Modeling Nonspecific Interactions at Biological Interfaces

Andrew D. White

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Shaoyi Jiang, Chair

David Beck

Valerie Daggett

Walter Pfaedtner

Program Authorized to Offer Degree:
Chemical Engineering

University of Washington

Abstract

Modeling Nonspecific Interactions at Biological Interfaces

Andrew D. White

Chair of the Supervisory Committee:
Professor Shaoyi Jiang
Chemical Engineering

Difficulties in applied biomaterials often arise from the complexities of interactions in biological environments. These interactions can be broadly broken into two categories: those which are important to function (strong binding to a single target) and those which are detrimental to function (weak binding to many targets). These will be referred to as specific and nonspecific interactions, respectively. Nonspecific interactions have been central to failures of biomaterials, sensors, and surface coatings in harsh biological environments. There is little modeling work on studying nonspecific interactions. Modeling all possible nonspecific interactions within a biological system is difficult, yet there are ways to both indirectly model nonspecific interactions and directly model many interactions using machine-learning. This research utilizes bioinformatics, phenomenological modeling, molecular simulations, experiments, and stochastic modeling to study nonspecific interactions. These techniques are used to study the hydration molecules which resist nonspecific interactions, the formation of salt bridges, the chemistry of protein surfaces, nonspecific stabilization of proteins in molecular chaperones, and analysis of high-throughput screening experiments. The common aspect for these systems is that nonspecific interactions are more important than specific interactions. Studying these disparate systems has created a set of principles for resisting nonspecific interactions which have been experimentally demonstrated with the creation and testing of novel materials which resist nonspecific interactions.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

DEDICATION

to my mother, who drove me 60 miles every Saturday one summer to learn to program,
thus leading me down the path to simulations.

Chapter 1

INTRODUCTION

The interactions which govern chemical processes may be broadly categorized into specific interactions and nonspecific interactions. Specific interactions are high activity interactions between two molecules. They are by far the most researched of the two interaction types. Some examples include designing functional peptides through screening libraries of peptides,[1] characterizization or design of enzymes possessing activity for a specific target,[2] and catalyst design where activity for a specific substrate is desired.[3] In contrast, nonspecific interactions are weak interactions for all potential targets of a molecule. These are generally considered a nuisance to experiments and as few nonspecific interactions as possible are desired. Nonspecific interactions are themselves the sum of many weak specific interactions. Examples of systems that emphasize nonspecific interactions include proteins which are stable in environments with many other macromolecules, materials which resist nonspecific binding,[4] and enzymes which are highly selective and rarely bind to non-targets. Despite their ubiquity in biology and chemistry, nonspecific interactions are generally overlooked. It is fundamentally a challenge to screen or study these ubiquitous and weak nonspecific interactions.

One key application of understanding nonspecific interactions is in the development of nonfouling materials, which resist the attachment of biomolecules and microorganisms. They resist binding from biological species, preventing fouling. Although many nonfouling materials, particularly zwitterionic polymers, have been used in applications,[4] it is still challenging to design synthetic drug delivery carriers matching native protein circulation time in the blood stream and implantable materials fully compatible with human tissue.

For example, the circulation half-life of albumin[5] in blood is still longer than any synthetic particles, even after state-of-the-art PEG-modification.[6] These are interesting applications because it is impossible to enumerate all the specific binding interactions which must be minimized. The problem may only be modeled by studying nonspecific interactions.

A second application area is maximizing activity in real systems. It is well understood how to maximize activity for enzymes or catalysts when only substrate is present. However, it is often the nonspecific activity that prevents *in vivo* applications because nonsepcific activity competes with specific activity.[7] For example, antibodies exist which may detect cancer biomarkers today, yet most biosensors fail at detecting cancer biomarkers because whole blood contains so many nonspecifically binding components.[8] Motivated by these important applications, in this work I present three techniques to model nonspecific interactions applied to three systems. In Chapter 2, atomistic molecular simulations are applied to design nonfouling materials. In Chapter 3, bioinformatics are applied to model nonspecific interactions on protein surfaces and the nonspecific stabilization of protein folding from molecular chaperones. In Chapter 4, structure-property relationships and motif models are applied to separate specific from nonspecific effects in high-throughput screening experiments.

