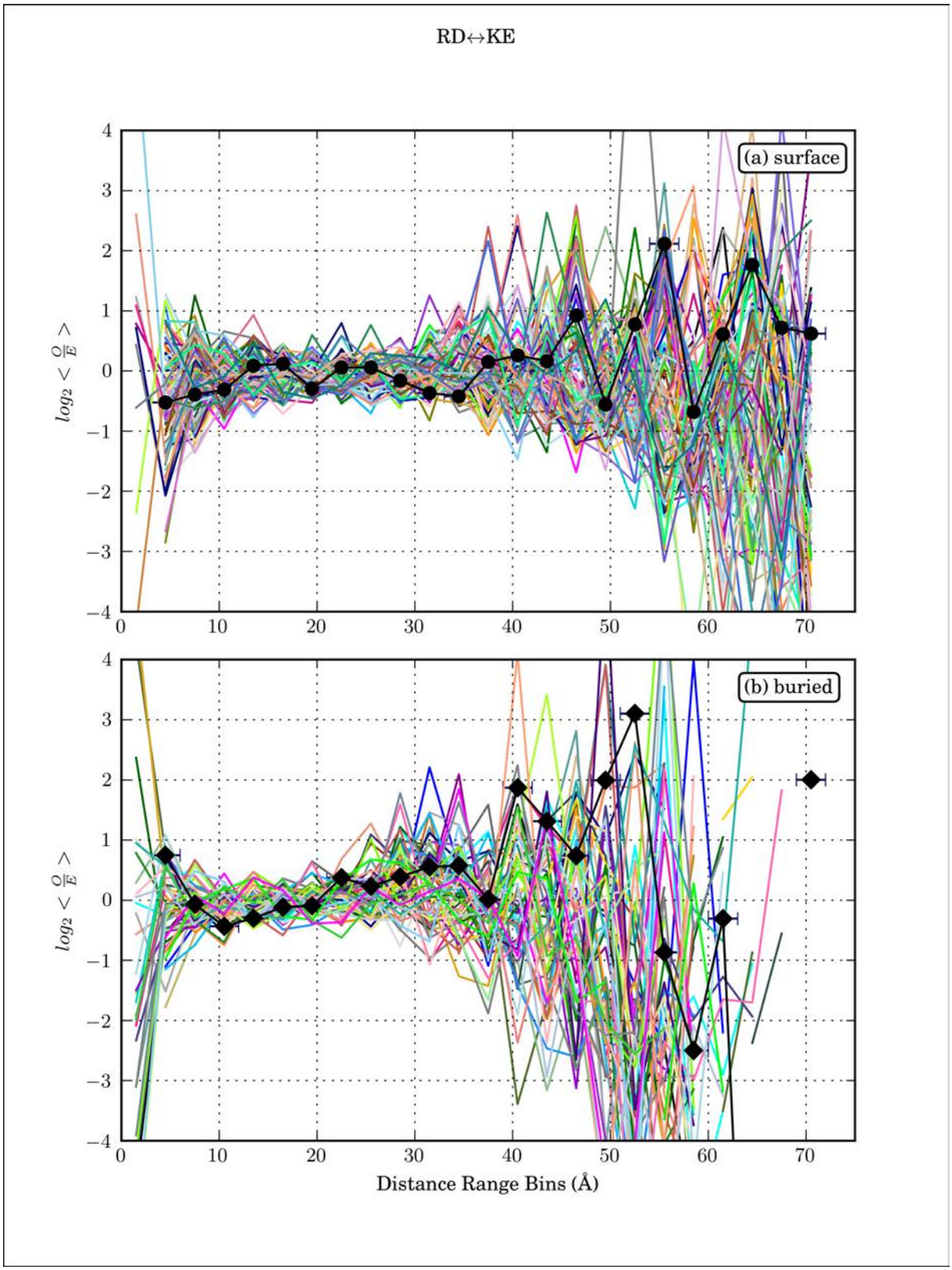

Figure 5.11: Co-Substitution propensity RD ↔ KE at distance increments of 3 Å, showing the average for 78 Pfam families, with 99 bootstrap lines.

5.5 Discussion

5.5.1 Discussion of methodology

In the methods section two different methodologies were described, which were used to determine the propensity for co-substitution types to occur in the data. The first method attempted to search every pair of sequences from a Pfam family (that had been filtered as described in the methods) for the occurrence of any and all co-substitutions from a list of co-substitution types. On completion of searching a given sequence pair, the program would then perform the statistical analysis to determine the $\frac{O}{E}$ for each observed co-substitution type in that sequence pair. The second method broke the operation of the first method into two parts. Firstly the multiple sequence alignment was searched for the location of specified co-substitution types, which was stored in a MySQL database. Secondly the location data of each co-substitution type was retrieved from the database, which was used to determine the $\frac{O}{E}$ for each pair of sequence in a given Pfam family, where the co-substitution type was observed. The results presented in the Section 5.4, were produced using the second method.

Both methods were successful in their operation. However the first method was not optimal for performing bootstrap analyses as well as the analysis of the regular data. Having completed the operations necessary to produce results for the regular data, it would have to repeat the entire operation for each bootstrap analysis. This made the method prohibitively slow and so the decision was made to make efforts to reduce the number of operations that the program would need to complete in order to produce results inclusive of bootstrap analyses. By searching for and storing the location data for a given set of co-substitution types in each Pfam alignment, the program would only have to request the location data for each co-substitution type in a family, to calculate $\frac{O}{E}$. As such the number of times a pair of sequences would need to be searched for a co-substitution type was reduced to one. Yet, the amount of storage space required for a small set of co-substitution types was extremely large, and the time taken to calculate $\frac{O}{E}$ for a small set of co-substitution types was considerable, both leaving scope for substantial improvement.

At present there are three challenges that need to be addressed. First an issue concerning

the statistical significance of any results produced by the method. Addressing this requires consideration of the data selection steps taken and anything that might be done to increase the amount of data being analysed. Secondly an issue of disk space to store location data for co-substitution events in the sequence data. Finally the issue of the time required to complete an $\frac{O}{E}$ analysis for a given Pfam family. Addressing these challenges in the first instance, should facilitate further development of the project, to consider a greater number of Pfam families and more co-substitution types.

Issues of statistical significance and data selection

The individual trend lines for each Pfam family in Figures 5.4 and 5.5 show considerable variation between families. This indicates that the available data for the co-substitution type RD $\leftrightarrow$ KE was sparse. The Figures 5.8, 5.9 and 5.11 of the individual lines for bootstrapped data, suggest that increasing the amount of data being analysed will improve the reliability of the results.

The number of sequence-pairs in which a given co-substitution type is present in the sequence alignment for a given Pfam family, contributes to the variation observed between individual Pfam families in Figures 5.4 and 5.5. Therefore, when considering the statistical significance of the $\frac{O}{E}$ for a given co-substitution type, it is necessary to consider the number of sequence pairs in a sequence alignment between which the co-substitution was observed. This will likely place a considerable constraint on the number of co-subsitution types for which the analysis could be completed, with existing available data.

However, the limiting factor on the number of Pfam families considered, resulted from the decision to segregate co-substitution events on the basis of solvent exposure. The set of data was originally selected for the solvent exposure analysis discussed in Chapter 4. In the context of that analysis, the use of the PiQSi database and the use of protein domains in the biological unit was necessary to ensure that solvent exposed regions, were indeed solvent exposed and did not include the interfaces between units in the protein complex. In the context of the co-substitution analysis presented here, that imperative is no longer applicable. In fact, the most straightforward approach to increasing the amount of data in the co-substitution analysis, would be to source

the reference structures from the PDB and concentrate exclusively on the buried regions of the protein. The surface regions of the proteins are the most likely to be affected by environmental considerations, while the protein interior is less likely to be. Therefore concentrating on the buried regions of proteins would allow for all Pfam families which have structural examples in the PDB (without concern for the cellular environment), and have a minimum number of sequences to be investigated. However, the removal of homo-oligomers would still be necessary, though it would be possible to remove homo-oligomer sequences from the Pfam alignment, if it was necessary to get more data. Taking these steps could significantly increase the amount of data being analysed, and could potentially increase the available data for each co-substiton type, which would result in increased confidence for the results produced.

Technical challenges of data storage

Data storage poses a significant barrier if the statistical analysis of the pre-searched co-substitutions were to be conducted in a single step on all available data. The location data alone for 10 co-substitution types from the 97 Pfam families (46 Eukaryota and 51 Prokaryota) considered, approached 250 GB.

Extending the method to a larger set of Pfam families, e.g by considering all those for which structural exemplars exist, which number over 5,000, would pose a significant technical challenge. Storing location data for the same 10 co-substitution types in 5,000 Pfam family sequence alignments could require more than 12,000GB. Needless to say investigating the entire set of possible co-substitution types would require storage facilities that exceed the technological capabilities of currently available storage solutions. It should be noted that this implies trying to record and store the data as was done with the second method described in Section 5.2.3 Though it may be possible to make huge optimisations, it remains a technical challenge. Avoiding this challenge is the most parsimonious approach. As the storage of the location data was done to speed up the bootstrap analysis, then storage is something that could be left out if a decision were made not to bootstrap.

Technical challenges of optimised searching

The time required to search for a set of 10 co-substitution types in the largest Pfam family included in the selected data (roughly 22 million pairs of sequences), using the Python index function, was less than 3 days. The search was performed on a single computer-cluster node with 8GB RAM using a single core from a 2.4GHz AMD Opteron CPU. This does not suggest a significant impediment to the current method, however any reduction in operation time would always be useful.

A possible optimisation to the method would be to use the Python index function to search the columns of the alignment. Rather than searching the sequence pairs, it is possible to search columns from the sequence alignment and locate sequence-pairs with the co-substitution type; this could reduce the number of sequence comparisons that need to be made, reducing the search time necessary to locate each co-substitution type in a family. If the set of co-substitution types were to be expanded to include all 22,155, this optimisation could make it more feasible to complete the search in a reasonable time. However, this would only be optimal if the number of sequence pairs was greater than the number of column pairs in the alignment. The computer science literature may contain other more efficient methods.

However, reducing the number of co-substitution types being searched for would also reduce the overall run-time. The question needs to be addressed as to whether it is entirely necessary to search for all possible co-substitution and conservation types. In the development of a Bayesian method for structure prediction it might be sufficient to have only a small number of co-substitution types to generate a prior, since e.g. co-substitutions that conserve electrostatic attraction may behave in a similar fashion to each other and be representable by one prior. Though selecting co-substitution types on the basis of their general prevalence in the data would be helpful, doing so would require knowing *a priori* the frequency of each co-substitution type in the data. In the absence of this information, substitution matrices, which estimate substitution propensities on the basis of physico-chemical properties, can provide an indication of the prevalence of a given substitution type in general. These could be used to remove the substitution types which are less likely to occur and thus inform the selection of substitution types to

pair together for statistical analysis.

Technical challenge of time required for $\frac{O}{E}$ analysis

The largest bottle-neck with respect to run time has been the $\frac{O}{E}$ analysis for each co-substitution type in each Pfam family, using the location data. The analysis for the co-substitution type (RD $\leftrightarrow$ KE) was loaded on to a computer cluster. The analysis of the filtered alignments for Pfam families was loaded on a single CPU core for each family. The analysis including the 100 bootstraps for the Pfam family with the greatest number of sequence pairs (c. 22,000,000) for a singe co-substitution type, failed to complete within 6 weeks, at which time a technical fault terminated the program.

The most parsimonious solution to this bottleneck is to make the operation more parallel. Currently individual Pfam families are being assigned to a single CPU core for analysis. if the sequence pairs for each family were distributed across multiple CPU cores and computer cluster nodes, this could greatly reduce the overall run time for the analysis.

The challenge of a co-substitution analysis with more data

The current state of the project poses some immediate considerations, detailed earlier. However, if the project were to encompass a greater amount of data and consider a greater number of co-substitution types, then a re-framing of the problem could produce a more complete solution and result.

Consider two parameters set by the problem: the number of co-substitution types and the number of available Pfam domains with structural exemplars. There are 22,155 possible co-substitution types and there are over 5,000 Pfam domains for which structural exemplars are available, this number will invariably increase with time. These set the current maximum conceivable data that could be considered. Though a selection of Pfam families based on the number of sequences they contain, would likely reduce this somewhat. With the analysis method in its current form, analysis of the propensity for all co-substitution types in 5,000 Pfam families would require that the analysis be performed over 100 million times. Storing the location data would literally be impossible and analysis time would be prohibitive with the currently avail-

able resources. Therefore reconsidering the work-flow in the context of significantly larger data considerations, is necessary. Which, due to time considerations it was not possible to do. Discussed here is an alternative approach to the search and calculation of $\frac{O}{E}$ for each co-substitution type, which could address the run-time issue.

The two challenges that need to be addressed in re-framing the problem are, computational time and storage requirements. Any reworking of the method must improve on both these concerns to offer a more complete solution.

Firstly, it should be noted that the pre-processing step, of filtering the sequence alignment data and generating the alignment-sequence-to-structure map is a fairly rapid process which completed in under 8 hours cumulatively on a 2 year old Linux workstation with an Intel Core i7 CPU and 16GB RAM, for all 97 Pfam families. This data selection step could be distributed across multiple CPU cores and as such would not pose a bottle neck in the analysis process.

The question was raised in Section 5.5.1 as to the need for calculating $\frac{O}{E}$ for all co-substitution types. However, the question does not considered the actual co-substitution types that are present in the data. Even though there are 22,155 possible unique combinations of co-substitution types, only a subset are likely to occur in any given Pfam alignment. Therefore as a first optimisation step, it could be useful to only analyse co-substitution types which are present in a given Pfam family.

The most direct approach to retrieving an assessment of co-substitution types in a selected data-set, would be to redirect the emphasis to recording the frequency of all observed residue-pairings in a given pair of sequences, for a given Pfam family. Thus determining the $\frac{O}{E}$ with respect to distance for all co-substitution types in a given sequence pair - which will either be co-substitution events or conservations. This does not require the storage of the location data; the frequency of each observed residue pairing would need to be recorded, though only temporarily until the $\frac{O}{E}$ for each has been determined unless it is wanted for further analysis or assessment. This can be achieved with the slight modification to the source code used for the first method described in Section 5.2.3.

However, determination of the sequence-weighting for given co-substitution types, does

require a record of the sequence pairs in which each co-substitution type was observed. This could pose problems with respect to storage requirements, but the general considerations of the storage requirements would be greatly reduced as specific location data would no longer be required and so the impossibility of the challenge will have been removed.

The analysis is of a type which can be referred to as an embarrassingly parallel problem. A short coming of the current method, as stated earlier, has been the use of a single CPU core for each Pfam family. This has allowed for the analysis of smaller families to complete and thus leave computing resources idle, while another CPU core continues to process a family with many sequences. There is a need to amend the work-flow to increase the use of multiple resources in parallel. It would not require a significant effort to alter the work-flow such that the analysis of the data from a single Pfam family is distributed to multiple CPU cores. Thus determining the $\frac{O}{E}$ with respect to distance for all observed co-substitution types in a pair of sequences, from a given family could be achieved in a reduced time scale, even if bootstrapping was performed; which would be worked into the distributed work-flow.

How much faster this alternative work-flow could be, would need to be investigated, but it has the potential to be considerably faster at determining $\frac{O}{E}$ for all possible co-substitution types in a given data-set. It would also address a slightly different question, at the same time: *'which are the most prevalent co-substitutions types in the available data?'* This is not the same question addressed by similarity/substitution matrices, which only considers the individual substitutions. Additionally this could make it possible to side-step the potential need to determine co-substitution types of interest from substitution matrices; allowing for an evaluation of the significance of the co-substitution types based on their prevalence in the data. Additionally the work-flow could be used to produce a "co-substitution matrix" to complement existing substitution matrices.

5.5.2 Discussion of Results

The plots shown in Section 5.4 present the $\frac{O}{E}$ data for the co-substitution type RD $\leftrightarrow$ KE. The plots, were separated into co-substitutions on the surface residues and those which were buried.

This was done for sequences from Eukaryota and Prokaryota separately and the two Eukaryota and Prokaryota data sets were later merged. Data for the bootstrapping was also shown.

The plots of the average bootstrap line and of the individual bootstrapped lines, have data points in the distance range-bin 0-3 Å. This may appear curious, considering the inter-residue distances measured was between C_β atoms; inter-atomic distances are normally measured from the centre of atoms, and the radius of carbon is generally accepted to be 1.7 Å, thus two carbon atoms, in direct contact should have an inter-atomic distance of 3.4 Å. Anything less than 3.4 Å would suggest an overlap of van der Waals radii. As such, direct contact between residues occur in the distance range-bin 3-6 Å. We have included in our data glycine; the inter-atomic distance measured between a glycine in direct contact with any other residue-type, will be between a hydrogen atom and a carbon atom, which is less than 3.0 Å, putting any of these distances in the distance range-bin 0-3 Å. The bootstrap data is a random reassignment of the observed distances in the reference structure, with replacement, therefore the distance between glycine and a contacting residue can be assigned multiple times to other residue pairs. It should be noted that any co-substitution event observed within the range 0-6 Å can be said to be in direct contact. Others have considered residues at 8 Å apart to be in direct contact, but these measure the distances between C_α atoms [76].

The separation of the data into two distinct populations of co-substitutions between buried residues and those occurring on the surface allows for a comparison of their respective behaviour, for the first time. In the individual plots for Eukaryota and Prokaryota, it can be observed that there is a difference in their behaviour. Firstly, if considering the plots in Figure 5.4 and 5.5, showing the data for each family individually, the distribution of the data points for the surface are different to the data points for the protein interior. The buried data is more concentrated for shorter range interactions compared to the surface. The surface has inter-residue interactions in excess of 60 Å while the buried data does not. This is to be expected as residues on the surface can obviously be farther apart. However visual inspection of the plots suggest that the co-substitution events observed in the buried data are more concentrated below 40 Å, while the surface data is less concentrated.

Direct contact between residues would cover distances from 3-6 or 3-7 Å, which covers data in three range-bins: 0-3, 3-6 and 6-9 Å. On the graphs it would be the first two or three data points that would indicate direct contact or very close proximity between residues. When considering the co-substitution type for which data is presented: at close distance RD and KE will both form salt-bridges. Both would need to be involved in salt bridges to neutralise their respective charges, to be in the buried state. However on the surface the presence of a salt bridge could interfere with the flexibility or movement of secondary structure units; given that some structures exhibit a 'breathing' like motion which could be prevented by the presence of a salt bridge. Inspection of the $log_2\langle\frac{O}{E}\rangle$ data shown in Figures 5.6, 5.7 and 5.10, suggests a difference of behaviour between the surface and the buried data for direct contacts. Though it should be noted that the error bars on graphs and the bootstrap data indicate that the data is too sparse to provide statistically signficant results. What possible signal is visible in the graphs currently will need to be investigated further with a larger data set. Each of the graphs indicate that on the surface there may be a slight dis-preference for seeing this co-substitution pattern for contacting residues, although it is close to having no preference/dis-preference. While the buried data for Eukaryota indicates a dis-preference, the prokaryota and the merged data indicate a preference.

A further difference between the two populations is visible in the buried data for all the plots. There appears to be an indication that there is a preference for the co-substitution to occur at longer non-contact separations in the buried data. The $log_2\langle\frac{O}{E}\rangle$ is greater than zero for an extended period and there is a distinct point visible for the distance range-bin 39-42 Å in the eukaryotic data and in the range 39 - 45 Å for prokaryotic data and in the merged data taken from 78 eukaryotic and prokaryotic Pfam families, shown in Figure 5.10. This preference is visible in all of them and is not apparent in the bootstrap data. The bootstrap data shown in Figures 5.8, 5.9 and 5.11 show this region to be the beginning of the sparse data, which suggests that it could be a random signal arising from insufficient data. The error bars for the same regions also indicate great varitation between Pfam families, indicating that more data is required to determine the true strength of the observed signal. However, its presence in Eukaryota and Prokaryota, as well as different size samples of the merged data, suggest that

the preference indicated is genuine. The implication of this result, is that there appears to be a distance effect on the co-substitution type RD↔KE at distance of about 40 Å which is roughly 4 times greater than random.

Although non-contacting correlated mutations are seen, the literature tends to focus on predicting contacts. Therefore the long distance signal is quite surprising. If we consider the implications of the observation: charged residues in the buried state are expected to be in salt bridges since it is a hydrophobic environment. Yet, it would appear that these residues are still capable of interacting with each other over this distance. The dielectric constant of the protein interior is thought to be significantly less than that of water, which would allow for charged residues to affect each other through the protein. This begs the question as to whether the $HSEu_{13}$ value determined to delimit residue burial is including residues which are still just on the surface and thus not in salt-bridges or if something else might be responsible, e.g. the dipole of a salt-bridge, or some effect associated with the folding process. To investigate this would require repeating the analysis with different $HSEu_{13}$ values as a cut off, or separating the data into three solvent exposure regions: surface, not-quite-buried and buried. This requires further consideration.

However, a discussion on the implication of the result is predicated on an assumption that the result is in fact genuine and statistically significant. Though it appears that a signal does exist and it is not simply random noise in the data, the signal appears weak and assessing the statistical significance of the result remains a challenge. The points of interest, indicative of a long-range interaction in the buried data, occur in a region where the available data is beginning to be sparse. The question as to whether the statistical significance of individual $\frac{O}{E}$ points, i.e. the value for specific distance range-bins, can be measured in isolation to all other points for that data set or if they must be viewed in the context of all other points, is unclear. Thus, it is difficult to say that a single point that sits outside of the bootstrap data, is statistically significant or not, as is the case for the point for the distance range-bin 39-42 Å in the buried data.

Yet this question really only arises due to the uncertainty over the amount of data necessary to ensure confidence in the results. The plots showing the individual bootstrap lines for

eukaryotic and prokaryotic data sets in Figures 5.8 and 5.9, has a greater variation between bootstrap lines than the bootstrap data for the merged data for the 78 Pfam families with individual bootstrap lines shown in Figure 5.11. The merged bootstrap data is smoother than the plots in Figures 5.8 and 5.9, which suggests that increasing the number of Pfam domains and the number of sequence pairs per co-substitution type being the analysed will improve the statistical significance of the results. This is supported by vertical error bars in the plots, which are the standard deviation of the individual Pfam families, for each range-bin. These show that there is some wide variation in the data between families; as such it is necessary to consider expanding the number of families being analysed.

Another consideration is apparent on examination of the plots of $log_2\langle\frac{O}{E}\rangle$ for individual Pfam families, shown in Figures 5.4 and 5.5. A great degree of variation between families exists in the plots. Each line represents the occurrence of the co-substitution type between sequence pairs in a single sequence alignment, i.e. Pfam family. The number of sequences in a sequence alignment and the number of sequence-pairs between which a co-substitution event is observed, have implications on the significance of the result for that family. It is expected that with a sufficiently large number of families, the uncertainty introduced by lower numbers of sequences and observations of a given co-substitution type in a family, will be compensated for as the number of families investigated increases. However, [29] set a limit on the minimum number of sequences in their data selection method, to 1,000 sequences in a given sequence alignment. Here, a limit was set at a minimum of 400 pairs of sequences, which is significantly fewer individual sequences than [29], and was a consequence of all other selection criteria having whittled down the number of the usable Pfam families to such a small number.

Therefore, two considerations that must be addressed are, firstly the number of Pfam families being investigated needs to be increased. Secondly a limit must be set for the minimum number of sequence-pairs being considered in any multiple sequence alignment being used for the analysis for a given co-subsitution type. This is achievable, but requires some consideration in at the data selection stage of the work-flow. However, this would significantly reduce the number of usable families in the current data.

Finally, the analysis was run for more than one co-substitution type, a total of 10 were searched for however the analysis was cut short by a technical fault. Thus the analysis did not complete for all the co-substitution types. In Appendix H, a plot is shown of the co-substitution of IL $\leftrightarrow$ LV, in the Eukaryota. This has not been included in the results section here because the analysis was not completed. As with the RD $\leftrightarrow$ KE data, it is apparent that there is not enough data to make any strong claims about the behaviour of the co-substitutions with repsect to distance.

5.5.3 General Discussion

In the literature, co-evolution analyses attempting to predict inter-residue contacts consider the distance between C_α atoms to measure inter-residue distances. The method developed for this thesis is different because it considers the distance between C_β atoms. This has the advantage that it can define contacts between side-chains rather than between residues as is the case with inter-C_α distances; the distinction is subtle but the latter cannot distinguish between a backbone contact and a side-chain contact, whereas the former can define clearly a side-chain contact.

A short-coming of co-evolution analyses attempting to predict direct contacts, beyond the low success rate in predictions, lies in the statistical significance of the available data for direct contacts. In the protein structure the number of direct contacts will be significantly fewer than non-contacts. Consequently the majority of information that can be analysed is between non-contacts. These methods rely on a binary distinction between contact and not-contact. The method developed here provides a more flexible approach, offering a less rigid distinction between contact and non-contact; additionally providing insight into long-range interactions, and providing the possibility of including the uncertainty associated with a predicted inter-residue distance via a probability distribution function.

As extensive efforts have been made to keep the data-sets uniform to adhere to a notion of statistical rigour, the amount of data from which the results have been derived could impact the statistical significance of the result. The bootstrapping used can only provide a comparison of our results with respect to random. In the presented results an additional step which could be

used as an indication of the significance of the results, would be to average over a randomly selected subset of Pfam families and plot the $log_2\langle\frac{O}{E}\rangle$ data. Repeating this process up to 100 times and determining the correlation co-efficient between each plot, would allow us to determine which trends are generalised in the data. This has not been done due to time considerations.

The question that is of greatest consideration remains: *'is this analysis worth the effort?'*. The benchmark against which such an assessment can be made lies with the usefulness of the result to make predictions. Conventionally this means predictions of protein structure. With the $P(d|s)$ data generated by Mr. Welland, combined with the $\frac{O}{E}$ data generated by the analysis, it is possible to start making structure predictions, with some additional software development. Though it would be necessary to analyse the $\frac{O}{E}$ of more co-substitution types. A successful structure prediction or even a significant improvement in prediction would suggest that the analysis is worth the effort.

5.6 Future development

The co-substitution analysis presented in this chapter, would benefit from a redesign of the software work-flow, as discussed earlier. Combined with a modified data selection, a more complete picture of the propensities of co-substitutions with respect to distance can be achieved. Further the production of co-substitution matrices to complement substitution matrices, can also be compiled

A useful feature of the $\frac{O}{E}$ function is its relationship with Baysian statistics. Combined with the $P(d|s)$ determined by Mr. Welland, the results from the co-substitution analysis can be used in the prediction of protein structures. This requires further investigation and development but is the next part of this project that will be developed.

Finally, the method will be applied to investigate protein-protein interaction and potentially it will be used to predict these interactions.

5.7 Acknowledgements

Mr. Matthew Welland, was an undergradute student in his final year. He worked on calculating $P(d|s)$ from the protein structures that had been mapped to the multiple sequence alignments used for the co-substitution analysis. His work made possible the data to generate figures 5.2 and 5.3. He created both those figures for at the request of Mr. Bhima Auro, for this chapter.

Chapter 6

Conclusions

The main objective of the thesis has been the development of a statistical framework to investigate co-substitution events. This has been achieved. Each of the above mentioned tools, observations and results collectively contribute to the statistical framework developed.

The analysis undertaken has revealed differences between the co-substitution propensity of RD $\leftrightarrow$ KE on the protein surface compared to the interior. Additionally there appears to be a difference in co-substitution behaviour, for the same co-substitution, between Eukaryota and Prokaryota, though this requires further investigation.

The method that has been developed is capable of determining the propensity for co-substitution events to occur with respect to distance, for a set of specified co-substitution types. Steps need to be taken to increase the amount of data analysed to improve the statistical significance of the results. With some further development the method could be expanded to investigate large sets of data and many more co-substitution types.

Continuation of this work will likely lead to new insights into the process of protein evolution, as well as insights into the physical forces that maintain protein structure and act as evolutionary selection pressures. On the road to this final objective a number of developments were needed, each of which is interesting and publishable in its own right.

Sequence pair weighting The first novel solution of this work was the development of the method for weighting pairs of sequences described in Chapter 2. Though there are a number of methods described in the literature for weighting individual sequences in a sequence align-

ment, e.g. the Henikoff & Henikoff method, these do not lend themselves to weighting pairs of sequences; our method presented addresses this need.

Database The second novel solution of this work was the development of the database which merged the SwissProt/UniProt, Pfam and PDB/PiQSi databases. The decision to use a highly selective criteria for the data used throughout this project, required a reliable and easily repeatable method of selecting data. Unfortunately no readily available solution existed and thus development of this database became necessary. Given the breadth of the data incorporated into it and its modular design, the scope of its use far exceed its application in this thesis. With some further development and community support this tool could become an invaluable to researchers working with proteins in numerous fields.

Solvent Exposure Though initially motivated by a lack of an accepted measure of solvent exposure to delimit residue burial, in the literature, the investigation into the solvent exposure propensities for the twenty protienogenic amino acids has yielded some interesting and even unexpected results.

Having shown that $HSEu_{13}$ is an appropriate measure of HSEu to use, the comparative investigation between $HSEu_{13}$ and ASA has provided interesting insight into the relationship between the two measures. Firstly it was shown that a linear relationship between ASA and HSEu exists, for each residue type. Secondly it was shown that 26 $HSEu_{13}$ (+/- 2) is a universal value where amino acids no longer have solvent accessible surface area. The reasons for the latter and its implications are yet to be fully understood.

The close correlation between ASA and $HSEu_{13}$ show that $HSEu_{13}$ can be used instead of ASA as measure of solvent exposure. The advantages that HSEu has over ASA are, firstly that it can describe the distribution of residues from the protein surface to the protein interior, while ASA is limited to the solvent accessible surface. Secondly, because the measure of HSEu is only concerned with an integer count of C_α atoms in the direction of the side chain, it is less obviously affected by the differences in residue size. This makes it much easier to work with when performing complex statistical analyses.

A comparison of the pairwise correlation between residue types for solvent exposure and

substitution matrices has shown a correlation. This requires further investigation with different statistical tools.

Finally, the objective of the solvent exposure analyses presented in the thesis, was to determine if a statistically defined measure to distinguish the amino acid compositional change between the surface and the protein core, could be found. It was found that a value of 20 $HSEu_{13}$ is a very close approximation to such a value.

Appendix A
Further development of statistical methods

Included here is an **incomplete** treatment of the different approaches to considering the ways in which $\frac{O}{E}$ can be calculated. It was intended for the main thesis as a part of Chapter 2, however the detail was considered unnecessary. This may provide further insight into the development of the statistical analysis developed for the co-substitution analysis. This should be treated more like notes than an extensive and thorough working of the ideas.

A.1 Application of OE-Ratio

A.1.1 Application to Co-Substitution

Defining Substitutions and Co-Substitutions

A substitution event is defined in the context of this work as: the case when two homologous sequences are compared, it is observed at the same position in the first sequence some residue x is present, while in the second sequence some residue u is present. This is illustrated in Table A.1

Table A.1: A Substitution: in column i, of sequence l, the residue is x while in sequence l it is residue u. Through the course of evolution, the residue at position i has been substituted from x to u or vice-versa as it is difficult to determine temporal events from a sequence alignment.

Sequences	Columns					
	1	2	...	...i...	...	N
k	T	R	...	x	...	L
				↓		
l	E	R	...	u		R

The subject of investigation in this work is the co-substitution of amino acids between homologous sequences. Co-substitutions can be defined as two substitutions taking place at two different positions in the sequence in concert with each other, as shown in Table A.2

Table A.2: A Co-Substitution: in sequence k at positions i and j respectively residues x and y are found, while at the some positions in sequence l, residues u and v are found respectively. The investigation is concerned with determining the statistical propensity of these events to occur at different euclidean distances within the protein structure.

Sequences	Columns					
	1	2	...i...	...	...j...	N
k	T	R	...x...	...	...y...	L
			↓		↓	
l	E	R	...u...	...	...v...	R

Consider sequence-k and sequence-l. In column i it is observed that residue $x \rightarrow u$ and in column j it is observed $y \rightarrow v$. Where $i \neq j$ and $i > j$ to avoid over-counting. It should be noted that $x \rightarrow u \equiv u \rightarrow x$ and similarly for $y \rightarrow v$.

The following section is a development of a statistical analysis method to determine the propensity for co-substitution events to occur when separated by different physical distances.

Defining The Co-substitution Analysis

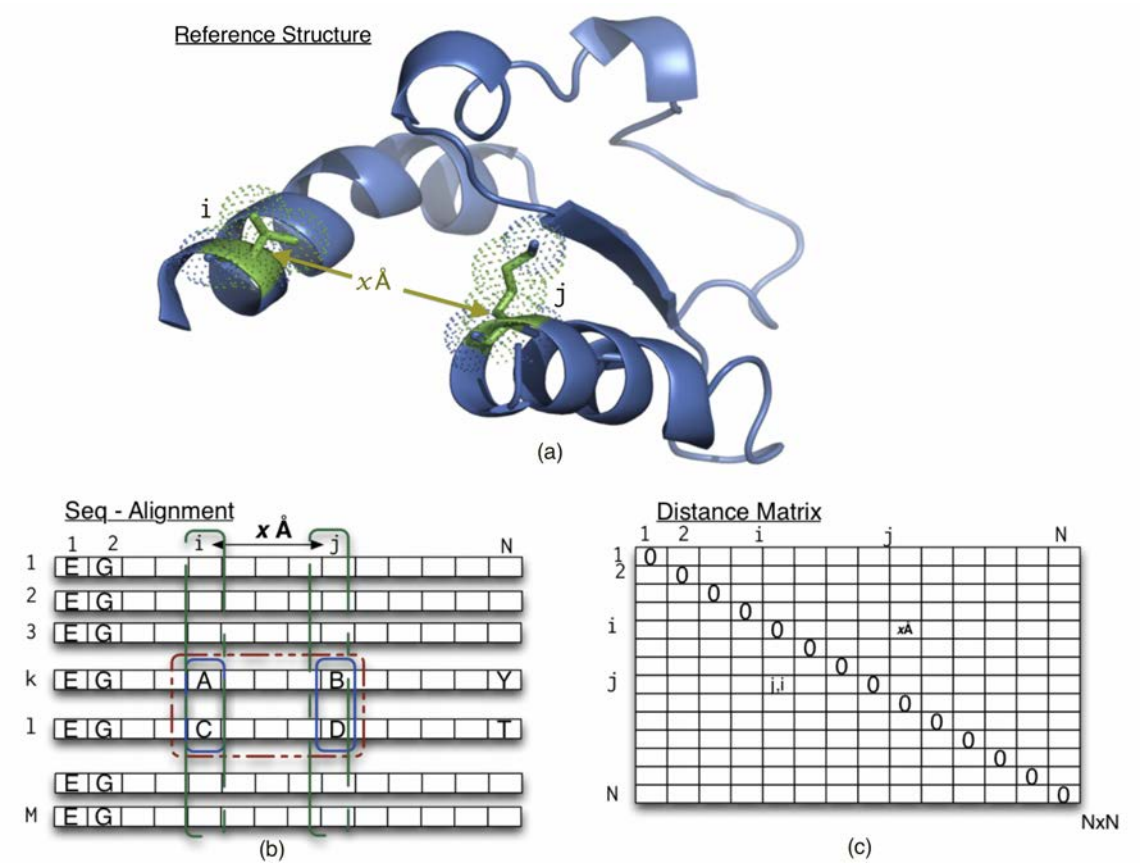

Figure A.1: Capturing the distance information for co-substitution event AB ↔ CD: (a) is a segment of tertiary structure with the physical spearation of x Å between two residues i and j highlighted. (b) is a sequence alignment of homologuos sequences, for which the structure segmenant in (a) is a representative structure. The columns i and j are aligned to the positions i and j in the strucutre. (c) is a distance matrix, which is used to store all inter-residue distance from the structure shown in (a). The inter-residue distnaces in the distance matrix are used as the physical distances between columns in the sequence alignment shown in (b).

The notation used here:

$$d \in D = \{\text{all inter-residue distances}\}$$

d is an inter-residue distance (measured between the C_β atoms) and D is the set of distances, given the set of all inter-residue distances (measured between the C_β atoms) retrieved from the structure.

$$c \in C = \{\text{all co-substitutions}\}$$

c is any two pairs of aligned positions in the two sequences being considered, and C is the set of all possible such pairs. No special distinction is made between conservations and substitution.

- $\sum_d$ is the sum with respect to a specified distance.

- $\sum_D$ is the sum with respect to all distances.

A.1.2 Calculating the Probability of a Co-Substitution Event

Below follows a discussion of how to calculate $P(c)$ the probability of a co-substitution event c.

Before we derive a general statistical method which we can use to determine the relationship between co-substitutions and distance. Let us start by considering a single sequence and address the question "what is the probability of finding xy?" Where xy represent any unique-pair of residues in the sequence.

BBBAAC →sequence-1

There are two possible ways of calculating this:

1. Pick x and then y or
2. Pick y and then x.

As we are randomly selecting residues from a sequence, let us denote the occurrence of xy or yx as O_{xy}

$$P(xy) = P(x)P(y|x) = \frac{\sum x}{N} \bullet \frac{\sum y}{N-1} \tag{A.1}$$

$$P(yx) = P(y)P(x|y) = \frac{\sum y}{N} \bullet \frac{\sum x}{N-1} \tag{A.2}$$

Thus $P(O_{xy}) = P(xy) + P(yx)$:

$$P(O_{xy}) = P(x)P(y|x) + P(y)P(x|y) \tag{A.3}$$

$$P(O_{xy}) = \frac{\sum x}{N} \bullet \frac{\sum y}{N-1} + \frac{\sum y}{N} \bullet \frac{\sum x}{N-1} \tag{A.4}$$

$$P(O_{xy}) = 2\frac{\sum x \bullet \sum y}{(N^2 - N)} \tag{A.5}$$

which is equivalent to:

$$P(O_{xy}) = \frac{\sum xy}{\sum \mathbf{xy}} \tag{A.6}$$

where $\sum$ **xy** = total number of combinations of residues (all possible residue pairs).

This can also be derived by considering this as a combinatorial problem. Recall our sequence:

BBBAAC →sequence-1

Below is a table which shows all the pair-wise interactions that could occur in this sequence:

	B	B	B	A	A	C
B		↔	↔	↔	↔	↔
B	↕		↔	↔	↔	↔
B	↕	↕		↔	↔	↔
A	↕	↕	↕		↔	↔
A	↕	↕	↕	↕		↔
C	↕	↕	↕	↕	↕	

The total number of possible pairs that could be found in a sequence of length is

$$(N-1)+(N-2)+.....+2+1=\frac{(N-1)\bullet N}{2}=\frac{(N^2-N)}{2} \tag{A.7}$$

The number of ways of picking any pair xy or yx is $\sum x \bullet \sum y$. In the sequence, each A forms a pair with every B, assuming that BA = AB, this gives:

$$P(O_{xy})=\frac{\sum x \bullet \sum y}{\frac{(N^2-N)}{2}} \tag{A.8}$$

$$P(O_{xy})=2\frac{\sum x \bullet \sum y}{(N^2-N)} \tag{A.9}$$

Now let us consider a second sequence, where we define any pair of residues uv. We derive the probability of any pair in the new sequence $P(O_{uv})$:

BBAACA →sequence-2

We have shown how to arrive at $P(O_{xy})$ and so arriving at $P(O_{uv})$ must be the same:

	B	B	A	A	C	A
B		↔	↔	↔	↔	↔
B	↕		↔	↔	↔	↔
A	↕	↕		↔	↔	↔
A	↕	↕	↕		↔	↔
C	↕	↕	↕	↕		↔
A	↕	↕	↕	↕	↕	

The probability $P(O_{uv})$ can be derived as above, to give:

$$P(O_{uv})=2\frac{\sum u \bullet \sum v}{(N^2-N)} \tag{A.10}$$

similarly:

$$P(uv)=\frac{\sum uv}{\sum \mathbf{uv}} \tag{A.11}$$

Now we consider the probability of xy in sequence-1 and the probability of uv in sequence-2. We have two sequences:

BBBAAC →sequence-1
BBAACA →sequence-2

Because these two sequences and their distributions are independent of each other, the joint probability of xy and uv is simply:

$$P(O_{xy}, O_{uv}) = 2\frac{\sum x \bullet \sum y}{(N^2 - N)} \bullet 2\frac{\sum u \bullet \sum v}{(N^2 - N)} \tag{A.12}$$

which is equivalent to:

$$P(xy, uv) = \frac{\sum xy}{\sum \mathbf{xy}} \bullet \frac{\sum uv}{\sum \mathbf{uv}} \tag{A.13}$$

The term $P(O_{xy}, O_{uv})$ considers the probability of xy and uv as independent events. If we return to the definition of a substitution event in section A.1.1, we define it as the event where $x \rightarrow u$ in an aligned position in sequences k and l. In the context of a sequence alignment, this would be when we consider a column in the alignment and examine two residues in the column, each one from different sequences. The substitution event $x \rightarrow u$ would represent an event such as E →D. In the derivation above, we do not consider the likelihood of this event directly. In fact we are ignoring the propensity of $x \rightarrow u$ altogether.

As we are trying to approach the probability distribution of a co-substitution event, we take the E in the first sequence, and pair it up with a second residue, say T. We then consider how frequent the pairing of E and T are in the first sequence. Following this, we look at some second sequence where we are interested in D and Y (for example). The analysis derived above would solve for the independent composition of both sequences and tell us how likely it would be to randomly select E paired with T in the first sequence and then how likely it would be to randomly select D paired with Y in the second sequence. There is no relationship defined between the pairs ET and DY.

We are ultimately interested in the probability of a co-substitution event, given the two sequences, in this case P(E →D; T →Y) or more generally P($x \rightarrow u$; $y \rightarrow v$). We define the probability of a co-substitution event $P(c)$:

Where c is a specific co-substitution, e.g. $xy \rightarrow uv$, whereas C is the set of all aligned pairs of residues from two sequences: (**xy →uv**)

If we consider sequences 1 & 2 above, the set C would be:

$$C = \{(BB, BB), (BB, BA), (BA, BA), .., (AA, AC), (AC, CA)\} \tag{A.14}$$

The example given, is useful, simply because it shows that it does not matter, if we consider xy,uv or xu,yv.

If we now consider the number of pair-wise interactions that occur, by aligning sequences 1 & 2 together, in a table as we did above for each sequence independently:

k		B	B	B	A	A	C
	l	B	B	A	A	C	A
B	B		↔	↔	↔	↔	↔
B	B	↕		↔	↔	↔	↔
B	A	↕	↕		↔	↔	↔
A	A	↕	↕	↕		↔	↔
A	C	↕	↕	↕	↕		↔
C	A	↕	↕	↕	↕	↕	

$$P(c) = 2\frac{\sum xu \bullet \sum yv}{(N^2 - N)} \tag{A.15}$$

By using the rules of probability we can also say:

$$P(c) = \frac{\sum c}{\sum C} = \frac{\sum xu}{\sum \mathbf{xu}} \bullet \frac{\sum yv}{\sum \mathbf{xu}} = \frac{\sum xy \rightarrow uv}{\sum \mathbf{xy} \rightarrow \mathbf{uv}} \tag{A.16}$$

$$P(xy, uv) = \frac{\sum xy}{\sum \mathbf{xy}} \bullet \frac{\sum uv}{\sum \mathbf{uv}} \tag{A.17}$$

$$P(c) \neq P(xy, uv) \tag{A.18}$$

A.1.3 Discussion of predictions

Three ways to do predictions

We have a choice at this point, to use the distribution of co-substitution events in the sequence alignment, with respect to distance. Alternatively we can use the distribution of inter-residue distances between pairs of residues in the the sequence alignment. The two approaches allow for subtly different analyses to be performed with the same data-set.

Approach 1

If we consider the initial development in the derivation of $p(c)$ in section A.1.2, we started by discussing $p(xy)$, i.e. the probability of any two residues being selected at random. This was, of course, not constrained by distance in any way. We were considering the entire sequence of residues. The term $p(xy)$ intrinsically captures data pertaining to the composition of the sequence. As $p(xy)$ does not consider two sequence it cannot be thought of as a co-substitution. However it can be used in a Bayesian prediction, as per the discussion at the start of this section. If we consider the sequence-1 we used earlier:

BBBAAC →sequence 1

If xy = "BA" there are a total of 6 possible interactions between B & A, i.e they can be paired 6 times out of a total of 15 possible pairs that can be made from the residues in the sequences. Thus:

$$p(xy) = \frac{6}{15} = \frac{2}{5} \tag{A.19}$$

If we want to know how likely two pairs of residues are to be apart we can simply reformulate the $\frac{O}{E}$ functions defined in section A.1.1 Such that:

$$P(c) = P(xy) \tag{A.20}$$

i.e the probability of a given pair in the whole sequence. This would be the Expected defined in Equation A.42.

The observed defined in Equation A.40 would become:

$$P(c|d) = P(xy|d) \tag{A.21}$$

i.e the probability of a given pair of residues to be a certain distance apart, in a given protein.

A weighted average over a group of Pfam families would offer insight as to whether certain residues prefer or need to be a certain distance apart. For example, Argenine and Argenine,

Argenine and Histidine or Glutamic Acid and Aspartic Acid. If we now refer back to the prediction step described earlier:

$$P(d|c) = \frac{P(c|d) \bullet P(d)}{P(c)} \tag{A.22}$$

We can see that we have a term for $P(c|d)$ and $P(c)$, and the term $P(d)$ would be equivalent to a 'prior' in Bayesian analysis, which can be generated by taking a weighted average of a set of "training" distance matrices. Therefore we can solve for $P(d|c)$ by:

$$P(d|c) = \frac{P(xy|d) \bullet P(d)}{P(xy)} \tag{A.23}$$

This in truth is a simple statistical potential derived from a combination of sequence-alignment data and structure data.

Approach 2

Now if we want to add another constraint where we observe in one sequence two residues at a given distance d apart, and in a homologous sequence we see another pair of residues also d apart, we can expand the above approach to include two sequences. In this approach we treat the two sequences as independent of each other and do not fix a relationship between residue positions by specifying an alignment of said positions.

Say we consider our two sequence 1 & 2 from earlier:

BBBAAC sequence 1
BBAACA sequence 2

We see several pairs in both sequences:

BB – BB
BA – BA
BA – BC
:: – ::
BA – AC
:: – ::

We can now reformulate $P(c)$ to accommodate two sequences:

$$P(c) = P(xy, uv) = P(xy).P(uv) \tag{A.24}$$

If we consider xy = BA and uv = AA.

From sequence 1 we get:

$$P(xy) = P(BA) = \frac{6}{15} = \frac{2}{5} \tag{A.25}$$

From sequence 2 we get:

$$P(AA) = \frac{2}{15} \tag{A.26}$$

Thus:

$$P(xy, uv) = P(BA) \bullet P(AA) = \frac{2}{5} \bullet \frac{2}{15} = \frac{4}{75} \tag{A.27}$$

This reflects the likelihood of randomly selecting BA from sequence 1 and AA from sequence 2, and is a reflection of the composition of both sequences. However this does not include any information about the positions of BA and AA with respect to each other in the sequences.

For completeness, we now consider the observed for $P(xy, uv)$: $P(xy, uv|d)$. Figure A.2 shows that in the calculation of $P(xy, uv|d)$ we would not be considering any information regarding the alignment of xy with uv at a given physical distance apart.

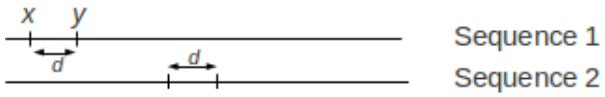

Figure A.2: The term $p(xy, uv|d)$ reflects the joint probability for two pairs of residues in two different sequence to be the same distant apart, but does not consider their positions. Thus $P(xy, uv|d)$ does not reflect the probability that xy and uv are equidistant and aligned with each other.

Compared to simply considering $P(xy)$, this approach has been considered as an alternative method for studying the propensity for co-substitution events to occur at a given distance apart. However, as there is no explicit alignment of residue pairs from both sequences, we don't have any measure of the propensity for $xy \rightarrow uv$. Though, this approach can also be used in prediction of inter-residue distance by including the additional information into the prediction equation:

$$P(d|c) = \frac{P(xy, uv|d) \bullet P(d)}{P(xy, uv)} \tag{A.28}$$

It just relies purely on the intrinsic propensity for residues to be a certain distance apart. The additional grouping of residues from two sequences, considers the unbiased joint distribution of residues in two sequences.

Approach 3

Finally, we consider the case where $P(x \rightarrow u : y \rightarrow v)$, the probability of finding xy in sequence 1 and uv in sequence 2, such that they are aligned with each other, as discussed in the last step of the derivation in section A.1.2.
Let us consider sequence 1 & 2 again.

BBBAAC sequence 1
BBAACA sequence 2

Earlier we considered $P(xy, uv) = P(BA, AA)$.
We now consider $P(x \rightarrow u : y \rightarrow v) = P(B \rightarrow A : A \rightarrow A)$.
This occurs once out of fifteen possible paired substitution events, thus:

$$P(c) = P(x \rightarrow u : y \rightarrow v) = P(B \rightarrow A : A \rightarrow A) = \frac{\sum(B \rightarrow A : A \rightarrow A)}{\sum(\mathbf{x} \rightarrow \mathbf{u}\text{: } \mathbf{y} \rightarrow \mathbf{v})} \tag{A.29}$$

$$P(B \rightarrow A : A \rightarrow A) = \frac{1}{15} \tag{A.30}$$

This method constitutes an explicit co-substation analysis. Using the earlier derivation for the prediction step, prediction of inter-residue distances is possible following a small reformulation.

The Expected:

$$P(c) = P(x \to u : y \to v) \tag{A.31}$$

The Observed:

$$P([x \to u : y \to v]|d) \tag{A.32}$$

The inter-residue distance between co-substitution site:

$$P(d|c) = \frac{P([x \to u : y \to v]|d) \bullet P(d)}{P(x \to u : y \to v)} \tag{A.33}$$

A.1.4 The Application of $\frac{O}{E}$ to Co-Evolution

There are three methods which are mathematically equivalent, by which an Observed Distribution and the Expected Distribution can be calculated which result in identical $\frac{Observed}{Expected}$ values.

$\frac{O}{E}$ **Method 1:** The first method presented here, considers the distribution of distance with respect to co-substitution events.

$$\text{Observed}_1 = P(d|c) \tag{A.34}$$

$$O_1 = \frac{\sum_d c}{\sum_D c} \tag{A.35}$$

This measure of the Observed represents the probability of a distance-d given a specific co-substitution event. It is the conditional probability of a distance, given that we observe a co-substitution c. Alternatively, it represents the proportion of all co-substitutions c, found to occur at distance d.

$$\text{Expected}_1 = P(d) \tag{A.36}$$

$$E_1 = \frac{\sum_d C}{\sum_D C} \tag{A.37}$$

This measure of the Expected, represents the intrinsic bias for any two residues to be a some distance-d apart. It is ignorant of the amino acid composition of the protein and is only concerned with the probability of some distance d to exist between any two points in the structure. Alternatively it represents the proportion of all amino acid-pairs in the protein that are separated by a given distance. This can be calculated without referring to multiple sequences, but to a single "reference sequence" matched to a representative structure for all sequences being considered.

The null hypothesis for this method states: 'distance is not dependant on co-substitution.'

$$\frac{O_1}{E_1} = \frac{P(d|c)}{P(d)} \tag{A.38}$$

$\frac{O}{E}$ **Method 2:** The second method presented here considers the distribution of co-substitution events with respect to distance.

$$\text{Observed}_2 = P(c|d) \tag{A.39}$$

$$O_2 = \frac{\sum_d c}{\sum_d C} \tag{A.40}$$

This Observed represents the probability of a co-substitution event, given a specific distance-d. It is the conditional probability of the co-substitution given a distance-d. Alternatively, it represents the proportion of all co-substitutions separated by distance d, which are c.

$$\text{Expected}_2 = P(c) \tag{A.41}$$

$$E_2 = \frac{\sum_D c}{\sum_D C} \tag{A.42}$$

This measure of the Expected, represents the intrinsic bias in the data for a specific co-substitution event to occur. It is ignorant of distance, but is concerned with the total number of possible pairings of aligned residue positions in the alignment of the two sequences being considered. i.e. If there were no distance bias in the distribution of co-substitution events then c should occur in each distance bin proportional to its existence in the set C.

$$\frac{O_2}{E_2} = \frac{P(c|d)}{P(c)} \tag{A.43}$$

$\frac{O}{E}$ Method 3:

$$\text{Observed } 3 = P(c, d) \tag{A.44}$$

$$O_3 = \frac{\sum_d c}{\sum_D C} \tag{A.45}$$

This is the joint probability distribution of c and d.

$$E_3 = P(c)P(d) = \frac{\sum_d C}{\sum_D C} \bullet \frac{\sum_D c}{\sum_D C} \tag{A.46}$$

Deriving equivalence of the three methods

This derivations assumes a dependence of c on d or vice-versa.

$$P(c, d) = P(c|d)P(d) = P(d|c)P(c) \tag{A.47}$$

$$\therefore P(c|d)P(d) = P(d|c)P(c) \tag{A.48}$$

$$O_2.E_1 = O_1.E_2 \tag{A.49}$$

$$\frac{O_1}{E_1} = \frac{O_2}{E_2} \tag{A.50}$$

from equations A.45, A.47, A.48 and A.49:

$$P(c, d) = O_3 = O_2.E_1 = O_1.E_2 \tag{A.51}$$

from equations A.37, A.42 and A.46:

$$E_3 = P(c)P(d) = E_1.E_2 \tag{A.52}$$

Thus:

$$\frac{O_1}{E1} = \frac{O_2}{E_2} = \frac{O_3}{E_3} \tag{A.53}$$

A.1.5 Discussion of OE-ratio for Co-substitutions and usefulness in predictions

The great historical interest in correlated mutations - represented by the roughly 20,000 or so results from a PubMed search for 'correlated mutations' - has been driven by the prospect of being able to perform useful protein structure predictions. Ultimately it would be a great step forward in protein-structure prediction using Bayes Theorem if we could use this work for predicting inter-residue distances in unsolved protein structures.

If we start with the relationship between each of the three possible methods of doing the Observed over Expected analysis, defined in section A.1.2:

$$\frac{O}{E} = \frac{P(d|c)}{P(d)} = \frac{P(c|d)}{P(c)} = \frac{P(c,d)}{P(c)P(d)} \tag{A.54}$$

Let us consider method 1 and method 2 from section A.1.2:

$$\frac{P(d|c)}{P(d)} = \frac{P(c|d)}{P(c)} \tag{A.55}$$

With a little bit of algebra we can derive a relationship between the intrinsic bias of the distance-matrix, $P(d)$ and the data that is purely in the sequence alignment:

$$\frac{P(d|c) \bullet P(c)}{P(c|d)} = P(d) \tag{A.56}$$

If however we wanted to know specifically the likely "range-bin" of a pair of residues, we could build a training set to determine the $P(d)$ and solve for the $P(d|c)$:

$$P(d|c) = \frac{P(c|d) \bullet P(d)}{P(c)} \tag{A.57}$$

and we determine $O_2 = P(c|d)$ and $E_2 = P(c)$

Therefore if we also determine P(d) we , then we can determine $P(d|c)$ the probability of a given co-substitution c occurring at each distance d. Where $P(d)$ is defined in equation A.37 as $E_1 = \frac{\sum_d C}{\sum_D C}$.

Appendix B
Sequence Weighting Proof

We have a set of sequences:

$$x_i \in S \text{ for} i = 1, 2, ..., n$$

Define the function $d(x_i, x_j)$ to be the sequence difference between x_i and $_j$. Clearly:

$$d(x_i, x_j) = 0$$

Denote the Henikoff and Henikoff weighting of a sequence by $h(x_i)$. Now define the weighting given to a sequence pair x_i, x_j to be:

$$W(x_i, x_j) = h(x_i) \bullet h(x_j) \bullet d(x_i, x_j)$$

Thus if we have all $x_i \in S$ either completely identical or completely different (in every sequence position) to one another, then let the number of types of sequences by $m \leq n$ and the set of sequence in each type to be:

$$t_p \subseteq S, t_p = d(x_i, x_j) = 0 \forall x_i, x_j \in t_p$$

for

$$p = 1, 2, ..., m$$

Then, if

$$x_i \in t_p$$

so

$$h(x_i) = \frac{1}{m \dot{|} t_p \mid}$$

Therefore, if

$$x_i, x_j \in t_p$$

then

$$W(x_i, x_j) = 0$$

. However, if

$$x_i \in t_p, x_j \in t_q, p \neq q$$

then

$$d(x_i, x_j) = 1$$

and:

$$W(x_i, x_j) = \frac{1}{m \bullet \mid t_p \mid} \bullet \frac{1}{m \bullet \mid t_q \mid} = \frac{1}{m^2 \bullet \mid t_p \mid \bullet \mid t_q \mid}$$

Now, if we consider all the pairings of one group t_p and another t_q $p \neq q$), then the sum of the weightings is:

$$\sum_{x_i}^{t_p} \sum_{x_i}^{t_q} W(x_i, x_j) = \mid t_p \mid \bullet \mid t_q \mid \bullet W(x_i, x_j) = \frac{1}{m^2}$$

Note that this is if it were not for the multiplication by $d(x_i, x_j)$ this would hold true for:

$$\sum_{x_i}^{t_p} \sum_{x_i}^{t_q} W(x_i, x_j)$$

and so the sum of all weightings would be

$$\sum_{x_i}^{S} \sum_{x_i}^{S} W(x_i, x_j) = 1$$

So we need to noramlise the actual weighting by $(1 - \sum_{p=1}^{m} \sum_{x_i}^{t_p} \sum_{x_i}^{t_q} W(x_i, x_j)^1$

Appendix C
Data selection & Database Development

Note: This appendix was prepared as a chapter for a progress report during my PhD. This should only be used as notes and guidence to the approach that went into the development of the database.

C.1 Introduction

Amino acid composition of proteins varies between sub-cellular locations [86]. This has been exploited to predict cellular locations of proteins [86, 87]. The effect of these differences has not been considered. Though the physics of the folding process don't differ between cellular locations, the differences between each environment will most likely affect the folding process. For example changes in pH will change the likely protonation state of Histidine, affecting its charge and its likely position in a structure. Similarly changes in redox potential will affect cysteines propensity to form disulphide bonds. Therefore, when performing a rigorous analysis of amino acid propensity to be solvent accessible or to be involved in a correlated substitution with other amino acids, we should not ignore these considerations.

C.2 What is being attempted

C.2.1 Requirements

We have a requirement for a method to make a selection of PDB structures based on a) cellular location, b) organism, c) taxonomy and d) classification (Prokaryotic, Eukaryotic, Archaea). A further requirement is being able to determine the number of domains present in a given protein structure. Within a protein structure we need to know where a domain starts and ends in the amino-acid sequence. When completed this method will be applied to the PiQSi database of quaternary structures.

It is possible to get specific data about proteins from a variety of on-line databases, such as the PDB, Pfam, SCOP, CATH, UniProt, to name but a few. However there is no single unified resource which collates the data from all these resources in a single place, although the PDBe offers a cross reference section for each structure that point to *UniProt, SCOP, CATH, Pfam and InterPro*.

Using the information from PDBe linking a PDB structure to entries in other on-line databases, it could be possible for a small number of protein structures to make the kind of selections detailed above. However each step would involve manually researching each structure's reference in the other databases. This would make selecting data from the entire PDB based on a classification set in another database virtually impossible. For example, if there were a need to determine all PDB structures which are exclusively eukaryotic and exclusively found in the nucleus of cells, this would not be straightforward to achieve using the on-line resources. If a further step is added of requiring all known domains in the subset of structures just selected, this would involve a further lengthy process which the on-line resources were not designed to accommodate. For this project it is necessary to have an automated process of getting data that is present in more than one on-line database.

To meet this requirement it was decided to use an SQL database. This is because a relational database offers great flexibility and versatility in selection of data. Relational databases were designed to meet the needs that arise from the type of problem being addressed here - which is to create a relationship between the data stored in several different on-line databases. A choice had to be made between database platforms of which there are several. The two primary candidates considered (because they were both free and open source) were: SQLite and MySQL. SQLite

stores the database in a file and does not run as a server - the data is stored in a text file which can be interrogated by an SQLite client program. It has advantages of portability and in certain conditions improved speed, but lacks support for foreign keys (described below, in conjunction with primary keys). MySQL on the other hand offers the possibility of remote access to the database and a robust platform to store the data in and so was selected. Both solutions offer the possibility of automated interrogation of the database, using a scripting/programming language such as Python or Perl. They can both be accessed directly using a client program and allow for the automation of data access.

A brief overview of databases

A database is a collection of tables, where data is stored and rows and categories are defined by columns. Tables can be joined together by matching categories. There are special categories which can be used for this purpose called PRIMARY KEYS.

A primary key is a category where it is certain that each entry is going to be unique. This can be a category of data such as a unique code or integer assigned to each entry, or a category which is known to have no repetition present. The primary key of one table can be used as the primary key of another table and in this context is referred to as a FOREIGN KEY in the second table. The foreign key ensures that the entry in the second table is correctly linked to the data in the first table.

C.3 Methods

The PDB contains the structure data that we are using and we chose the Pfam-A database to provide the data for protein domain families. Pfam-A is manually curated and contains sequence alignments for each family, which is useful for the co-evolution analysis . Neither of these databases offers a straightforward way to make selections on the bases of the criteria set out above (cellular location, organism, taxonomy, classification). This information is available in the UniProt database, of which there are two versions, the UniProt/Trembl and UniProt/SwissProt. The former is automatically curated and the latter is manually curated. We chose the latter as we need the most reliable data available for our analyses. Therefore we have selected three on-line databases to provide the data for our database: The PDB, Pfam and UniProt/SwissProt.

The objective is to create a relational database, which allows for a three way cross referencing between PDB, Pfam and UniProt/SwissProt. Pfam-A contains cross reference data to PDB and UniProt, while SwissProt contains cross reference data to PDB and Pfam. The Pfam-A data is not restricted to SwissProt and it was not known how well the two agree with each other, before this work was done. To overcome any discrepancies between them, the intention is to use all three sources of data to build a three way cross reference by checking each one against the other.

The data was downloaded from the three different databases/data-banks and was initially treated separately. What follows is a description of how the data from the Pfam-A, SwissProt and PDB was populated in the database. The following sections cover these steps:

1. **Pfam-A:** Load the data into the database and build a three way cross-reference between Pfam, PDB and UniProt, using the data available in Pfam-A.

2. **SwissProt:** Load the data into the database and build a three way cross-reference between SwissProt, PDB and Pfam, using the data available in SwissProt.

3. **PDB:** There is only cross reference data to UniProt in the PDB header. This was loaded into the database because it can be used to fill in missing information.

It appears that the Pfam-A (step 1 above) and SwissProt (step 2) sections are repetitions of each other. This is true, however it is a case of applying the same method to two different sets of data. This made it possible to highlight discrepancies between the two and to include data from one that was not in the other.

C.3.1 Pfam-A

Gathering the data available in Pfam-A to build a table in the database with the Pfam-A three way cross reference was accomplished in several steps. The first was to find where the data was stored. I settled on the Pfam-A single text file in FASTA format (see Glossary). This file contains a list of Pfam ID accompanied by a UniProt ID (this is a UniProt entry, and is not restricted to SwissProt or Trembl) the start and end points of Pfam the domain in the UniProt sequence and the domain sequence for that UniProtID. Two other files were also chosen, which appeared to be text files of tables from the Pfam-A SQL database used by the Pfam group. These files contained references between what appeared to be primary/foreign keys for the Pfam SQL database, which made it possible to build a cross reference between Pfam-A and the PDB. This data was joined to the data from the FASTA file, to produce the three way cross reference between Pfam-A, UniProt and PDB. This is the cross reference based on the annotations made by the Pfam development team.

[In the previous paragraph I mention a Pfam-SQL server. From the available information on the Pfam FTP server, it appears that the Pfam group host Pfam on an SQL server - this would make complete sense given the amount of information they are working with. When I refer to the Pfam-SQL server it is that server I am referring to.]

I later discovered a single 12 GB text in Stockholm format (see Glossary) file which contained all the data from Pfam-A including the sequence alignments for each Pfam family. This information will be needed for the co-evolution part of this project. Using this file to populate the MySQL database would have been a better approach, one which I may use later if necessary.

Pfam-A data

Pfam-A version 24.0 (the current version is 25.0 and our database could be updated) from the FASTA file was loaded into a single table called *pfam*, in the MySQL database. This was done using the Python module MySQLdb and the BioPython module SeqIO.

Qualifying the table the number of unique IDs is:

- **Pfam-A (ver. 24.0):** 11,912
- **UniProt:** 3,605,941

This was determined by running the following SQL queries (for information on SQL please visit www.mysql.com):

```
SELECT COUNT(DISTINCT UniProtID)FROM pfam
```

This agrees with the Pfam website, demonstrating that the parsing and populating of the database had been completed without errors.

Data for Pfam to PDB referencing

The Pfam-A is hosted as an SQL database. It is possible to download component parts of that database (which are likely tables in that database) as plain text files. There appear to be two key categories in the Pfam-A SQL tables, to ensure the correct relationship between entries in each table of their database, these are: pfamseq_acc and auto_pfamseq. I found two text files containing firstly a direct relationship between a Pfam ID, UniProt ID and the two keys, pfamseq_acc and auto_pfamseq. I found a second file which contained the residue data for each Pfam sequence in Pfam-A in a given PDB structure. This provided a map between pfamseq_acc and auto_pfamseq to a PDB id. Using the data in these two files I was able to build a cross reference between Pfam IDs, UniProt IDs and PDB IDs. This was done by parsing the data in both the text files into their own tables in the database.

The table created for the PDB to *pfamseq_acc* and *auto_pfamseq* data was called *PfamAcc_to_PDB_map* and contained the following categories of data:

- PDB id
- Chain,
- auto_pfamseq
- pfamseq_acc,
- pfamseq

The table created for *Pfam* to *pfamseq_acc* and *auto_pfamseq* was called *Pfamid_to_PfamseqAcc* and contained the following data:

- Pfam id
- pfamseq_acc
- UniProt id
- auto_pfamseq.

The Pfam-A cross reference to UniProt and PDB

The two tables were joined together to create a new table with the intersection between Pfam ID, UniProt ID and PDB, with the following SQL command:

CREATE TABLE PDB_Pfam_UniProt_Map SELECT pfpm.auto_pfamseq, pfpm.pdb_id, pfpm.chain, pfpf.pfamid, pfpf.UniProtID, pfpf.pfamseq_acc FROM PfamAcc_to_PDB_map AS pfpm JOIN Pfamid_to_PfamseqAcc_Map AS pfpf WHERE pfpf.auto_pfamseq = pfpm.auto_pfamseq AND pfpf.pfamseq_acc = pfpm.pfamseq_acc

This produced a table called *PDB_Pfam_UniProt_Map* with the following categories:

- auto_pfamseq
- pdb_id
- pfamid

- UniProtID

- chain

- pfamseq_acc

To see how many of each unique IDs are present in the map the following queries are run:

SELECT COUNT(DISTINCT pfamid) FROM Pfam_UniProt_PDB_map_pfm
SELECT COUNT(DISTINCT pdb_id) FROM Pfam_UniProt_PDB_map_pfm
SELECT COUNT(DISTINCT UniProtID) FROM Pfam_UniProt_PDB_map_pfm
SELECT COUNT(DISTINCT UniProtID) FROM Pfam_UniProt_PDB_map_pfm p WHERE p.UniProtID IN (SELECT UniProtID FROM UniProt)

The first three queries determine the number of unique entries in each of the categories. The last query checks the number of UniProtIDs present in the Pfam-A which are also present in SwissProt. The results of the queries are summarised in table C.1.

Table C.1: Summary of cross reference data available in Pfam-A. The table shows the number of entries from each database, for which there is reference data for both the other databases.

Database	Number of entries
Pfam	5,580
PDB	53,748
UniProt	19,451
SwissProt	12,860
PDB (Sprot)	41,139
Pfam (Sprot)	4,347

These results tell us that Pfam-A has cross reference data between UniProt, Pfam and PDB, for: 5,580 domains, present in 53,748 known crystal structures, which represent 19,451 UniProt entries. Of the 19,451 UniProt entries present in the cross reference data. The last three rows contains the cross reference data for which SwissProt entries exist. The number of UniProt entries present in SwissProt is 12,860, which is represented by 41,139 structures, and has 4,347 pfam-domains.

C.3.2 UniProt/SwissProt

From the Pfam-A data it has been possible to construct a three way relationship between Pfam, PDB and UniProt. However there are two considerations here, firstly that we had decided to use Pfam-A, the PDB and UniProt/SwissProt, while the UniProt data in Pfam-A is not restricted to SwissProt. The second consideration is that the cross reference data has been built exclusively from Pfam-A and it is possible to construct a three way relationship between SwissProt, PDB and Pfam using the data in SwissProt. This makes it possible to cross check the data in Pfam-A against the data in UniProt/SwissProt and not rely entirely on one group's methods.

The SwissProt database is available for download as a single text file. This was downloaded and a parser was written to extract data from the file and put it in a table in our MySQL database. The parser excluded all entries in SwissProt for viruses, because in our context virus proteins could be in any or many organelles. A section in each entry contains cross references to other databases, including Pfam and the PDB.

Taking advantage of the relational database, the data from SwissProt was distributed across three different tables. The cross reference information for Pfam was stored in its own table, called *uPfam_data* in the MySQL database. Similarly cross reference data for the PDB was stored in its own table called *pdb_data*. The third table contained a selection of the other information included in each SwissProt entry, considered useful to our objective. This data was stored in a table called *UniProt*.

The selection of the other information from each UniProt/SwissProt entry was made based on the amount of additional information it would provide and how much easier it might make selecting data from the database. The following categories were selected as for the *UniProt* table:

- UniProt ID
- UniProt accessions
- cellular location
- sub-unit
- taxonomy organelle
- organism
- description
- comments
- key words
- sequence

Some of this information exists in its own section in each entry, some of it had to be parsed from either the comments, description or key words section of the entry. The *UniProt* table is the main table for the SwissProt data in our database.

The information was selected to be fairly general but, specific selections were made to address the criteria set earlier. Namely, cellular-location, taxonomy, organelle and organism. The comments, keywords and description sections were included entirely, as it was thought that this would make the database more generalised. Using either regular expressions or parsing of text in Python or Perl, it is possible to use the information in these section to address questions completely unrelated to this project.

The parser, when populating the table used the entry number in the text file (acquired by counting), as the PRIMARY KEY for the entry, and was labeled **id**. This was used to the PRIMARY KEY/FOREIGN KEY for the two other tables.

The *UniProt* table has a total of 509,917 entries. There are 519,348 entries in the whole of SwissProt, however we removed all entries for viruses and so we have roughly 9,500 less entries.

The UniProt/SwissProt cross reference to PDB

The cross reference data for the PDB was stored in a table called *pdb_data*. To make sure that each PDB entry was correctly linked to the *uniprot* table, the PRIMARY KEY **id** from that table is used as the PRIMARY KEY in the *pdb_data* table. The total number of unique entries in the table for each ID is:

- **PDB-ID:** 51,030
- **UniProtID:** 15,668

This result indicates that in SwissProt there are only 15,668 entries for which structures have been solved. There are 51,030 crystal structures, this suggests that many proteins have been crystallised several times (exclusive of virus entries in SwissProt).

The UniProt/SwissProt cross reference to Pfam

The cross reference data for Pfam was stored in a table called *uPfam_data*. Again to make sure that the sure that each entry was correctly linked to the *uniprot* table, the **id** was used as the PRIMARY KEY. The total number of unique entries in the table for each ID is:

- **Pfam-IDs:** 8,436
- **UniProt-IDs:** 477,161

This falls quite short of the total number of Pfam families in Pfam-A. This is at least in part likely to be a result of this version of SwissProt having been released prior to the version of Pfam-A being used. It will be possible using the method proposed later, to use the cross reference data from Pfam-A to fill in the gap.

This result indicates that there are 477,161 proteins in SwissProt which have known Pfam domains. This is exclusive of any entries with viruses.

The UniProt/SwissProt cross reference to PDB & Pfam

Using the data in the two cross reference tables, with an SQL query it was possible to join the two ables together to construct the three way relationship between SwissProt, Pfam and PDB, present in the SwissProt data. This was done with the following SQL command:

```
CREATE TABLE UniProt_PDB_Pfam_map_uprt SELECT
pdb.pdbid,pfm.pfamid,pdb.UniProtID, pdb.chains, pdb.Seq_start, pdb.Seq_end FROM
pdb_data AS pdb JOIN uPfam_data AS pfm WHERE pdb.id = pfm.id
```

To see how many of each unique IDs are present in the map the following queries are run:

```
SELECT count(distinct pdbid) FROM UniProt_PDB_Pfam_map_uprt
SELECT count(distinct pfamid) FROM UniProt_PDB_Pfam_map_upr
SELECT count(distinct UniProtID) FROM UniProt_PDB_Pfam_map_uprt
```

The results of those queries are shown in Table C.2 below.

To summaries, these results tell us that there are 15,077 unique UniProt-IDs in SwissProt for which there are are 50,006 known crystal structures in the PDB, and there are 4,493 Pfam domain families present in those structures.

Table C.2: Summary of cross reference data available in SwissProt. The table shows the number of entries from each database, for which there is reference data for both the other databases

Database	Number of entries
Pfam	4,493
PDB	50,006
UniProt/SwissProt	15,077

C.3.3 PDB

There are currently about 73,000 structures stored in the PDB. However these are not all unique, many protein structures have been resolved several times - for example HIV-1 protease has been resolved approximately 200 times in the last 25 years.

The PDB has a considerable amount of information stored in the header of each PDB file. Where each PDB file contains information about a single resolved structure. Information stored in the header of the file includes among other things a reference to the UniProt database. However there is not usually a reference to Pfam in the header.

As this project is concerned with structural considerations in protein co-evolution and solvent accessibility, the entire PDB database has been downloaded and stored (at the time of download there were only about 72,000 entries in the PDB). The database consists of individual text files in a standard PDB format. A number of different programs have been written in Python for other parts of this project, to interrogate the data stored in these files. Some of the code written was adapted to extract cross reference data from the PDB header. For the purpose of data selection I followed the reasoning, that a) relying on a single group's method for referencing other databases is unreliable and b) more information in the database makes it more versatile, thus the UniProt reference was parsed from the header of each PDB entry. The information was stored in a single table called *pdb_cross_ref*.

There are 67,251 unique PDB IDs present in this table. This means that of the roughly 72,000 PDB files, nearly 5,000 of them may not have any cross reference information to UniProt, though a thorough investigation of what happened with the remaining files still needs to be done.

There are 25,685 unique UniProt IDs in the table (this, like with Pfam-A is not restricted to SwissProt). This means that the *pdb_cross_ref* table contains 67,251 structures representing 25,685 UniProt entries. This is about 10,000 more than is present in the by UniProt/SwissProt. Table C.3, contains a summary of the cross reference data in the PDB to UniProt.

Table C.3: Summary of the number PDB structures for which cross reference to UniProt could be found in the header. This may not be complete and requires further investigation.

Database	Number of entries
PDB	67,251
UniProt	25,685

To determine how many of the UniProt IDs found in the PDB are present in SwissProt the following query is run:

```
SELECT COUNT(DISTINCT up.UniProtID) FROM UniProt_pFam_PDB.pdb_cross_ref up
    WHERE
```

up.UniProtID IN(SELECT UniProtID FROM UniProt)

To determine how many PDB-IDs have a UniProt-ID present in SwissProt the following query is run:

SELECT COUNT(DISTINCT up.pdb_id) FROM UniProt_pFam_PDB.pdb_cross_ref up
WHERE
up.UniProtID IN(SELECT UniProtID FROM UniProt)

Table C.4 contains a summary of the number of PDB to SwissProt cross references, which is the result of the two SQL queries above.

Table C.4: The number of PDB entries with cross references to UniProt entries in SwissProt.

Database	Number of entries
PDB	44,554
UniProt	13,157

To get an idea of the number of entries for which discrepancies exist between SwissProt and the PDB, we can compare the data in the *pdb_data* table (described earlier) – which contains the SwissProt cross reference data to the PDB – with this table to determine how many entries in the *PDB_cross_ref* table match the SwissProt table. This would be done by joining the two tables using the following SQL command (the results for this query were saved in their own table):

SELECT pcr.pdb_id, pd.pdbid,pcr.UniProtID, pd.UniProtID FROM pdb_cross_ref AS pcr JOIN pdb_data as pd WHERE pcr.UniProtID = pd.UniProtID AND pcr.pdb_id = pd.pdbid

This checks that UniProt IDs AND the PDB IDs are both matched. The results of this query were stored in their own table to check the numbers against the UniProt/SwissProt cross reference. The number of returs for this query was 55,505.

To check how many unique PDB-IDs were present in table produced by the above query, this was done with the following query:

SELECT count(distinct pdb_id) FROM pdb_uniprot_cross_ref_check

Returns 41,989 distinct PDB-IDs.

Similarly the following query is used to check how many unique SwissProt entries are present:

SELECT count(distinct UniProtID) FROM pdb_uniprot_cross_ref_check

Returns 12,482 distinct UniProt-IDs

This result indicates that there are only 12,482 entries in SwissProt that are represented in the PDB, in a total of 41,989 structures. This suggests that just over half the PDB is represented in SwissProt.