Chapter 2

DESIGN OF NONFOULING PEPTIDES THROUGH MOLECULAR SIMULATIONS

2.1 Molecular Mechanism of Nonfouling: Zwitterions and Polymers

In this chapter, the design self-assembling nonfouling peptides is described. Peptides are short oligomers of amino acids. Due to the choice of 20 amino acids and choice of length of peptides, there is an enormous design space. It is important to have a strong understanding of what causes nonfouling in order to narrow the design of nonfouling peptides. In this chapter I consider three steps which result in experimentally successful nonfouling peptides. In Section 2.2, I revisit the design criteria for nonfouling materials and address remaining questions to narrow the possible amino acids to four charged amino acids: E, D, R and K. These are shown in Figure 2.1. In Section 2.3, I study the interactions between these four charged amino acids to determine which have minimal self-interactions and maximum hydration. Finally, in Section 2.4 I describe the modeling and experimental work to incorporate the final nonfouling peptide design into a self-assembling peptide.

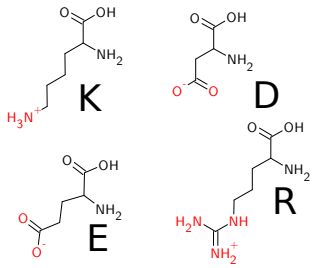


Figure 2.1: The four charged amino acids considered for nonfouling peptides. Side-chain atoms are highlighted in red.

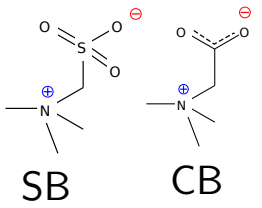


Figure 2.2: SB is sulfobetaine, a commonly used nonfouling zwitterion. CB is carboxybetaine and is quite similar in structure to glycine betaine. CB and SB zero adsorption from undiluted blood, providing some of the best nonfouling performance.

Fouling affects fields from medical device coatings to maritime coatings.[9] Anti-fouling materials have been studied for a number of years[10–12]. Well accepted materials, such as poly(ethylene glycol)(PEG) and recently recognized materials, such as poly(carboxy betaine) (pCB) and poly(sulfobetaine) (pSB) are used as nonfouling materials in many environments.[4] The relevant side-chains of these polymers are shown in Figure 2.2. Initial studies on anti-fouling and the success of PEG hypothesized that the resistance of fouling is due mostly to a polymer physics effect.[13] There exists a rich description of theory for tethered polymers and their resistance to adsorption.[14] One of the key conclusions is that the excluded volume is one of the most important parameters. Practically, the excluded volume may be increased either by increasing molecular weight or the solubility of the polymers. In water, this means ensuring a polymer is hydrophilic and has a high molecular weight. When considering multiple chains, maximizing surface density also improves the physical resistance to adsorption.

Surprisingly, it is also possible to resist protein adsorption with self-assembled monolayers (SAMs).[15] SAMs have no flexibility and thus no excluded volume effect. This demonstrates that there must be another mechanism for resisting protein adsorption. Research from Ostuni *et al.*[16] tested many chemistries for protein adsorption to discover the structure-properties of chemical, as opposed to polymer, anti-fouling. They found that the compounds must be hydrophilic, contain only hydrogen bond acceptors, and be net neutral.[10] The net

neutrality is easily explained; most proteins have a net charge. The condition of being hydrophilic has been followed up with additional research.[11,12,17] The hydration of a hydrophilic surface creates a layer of a well structured water.[18] This water structuring then imparts an enthalpic penalty for proteins to penetrate the water and adsorb.[11] It is interesting that this makes hydration doubly important for polymers. Polymers must be highly soluble to create excluded volume and to have strong hydration.

Based on these principles, pCB has been recognized as being a highly hydrated and effective nonfouling material. Quaternary ammonium zwitterions are an elegant answer to the criteria for nonfouling. They are net neutral, but the strong charges create hydration. The quaternary ammonium contains no hydrogen bond donors and the carboxylate anion contains hydrogen bond acceptors.

Returning to the design of nonfouling peptides, we find there are no zwitterionic amino acids. It is possible to alternate between positively and negatively charged amino acids, but that introduces hydrogen bond donors from the amino positively charged amino acids (K – primary amine, R – guanidinium). Another solution is to use uncharged amide or hydroxyl groups which contain fewer hydrogen bond donors but are also less hydrophilic than charged groups. Thus, we are left with competing designs due to the hydrogen bond donor criterion.