This discrepancy exists in part because this method relies entirely on matching the ID from each database. Each UniProt entry (both in SwissProt and Trembl) has entries for a "UniProt accession code", which can be used as to link a sequence or structure to a UniProt entry which may have been amened or removed and replaced with a new entry/entries. To see if I could get more matches between the data from the PDB and SwissProt I used the UniProt accession data

in *pdb_cross_ref* to match it against the UniProt ID data in the *UniProt* table. This involved a couple of steps, firstly creating a new table which mapped UniProt ID data with UniProt accessions, using the data in the *UniProt* table - which contains SwissProt data only. This data was stored in a table called *UniProt_ID_to_accession*

Then using the relational database I joined the *pdb_cross_ref*, *pdb_data* and *UniProt_ID_to_accession* tables to find the number of PDB entries for which there was a mappable UniProt accession to a UniProt ID. This data was stored in a new table called *pdb_uprot_acc_cross_ref*.

SELECT COUNT(DISTINCT pdb_id) FROM pdb_uprot_acc_cross_ref

Returned 43,848 distinct PDB-IDs.

SELECT COUNT(DISTINCT UniProtID) FROM pdb_uprot_acc_cross_ref

Returned 12,572 distinct UniProt-IDs.

This decreases the number of of entries for which a discrepancy between the two exist. This result tells us that there are 43,848 PDB entries for which there is agreement between the SwissProt cross reference to the PDB and the PDB cross reference to UniProt. These structures represent 12,572 entries in SwissProt.

C.4 Discussion on the data

The three on-line resources, Pfam-A, UniProt/SwissProt and the PDB each have information which can be used to build relationships with either one or both of the other two. The number of entries in each database for which there is a mapping between them is given in table C.5 below.

Table C.5: A summary of the cross reference data available in each of the three on-line databases. The numbers for SwissProt do not contain entries for viruses.

Database	Map in Pfam-A	Map in SwissProt	Map in PDB
Pfam	5,580	4,493	N/A
PDB	53,748	50,006	44,554
UniProt	19,451	15,077	25,685
SwissProt	12,860	15,077	13,157

We have restricted ourselves to using data which is manually curated and as such is assumed to be correct. However we have taken the step of checking the data in each database against the other. This is a necessary step, but does run into some issues related to how well synchronised each group is with the other. That is to say, these groups are assumed to work in isolation from each other and therefore amendments to the data in one database, will not be known to a group working on another until a public update is made of the their database. As such there are mismatches between the databases' cross-references with each other. This can be understood when we investigate if the cross reference data in Pfam-A corresponds well with the cross reference data in SwissProt. The summary table shows that there are differences in the number of available cross-references.

To understand the discrepancies in the number of matches, let us consider how many of the Pfam-A cross references to UniProt can actually be found in UniProt/SwissProt. Using the

following SQL query, we can try to match the *pfam* table's UniProt references to the *UniProt* UniProtID category, to see how many matches we get:

SELECT p.pfamid, u.UniProtID, p.UniProtID FROM pfam p, UniProt u WHERE p.UniProtID = u.UniProtID;

This says "return the Pfam-ID from the *pfam* table, and the UniProt ID from both the *pfam* and *UniProt* tables, where the UniProt-ID in both tables are the same". This returned a total of 286,281 matches. Which is the number of entries in SwissProt (excluding entries for viruses) for which there are known domains in Pfam-A. This suggests that there are roughly 223,000 entries in SwissProt which are not referenced in Pfam-A.

However these numbers only give an idea of what the ceiling of possible Pfam-to-SwissProt matches is. What is actually of interest is to compare the mutual cross reference data in both UniProt/SwissProt and Pfam-A. To illustrate this, I ran the following query:

SELECT DISTINCT p.pfamid, up.pfamid, p.UniProtID, up.UniProtID FROM pfam p, uPfam_data up WHERE up.UniProtID = p.UniProtID

This query looks identical to the one above, however it is different in that it has the *"up.pfamid"* as an additional field to be returned by the query. There is no restriction on either UniProtID or pfam-ID in each table coming up more than once. The DISTINCT qualifier removes repeat instances of the same combination of returned data.

The query returns a table which has a Pfam-id, from the *pfam* table and one from the *uPfam_data* table, and a UniProt-ID from each of the tables - the table was not stored in the database. Table C.6 below shows the top four entries returned that matched the conditions of the query.

Table C.6: An example of data returned when joining the Pfam table to the SwissProt cross reference table, for all cases where the UniProt-ID in both tables are the same

Pfam Pfam-ID	SProt-cross ref Pfam-ID	SProt UniProtID	Pfam UniProtID	row
PF00190	PF00190	11S2_SESIN	11S2_SESIN	1
PF00190	PF00190	11SB_CUCMA	11SB_CUCMA	2
PF02824	PF01926	128UP_DROME	128UP_DROME	3
PF02824	PF02824	128UP_DROME	128UP_DROME	4
PF01926	PF02824	DRG1_XENLA	DRG1_XENLA	5

The first thing to notice is that in row 3 the Pfam-id is not the same in the *Pfam* and *UniProt* tables. Row 3 suggests that there should be a reference in the Pfam family PF01926 to the UniProt entry 128UP_DROME, according to the SwissProt data. When we check the Pfam-A for a reference to 128UP_DROME, with the following query:

SELECT DISTINCT p.pfamid, up.pfamid, u.UniProtID, p.UniProtID FROM pfam p, UniProt u, uPfam_data up WHERE u.UniProtID = p.UniProtID AND up.UniProtID = u.UniProtID AND p.pfamid =''PF01926'' AND u.UniProtID = ''128UP_DROME''

There is no entry, where the UniProt-ID was the same as that in Row 3. Row 5 of Table C.6 is an example of the result we get. This indicates that there are cases where the data in one database points to an entry in the other, which is not supported by the reciprocal reference data. In this case, it means that the Pfam data-base does not have a record of the family PF01926 being present in the protein 128UP_DROME, while the UniProt database does record this family

as being present. To check the validity of this query, I ran it against the Pfam database on-line at www.pfam.org and searched for the UniProt-ID on the on-line UniProt database. This discrepancy exists on-line too.

Returning briefly to the results in Table C.5, it is apparent that there are similar numbers of entries in SwissProt and Pfam-A with references to the other and the PDB. However, what has not been shown is that all the entries in the two are identical. With our database, it is possible to merge the two data sets and thus improve the number of entries for which we have a three way cross reference data.

The issue raised above exists in matching the UniProt/SwissProt ID in the two databases with each other and also matching the Pfam-ID. One approach is to assume the database which has been most recently updated is the better reference, however that is no guarantee that the entry in question was updated prior to the latest release. The other approach is to try and determine if either one is correct, or both. The method proposed later tries to do this.

For the rest of this project, what is of greatest consideration is to be able to find quickly and easily the number Pfam-A domains present in a PDB structure and the amino acid positions where a domain starts and ends. The second consideration is being able to determine such things as the taxonomy, the cellular location and any other potentially useful information about the structure.

C.5 How the database has been used

It is clear from the last section that there are some discrepancies in the cross-references between each of the on-line databases. However this does not mean that it is not possible to build a data-set with the available information, it would just be smaller (though it could be larger if we could accurately merge the cross-reference data-sets) than might be possible without such stringent restrictions on the source of the data.

We had decided on using the PiQSi database [69] of quaternary structures for the solvent accessibility part of this project, because it would make our results more biologically relevant. Using the data in the MySQL database I created a table for all the PDB IDs present in the PiQSi database, with a selection of additional data from other parts of the database useful to the rest of this project.

C.5.1 The PiQSi database cross references

Using a combination of Python and SQL it has been possible to build a data-set. The PiQSi database has 13,371 unique PDB-IDs mat, derived from the PDB. These were stored in their own table in the database. This was then used to build the cross reference data to Pfam and SwissProt, using the tables described earlier. The cross reference was built first using the cross reference generated from SwissProt, and then again using the cross reference generated from Pfam.

Building the PiQSi cross reference from SwissProt data

Earlier a description was given of how a three way cross reference table was built from the data present in SwissProt. This is used here to match the PDB IDs from PiQSi to determine how much cross reference data was present in SwissProt. Firstly an additional table was created with the following SQL command:

CREATE TABLE PiQSi_UniProt_3way_map_data SELECT DISTINCT pq.pdb_id, up.pfamid, up.UniProtID FROM PiQSi_list AS pq JOIN UniProt_PDB_Pfam_map_uprt AS up WHERE

up.pdbid = pq.pdb_id

This command, creates a table for every entry in the SwissProt three way cross reference table with matching PDB IDs.

Determining PDB the number of PDB IDs To check how many unique PDB IDs are present in the PiQSi database the following query is run:

SELECT COUNT(DISTINCT pdb_id) FROM PiQSi_UniProt_3way_map_data;

This returned a total 11,570 PDB IDs, for which there is a both a SwissProt ID and a Pfam ID.

Determining the number of Pfam IDs The number of Unique Pfam-IDs present in PiQSi is determined by the following query:

SELECT count(DISTINCT pfamid) FROM PiQSi_UniProt_3way_map_data

This returns a total of 995 families from Pfam-A which appear in the PiQSi database.

The question in this case is how many of these are in Pfam-A. This is determined by the following query:

SELECT DISTINCT pfamid FROM PiQSi_UniProt_3way_map_data AS p3 WHERE p3.
pfamid
NOT IN (SELECT pfamid FROM pfam)

This returns 0. Which means all the PDB IDs in the PiQSi database, for which cross reference data exists in SwissProt, have domains in Pfam-A.

Determining the number of UniProt/SwissProt IDs To determine how many unique UniProt/SwissProt entries are represented in PiQSi the following query is used:

SELECT COUNT(DISTINCT UniProtID) FROM PiQSi_UniProt_3way_map_data

Which returns a total of 2,311.

Building the PiQSi cross reference data from Pfam data.

Using the three way cross reference table created from Pfam-A it was possible to build a 3 way cross reference table for PiQSi database, as I did above for the SwissProt data, with the following SQL query:

CREATE TABLE PiQSi_Pfam_3way_map_data SELECT DISTINCT pq.pdb_id,p.pfamid, p.UniProtID FROM PiQSi_list AS pq JOIN PfamPDBmap AS p WHERE p.pdb_id = pq.pdb_id

The PDB data Then the following query was used to determine how many unique PDB IDs were present in the table:

SELECT DISTINCT pdb_id FROM PiQSi_Pfam_3way_map_data

Which returns a total of 13,102.

The SwissProt Data The following query was used to determine how many UniProt IDs are present in the cross reference data:

SELECT DISTINCT UniProtID FROM PiQSi_Pfam_3way_map_data

This returned a total of: 2,670

However the above result has not been filtered for SwissProt, it contains all the UniProt IDS. To address this I checked the entries in the table using the following query:

SELECT DISTINCT UniProtID FROM PiQSi_Pfam_3way_map_data AS p3 WHERE p3.UniProtID IN (SELECT UniProtID FROM UniProt)

Which returned a total of: 2,200.

This query checks for the number of UniProt ID entries in the 3-way map which are also present in the UniProt ID entries in the SwissProt stored (somewhat confusingly in the *uniprot* table - which I will rename).

The above query was modified with the NOT IN condition:

SELECT DISTINCT UniProtID FROM PiQSi_Pfam_3way_map_data AS p3 WHERE p3.UniProtID NOT IN (SELECT UniProtID FROM UniProt)

Which returned a value of 470. This was to check that the total number of UniProt IDs matched the two conditions (IN and NOT IN), which it does

The Pfam-A data Next we determine how many Pfam families are present in the 3-way cross reference with the following command:

SELECT DISTINCT pfamid FROM PiQSi_Pfam_3way_map_data

which returns a value of: 1,120

C.5.2 Using the map

We nee to know the location of a Pfam domain in a structure from the PDB. In the data supplied both by SwissProt and Pfam-A there are start and end positions in the SwissProt sequence for the domains. However these do not follow a generalised or consistent rule. This can become problematic when trying to create automated systems which rely on the existence of generalisations in order to ensure accuracy. This was a problem encountered by the MSc. project student who worked on the co-evolution project before and she was unable to find a good work around. Rather than relying on the data present in the databases entirely, I have opted to use a different approach.

There exists a module for Python called Tre, which evaluates the Debye distance between two sequences, to determine how closely matched they are. An allowance of 10% mismatch was made in this case, because it was found that a threshold up to 20% did not improve the selection at all. This module is used to serve two functions: a) to ensure that a domain is present in a structure when the cross-reference data from the database says it should be, and b) to return the start and end position of the domain, if it is found in the structure.

A Python program was written which uses the maps developed in Section C.3 in conjunction with the Tre module (to deal with missing sequences or structures, and artifacts in the data), to build a table for the PiQSi. For each entry in the PiQSi database, the table contains the main data associated with the proteins in the quaternary structure:

- Pfam ID
- PDB ID
- UniProt ID
- UniProt keywords
- UniProt Comments

- UniProt Taxonomy
- Pfam Sequence
- seq_start
- seq_end

Using the Tre module in Python to cross check the Pfam sequence in the structure, was done in the following steps:

1. Extract the amino acid sequence from the atomic structure data section in the PDB file.
2. Interrogate the MySQL database and return the cross reference data already parsed from the SwissProt, Pfam and PDB. The program checks all the matches returned.
3. Using the Python module Tre, match the Pfam sequence in the PDB sequence and return the start and end points of the sequence.
   (a) IF there is no match for the given Pfam sequence, try to match any Pfam sequence from that Pfam family to the PDB structure and return the start and end points of that sequence.
   (b) ELSE return none
4. If a match has been found, insert the PDB-ID, the Pfam-ID, the SwissProt-ID and the start and end points of the Pfam-domains into the PiQSi_pfam_swissprot table in the MySQL database

The Python program populated the PiQSi_pfam_swissprot table with a total of:

C.5.3 Making specialised selections

The reason for including the data from UniProt, such as the comments and key words, is that SQL allows for the matching of words in SELECT queries. For example, if we wish to select all SwissProt entries which are associated with membranes, the following query could be used:

SELECT $*$ FROM uniprot WHERE keyword LIKE ‘‘\%membrane\%’’ OR comments LIKE ‘‘\%membrane\%’’.

Combinations of different specialised selections can be used to make increasingly specific selections. For example, I was able to make a selection for all Pfam domains in PiQSi which are non-membrane, non-DNA binding and exclusively cytosolic. This resulted in a total of 192 Pfam domains being selected. This was done using the table generated by the Python code discussed earlier.

As there are a collection of different tables present in the database, with quite a large amount, even if incomplete, mapping data between Pfam-A, SwissProt and the PDB, it is possible to use SQL queries to make selections which would otherwise be very complicated to achieve. Using regular expressions in any programming language code that can be written to interrogate a MySQL database, or in SQL itself, there exists now the possibility of making selections of PDB or Pfam data which is conditional on its existence in UniProt/SwissProt and criteria set against the data held for the respective entry.

A limitation is that it is not possible to extract structure data directly from the database. To do this, I had to write a program which will extract a domain from a PDB file and store it in a separate file in PDB format. I have done this already for the PiQSi database and will do it for the entire PDB if it is necessary.

C.5.4 Possible improvements.

Beyond the obvious need to improve the consistency of naming of tables and categories in each of the tables, which will be cleaned up before this is made available to others, there are some possible steps which could make this a more useful tool.

Firstly, it could be useful to store the entire Pfam-A database in multiple tables, including sequence alignments and all comments and key words for each family. This would provide the prospect of setting selection criteria against both SwissProt and Pfam-A. Secondly the full PDB could be moved into the database. This would make it possible to select domains directly from within a PDB structure just by querying the database. This would also open up the possibility of very novel queries to be asked about the data which could support other types of statistical inquiries.

Another useful possibility would be to use the Tre module to resolve an discrepancies between each of the cross-references, generated from the different data-banks. The example given in Table C.6, could be resolved by using Tre to check the presence of the Pfam sequences in the SwissProt sequence. If a discrepancy is resolved in this way, then the relevant table can be updated accordingly.

C.6 Conclusions

The new database provides a method to select structural data based on criteria such as cellular location, taxonomy or pfam domain. The merger of data held in three different on-line databases/data-banks, SwissProt, Pfam-A and the PDB has made this possible.

The database has been used to build a cross reference for the PiQSi database of manually curated structures. From that cross reference a very careful selection of Pfam domains has been made. That will now be used for the solvent accessibility analysis and the subsequent co-evolution analysis. However the database, does have potential for many other uses. For example, the use of the Tre module in Python is used to check for the start and end points of a Pfam sequence in a PDB/PiQSi structure ensures that the domain is present and is correctly mapped to the structure. This module can also be used to resolve discrepancies which exist between the respective cross reference data.

Appendix D
Appendix: Naccess reference states

Table D.1: Torsion angles for all reference tripeptides used by Naccess [4]. The PDB reference files were supplied by Simon Hubbard, the *torsion.py* script included in the program LINUS [5] was used to calculate the torsion angles shown here.

Tripeptide	ϕ	ψ	ω	χ_1	χ_2	χ_3	χ_4
ALA	—	134.99	179.99	—	—	—	—
ALA	-139.98	134.97	-180.00	—	—	—	—
ALA	-139.99	—	—	—	—	—	—
ALA	—	134.99	179.99	—	—	—	—
ARG	-139.98	134.97	-180.00	-179.97	-179.98	-179.99	179.99
ALA	-139.99	—	—	—	—	—	—
ALA	—	134.99	179.99	—	—	—	—
ASN	-139.98	134.97	-180.00	-179.95	-179.25	—	—
ALA	-139.99	—	—	—	—	—	—
ALA	—	134.99	179.99	—	—	—	—
ASP	-139.98	134.97	-180.00	-179.95	0.67	—	—
ALA	-139.99	—	—	—	—	—	—
ALA	—	134.99	179.99	—	—	—	—
CYS	-139.98	134.97	-180.00	179.37	—	—	—
ALA	-139.99	—	—	—	—	—	—
ALA	—	134.99	179.99	—	—	—	—
GLN	-139.98	134.97	-180.00	-179.97	-179.97	179.98	—
ALA	-139.99	—	—	—	—	—	—
ALA	—	134.99	179.99	—	—	—	—
GLU	-139.98	134.97	-180.00	-179.97	-179.97	-0.02	—
ALA	-139.99	—	—	—	—	—	—
ALA	—	134.99	179.99	—	—	—	—
GLY	-139.98	134.97	-180.00	—	—	—	—
ALA	-139.99	—	—	—	—	—	—
ALA	—	134.99	179.99	—	—	—	—
HIS	-139.98	134.97	-180.00	-179.97	-89.91	—	—
ALA	-139.99	—	—	—	—	—	—
ALA	—	134.99	179.99	—	—	—	—
ILE	-139.98	134.97	-180.00	-60.83	-180.00	—	—
ALA	-139.99	—	—	—	—	—	—

Continued on next page

Table D.1 – *Continued from previous page*

Tripeptide	ϕ	ψ	ω	χ_1	χ_2	χ_3	χ_4
ALA	—	134.99	179.99	—	—	—	—
LEU	-139.98	134.97	-180.00	-179.97	60.03	—	—
ALA	-139.99	—	—	—	—	—	—
ALA	—	134.99	179.99	—	—	—	—
LYS	-139.98	134.97	-180.00	-179.97	-179.98	-180.00	-179.99
ALA	-139.99	—	—	—	—	—	—
ALA	—	134.99	179.99	—	—	—	—
MET	-139.98	134.97	-180.00	-179.97	-179.98	-179.99	—
ALA	-139.99	—	—	—	—	—	—
ALA	—	153.42	175.22	—	—	—	—
PRO	-68.18	161.87	175.59	0.01	6.41	—	—
ALA	-176.80	—	—	—	—	—	—
ALA	—	135.02	-179.96	—	—	—	—
PHE	-140.00	134.95	-180.00	180.00	-89.98	—	—
ALA	-139.97	—	—	—	—	—	—
ALA	—	135.03	-179.96	—	—	—	—
SER	-139.99	134.98	179.99	179.96	—	—	—
ALA	-140.01	—	—	—	—	—	—
ALA	—	134.98	-180.00	—	—	—	—
THR	-139.98	135.03	-179.96	-60.82	—	—	—
ALA	-140.03	—	—	—	—	—	—
ALA	—	135.00	179.96	—	—	—	—
TRP	-139.99	135.06	179.98	-66.87	59.80	—	—
ALA	-140.02	—	—	—	—	—	—
ALA	—	134.99	179.95	—	—	—	—
TYR	-140.02	134.98	-179.96	179.98	89.94	—	—
ALA	-139.96	—	—	—	—	—	—
ALA	—	134.96	179.99	—	—	—	—
VAL	-139.94	135.01	179.97	—	—	—	—
ALA	-140.00	—	—	—	—	—	—

Appendix E
Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii

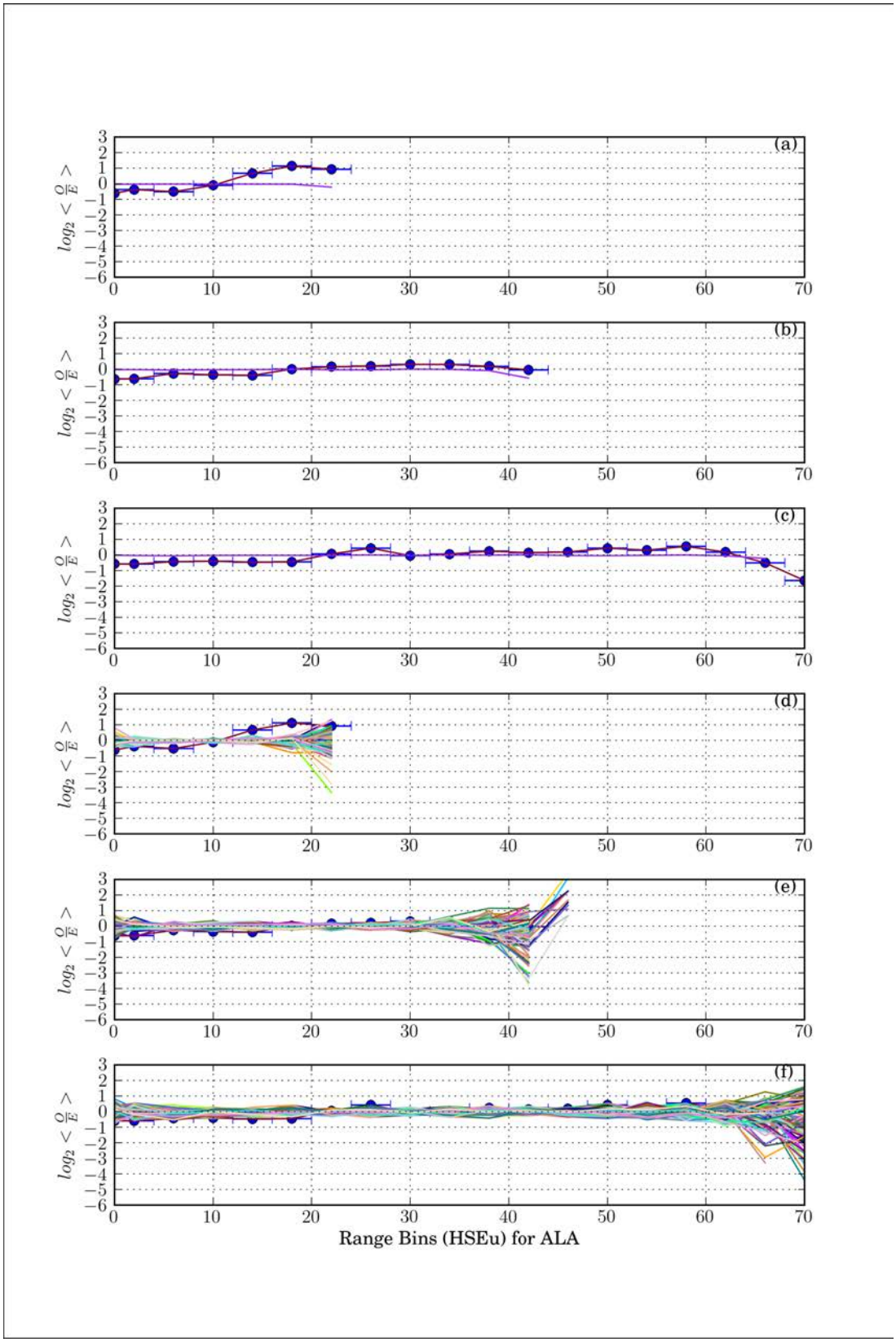

Figure E.1: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Ala: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

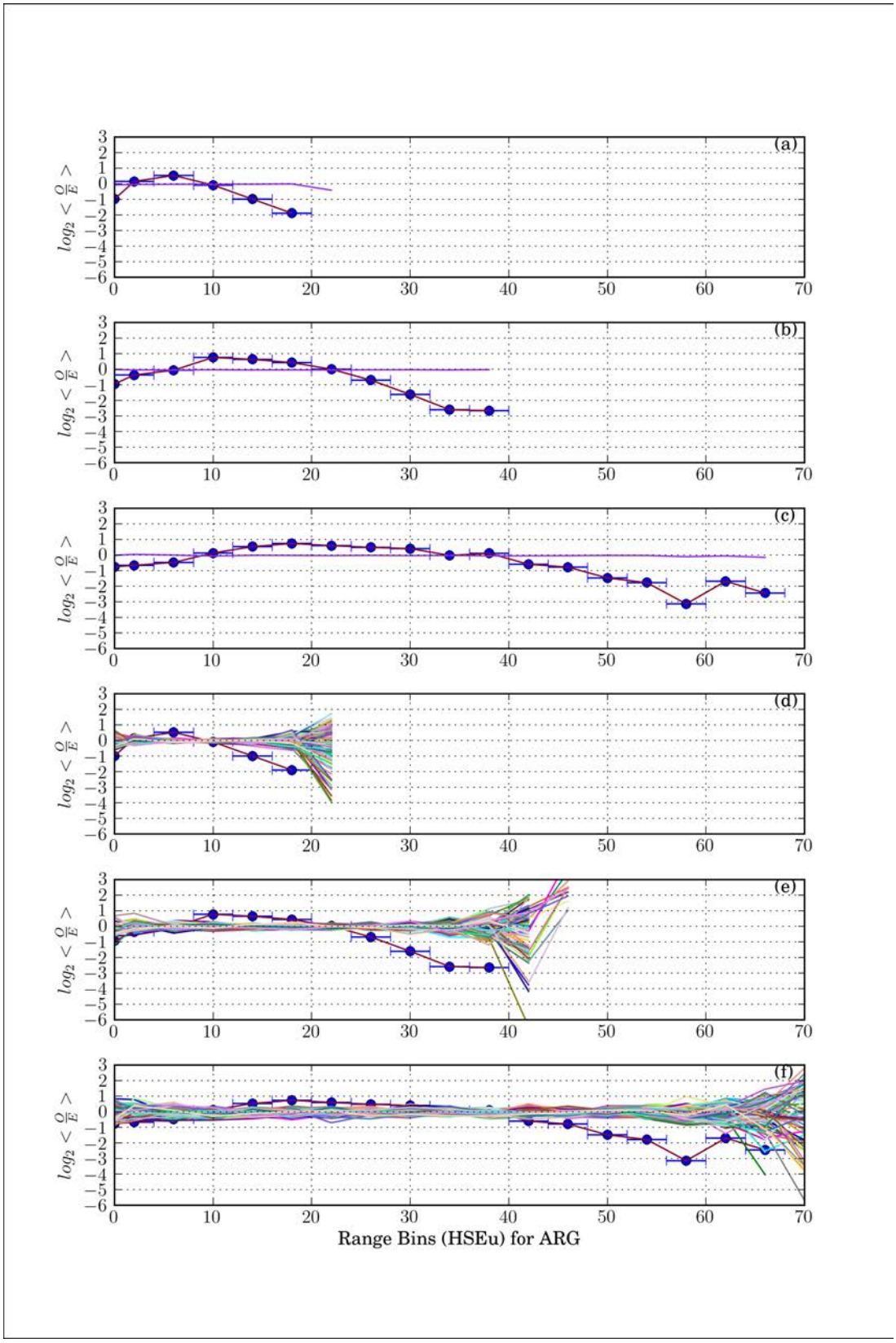

Figure E.2: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Arg: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

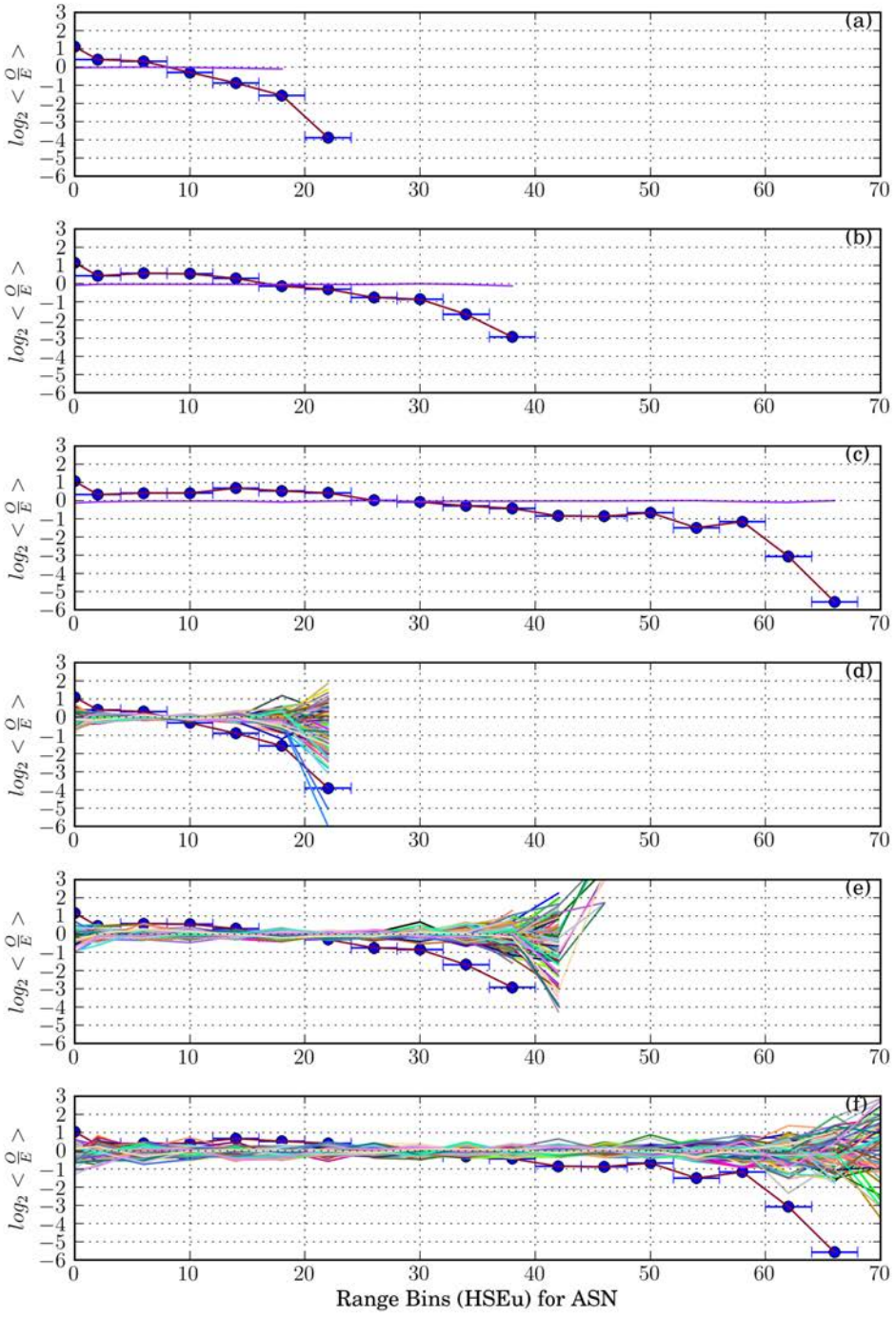

Figure E.3: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Asn: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

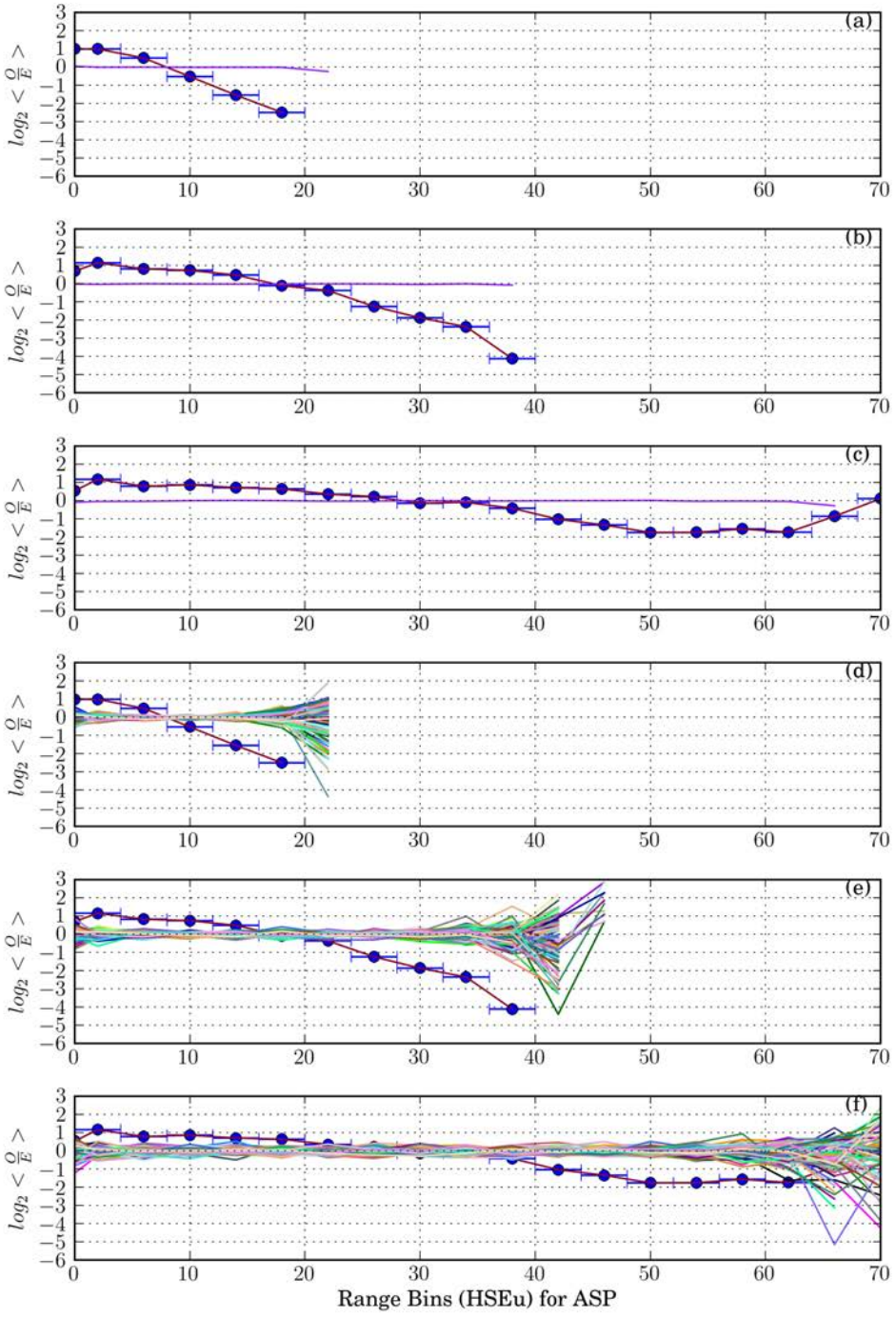

Figure E.4: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Asp: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

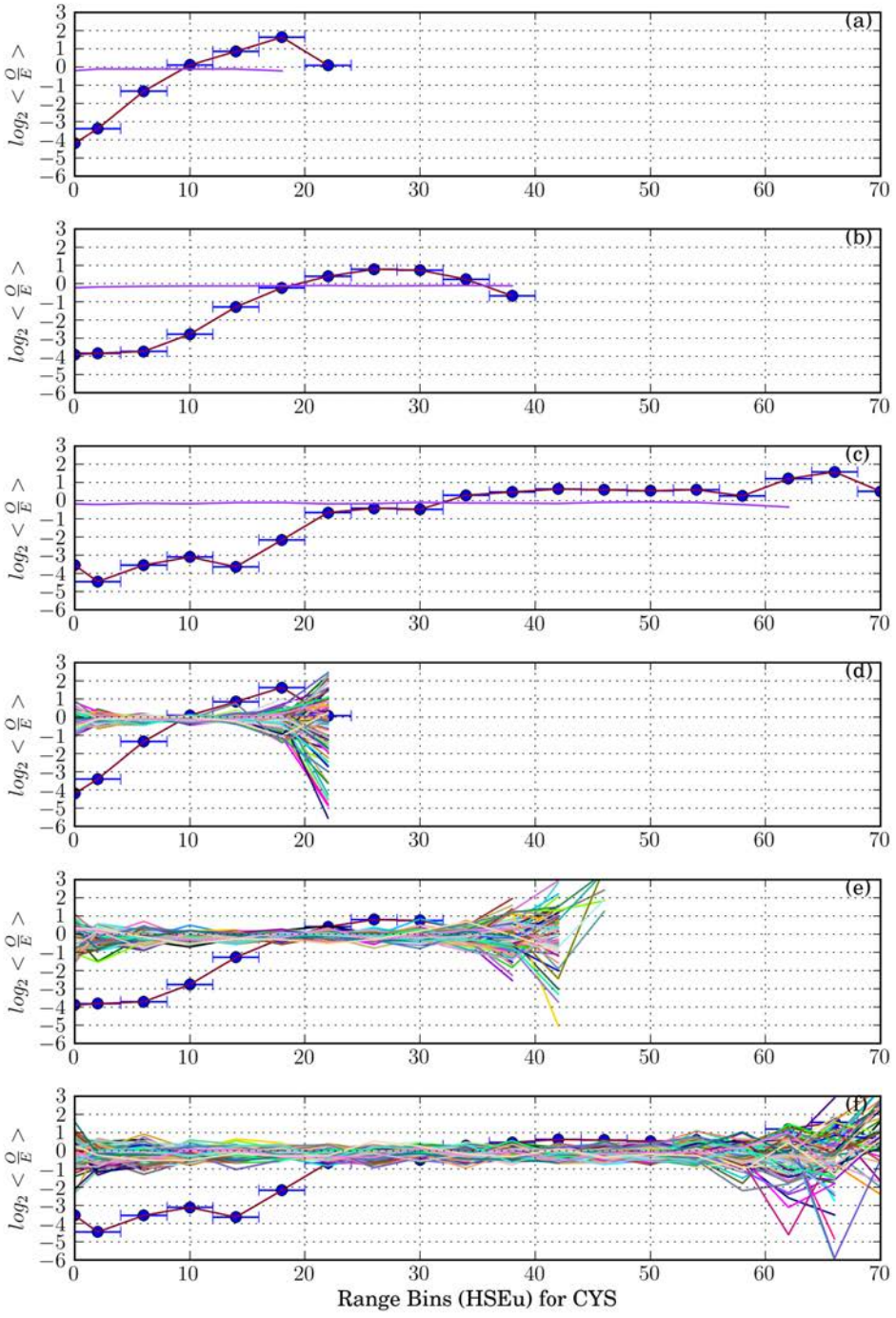

Figure E.5: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Cys: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

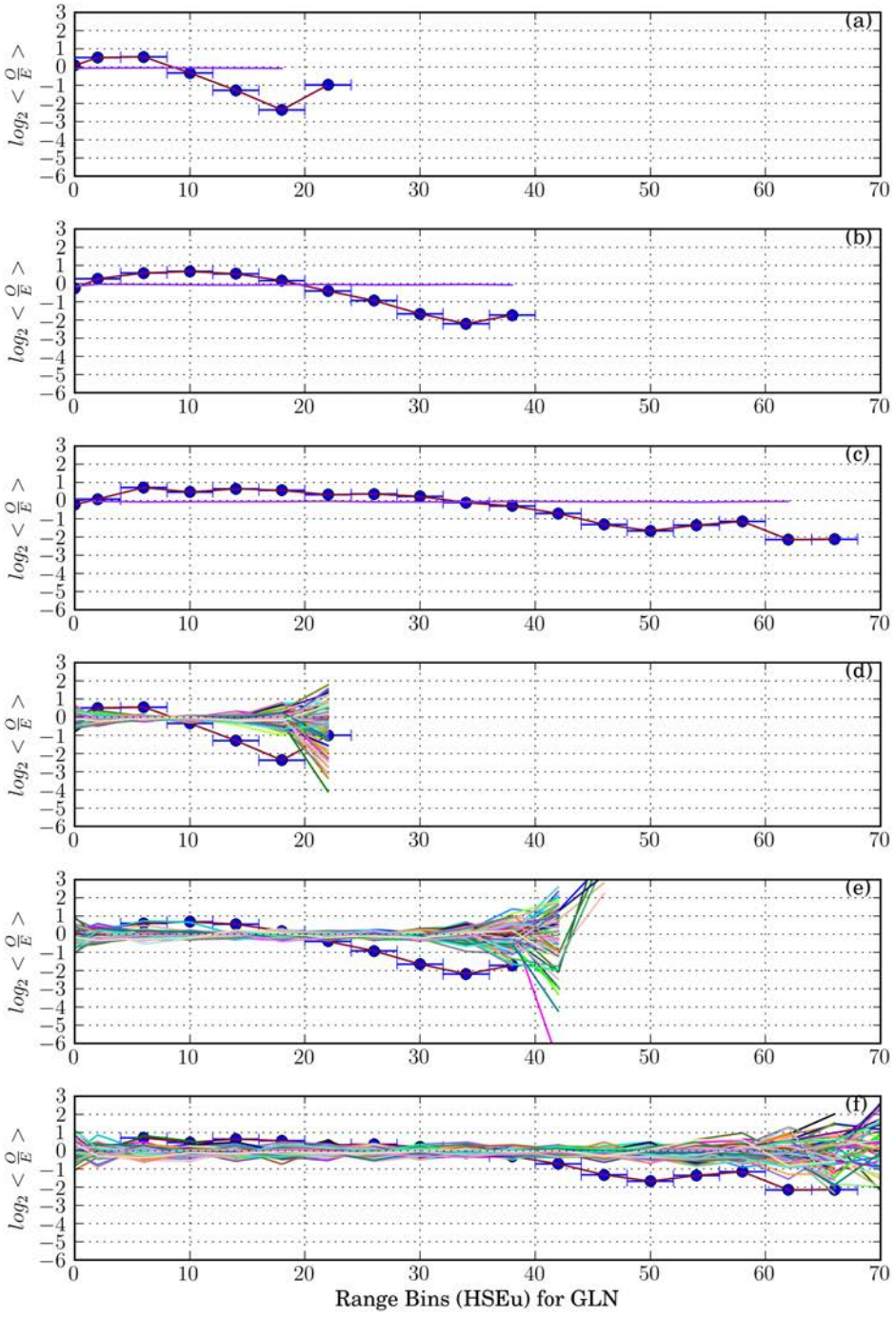

Figure E.6: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Gln: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

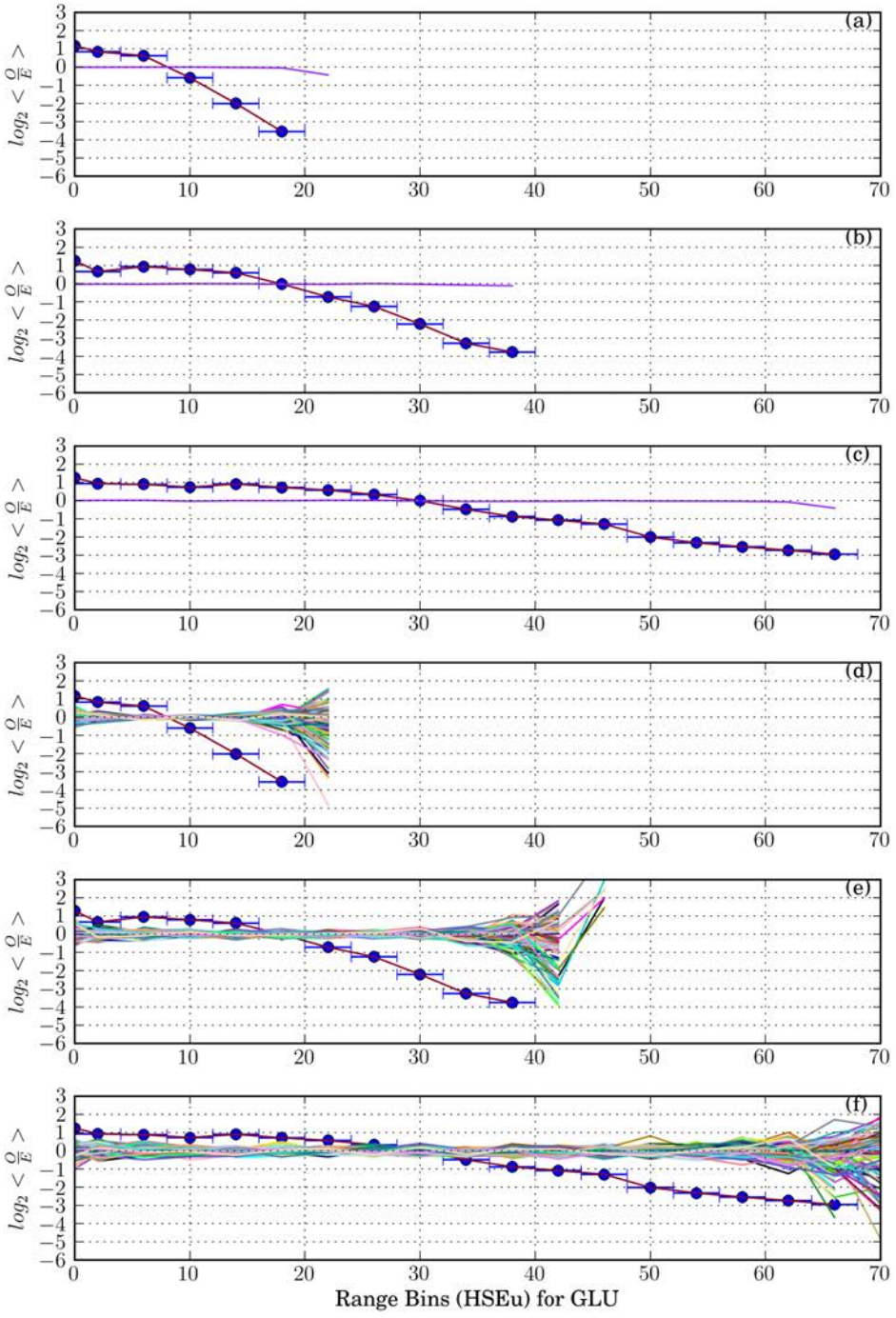

Figure E.7: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Glu: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

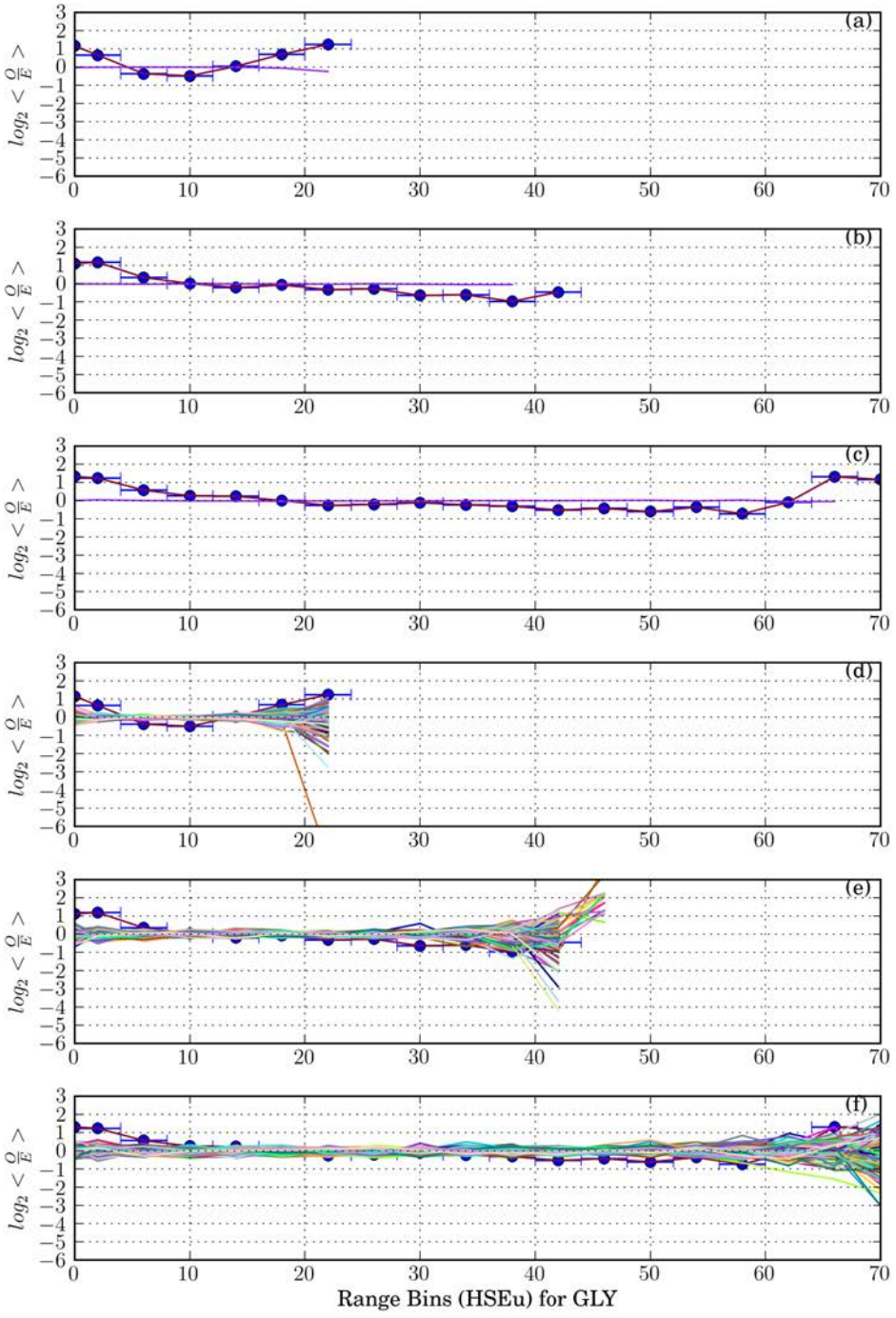

Figure E.8: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Gly: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

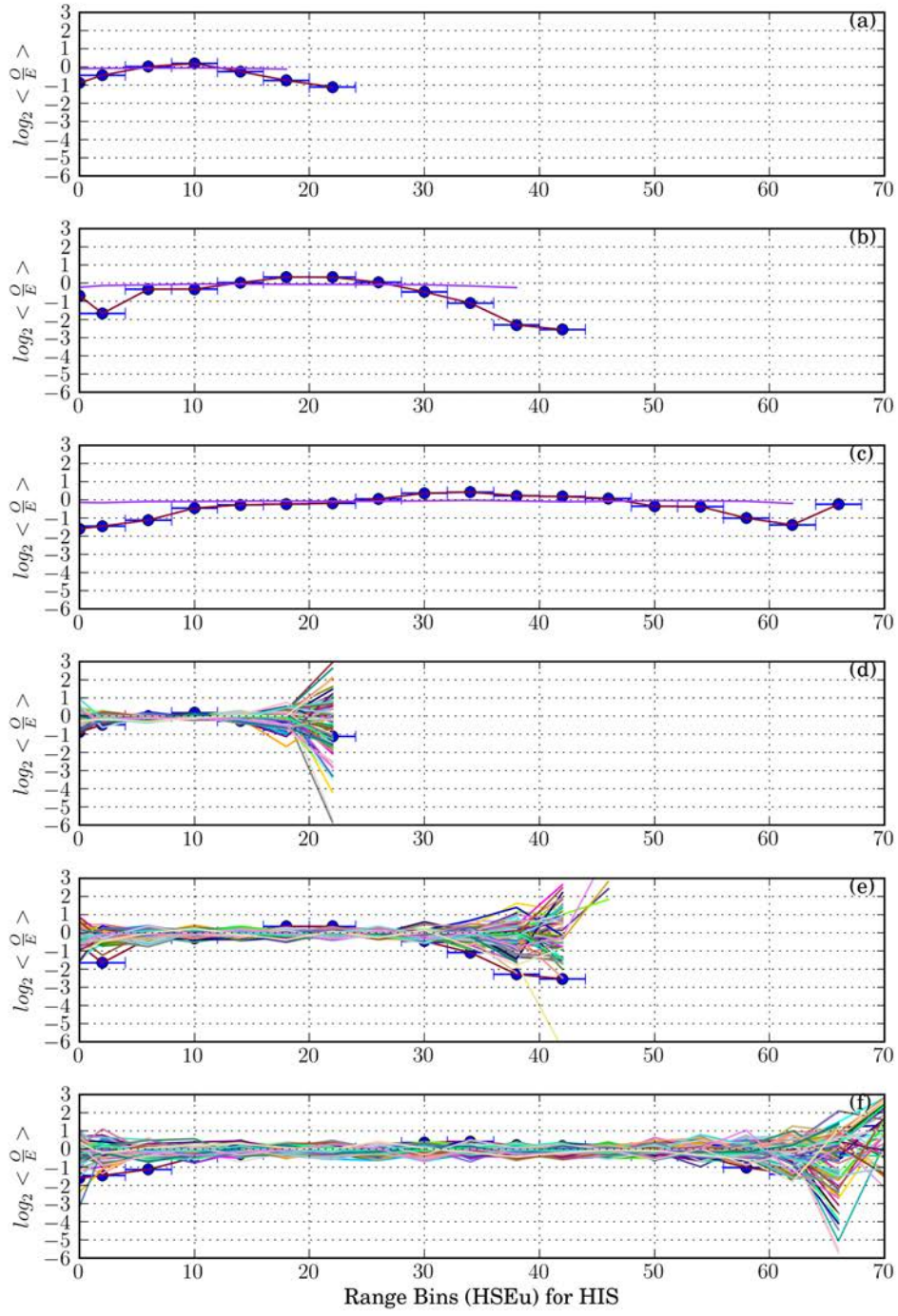

Figure E.9: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for His: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

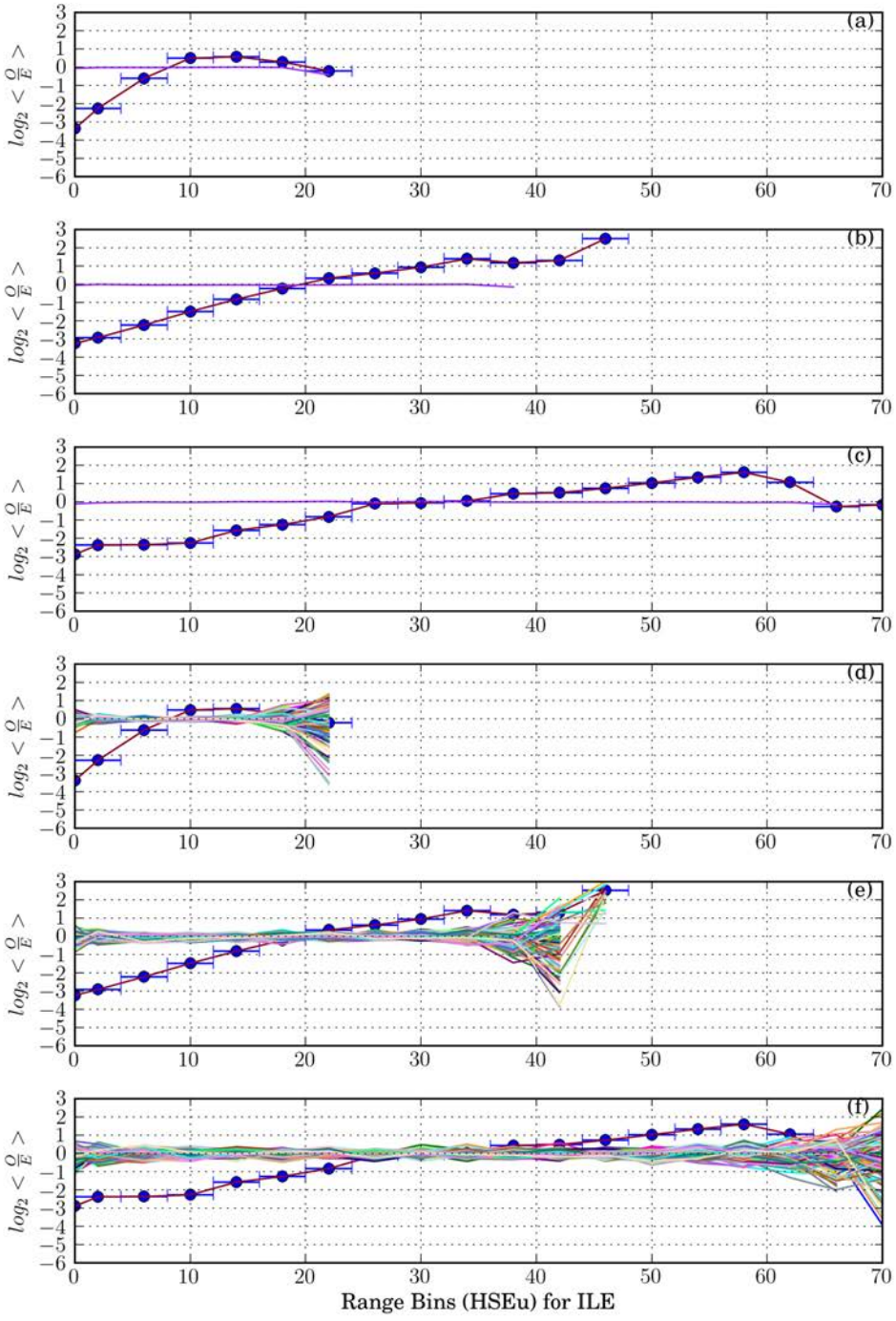

Figure E.10: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Ile: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

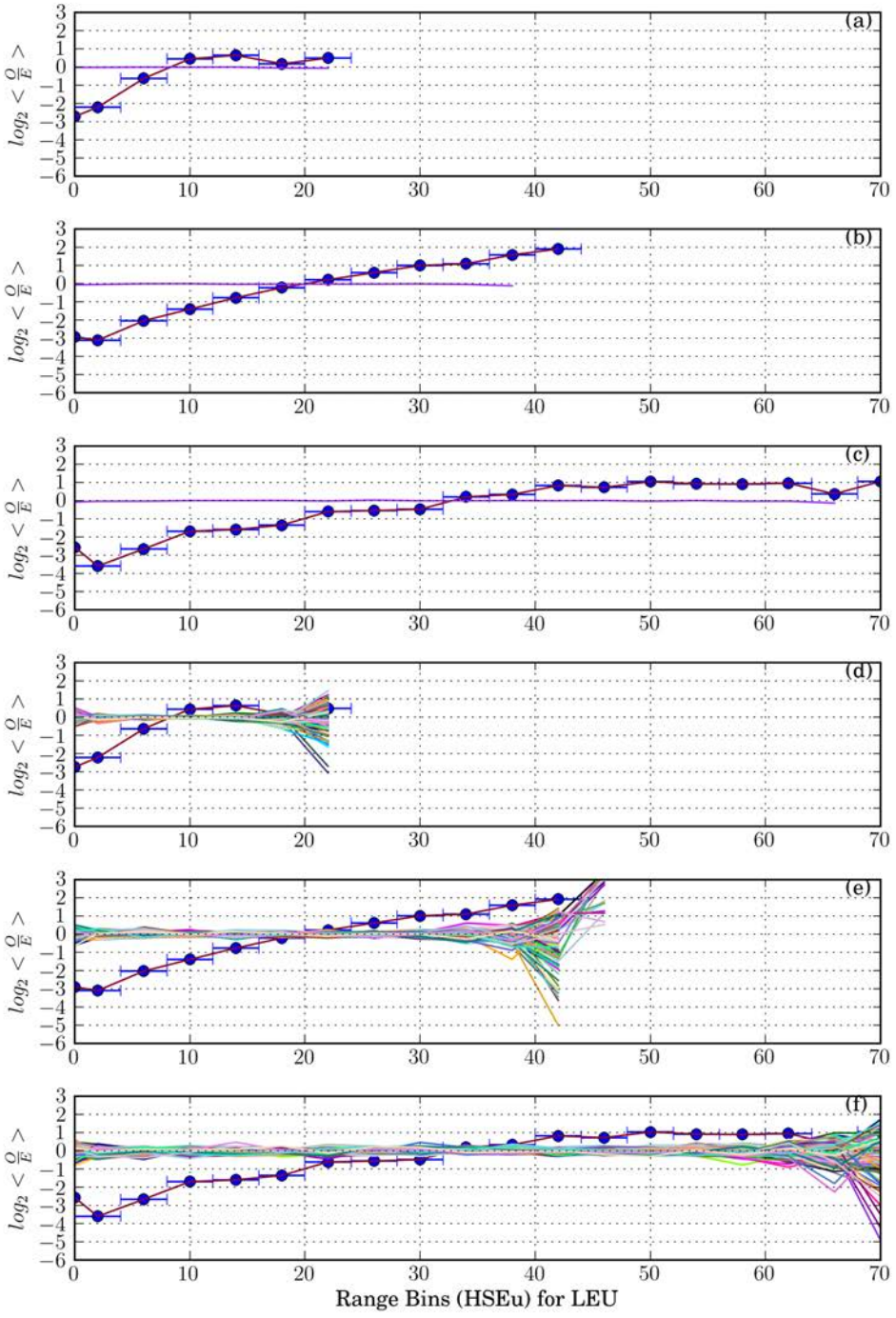

Figure E.11: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Leu: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

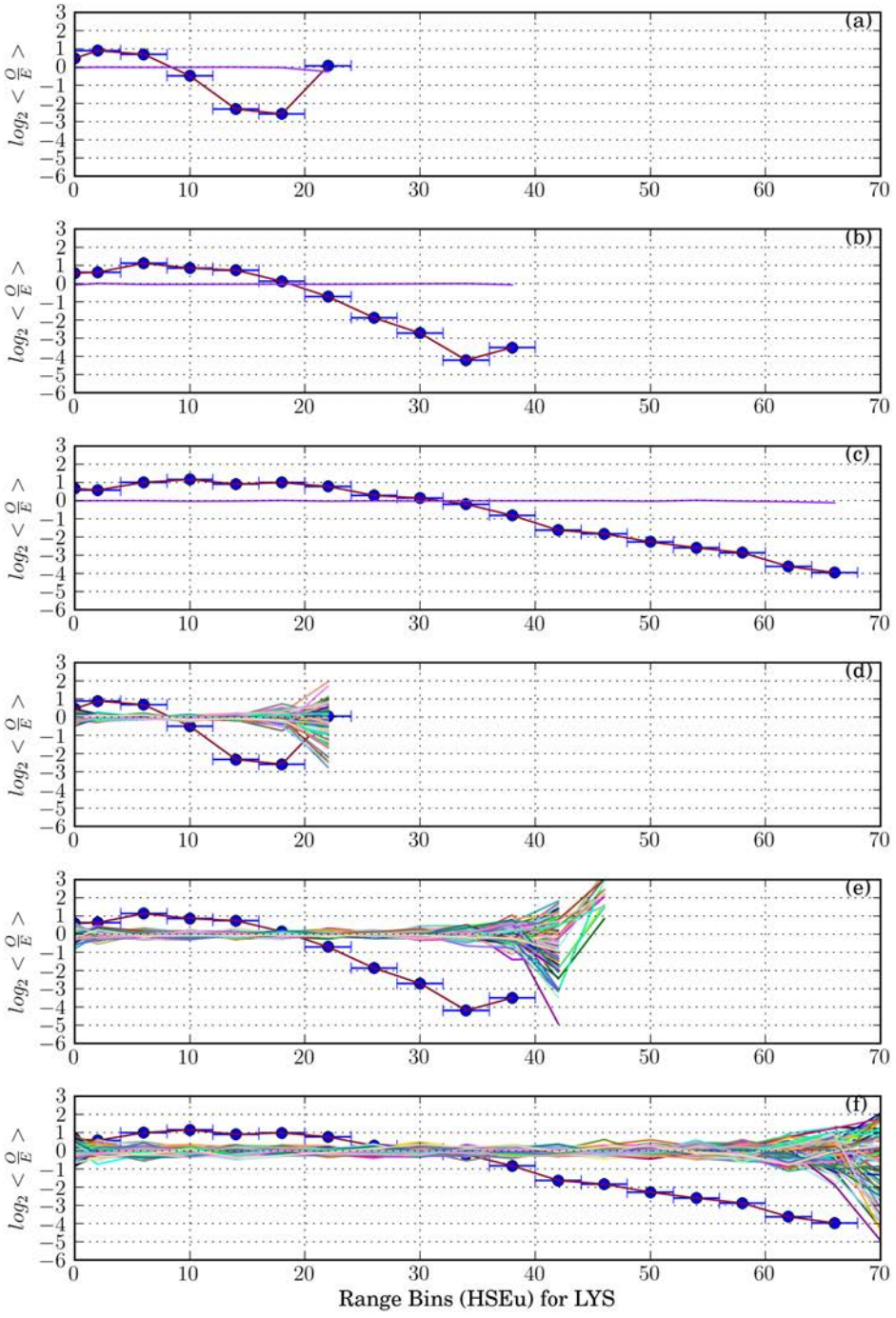

Figure E.12: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Lys: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

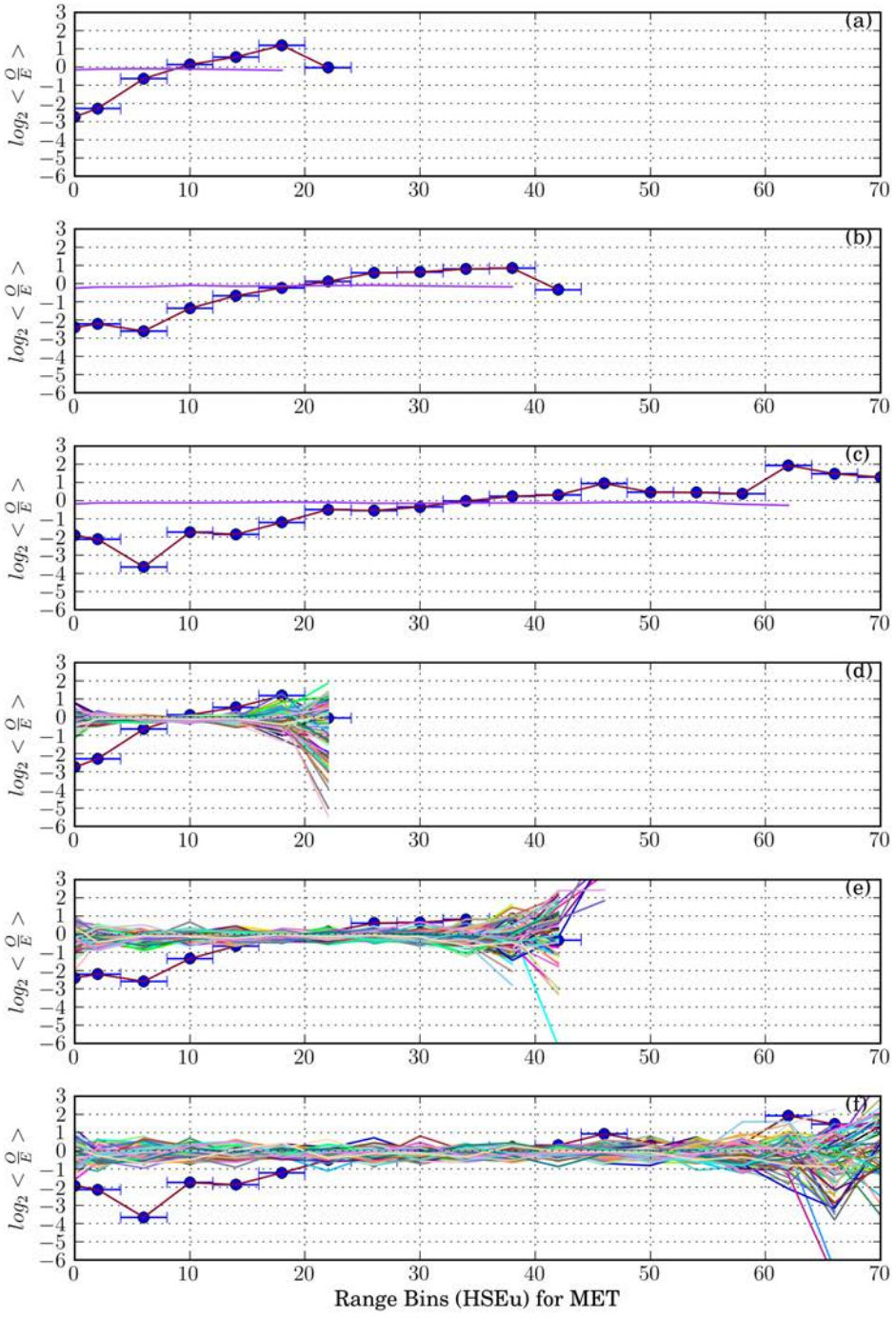

Figure E.13: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Met: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

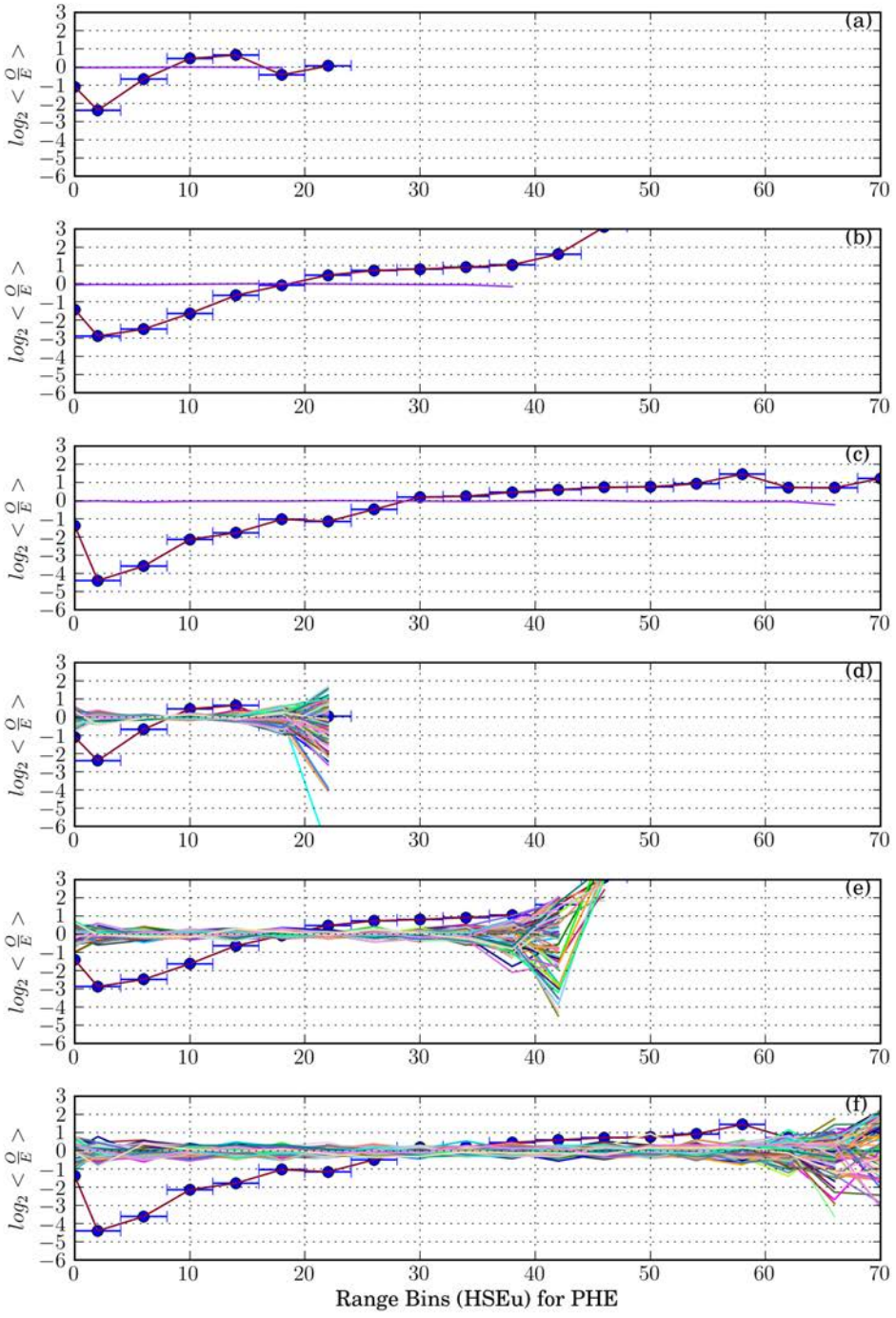

Figure E.14: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Phe: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

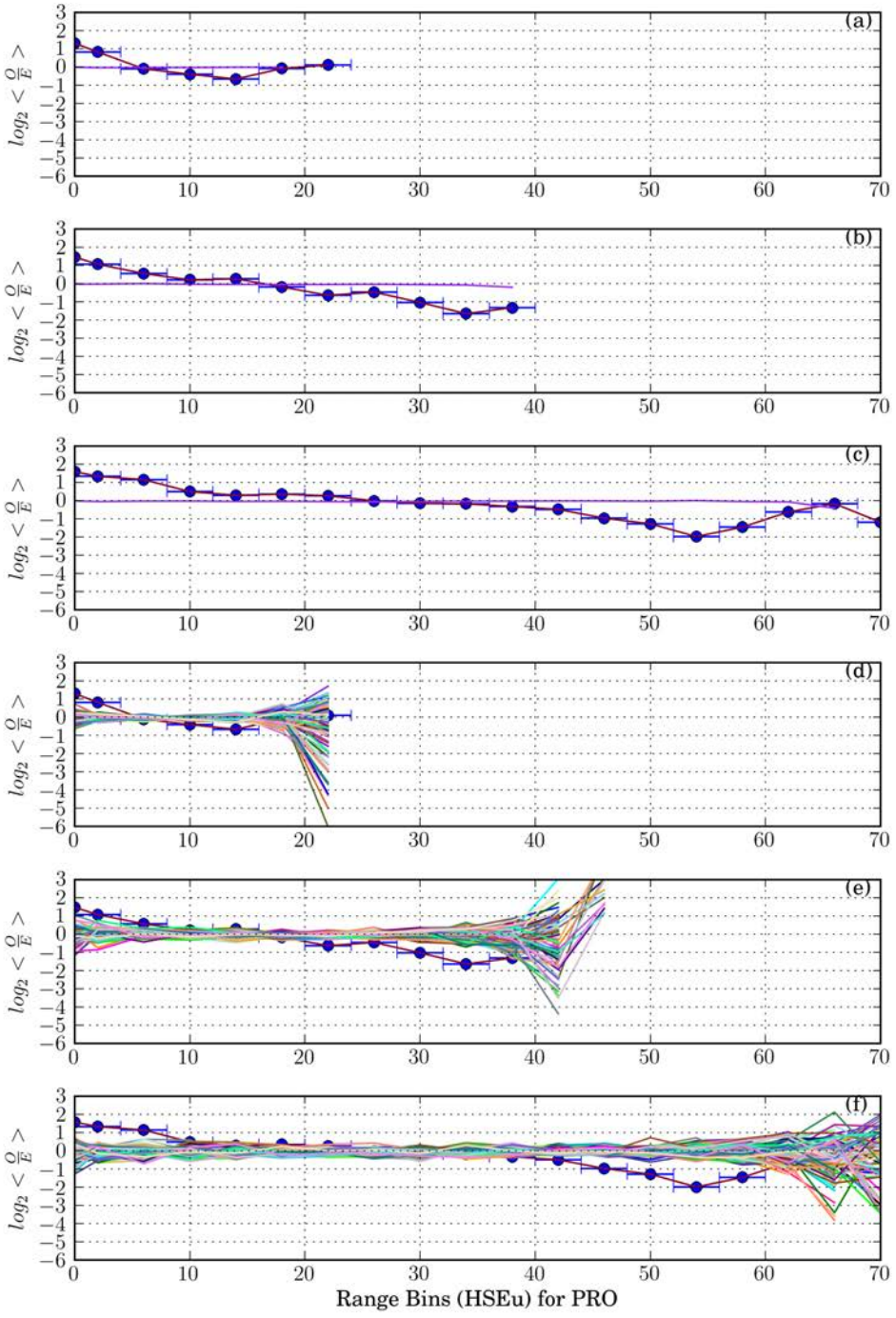

Figure E.15: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Pro: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