The no hydrogen bond donor criterion comes from experimental results on methylation of amine groups.[16] Adding methyl groups to a primary amine gradually increases the nonfouling performance. In the next section, I examine how this change from primary amines to quaternary amines effects hydration so that we may understand why hydrogen bond donors are detrimental to nonfouling and better choose a nonfouling peptide.

2.2 Choosing Amine Group: Primary vs Quaternary

2.2.1 Introduction

Molecular dynamics are used here to characterize the difference in hydration between glycine and two of its zwitterionic analogues: N,N-dimethylglycine and N,N,N-trimethylglycine

(glycine betaine). These three analogues contain primary, tertiary, and quaternary amines, respectively. Hydration is critical for nonfouling performance, and studying these the hydration of these three amine groups will help resolve the conflict between hydrogen bond donors and hydrophilicity. The hydration of dodecane and oligo(ethylene glycol) was studied for reference. Both structuring and dynamics of bulk and bound water were examined using a variety of properties and at multiple concentrations. Metrics, such as radial distribution functions and residence times, were used to characterize hydration. Also, we used more specialized metrics that can discriminate between subtle differences in hydration such as condensed phase order parameters, Voronoi tessellations, and multidimensional pair-pair correlation functions.

The molecules chosen for this study are dodecane, oligo(ethylene glycol), glycine, and two glycine analogues: dimethylglycine and trimethylglycine. The three zwitterions have a primary amine (glycine), a tertiary amine (dimethylglycine), and a quaternary amine (trimethylglycine). The differences among these three amines are rarely studied, especially their hydration. Dodecane is a hydrophobic alkyl chain. Alkyl self-assembled monolayers are known to make poorly hydrated fouling surfaces and serve as a reference. Nonfouling oligo(ethylene glycol) self-assembled monolayers are highly hydrated surfaces and are the most widely used nonfouling surfaces.[4] Oligo(ethylene glycol) is the hydrated reference in this work. Glycine is the simplest amino acid and one of the smallest zwitterions. Dimethylglycine is studied less often than the other molecules but has unique properties. For example, of all the glycine methylated analogues (glycine, sacrosine, dimethylglycine, trimethylglycine), dimethylglycine has the highest gas phase energy gap between the neutral and zwitterion forms.[19] Trimethylglycine has been found to have a diversity of uses in nature. It is a naturally occurring osmolyte, allowing the human pathogen *Listeria monocytogenes* to grow at high salt concentrations for instance.[20] It also encourages the growth of *Lactococcus lactis*[21] and increases strawberry plants' cold tolerance.[22] It is able to prevent protein aggregation,[23] and its effect as a cosolvent is often studied.[24] Trimethylglycine is also the nonfouling moiety in pCB, where it effectively prevents nonspecific protein adsorp-

tion.[25]

Both structure and dynamics are essential to understanding hydration. Structurally, water adopts two structural motifs at ambient liquid conditions: a tetrahedral (icelike) structure and an amorphous structure similar to gas-phase water[26,27]. In experiments, the OH bond stretching is recognized as an indicator of tetrahedral geometry.[28] In simulations, the water tetrahedral order parameter, q, can be used to quantify how tetrahedral water is.[29,30] Dynamically, water is slowed by nearly any solute, except certain ions[28,31–33]. The degree of slowing can measured by water self-diffusion coefficients and is an informative bulk property. Local properties are important as well, especially when understanding confined or volume localized hydration (e.g., the binding pocket of a protein). Local properties include hydrogen-bond lifetimes and residence times. These allow one to see how transient the water structure is and how strongly water is being held. A combined dynamics and structural approach is important when describing hydration.

The importance of understanding hydration is not unique to nonfouling. Hydration is gradually being recognized as integral for many phenomena, and simulation of explicit water is a near necessity in biomolecular simulations. Recent experimental research on the hydration monolayer around proteins, for example, has shown that polar groups on a protein adopt conformations to satisfy the tetrahedral orientation of water.[34] The removal of a hydration layer has been found to be essential for protein-protein association.[35] Hydration of enzymes is needed to characterize kinetics and reaction mechanisms.[36] All of these systems in which water is key necessitate a strong understanding of hydration and an ability to quantify it in simulations.