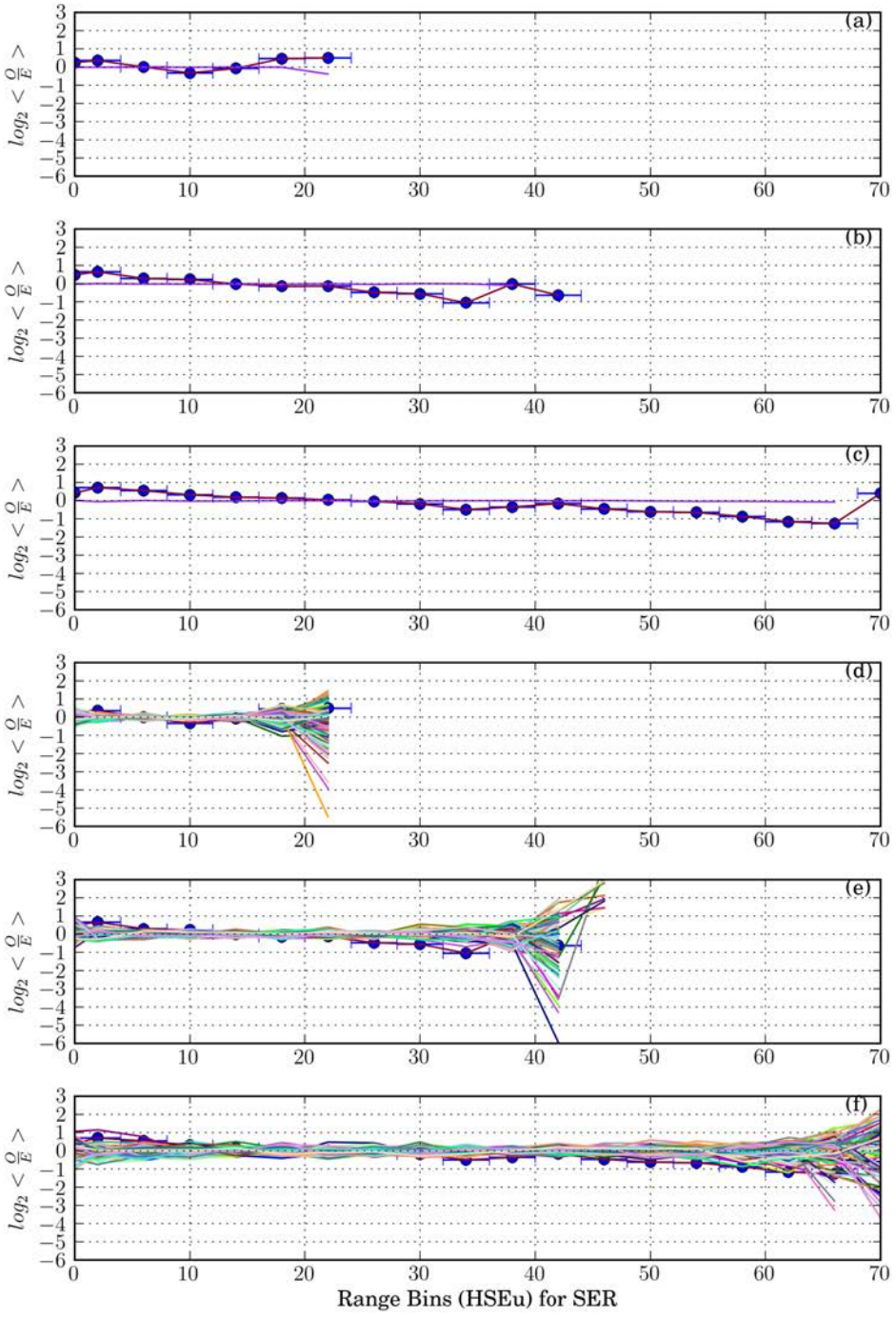

Figure E.16: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Ser: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

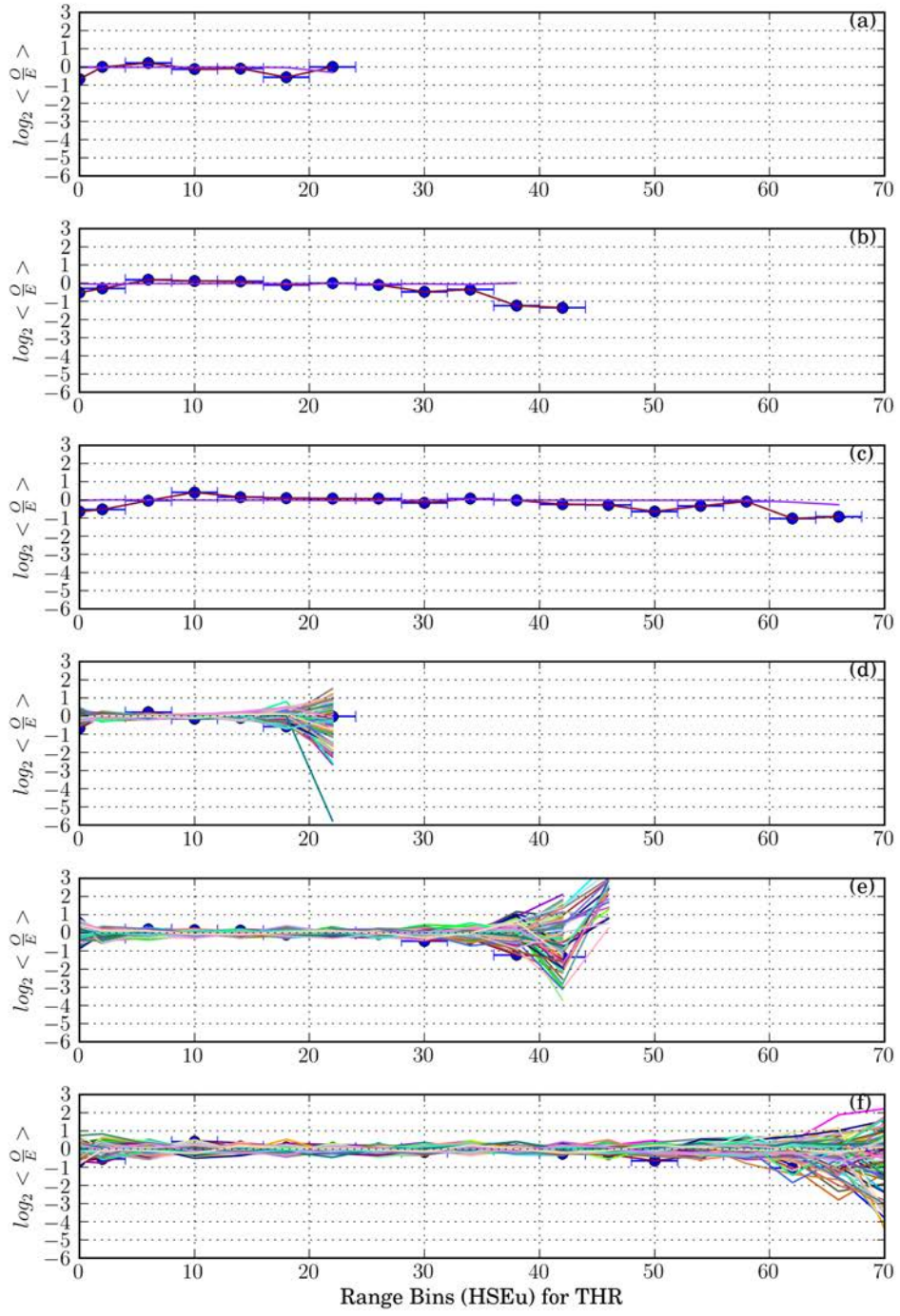

Figure E.17: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Thr: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

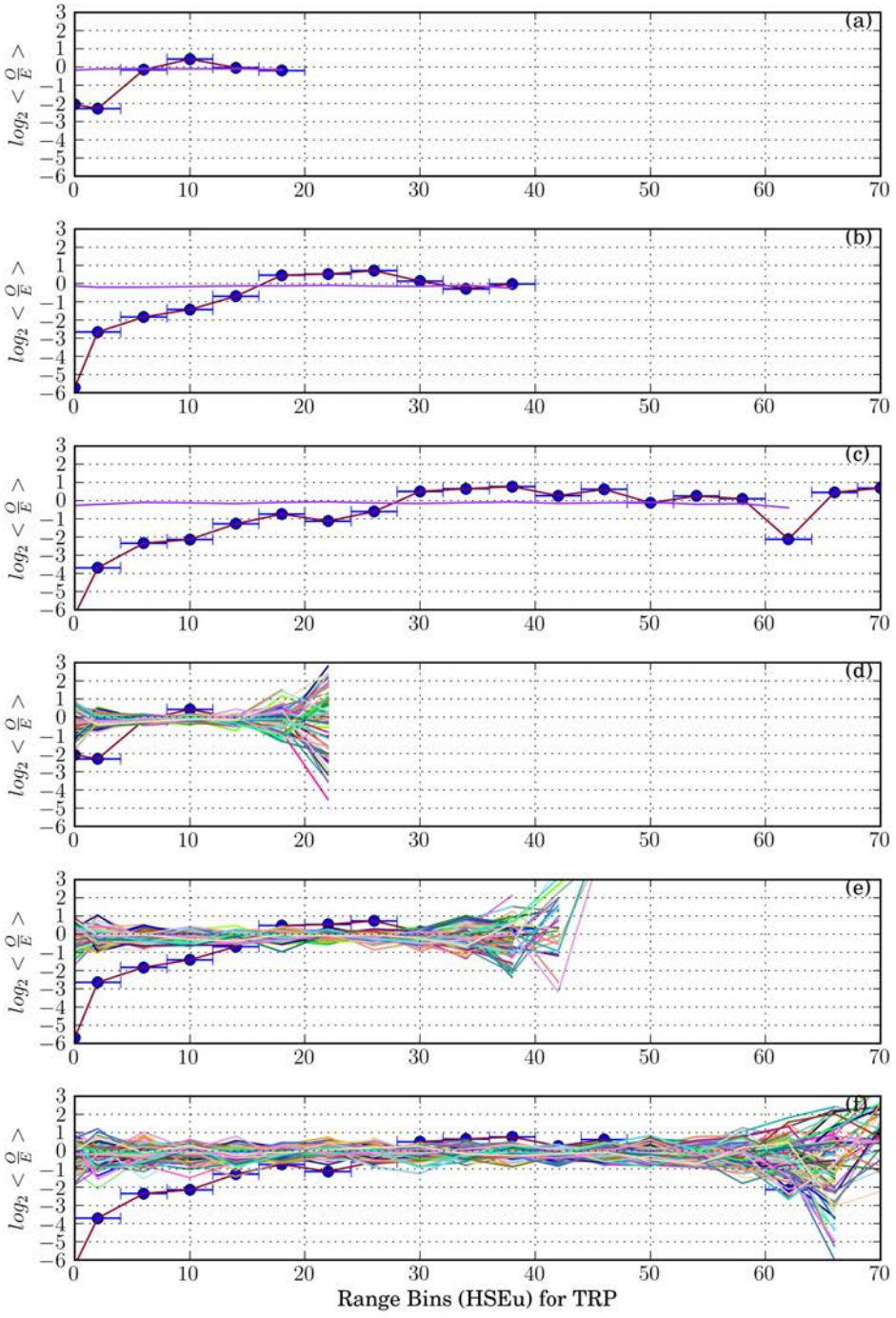

Figure E.18: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Trp: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

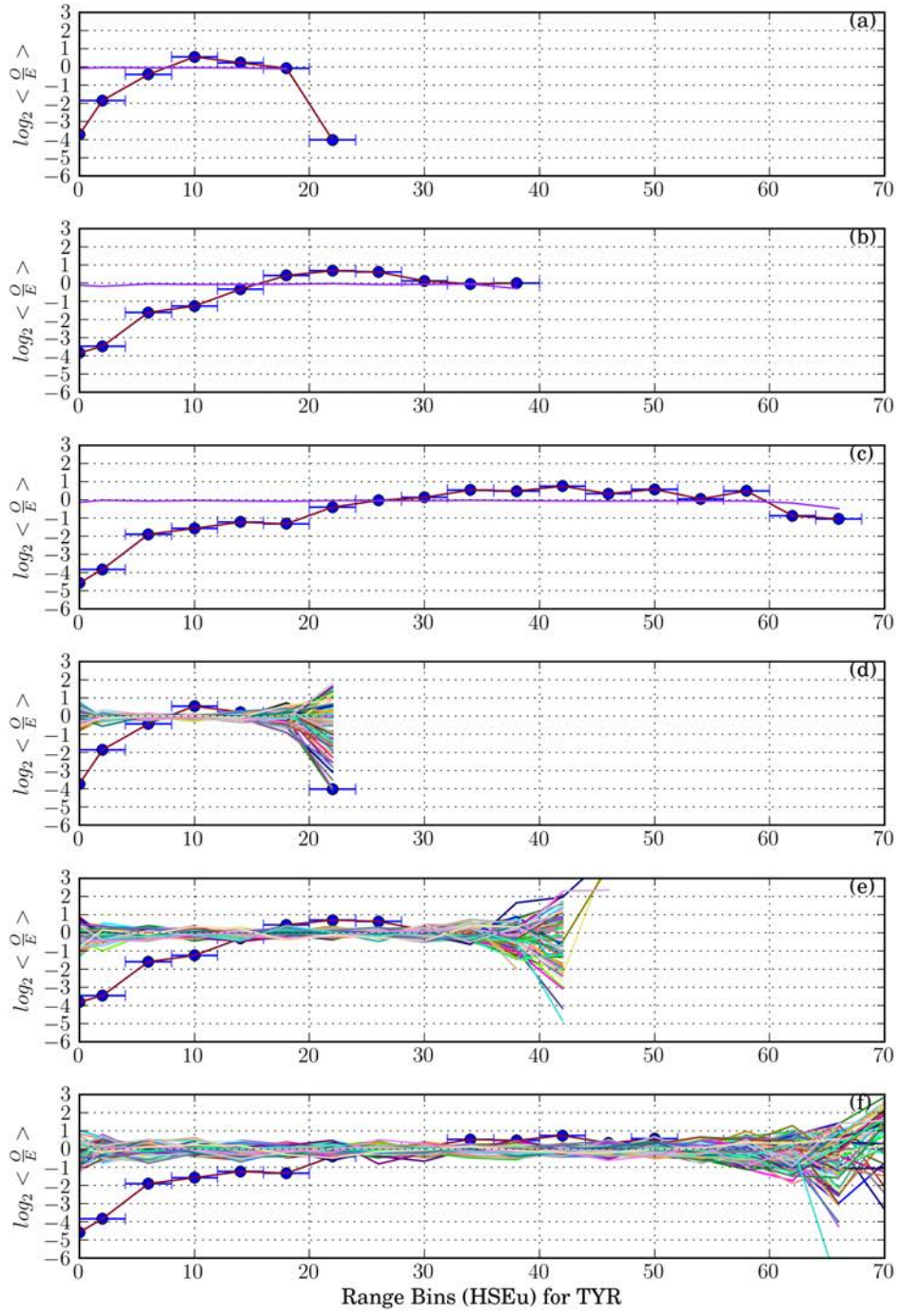

Figure E.19: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Tyr: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

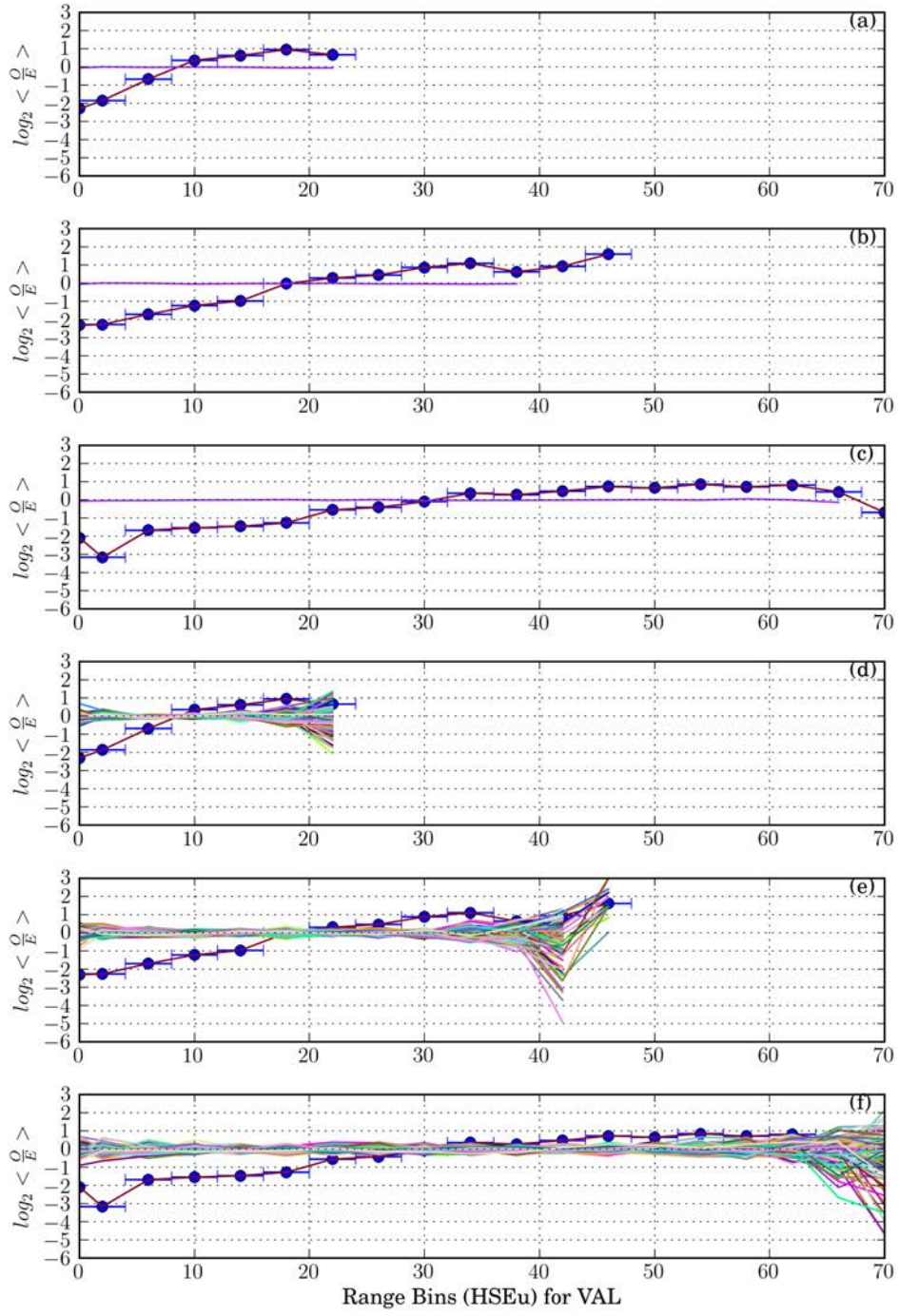

Figure E.20: Comparison of $log_2\langle\frac{O}{E}\rangle$ vs. HSEu for 3 different radii, with bootstrap data, for Val: The average $log_2\langle\frac{O}{E}\rangle$ for 100 bootstraps are shown in (a)with HSEu radius 10, (b) with HSEu radius 13, (c) with HSEu radius 16. The individual bootstrap lines are shown in (d) with HSEu radius 10, (e) with HSEu radius 13 and (f) radius 16 respectively.

Appendix F
Scatter Plots of $HSEu_{13}$ vs ASA

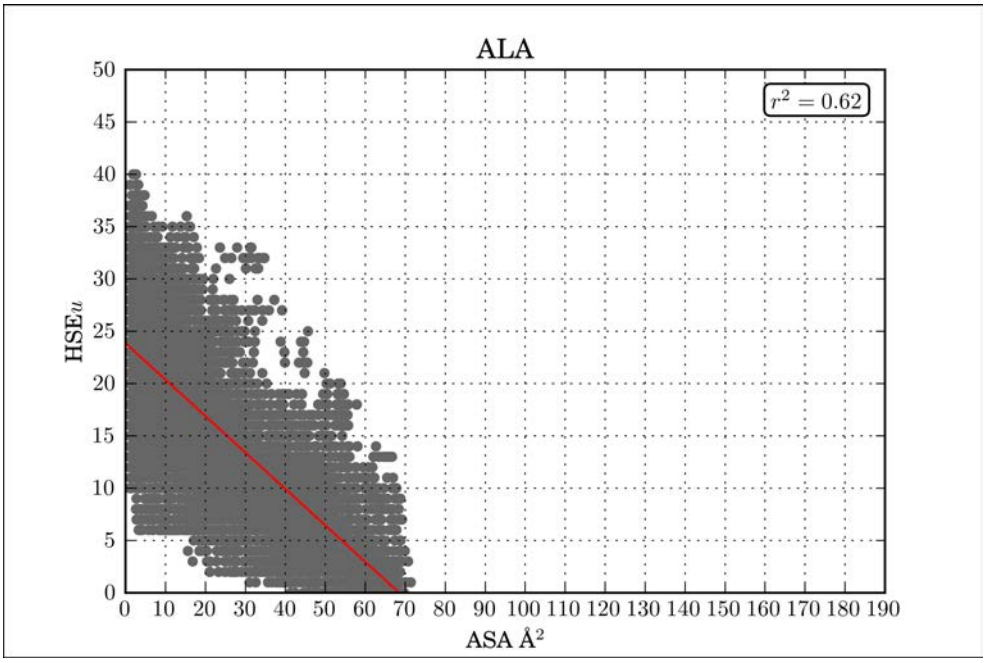

Figure F.1: Scatter plot of HSEu$_{13}$ vs Side chain ASA, for Alanine: Each point represents a single instance of the residue-type, for which both HSEu$_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between HSEu$_{13}$ and ASA for this residue type.

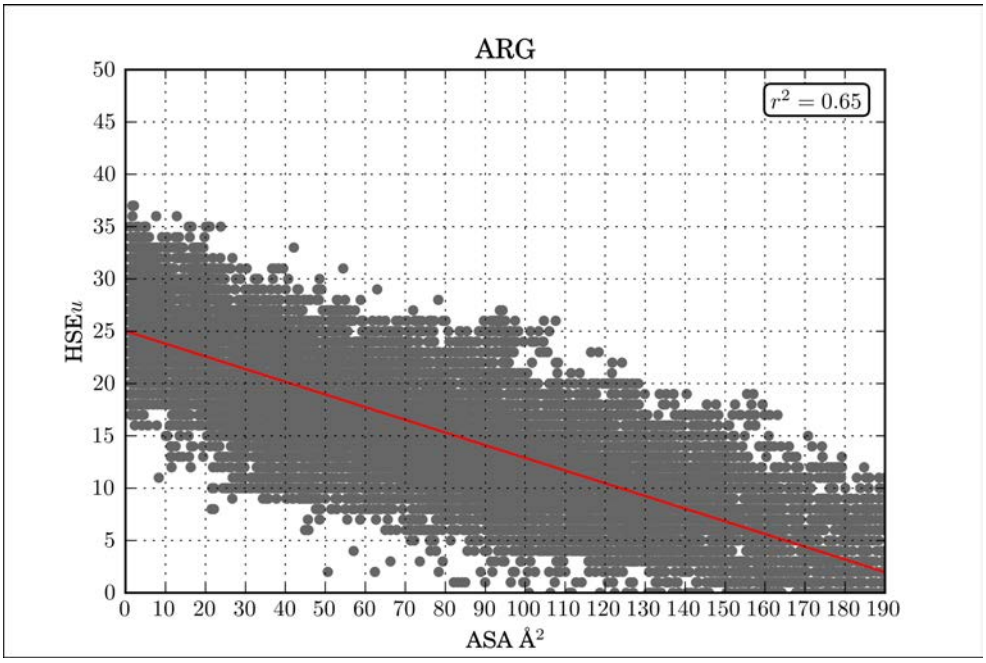

Figure F.2: Scatter plot of HSEu$_{13}$ vs Side chain ASA, for Argenine: Each point represents a single instance of the residue-type, for which both HSEu$_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between HSEu$_{13}$ and ASA for this residue type.

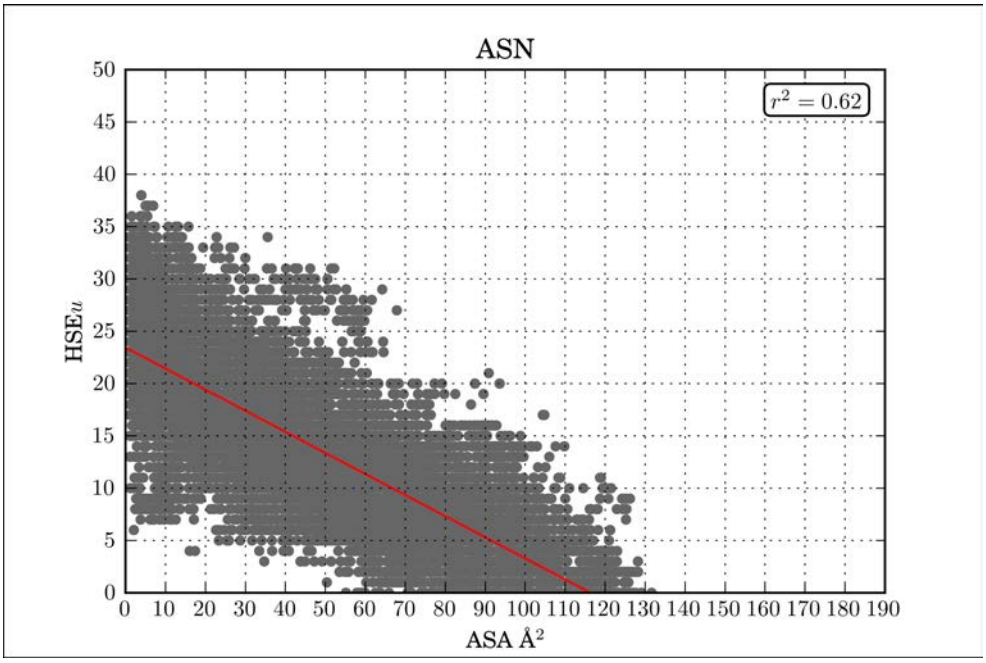

Figure F.3: Scatter plot of HSEu$_{13}$ vs Side chain ASA, for Asparagine: Each point represents a single instance of the residue-type, for which both HSEu$_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between HSEu$_{13}$ and ASA for this residue type.

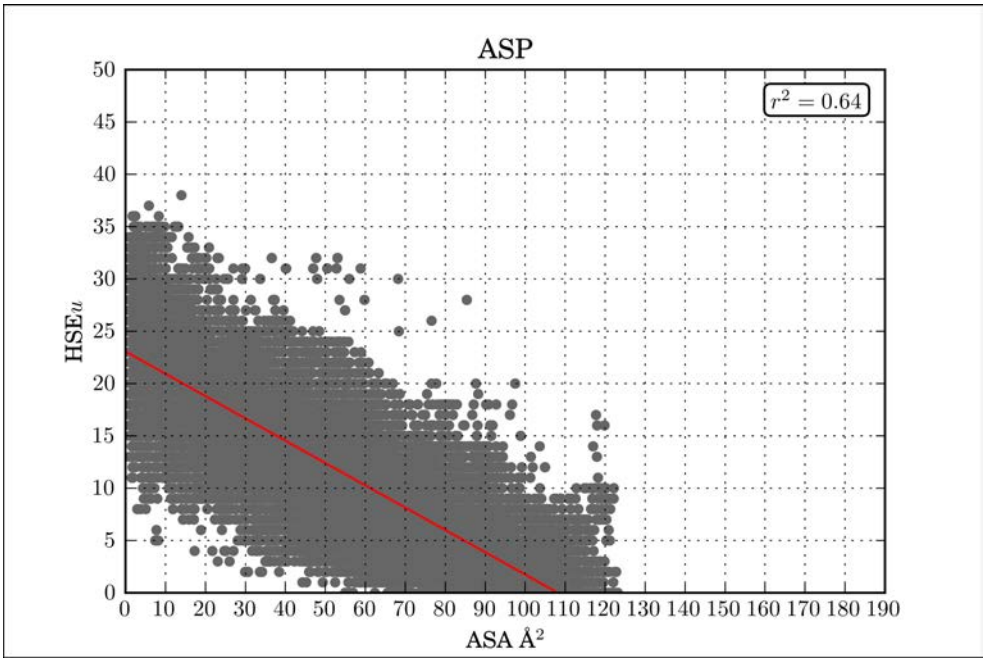

Figure F.4: Scatter plot of HSEu$_{13}$ vs Side chain ASA, for Aspartic Acid: Each point represents a single instance of the residue-type, for which both HSEu$_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between HSEu$_{13}$ and ASA for this residue type.

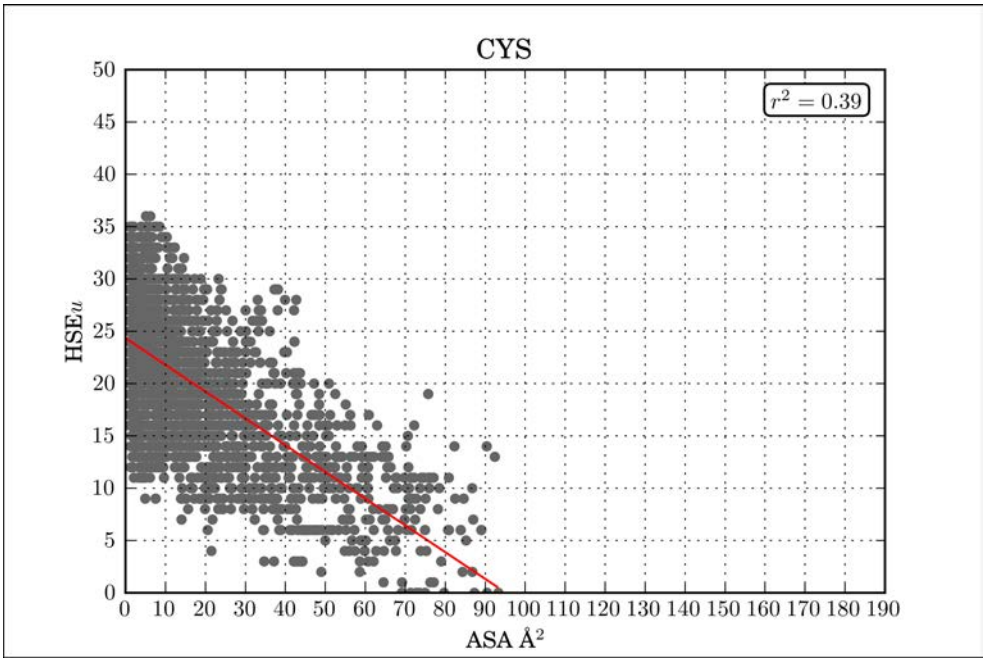

Figure F.5: Scatter plot of HSEu$_{13}$ vs Side chain ASA, for Cystine: Each point represents a single instance of the residue-type, for which both HSEu$_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between HSEu$_{13}$ and ASA for this residue type.

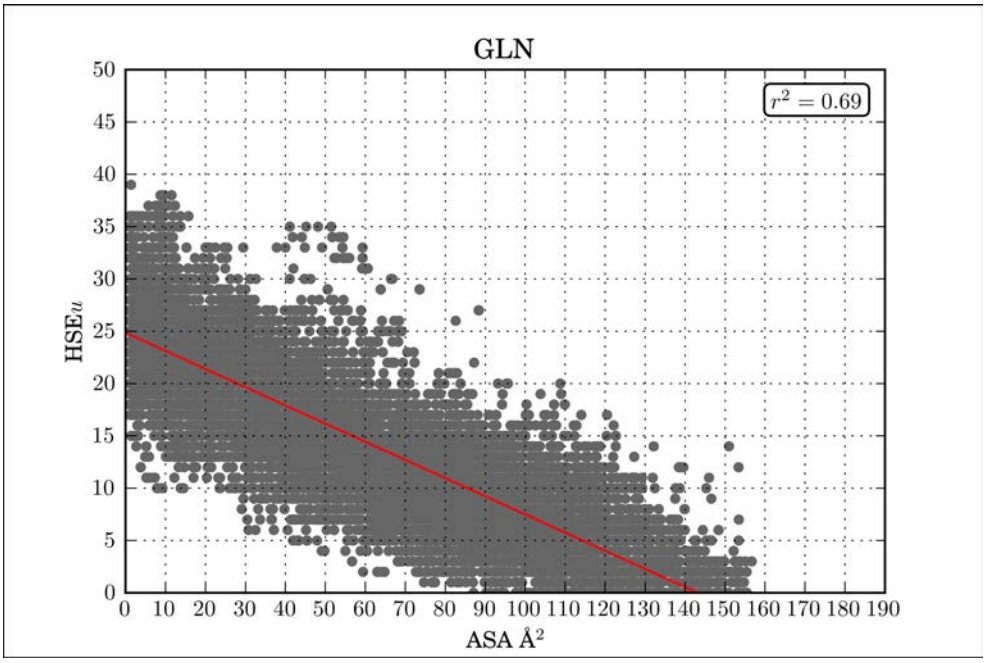

Figure F.6: Scatter plot of HSEu$_{13}$ vs Side chain ASA, for Glutamine: Each point represents a single instance of the residue-type, for which both HSEu$_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between HSEu$_{13}$ and ASA for this residue type.

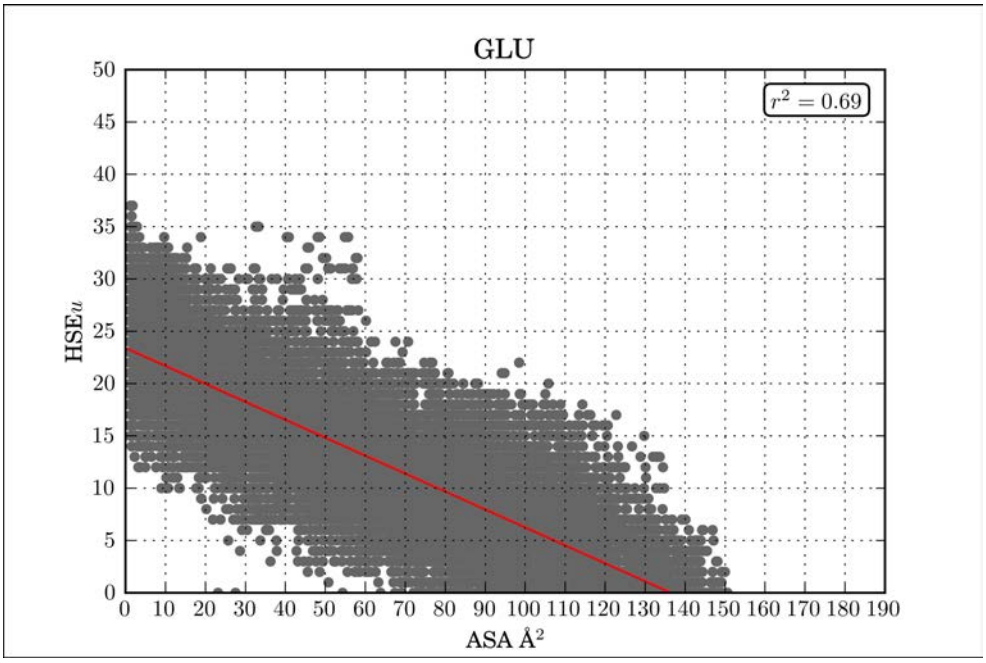

Figure F.7: Scatter plot of HSEu$_{13}$ vs Side chain ASA, for Glutamic Acid: Each point represents a single instance of the residue-type, for which both HSEu$_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between HSEu$_{13}$ and ASA for this residue type.

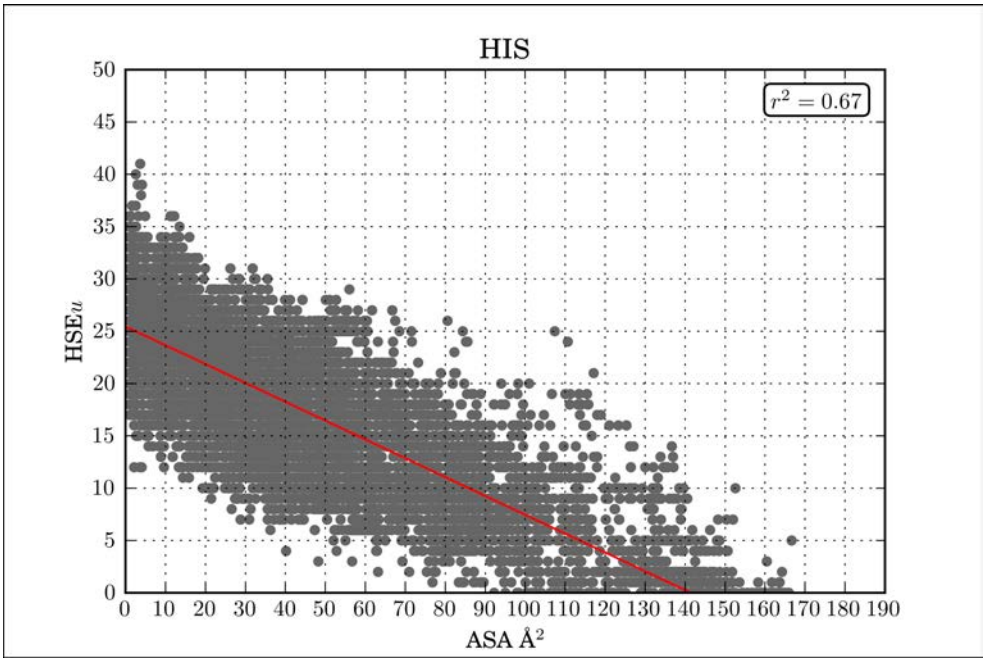

Figure F.8: Scatter plot of HSEu$_{13}$ vs Side chain ASA, for Histidine: Each point represents a single instance of the residue-type, for which both HSEu$_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between HSEu$_{13}$ and ASA for this residue type.

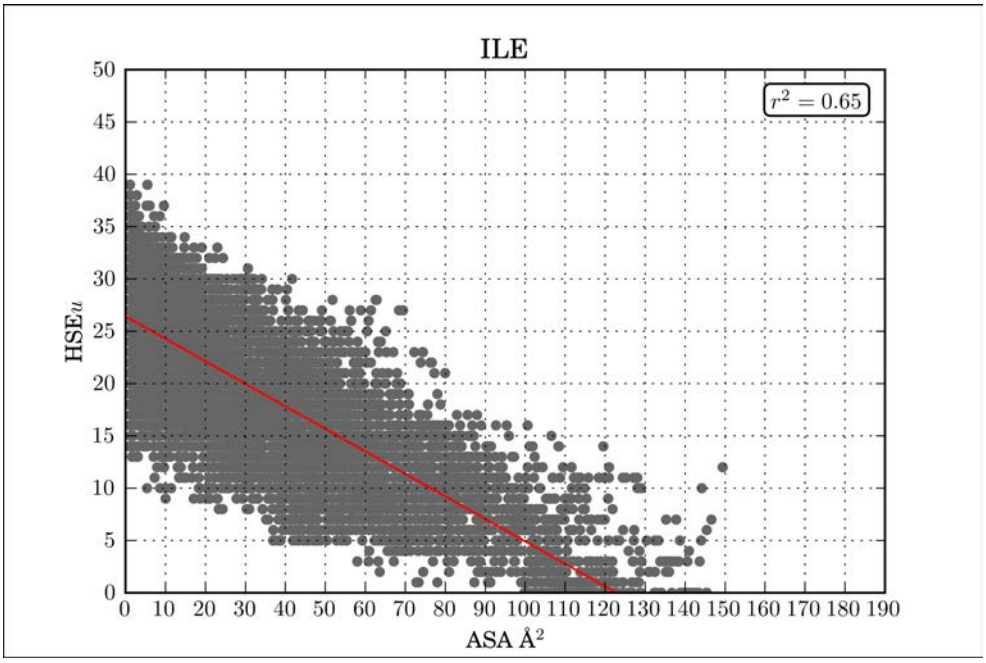

Figure F.9: Scatter plot of $HSEu_{13}$ vs Side chain ASA, for Isoleucined: Each point represents a single instance of the residue-type, for which both $HSEu_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between $HSEu_{13}$ and ASA for this residue type.

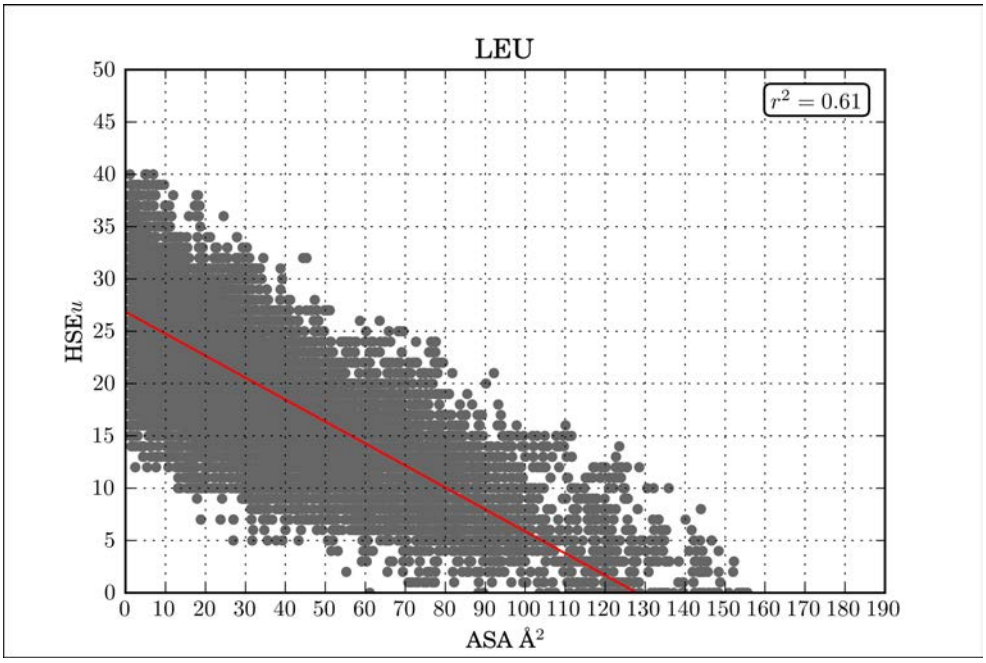

Figure F.10: Scatter plot of $HSEu_{13}$ vs Side chain ASA, for Leucine: Each point represents a single instance of the residue-type, for which both $HSEu_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between $HSEu_{13}$ and ASA for this residue type.

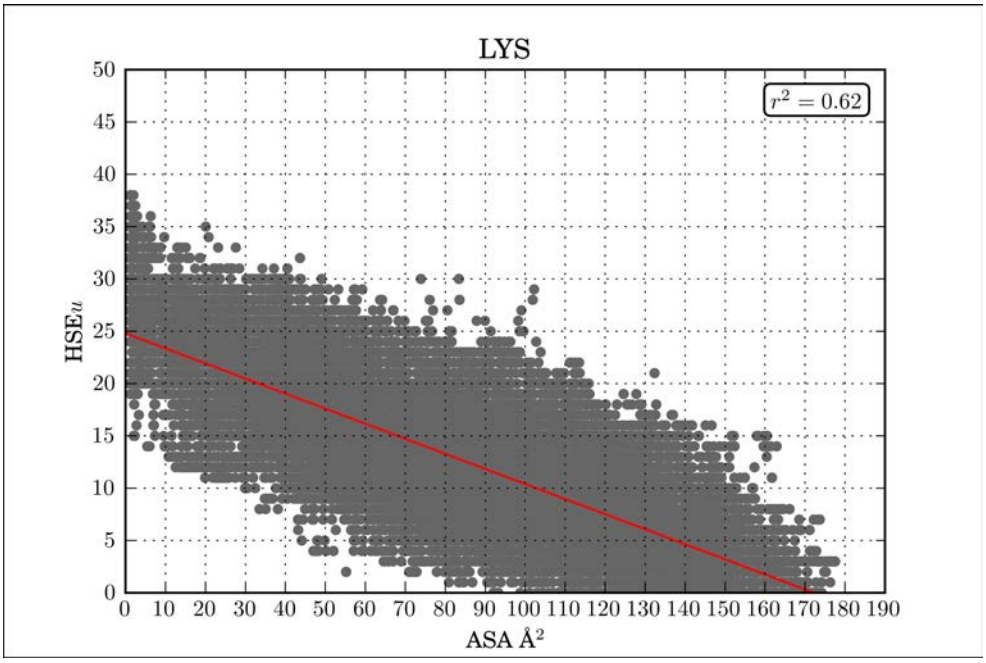

Figure F.11: Scatter plot of HSEu$_{13}$ vs Side chain ASA, for Lysine: Each point represents a single instance of the residue-type, for which both HSEu$_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between HSEu$_{13}$ and ASA for this residue type.

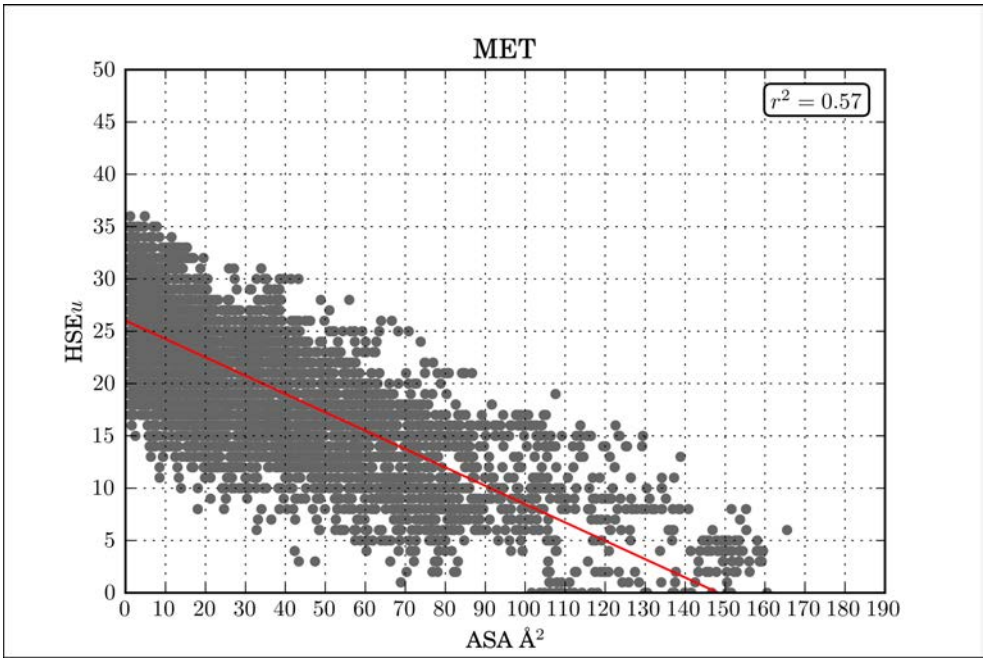

Figure F.12: Scatter plot of HSEu$_{13}$ vs Side chain ASA, for Methionine: Each point represents a single instance of the residue-type, for which both HSEu$_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between HSEu$_{13}$ and ASA for this residue type.

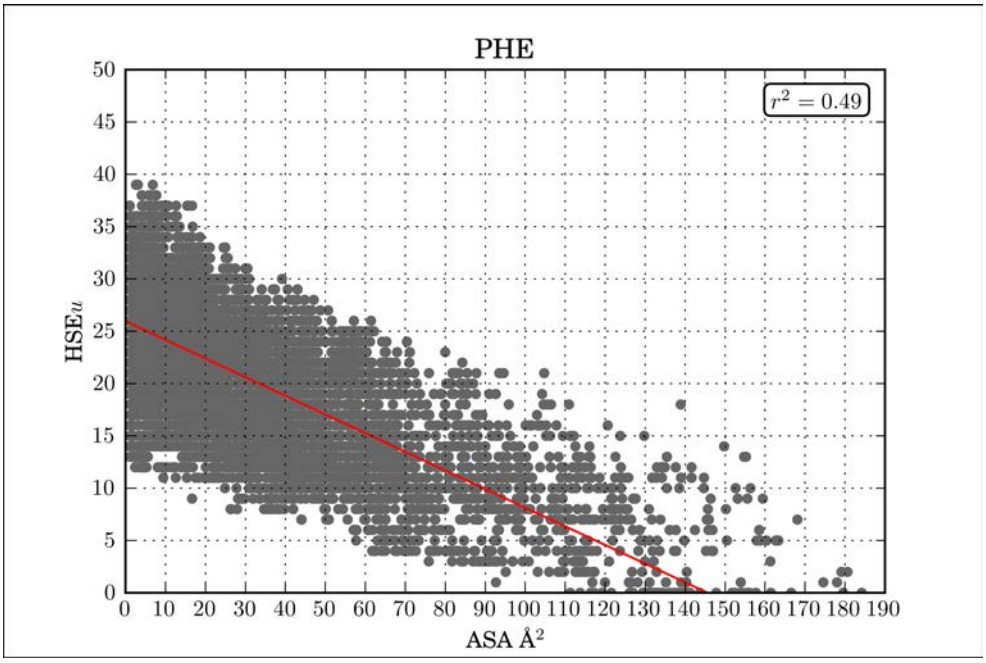

Figure F.13: Scatter plot of HSEu$_{13}$ vs Side chain ASA, for Phenylalanine: Each point represents a single instance of the residue-type, for which both HSEu$_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between HSEu$_{13}$ and ASA for this residue type.

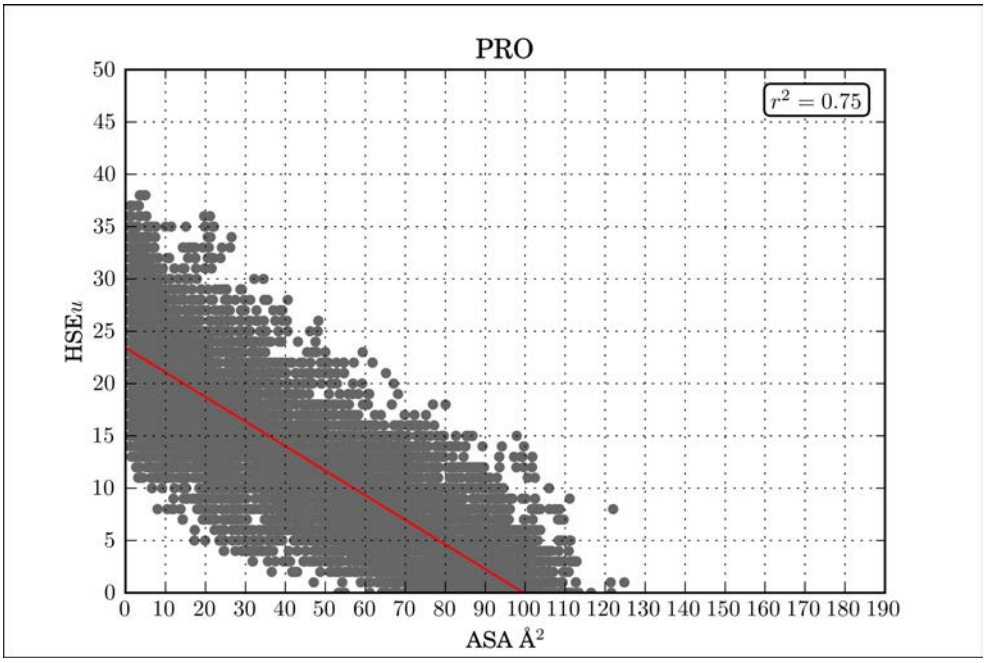

Figure F.14: Scatter plot of HSEu$_{13}$ vs Side chain ASA, for Proline: Each point represents a single instance of the residue-type, for which both HSEu$_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between HSEu$_{13}$ and ASA for this residue type.

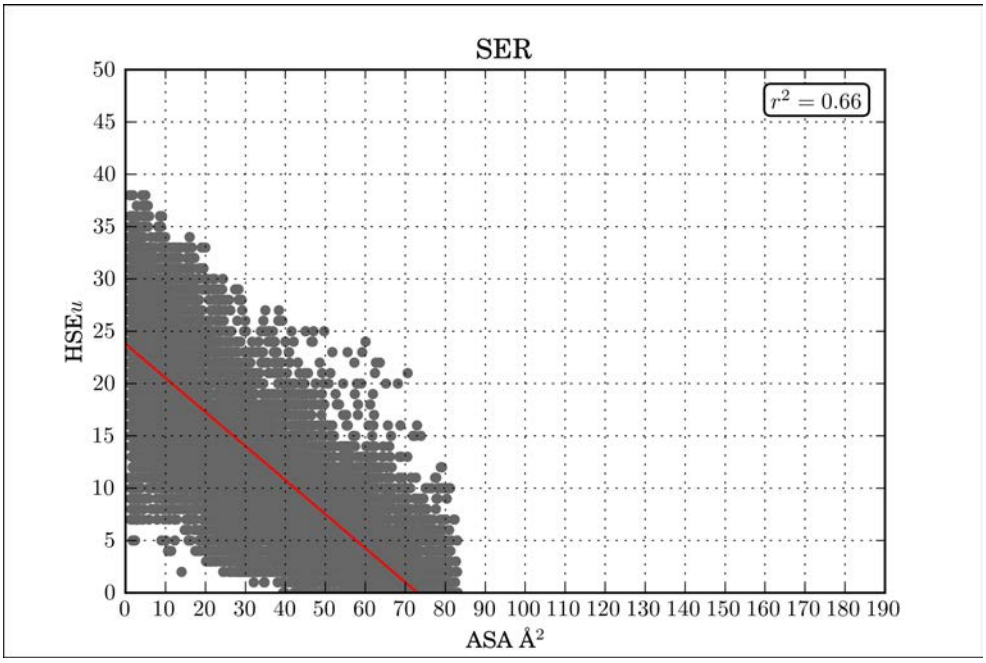

Figure F.15: Scatter plot of HSEu$_{13}$ vs Side chain ASA, for Serine: Each point represents a single instance of the residue-type, for which both HSEu$_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between HSEu$_{13}$ and ASA for this residue type.

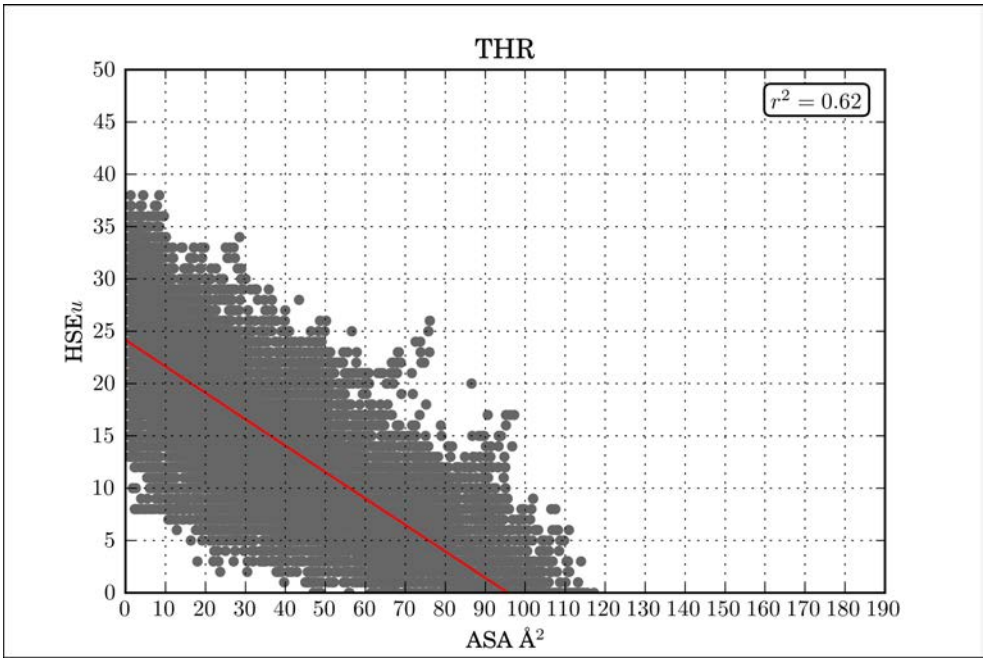

Figure F.16: Scatter plot of HSEu$_{13}$ vs Side chain ASA, for Threonine: Each point represents a single instance of the residue-type, for which both HSEu$_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between HSEu$_{13}$ and ASA for this residue type.

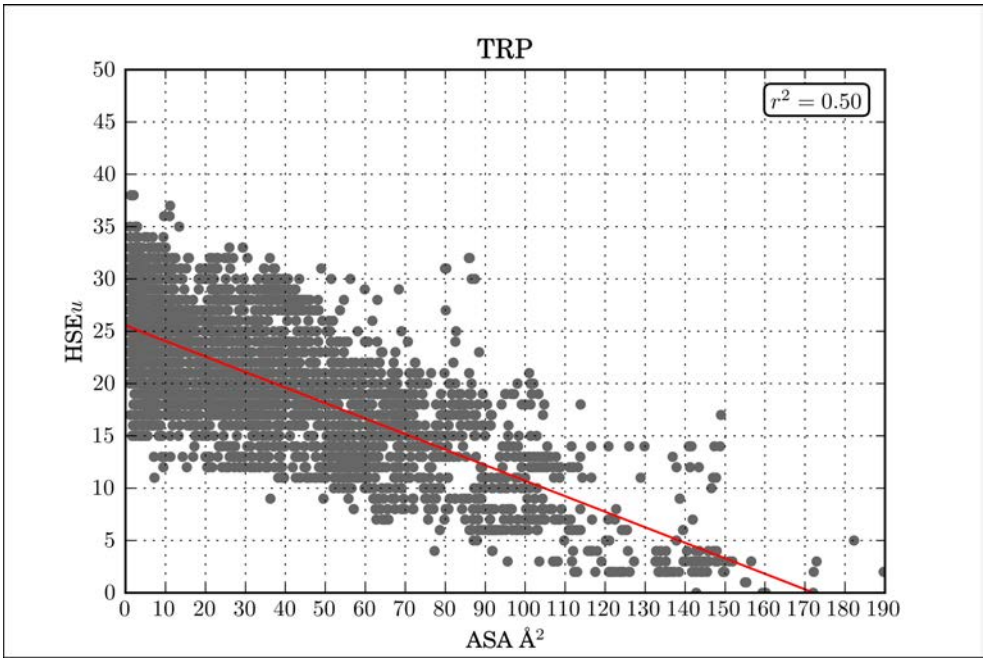

Figure F.17: Scatter plot of $HSEu_{13}$ vs Side chain ASA, for Tryptophan : Each point represents a single instance of the residue-type, for which both $HSEu_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between $HSEu_{13}$ and ASA for this residue type.

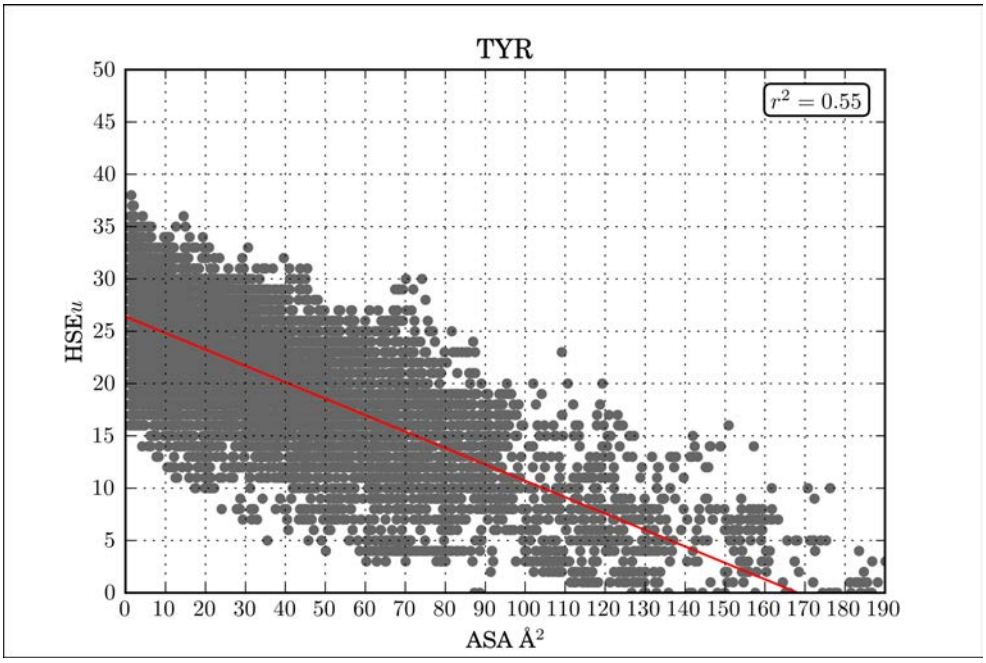

Figure F.18: Scatter plot of $HSEu_{13}$ vs Side chain ASA, for Tyrosine: Each point represents a single instance of the residue-type, for which both $HSEu_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between $HSEu_{13}$ and ASA for this residue type.

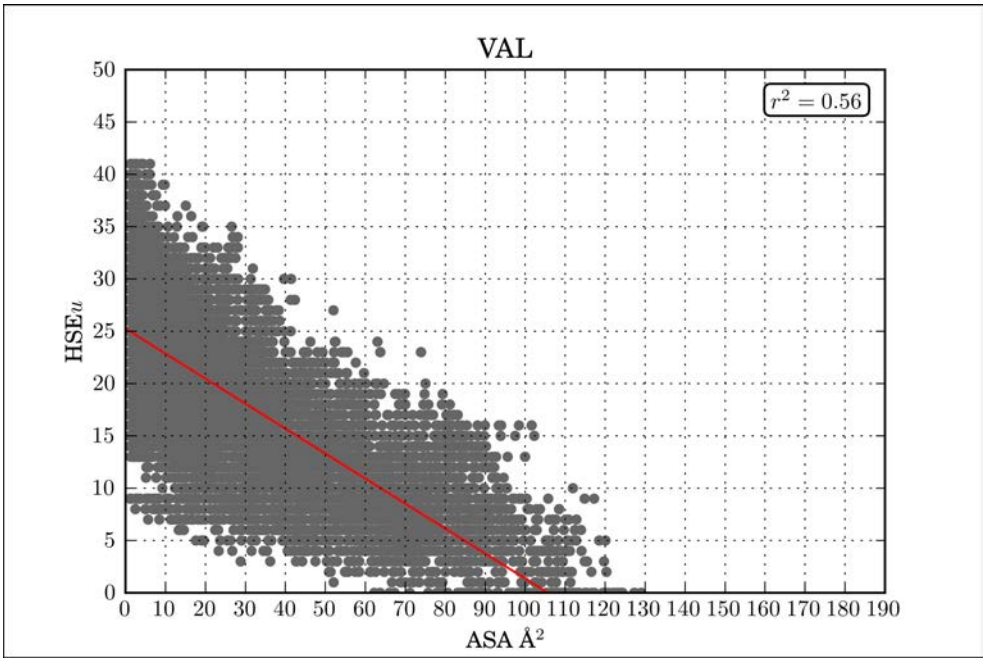

Figure F.19: Scatter plot of HSEu$_{13}$ vs Side chain ASA, for Valine: Each point represents a single instance of the residue-type, for which both HSEu$_{13}$ and the side chain ASA was measured. The red line represents the result of a linear regression analysis on all data points, and thus shows the linear relationship between HSEu$_{13}$ and ASA for this residue type.

Appendix G
Comparison of $HSEu_{13}$ and ASA Bootstraps For all Residue Types

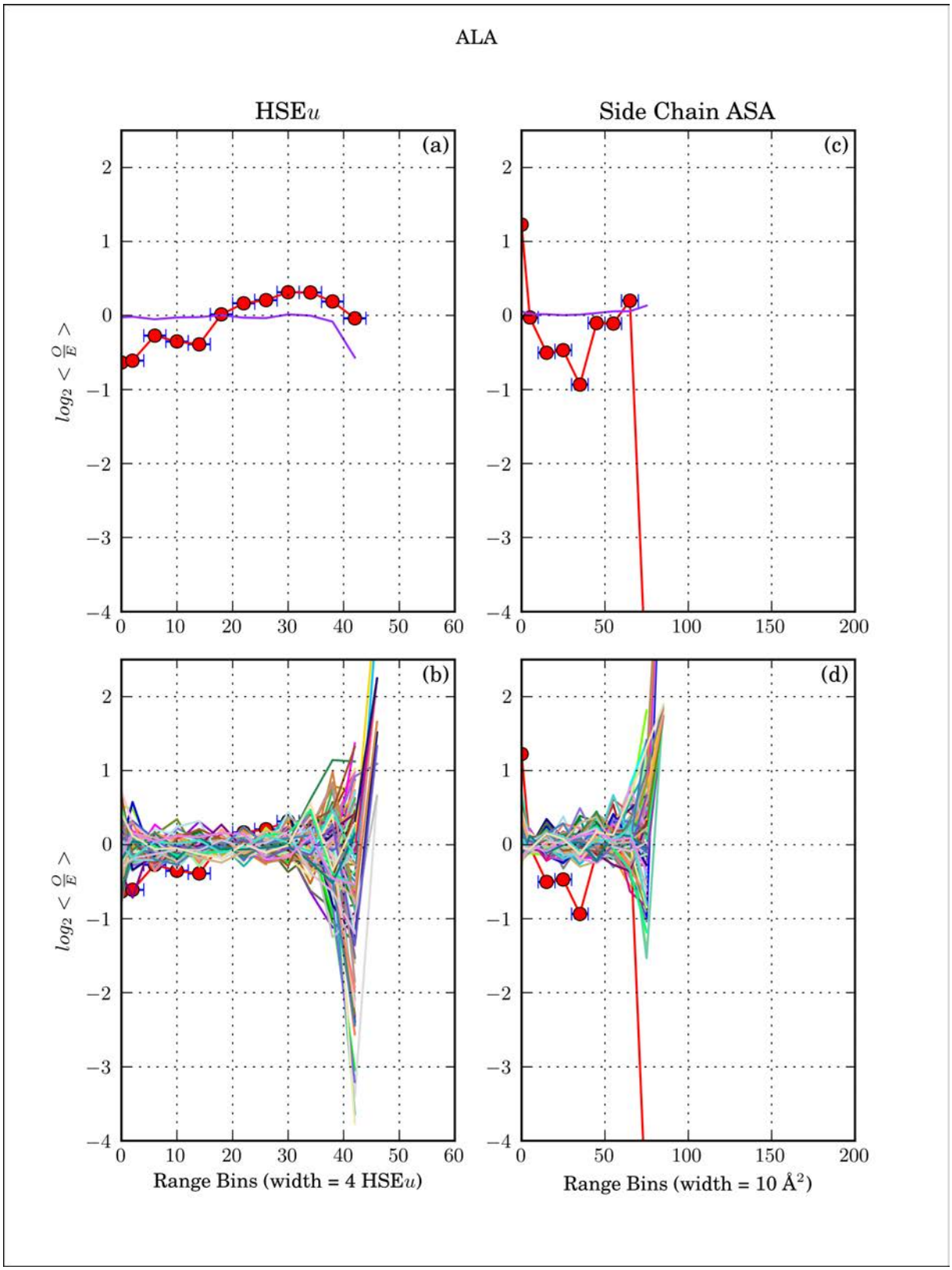

Figure G.1: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for Ala: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

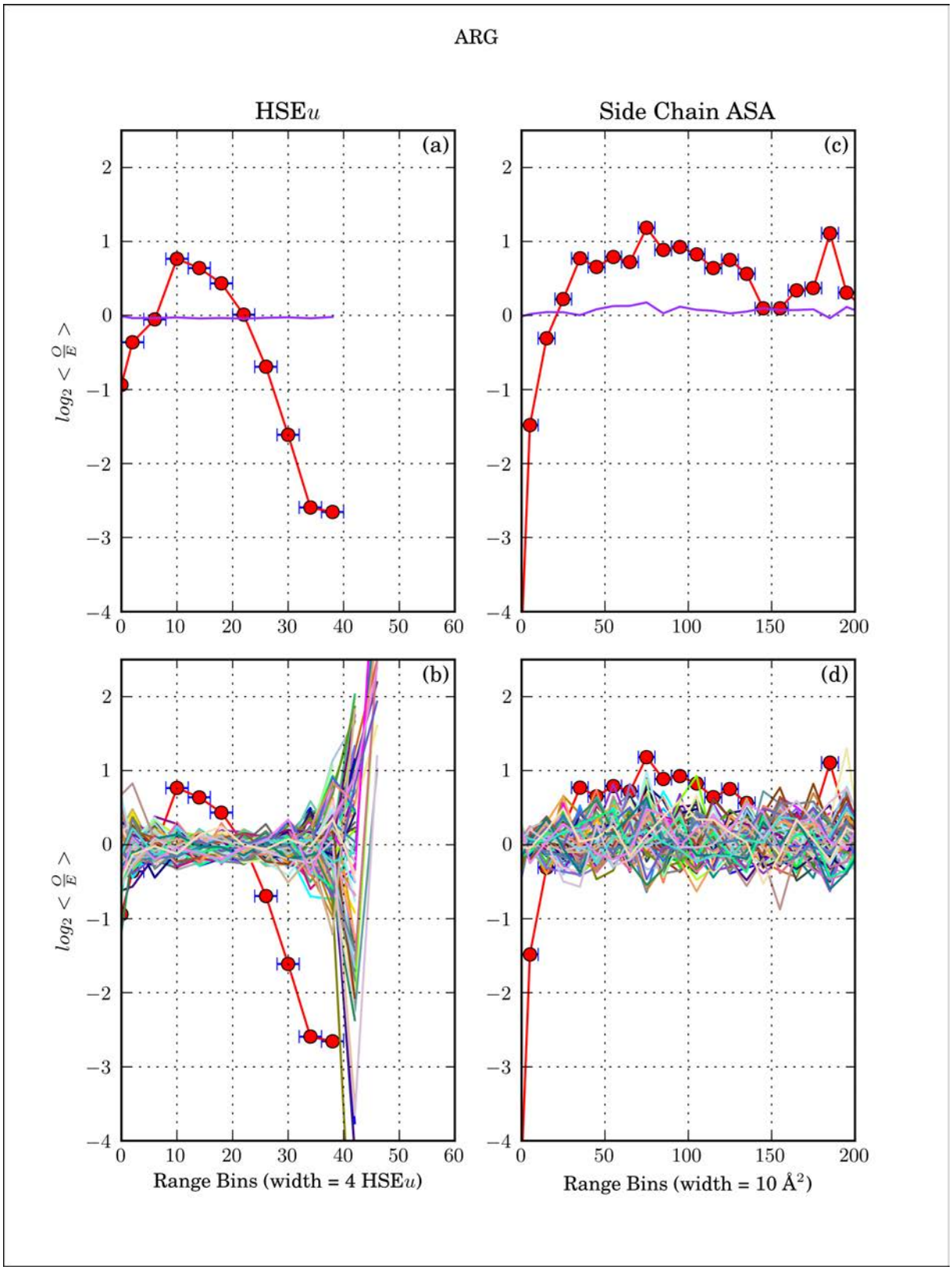

Figure G.2: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for Arg: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

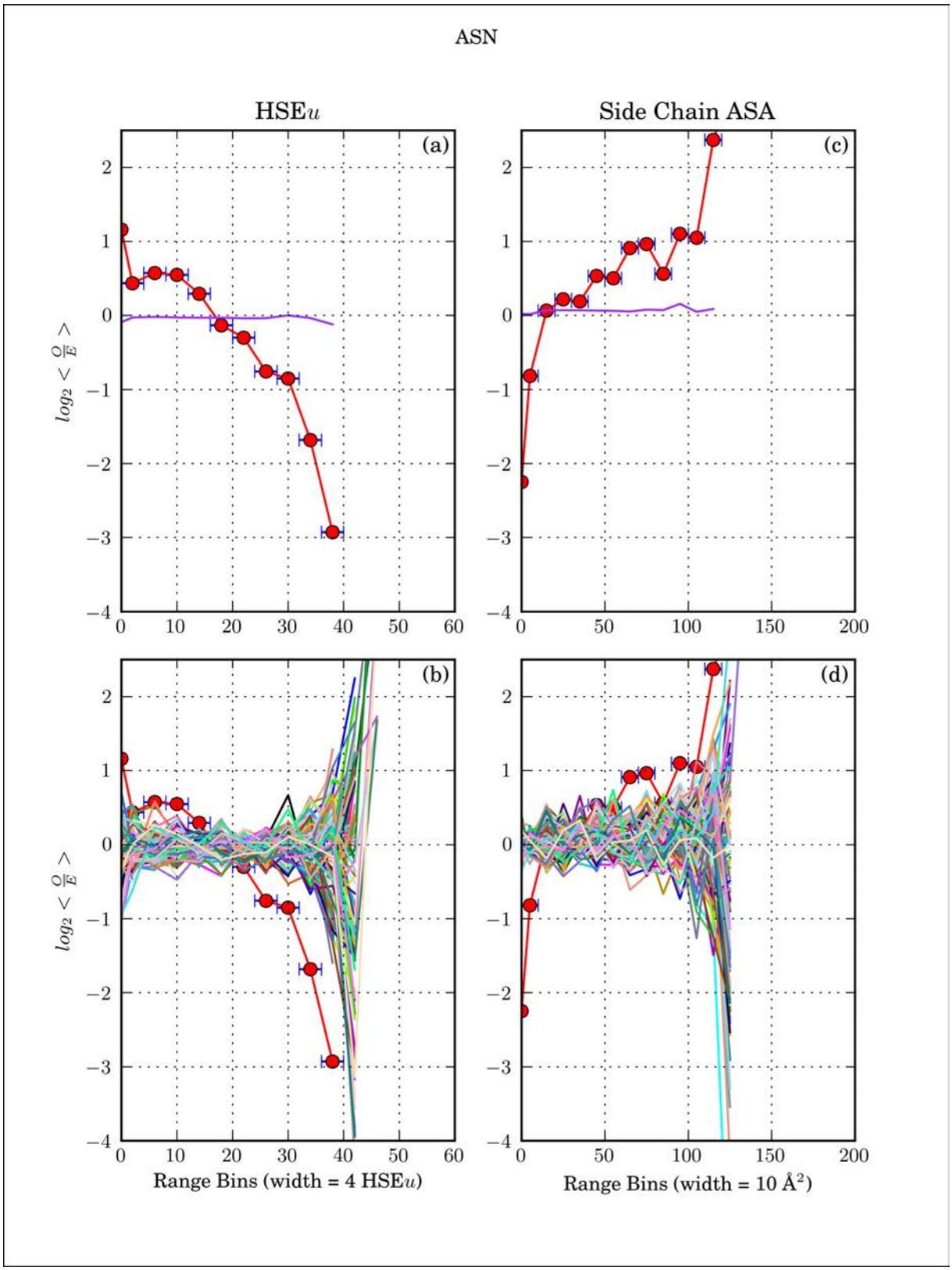

Figure G.3: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for Asn: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

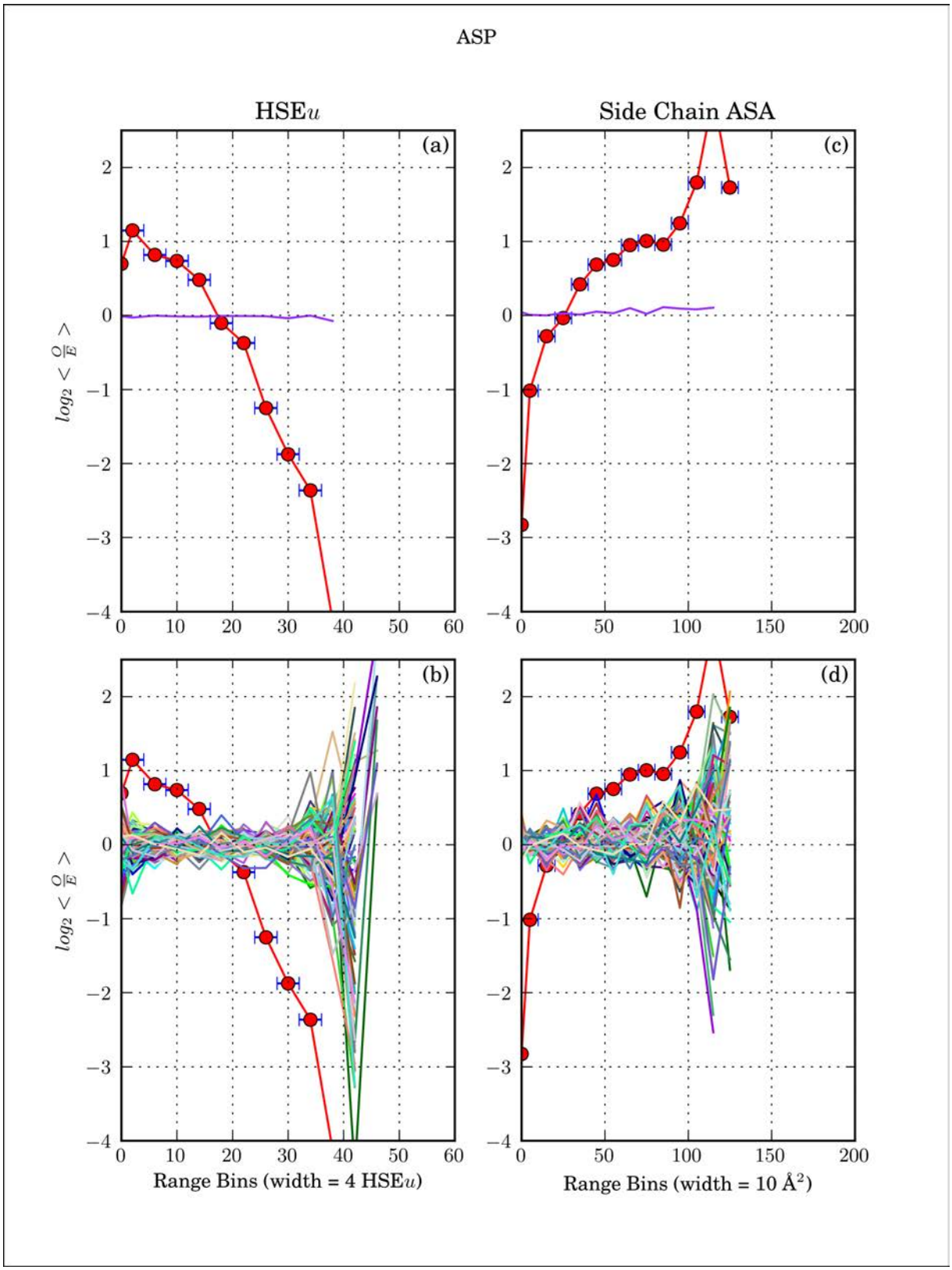

Figure G.4: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for Asp: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

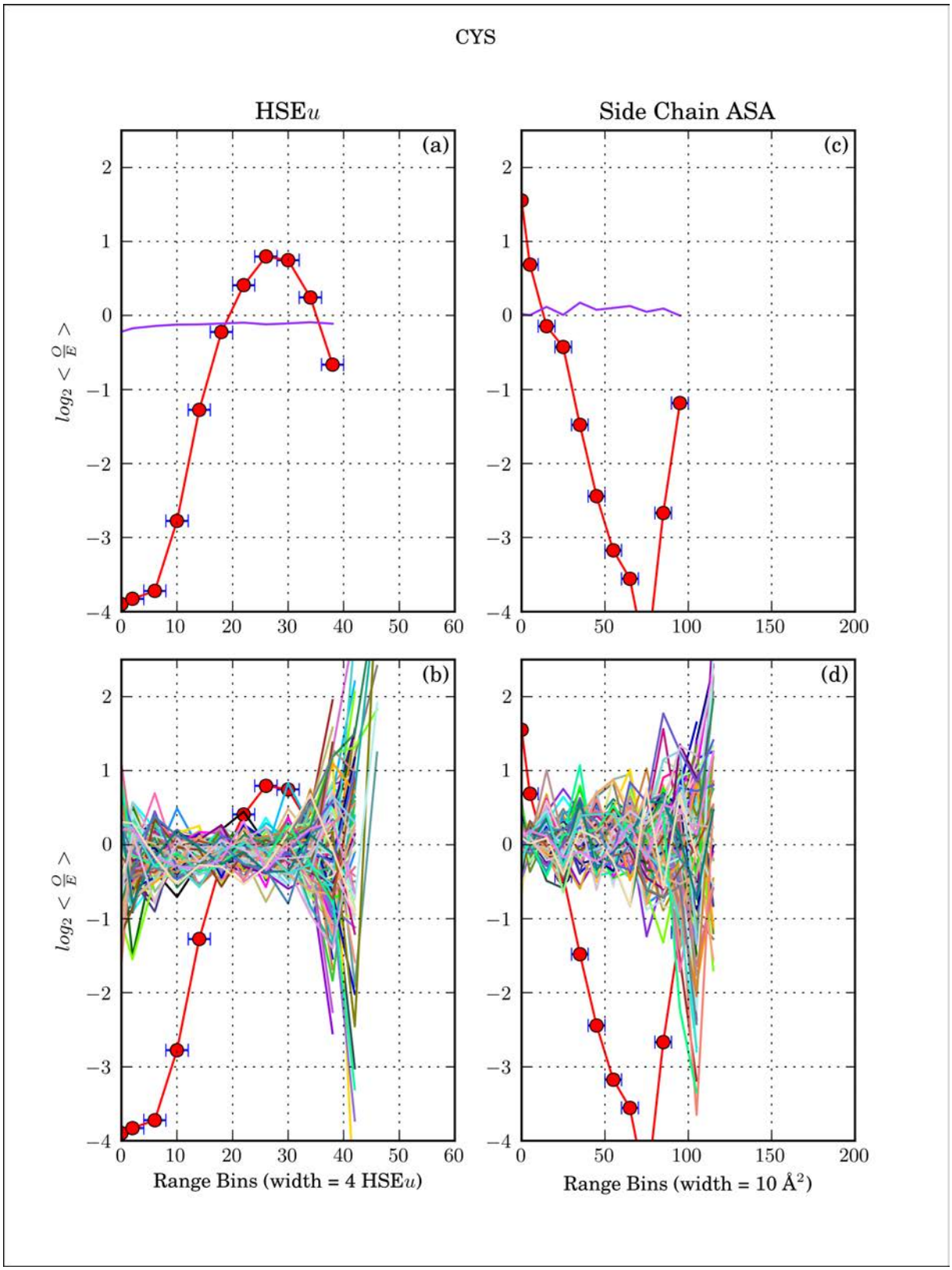

Figure G.5: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for Cys: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

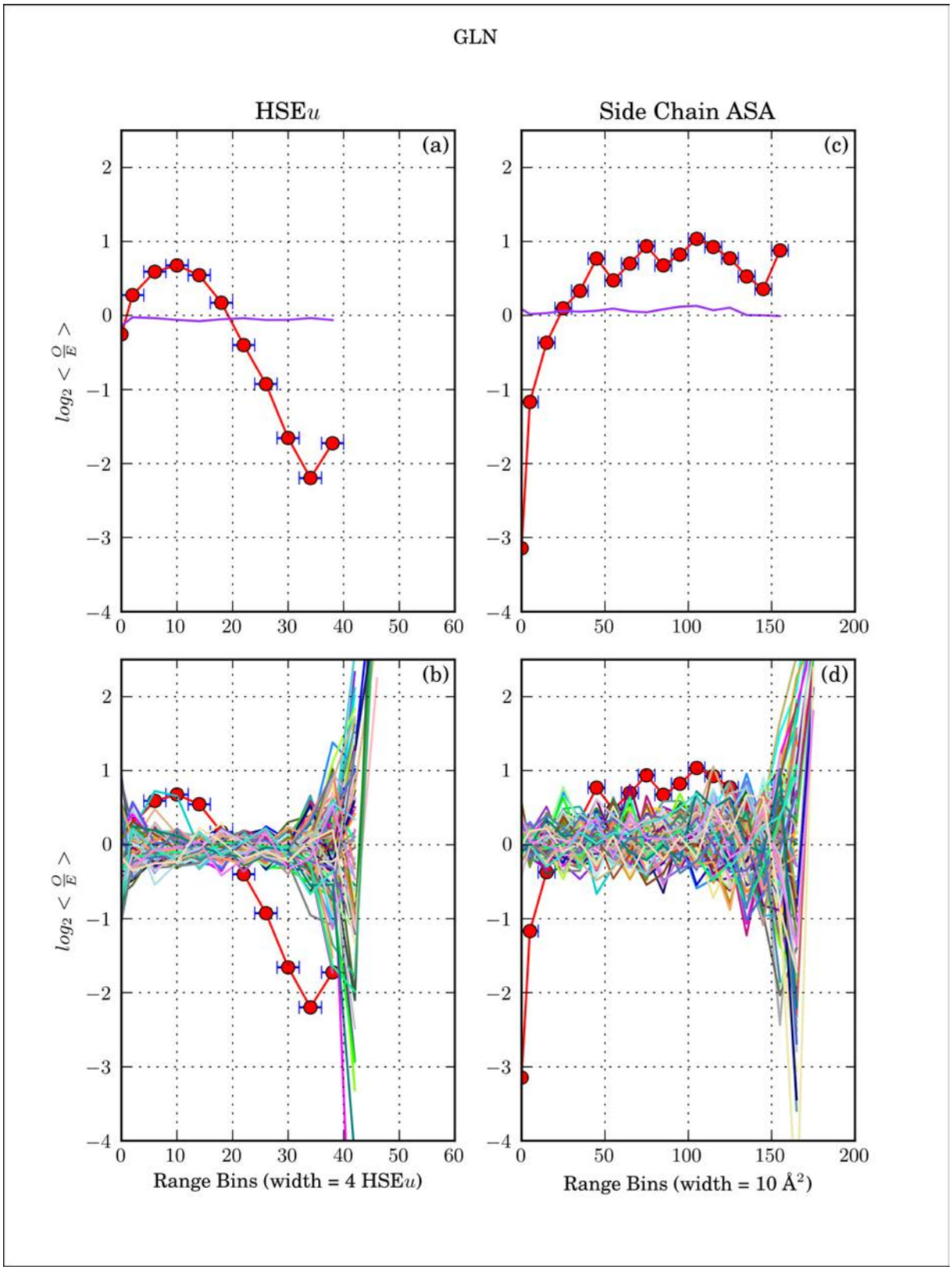

Figure G.6: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for Gln: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

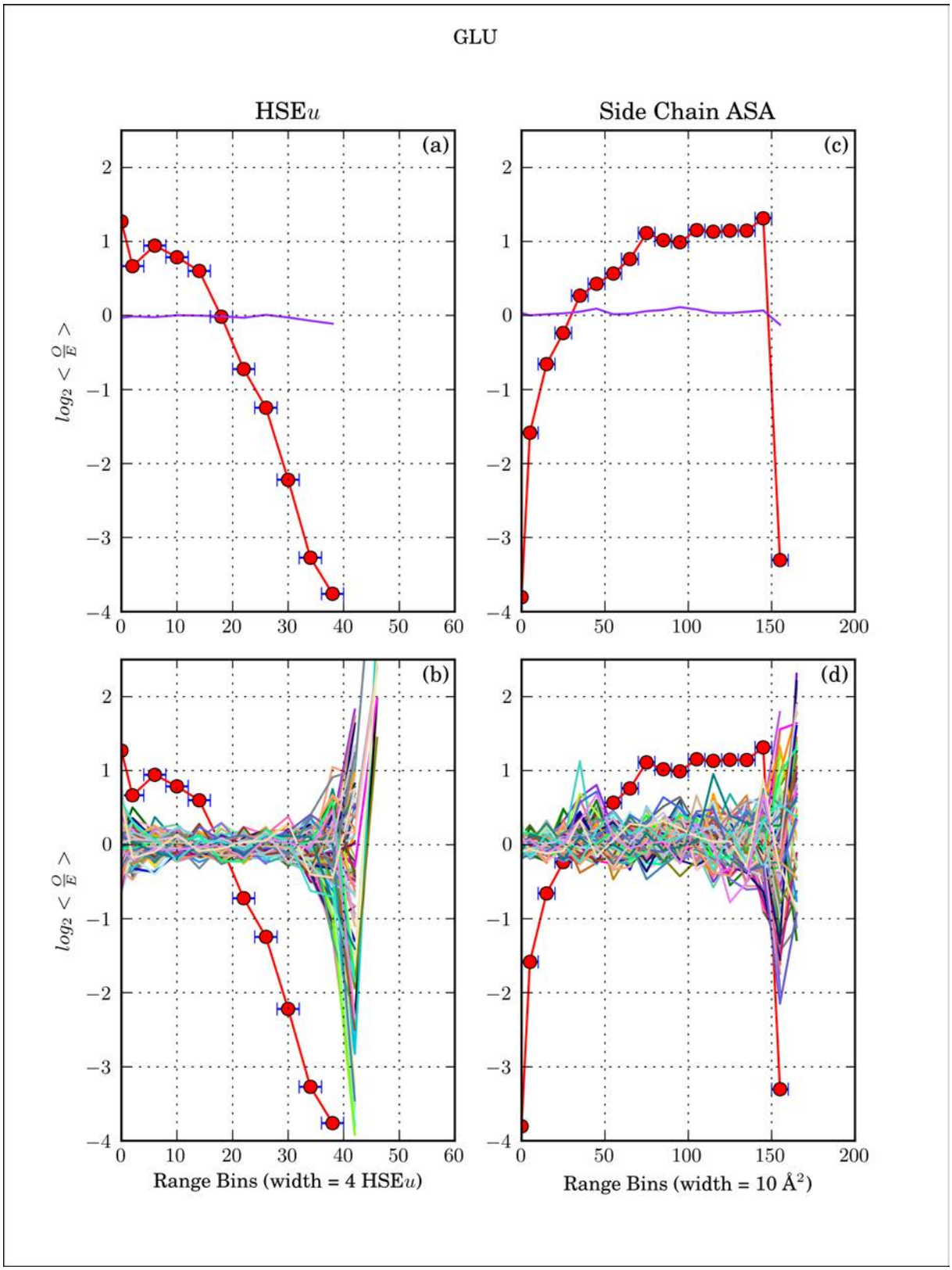

Figure G.7: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for Glu: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

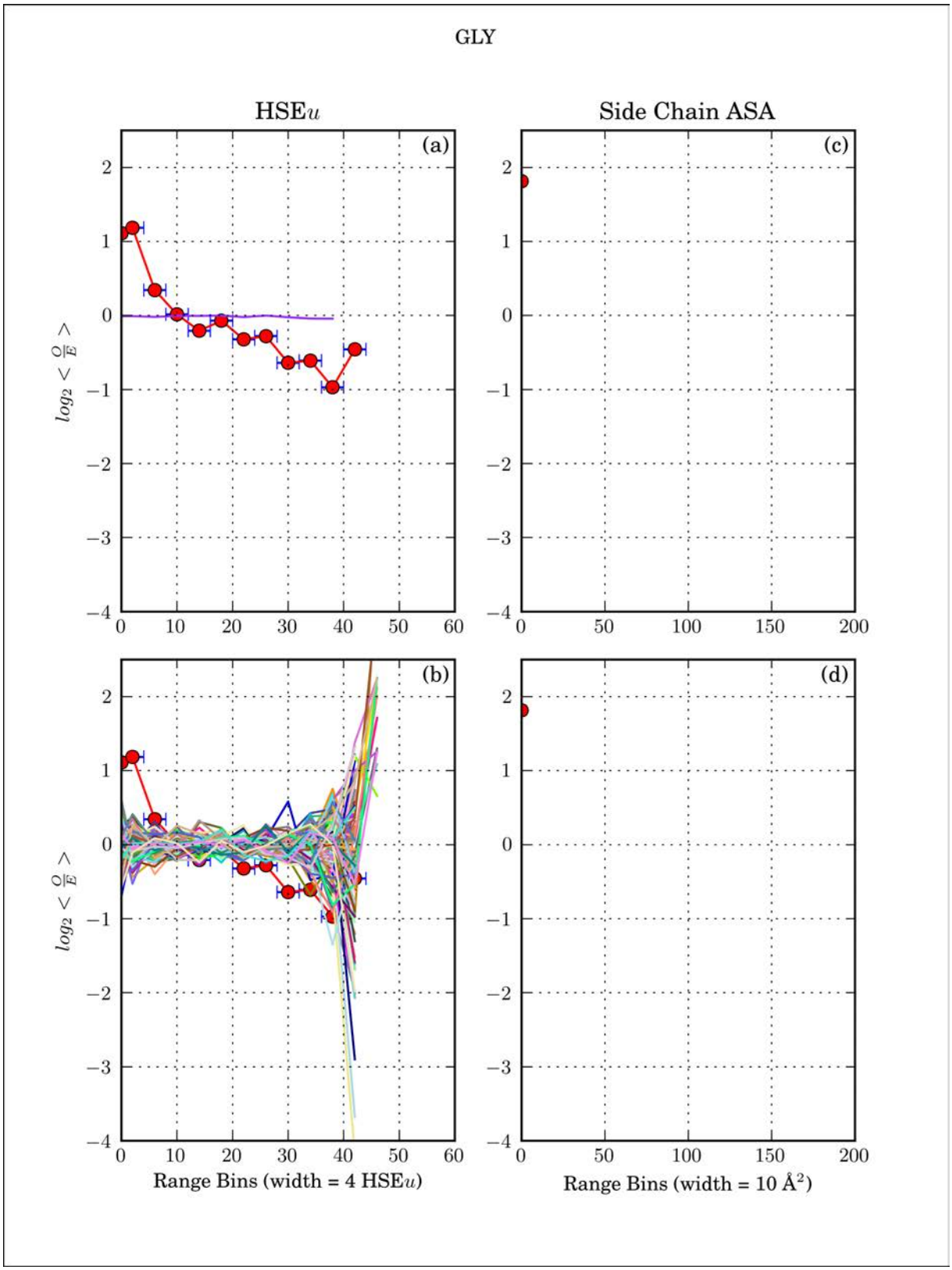

Figure G.8: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for Gly: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

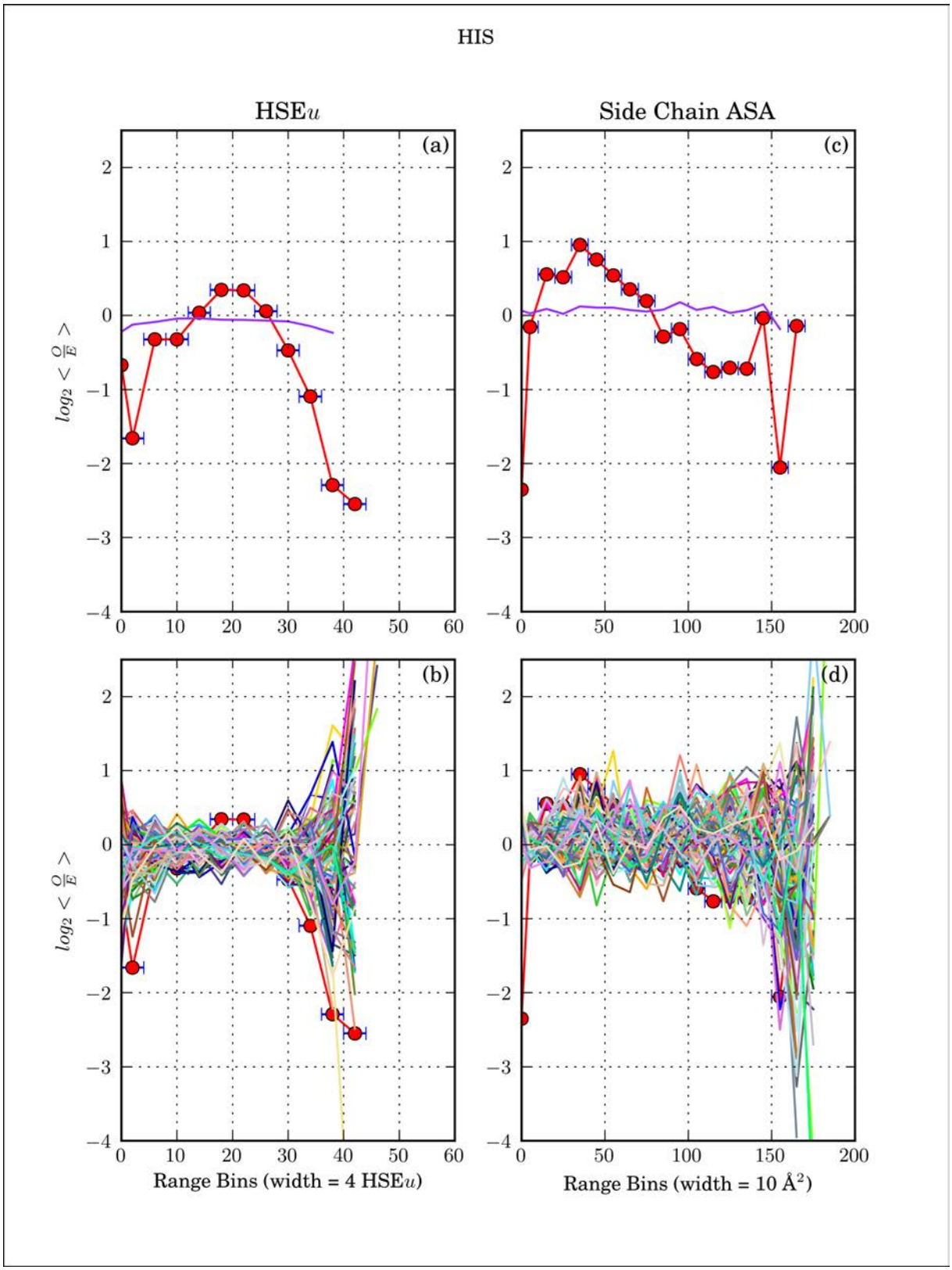

Figure G.9: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for His: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

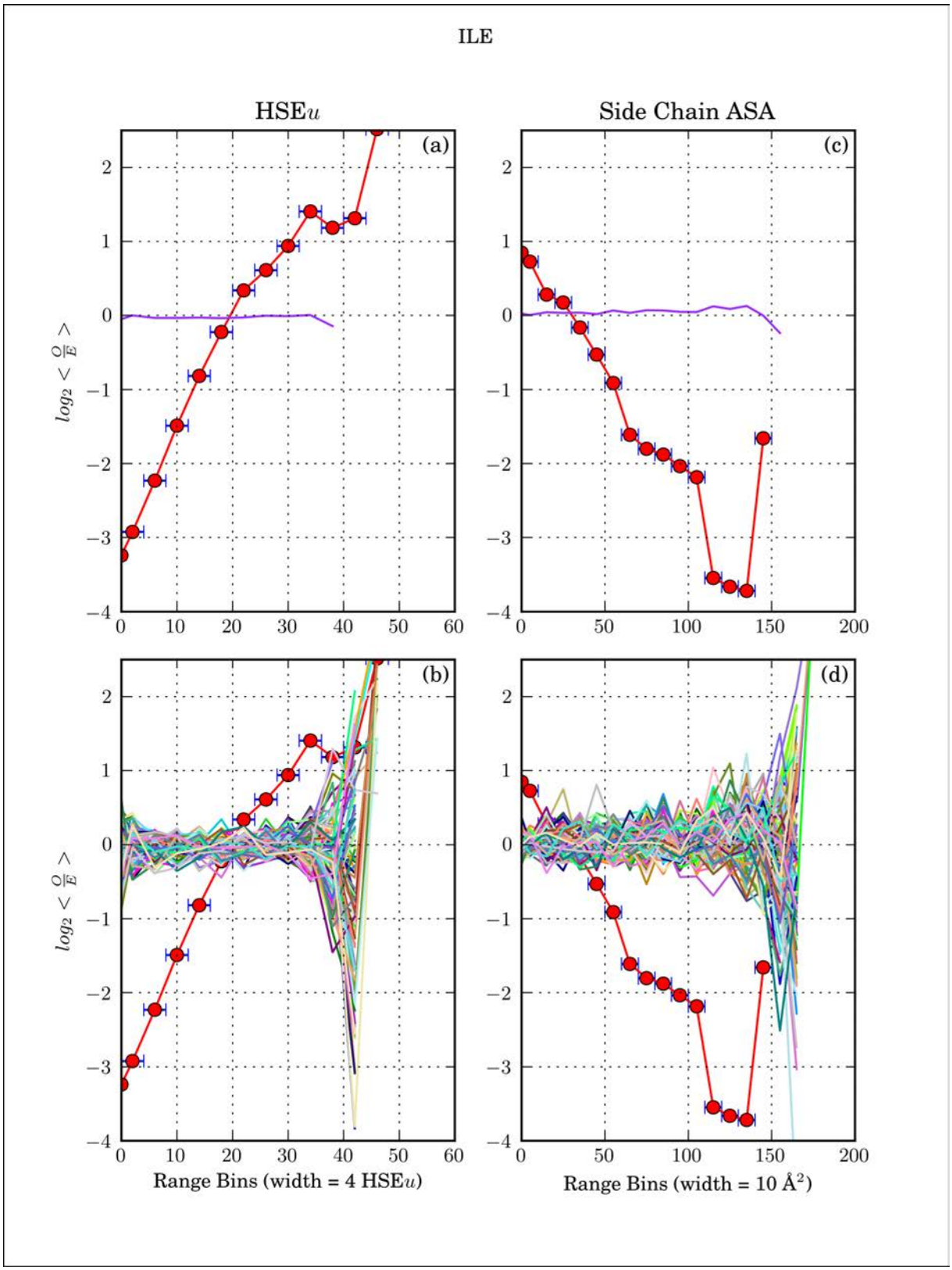

Figure G.10: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for Ile: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

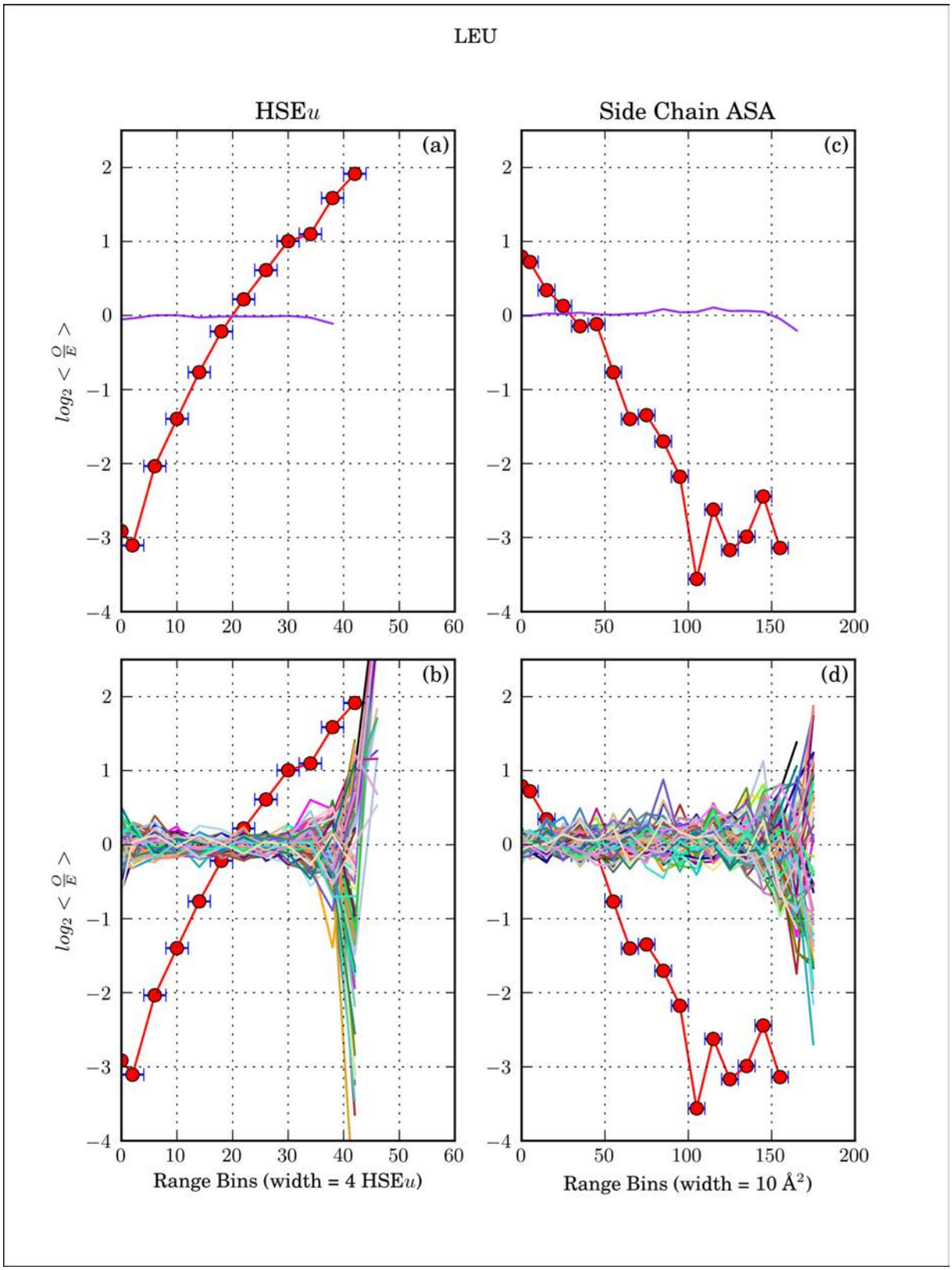

Figure G.11: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for Leu: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

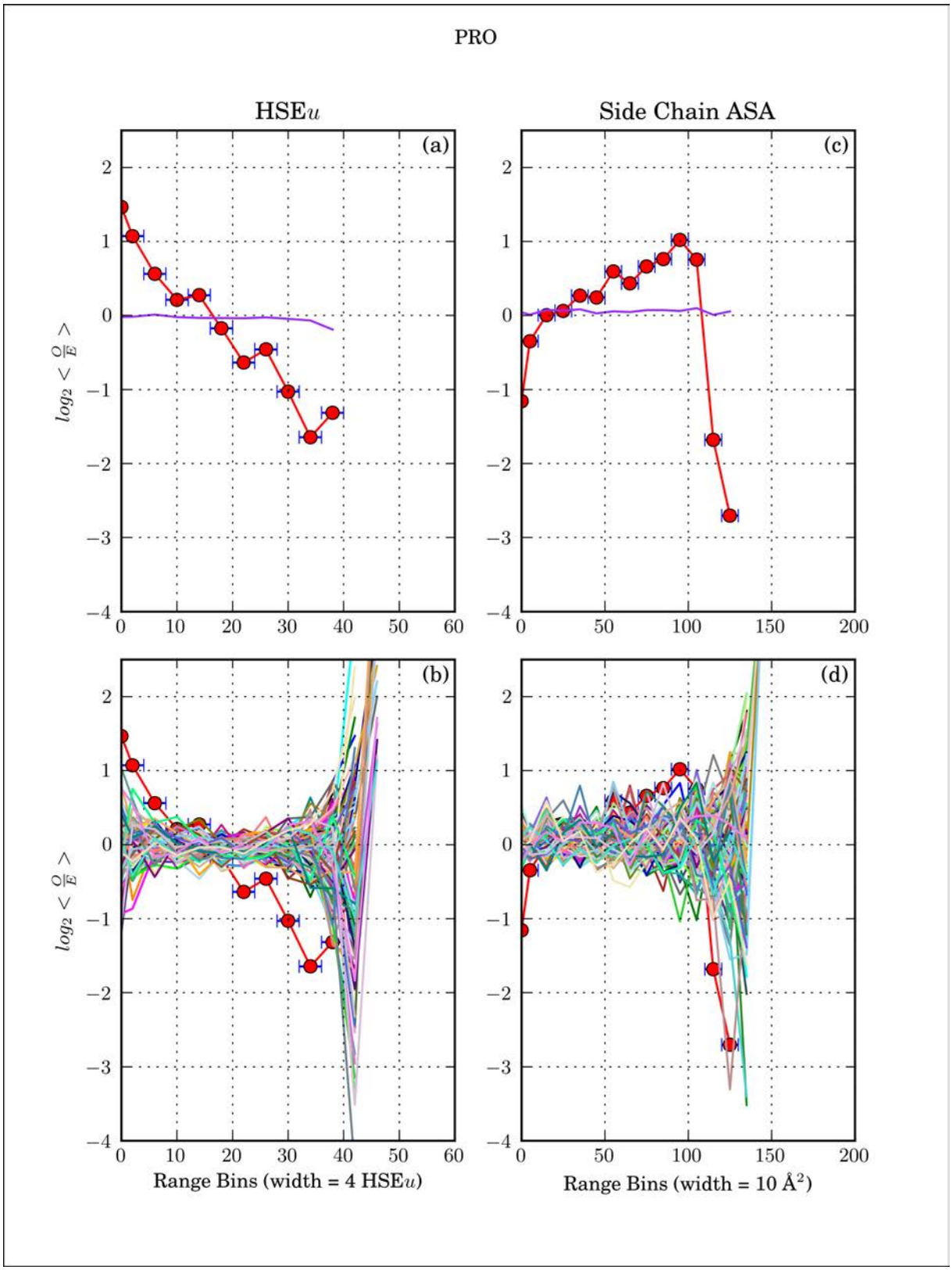

Figure G.12: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for Pro: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

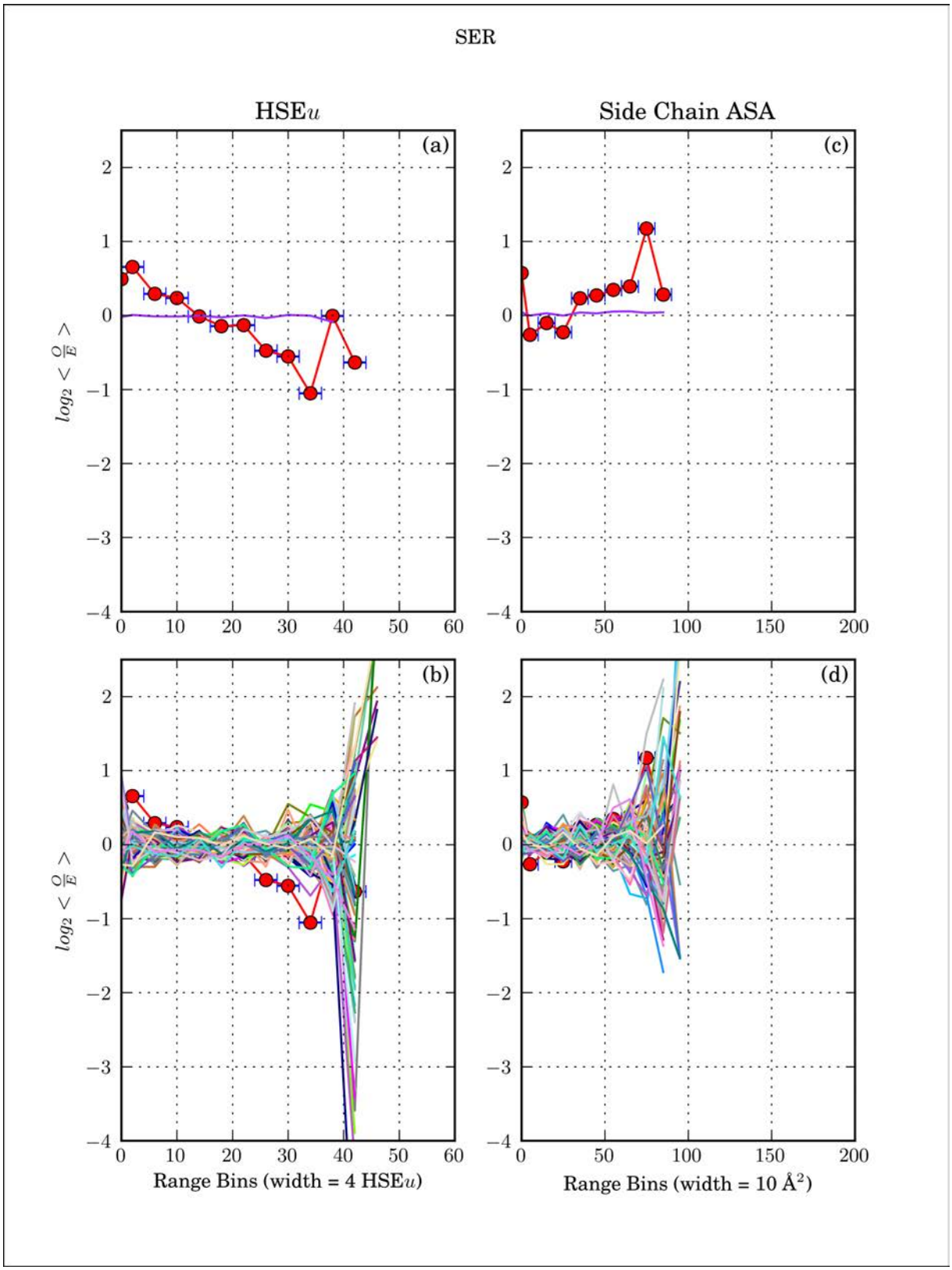

Figure G.13: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for Ser: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

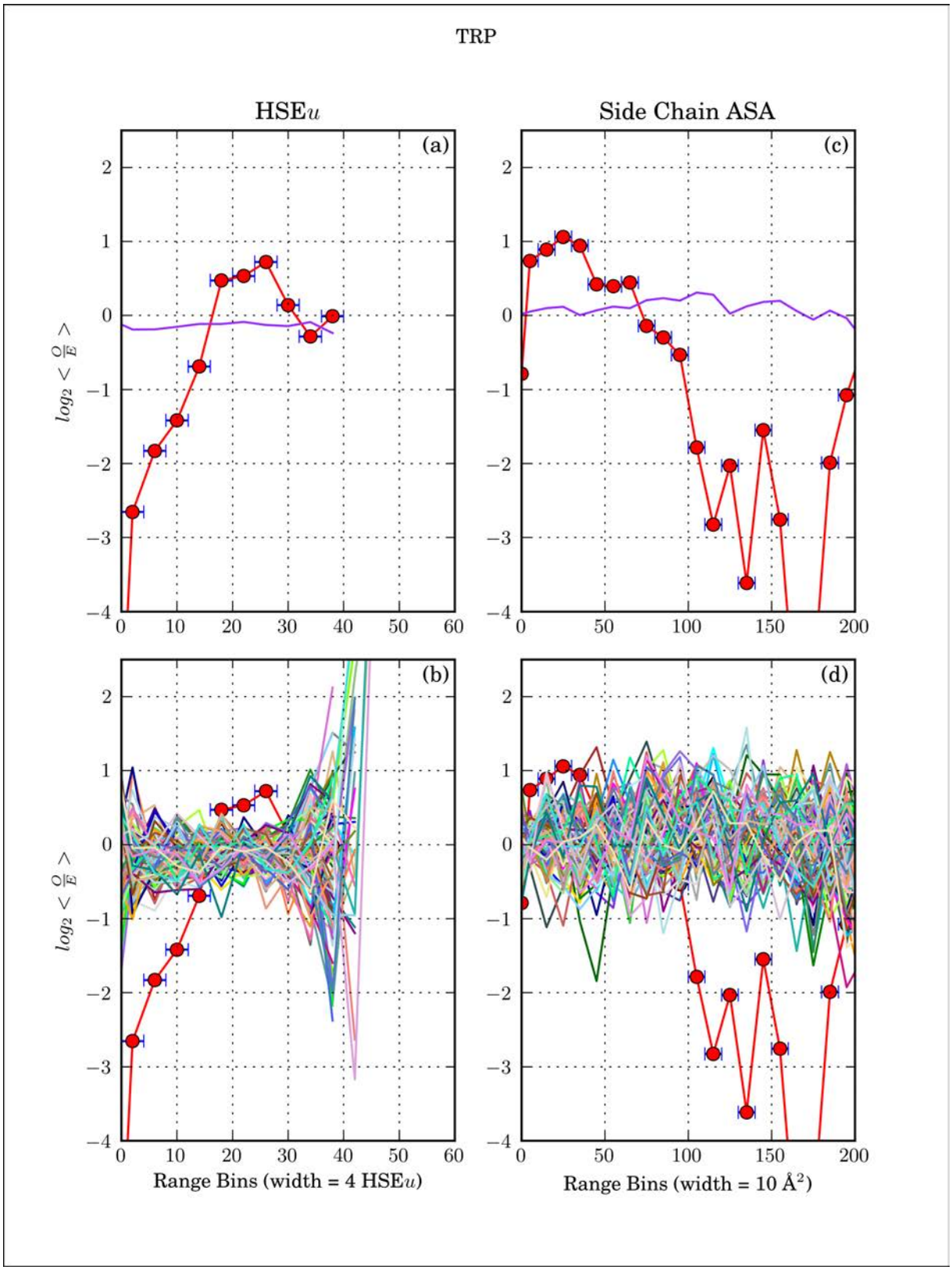

Figure G.14: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for Trp: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

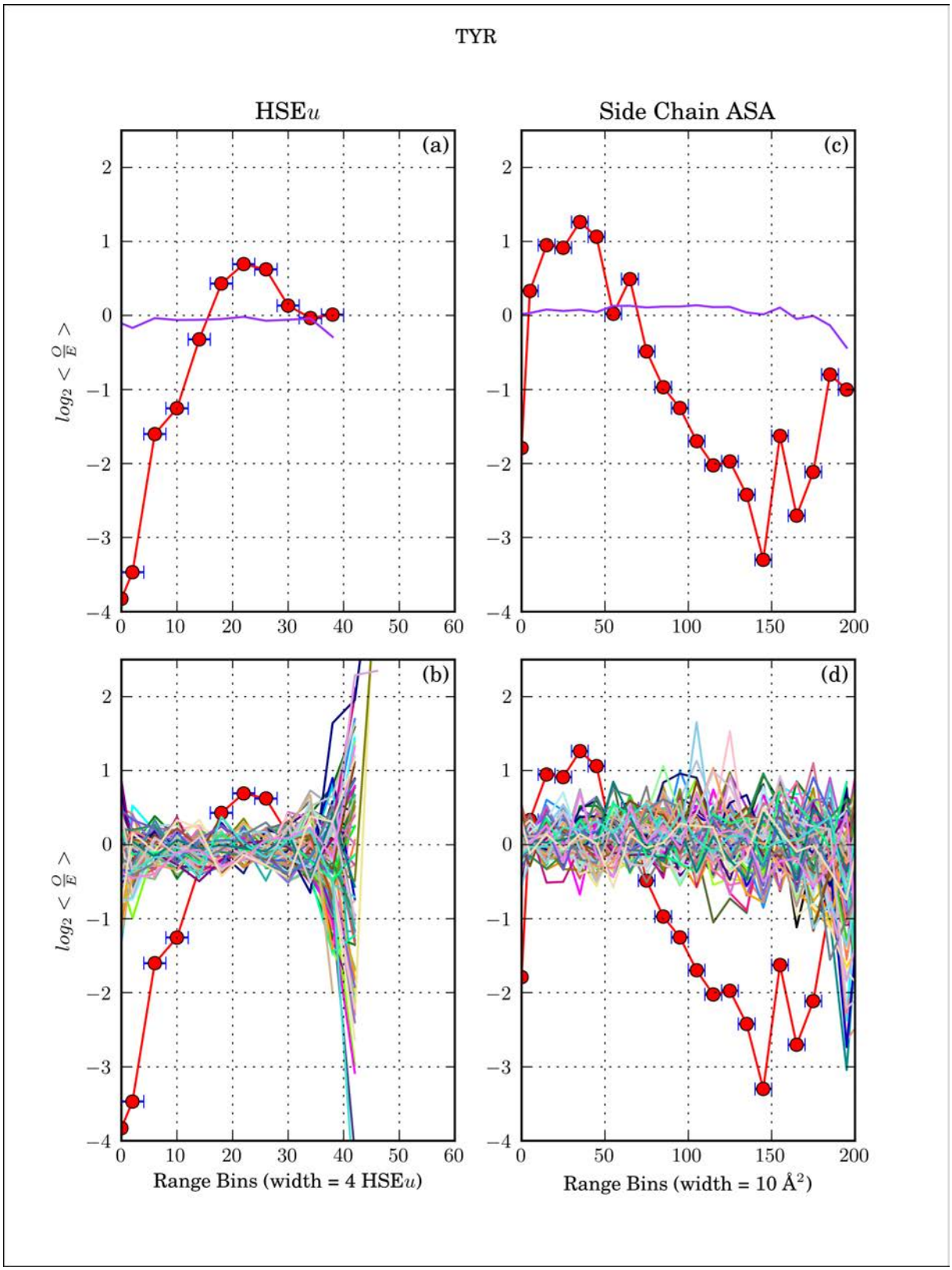

Figure G.15: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for Try: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

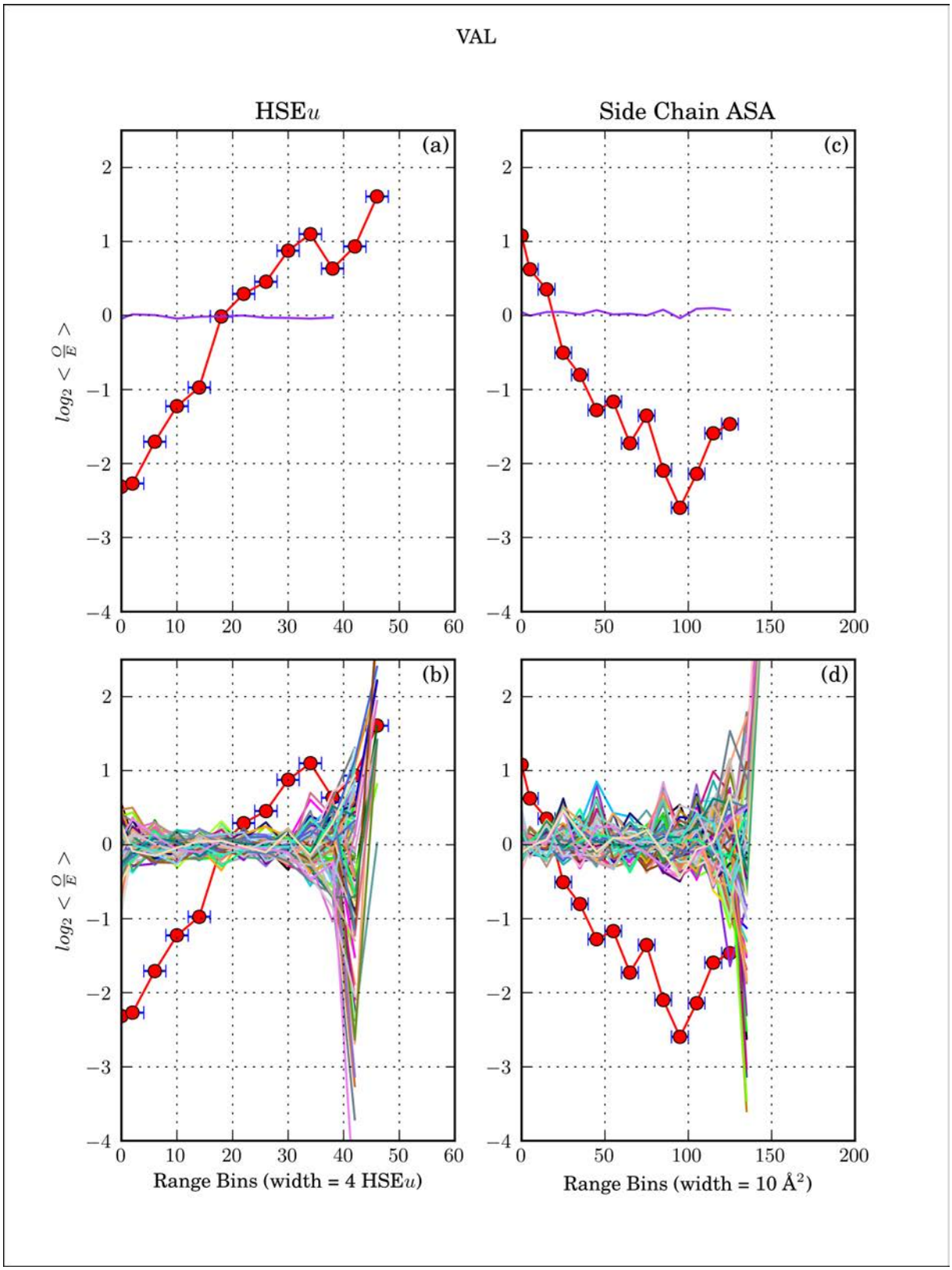

Figure G.16: Comparison of $HSEu_{13}$ and side chain ASA, with bootstrapping, for Val: (a) Plot for $HSEu_{13}$ with average line of 100 bootstraps. (b) Plot for $HSEu_{13}$ with 100 bootstrap lines. (c) Plot of side chain ASA with average line of 100 bootstraps. (d) Plot for ASA with 100 bootstrap lines.

Appendix H
Extra Co-Substitution Result

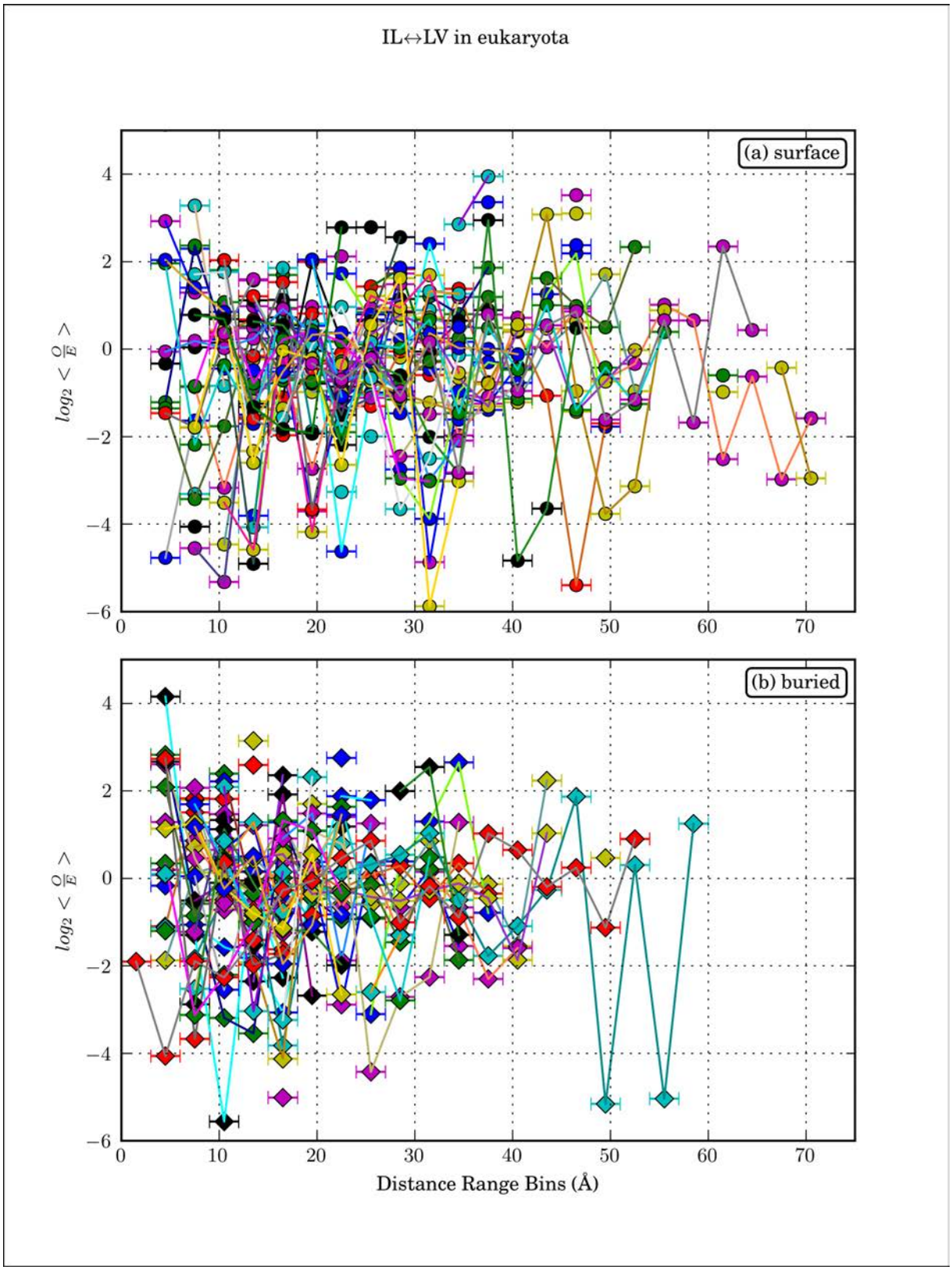

Figure H.1: The co-substitution propensity IL ↔ LV in individual Pfam families, derived from eukaryotic sequenes: The $log_2\langle\frac{O}{E}\rangle$ for each Pfam family has it's own line. The purpose of this plot is to show the distribution of the data across the Pfam families in which the co-substitution was observed.

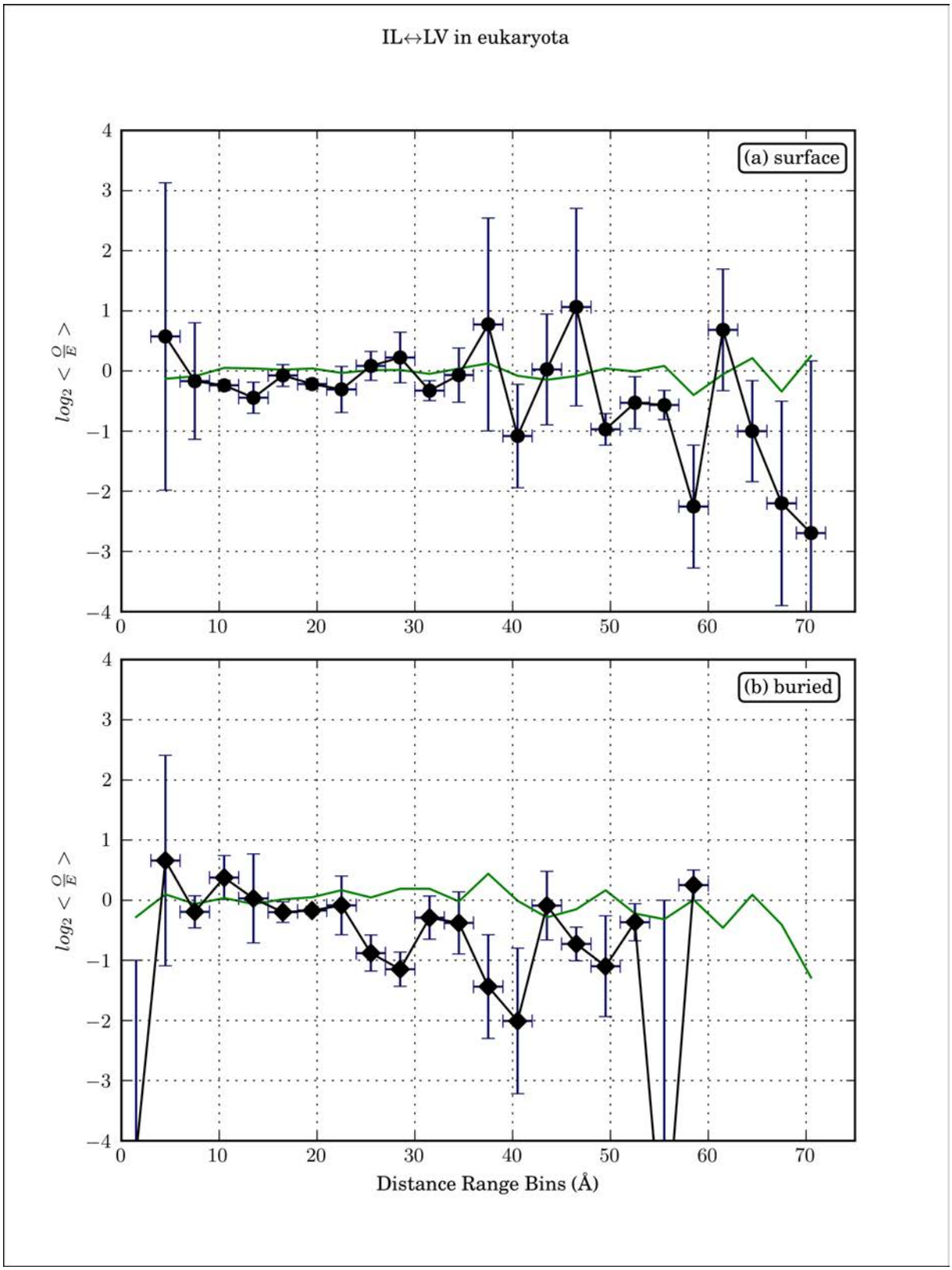

Figure H.2: The average co-substitution propensity IL ↔ LV derived from eukaryotic sequences: The black line with points show the average of 45 Pfam families and represents the independence of a 3 Årange-bin, indicated by the horizontal error bars. The vertical error bars are the log_2 of the standard deviation of $\frac{O}{E}$ for each Pfam family (shown in Figure H.1). The average $\frac{O}{E}$ of bootstrap analyses is shown in green.

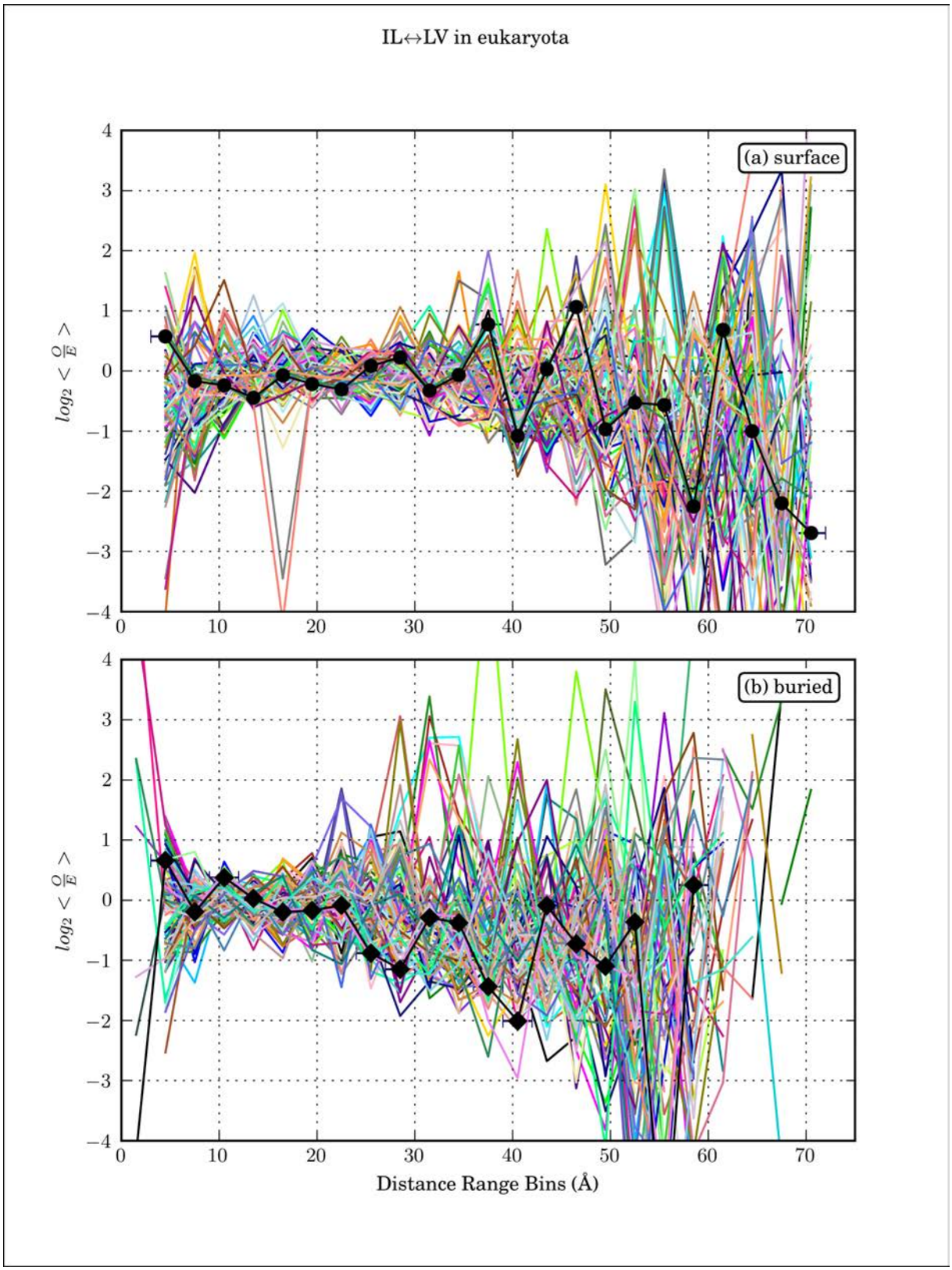

Figure H.3: Co-substitution propensity IL ↔ LV for bootstrap analyses, derived from eukaryotic sequences: The individual lines shown represent the average $\frac{O}{E}$ values calculated from a randomised distance matrix for each Pfam family. The Bootstrap data is incomplete , however it is included here to show the behaviour of the bootstrap data.

Appendix I
Contributions to Development of Solvent Exposure Analysis

The primary intellectual development and implementation of the study of solvent exposure, the development of the work-flow and working prototypes of most key python scripts (which provided early results that were presented at FEBS 2010) are the work of Mr. Bhima Auro (Mr. Bhima Auro van der Molen). As part of a final year undergraduate research project, Mr. John Le Brun completed the development and verified the outcomes of the code. This code was then used to perform an analysis of the full-residue ASA determined using the DSSP program and $HSEu_{13}$. John Le Brun's contribution towards the development and implementation of the project was valuable to arriving at a finished product. His most significant contribution was in developing the the bootstrapping code for the ASA data and the use of WhatIf to check structures. However Mr. Bhima Auro had performed all the research and development work for the project and guided John Le Brun to a finished project. After his final year project was finished, some code needed to be partially rewritten. All results presented here are from analyses preformed on the amended code, by Mr. Bhima Auro. The modification of code to perform the analysis of side chain ASA using Naccess instead of DSSP are the work of Mr. Bhima Auro; thus allowing analysis of side chain only ASA which would not be possible using DSSP. Table I.1 shows a breakdown of our relative contributions to the written code. Mr. Chinmay Kanchi provided some support in the early stages of the project toward learning the Python programming language and correcting programming mistakes and is thus also acknowledged in the table.

Table I.1: Contributions to software development. The relative contributions towards software development, for the solvent exposure analyses presented in this thesis.

File Name	Mr. Bhima Auro (%)	Mr. John Le Brun (%)	Mr. Chinmay Kanchi (%)
all dssp pdbout	100	0	0
loader files	50	0	50
bootstrap data	30	70	0
bootstrap line plotter	50	50	0
calculateExpectedAverage DSSP	0	100	0
calculateStatistics	100	0	0
dataHandler	100	0	0
extract pdb sequences	67	33	0
FindMaxDSSP	75	25	0
getGlobalDataCount	100	0	0
getSeqWeights per Pfam DSSP	40	60	0
getSeqWeights per Pfam HSE	40	60	0
line plotter	67	33	0
RangeDictionary	67	0	33
ResidueDictionary	67	0	33
residueTools	100	0	0
residueCounter	60	20	20
sequenceWeights	71	29	0
StrutureSegmentExtractor	60	40	0
UserInterface	50	50	0
whatiff scripts	0	100	0

List of References

[1] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences. 1992;89(22):10915–10919. Available from: http://www.pnas.org/content/89/22/10915.abstract.

[2] Gonnet G, Cohen M, Benner S. Exhaustive matching of the entire protein sequence database. Science. 1992;256(5062):1443–1445. Available from: http://www.sciencemag.org/content/256/5062/1443.abstract.

[3] Dayhoff MO, Schwartz RM, Orcutt BC. In Atlas of Protein Sequence and Structure. suppl National Biomedical Research Foundation. 1978;5:345–352.

[4] Hubbard SJ, Thornton JM. 'NACCESS', computer program. Department of Biochemistry Molecular Biology, University College London; 1993.

[5] Srinivasan R RGD. LINUS - A Hierarchiacal Procedure to Predict the Fold of a Protein. Proteins. 1995;22(2):81–99.

[6] Levinthal C. How to fold graciously. Mossbauer Spectroscopy in Biological Systems. 1969;Proceedings of a meeting held at Allerton House, Monticello, IL:22–24.

[7] Wolynes PG. Folding funnels and energy landscapes of larger proteins within the capillarityapproximation. Proceedings of the National Academy of Sciences of the United States of America. 1997 Jun;94(12):6170–6175. Available from: http://www.pnas.org/content/94/12/6170.abstract.

[8] Tiana G, Broglia RA, Shakhnovich EI. Hiking in the energy landscape in sequence space: a bumpy road to good folders. Proteins. 2000 May;39(3):244–251. Available from: http://view.ncbi.nlm.nih.gov/pubmed/10737946.

[9] Kresge N, Simoni RD, Hill RL. The Thermodynamic Hypothesis of Protein Folding: The Work of Christian Anfinsen. The Journal of Biological Chemistry. 2006 April;281(14):e11–e13.

[10] Anfinsen CB. Principles that govern the folding of protein chains. Science. 1973 Jul;181(96):223–230.

[11] Branden C, Tooze J. Introduction to Protein Structura - Second Edition. Garland Publishing; 1999.

[12] Kauzmann W. Some factors in the interpretation of protein denaturation. Adv Protein Chem. 1959;14:1–63.

[13] Chandler D. Interfaces and the driving force of hydrophobic assembly. Nature. 2005 Sep;437(7059):640–647. Available from: http://dx.doi.org/10.1038/nature04162.

[14] Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol. 1971 Feb;55(3):379–400.

[15] Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J Mol Biol. 1973 Sep;79(2):351–371.

[16] Richards FM. Areas, volumes, packing and protein structure. Annu Rev Biophys Bioeng. 1977;6:151–176. Available from: http://dx.doi.org/10.1146/annurev.bb.06.060177.001055.

[17] Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. Hydrophobicity of amino acid residues in globular proteins. Science. 1985 Aug;229(4716):834–838.

[18] Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJP, Chothia C, et al. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 2008 Jan;36(Database issue):D419–D425. Available from: http://dx.doi.org/10.1093/nar/gkm993.

[19] Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995 Apr;247(4):536–540. Available from: http://dx.doi.org/10.1006/jmbi.1995.0159.

[20] Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. The EMBO Journal. 1996;5(4):823–826.

[21] Li P, Goldman N. Models of Molecular Evolution and Phylogeny. Genome Research. 1998;8(12):1233–1244. Available from: http://genome.cshlp.org/content/8/12/1233.abstract.

[22] Halperin I, Wolfson H, Nussinov R. Correlated mutations: Advances and limitations. A study on fusion proteins and on the Cohesin-Dockerin families. Proteins-structure Function and Bioinformatics. 2006;63:832–845.

[23] Zwanzig R, Szabo A, Bagchi B. Levinthal's paradox. Proceedings of the National Academy of Sciences. 1992 Jan;89(1):20–22. Available from: http://www.pnas.org/content/89/1/20.abstract.

[24] Olmea O, Valencia A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Folding and Design. 1997;2.

[25] Olmea O, Rost B, Valencia A. Effective use of sequence correlation and conservation in fold recognition. J Mol Biol. 1999;293:1221–1239.

[26] Fariselli P, Olmea O, Valencia A, Casadio R. Prediction of contact maps with neural networks and correlated mutations. Protein Engineering Design & Selection. 2001;14:835–843.

[27] Fariselli P, Casadio R. Prediction of the number of residue contacts in proteins. Proc Int Conf Intell Syst Mol Biol. 2000;8:146–151.

[28] Fodor AA, Aldrich RW. Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. Proteins: Structure, Function, and Bioinformatics. 2004;56(2):211–221. Available from: http://dx.doi.org/10.1002/prot.20098.

[29] Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D Structure Computed from Evolutionary Sequence Variation. PLoS ONE. 2011 12;6(12):e28766. Available from: http://dx.doi.org/10.1371%2Fjournal.pone.0028766.

[30] Abascal F, Valencia A. Clustering of proximal sequence space for the identification of protein families. Bioinformatics. 2002;18(7):908–921. Available from: http://bioinformatics.oxfordjournals.org/content/18/7/908.abstract.

[31] Göbel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins: Structure, Function, and Bioinformatics. 1994;18(4):309–317. Available from: http://dx.doi.org/10.1002/prot.340180402.

[32] Lockless SW, Ranganathan R. Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families. Science. 1999;286:295–299.

[33] Dodge C, Schneider R, Sander C. The HSSP database of protein structure-sequence alignments and family profiles. Nucleic Acids Research. 1998 Jan;26(1):313–315. Available from: http://dx.doi.org/10.1093/nar/26.1.313.

[34] Bateman A, Birney E, Cerutti L, Durbin R, Ettwiller L, Eddy SR, et al. The Pfam Protein Families Database. Nucleic Acids Research. 2002;30:276–280.

[35] Thomas DJ, Casari G, Sander C. The prediction of protein contacts from multiple sequence alignments. Protein Engineering. 1996;9(11):941–948. Available from: http://peds.oxfordjournals.org/content/9/11/941.abstract.

[36] Eyal E, Frenkel-Morgenstern M, Sobolev V, Pietrokovski S. A pair-to-pair amino acids substitution matrix and its applications for protein structure prediction. Proteins. 2007 Apr;67(1):142–153. Available from: http://dx.doi.org/10.1002/prot.21223.

[37] Sel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nature Structural Biology. 2003;10:59–69.

[38] Kass I, Horovitz A. Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. Proteins. 2002 Sep;48(4):611–617. Available from: http://dx.doi.org/10.1002/prot.10180.

[39] Dekker JP, Fodor A, Aldrich RW, Yellen G. A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments. Bioinformatics. 2004;20(10):1565–1572. Available from: http://bioinformatics.oxfordjournals.org/content/20/10/1565.abstract.

[40] Pollock DD, Taylor WR, Goldman N. Coevolving protein residues: maximum likelihood identification and relationship to structure. Journal of Molecular Biology. 1999;287(1):187 – 198. Available from: http://www.sciencedirect.com/science/article/pii/S0022283698926018.

[41] Fares MA, Travers SAA. A Novel Method for Detecting Intramolecular Coevolution: Adding a Further Dimension to Selective Constraints Analyses. Genetics. 2006;173(1):9–23. Available from: http://www.genetics.org/content/173/1/9.abstract.

[42] Various. Pearson product-moment correlation coefficient. Wikipedia. 2011;.

[43] Pazos F, Valencia A. Protein co-evolution, co-adaptation and interactions. The EMBO Journal. 2008;27:2648–2655.

[44] Clarke ND. Covariation of residues in the homeodomain sequence family. Protein Science. 1995;4:2269–2278.

[45] Atchley WR, Wollenberg KR, Fitch WM, Terhalle W. Correlations Among Amino Acid Sites in bHLH Protein Domains: An Information Theoretic Analysis. Mol Biol. 2000;17:164–178.

[46] Atchley WR, Terhalle W, Dress A. Positional Dependence, Cliques, and Predictive Motifs in the bHLH Protein Domain. Journal of Molecular Evolution. 1999;48:501–516.

[47] Shackelford G, Karplus K. Contact prediction using mutual information and neural nets. Proteins-structure Function and Bioinformatics. 2007;69:159–164.

[48] Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. Cell. 2012 Jun;149(7):1607–1621. Available from: http://dx.doi.org/10.1016/j.cell.2012.04.012.

[49] Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nature Biotechnology. 2012 Nov;30(11):1072–1080. Available from: http://dx.doi.org/10.1038/nbt.2419.

[50] Moult J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. Proteins: Structure, Function, and Bioinformatics. 1995;23(3):ii–iv. Available from: http://dx.doi.org/10.1002/prot.340230303.

[51] Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci. 1994 Mar;3(3):522–524. Available from: http://www.proteinscience.org/cgi/content/abstract/3/3/522.

[52] Chakravarty S, Varadarajan R. Residue depth: a novel parameter for the analysis of protein structure and stability. Structure. 1999 Jul;7(7):723–732.

[53] Hamelryck T. An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. Proteins. 2005 Apr;59(1):38–48. Available from: http://dx.doi.org/10.1002/prot.20379.

[54] Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. Science. 1983 Aug;221(4612):709–713.

[55] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983 Dec;22(12):2577–2637. Available from: http://dx.doi.org/10.1002/bip.360221211.

[56] Richmond TJ, Richards FM. Packing of alpha-helices: geometrical constraints and contact areas. J Mol Biol. 1978 Mar;119(4):537–555.

[57] Greer J, Bush BL. Macromolecular shape and surface maps by solvent exclusion. Proc Natl Acad Sci U S A. 1978 Jan;75(1):303–307.

[58] Richmond TJ. Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect. J Mol Biol. 1984 Sep;178(1):63–89.

[59] Pedersen TG, Sigurskjold BW, Andersen KV, Kjaer M, Poulsen FM, Dobson CM, et al. A nuclear magnetic resonance study of the hydrogen-exchange behaviour of lysozyme in crystals and solution. J Mol Biol. 1991 Mar;218(2):413–426.

[60] Song J, Tan H, Mahmood K, Law RHP, Buckle AM, Webb GI, et al. Prodepth: Predict Residue Depth by Support Vector Regression Approach from Protein Sequences Only. PLoS ONE. 2009 09;4(9):e7072. Available from: http://dx.doi.org/10.1371%2Fjournal.pone.0007072.

[61] Tan KP, Varadarajan R, Madhusudhan MS. DEPTH: a web server to compute depth and predict small-molecule binding cavities in proteins. Nucleic Acids Research. 2011;39(suppl 2):W242–W248. Available from: http://nar.oxfordjournals.org/content/39/suppl_2/W242.abstract.

[62] Wikipedia. Bayesian inference — Wikipedia, The Free Encyclopedia; 2013. [Online; accessed 30-April-2013]. Available from: \url{http://en.wikipedia.org/w/index.php?title=Bayesian_inference&oldid=556416559}.

[63] Hammel PJBEA, O'Connel JW. Sex bias in graduate addmissions: Data from Berkely. Science. 1975;187:398–404.

[64] Henikoff S, Henikoff JG. Position-based sequence weights. Journal of Molecular Biology. 1994;243(4):574 – 578. Available from: http://www.sciencedirect.com/science/article/pii/0022283694900329.

[65] Fc B, Koetzle TK, Williams GJ, Meyer EE, Brice MD, Rodgers JR, et al. The Protein Data Bank: A Computer-based Archival File For Macromolecular Structures. J of Mol Biol. 1977;112:535+.

[66] Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, et al. UniProt: the Universal Protein knowledgebase. Nucleic Acids Research. 2004;32:115–119.

[67] Andrade MA, O'Donoghue SI, Rost B. Adaptation of protein surfaces to subcellular location. J Mol Biol. 1998 Feb;276(2):517–525. Available from: http://dx.doi.org/10.1006/jmbi.1997.1498.

[68] Yu CS, Chen YC, Lu CH, Hwang JK. Prediction of protein subcellular localization. Proteins: Structure, Function, and Bioinformatics. 2006;64(3):643–651. Available from: http://dx.doi.org/10.1002/prot.21018.

[69] Levy ED. PiQSi: protein quaternary structure investigation. Structure (London, England : 1993). 2007 Nov;15(11):1364–1367. Available from: http://dx.doi.org/10.1016/j.str.2007.09.019.

[70] M J Betts RBR. Amino acid properties and consequences of substitutions. In: M R Barnes ICG, editor. Bioinformatics for Geneticists. Chichester: Wiley; 2003. .

[71] Chen H, Zhou HX. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. Nucleic Acids Research. 2005;33(10):3193–3199. Available from: http://nar.oxfordjournals.org/content/33/10/3193.abstract.

[72] Zhang H, Zhang T, Chen K, Shen S, Ruan J, Kurgan L. On the relation between residue flexibility and local solvent accessibility in proteins. Proteins: Structure, Function, and Bioinformatics. 2009;76(3):617–636. Available from: http://dx.doi.org/10.1002/prot.22375.

[73] Miller S, Janin J, Lesk AM, Chothia C. Interior and surface of monomeric proteins. Journal of Molecular Biology. 1987;196(3):641 – 656. Available from: http://www.sciencedirect.com/science/article/pii/0022283687900386.

[74] Vriend G. WHAT IF: a molecular modeling and drug design program. J Mol Graph. 1990 Mar;8(1). Available from: http://view.ncbi.nlm.nih.gov/pubmed/91098170.

[75] Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–1423. Available from: http://bioinformatics.oxfordjournals.org/content/25/11/1422.abstract.

[76] Song J, Tan H, Takemoto K, Akutsu T. HSEpred: predict half-sphere exposure from protein sequences. Bioinformatics. 2008 Jul;24(13):1489–1497. Available from: http://dx.doi.org/10.1093/bioinformatics/btn222.

[77] Karchin R, Cline M, Karplus K. Evaluation of local structure alphabets based on residue burial. Proteins: Structure, Function, and Bioinformatics. 2004;55(3):508–518. Available from: http://dx.doi.org/10.1002/prot.20008.

[78] Yuan Z. Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. BMC Bioinformatics. 2005;6(1):248. Available from: http://www.biomedcentral.com/1471-2105/6/248.

[79] Kowarsch A, Fuchs A, Frishman D, Pagel P. Correlated Mutations: A Hallmark of Phenotypic Amino Acid Substitutions. PLoS Comput Biol. 2010 09;6(9):e1000923. Available from: http://dx.doi.org/10.1371%2Fjournal.pcbi.1000923.

[80] Singer MS, Vriend G, Bywater RP. Prediction of protein residue contacts with a PDB-derived likelihood matrix. Protein Engineering. 2002;15(9):721–725. Available from: http://peds.oxfordjournals.org/content/15/9/721.abstract.

[81] Yeang CH, Haussler D. Detecting Coevolution in and among Protein Domains. PLoS Comput Biol. 2007 11;3(11):e211. Available from: http://dx.plos.org/10.1371%2Fjournal.pcbi.0030211.

[82] Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. ClustalW and ClustalX version 2. Bioinformatics. 2007;23(21).

[83] Lee L, Leopold JL, Frank RL. Protein secondary structure prediction using BLAST and exhaustive RT-RICO, the search for optimal segment length and threshold. In: Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2012 IEEE Symposium on; 2012. p. 35–42.

[84] Reese MG, Lund O, Bohr J, Bohr H, Hansen JE, Brunak S. Distance distributions in proteins: a six-parameter representation. Protein Engineering. 1996;9(9):733–740. Available from: http://peds.oxfordjournals.org/content/9/9/733.abstract.

[85] Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to Algorithms, Third Edition. 3rd ed. The MIT Press; 2009.

[86] Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y. Automatic prediction of protein function. Cellular and Molecular Life Sciences CMLS. 2003;60(12):2637–2650. Available from: http://dx.doi.org/10.1007/s00018-003-3114-8.

[87] Höglund A, Dnnes P, Blum T, Adolph HW, Kohlbacher O. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. Bioinformatics. 2006;22(10):1158–1165. Available from: http://bioinformatics.oxfordjournals.org/content/22/10/1158.abstract.