To the best of our knowledge, there are no theoretical studies of the comparative hydration of the N-methylated glycine sequence (glycine, dimethylglycine, and trimethylglycine). The hydration differences between the primary, tertiary, and quaternary amine are also important to understand in other research areas. In this work, we attempt to answer these questions through simulations, which allow both bulk and local perspectives into hydration.

Atom	glycine	N,N-dimethylglycine	trimethylglycine
O	-0.842	-0.839	-0.842
N	-0.386	0.139	0.3914
C=O	0.900	0.914	0.883
CH2	0.022	-0.217	-0.122
C-N	NA	-0.292	-0.372
H-N	0.340	0.266	NA
H3-C	NA	0.139	0.165
dipole	13.02	13.756	14.142

Table 2.1: Partial charges in elementary charge units and dipole moments in Debyes for the three glycine analogues.

2.2.2 Methods

This research was done with molecular dynamics assisted by quantum chemistry. Molecular dynamics has the advantage of being able to simulate ensembles large enough for dynamic information (e.g., diffusion), while maintaining accuracy.[37] Five solutes were studied: dodecane, oligo(ethylene glycol) with 3 ethers, glycine, dimethylglycine, and trimethylglycine. The zwitterion structures are shown in Figure 2.3. Solutions of 800 water molecules and 3, 8, 16, 25, 32, and 45 solute molecules were used, for a total of six simulation systems for each zwitterion. Only 3, 8, and 16 solutes were simulated for the dodecane and oligo(ethylene glycol) simulations. This corresponds to 0.1 M to 3 M solutions. A run of only water molecules was done for comparison.

An OPLS-AA/Amber99 forcefield[38] was chosen for the molecular dynamics. The partial charges, when necessary, were supplemented with 6-31G(p,d) B3LYP quantum chemistry in the Gaussian 03 software.[39] The ESP (MK) electrostatic partial charge method[40] was used to derive the partial charges. This procedure was chosen because it best reproduced the

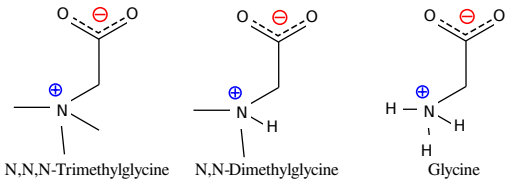


Figure 2.3: The three zwitterionic glycine analogues studied.

charges on parametrized moieties in OPLS and provided a way to consistently create partial charges. For example, MP2 gives a charge of 0.994 on the carboxylate anion carbon on glycine betaine, HF gives 0.8968, B3LYP gives 0.883, and OPLS-AA lists 0.700 for a general carboxylate anion. These charges are shown in Table 2.1. The TIP4P-Ew[41] water model was used because it has been shown to be one of the best models for reproducing experimental water structure.[42] It also closely matches experimental water self-diffusion.[41] Although no one water model is best, any constant systematic error should not affect conclusions drawn by comparing results among the different molecules. Additionally, force field dihedral angle minimums were verified using B3LYP quantum mechanics calculations.

The MD engine Lammps was used.[43] The initial configurations were generated with packmol[44] and Nosé/Hoover NPT[45] with the Martyna et. al equations of motion[46] was done for a 1.5 ns equilibration period and a 1 ns data collection period. Longer simulations of 50 ns had no effect on the results and the solute torsions, the slowest molecular motions, were examined to ensure sufficient sampling was accomplished. Particle-particle particle-mesh was used to eliminate cutoff artifacts of the electrostatics.[47]

The data analysis was done with a mixture of custom code and Lammps. The custom code was tested against VMD,[48] Lammps, or literature values. The interaction energy (E_i) was calculated and represents the sum of the nonbonded energy between two groups of atoms. In this case, E_i is between water and the solute. Self-diffusion coefficients were calculated with the mean squared displacement and Einstein equation.[49] The 95% confidence intervals were calculated by blocking the simulation into 100 ps regions as well as multiple

starting configuration simulations.

Coordination shells, $g(r)$, residence times, tetrahedral order, hydrogen bonding, and hydrogen-bond lifetimes were calculated. The coordination shells were determined from the minimum after the first peak in $g(r)$, giving r_{sh}. The coordination number, N_{sh} is given by $N(r_{\mathrm{sh}})$, where $N(r)$ is the number of molecules encapsulated by a radius r from a solute.

Residence times of solvent in the first coordination shell of the solute were calculated using an autocorrelation function. Each coordination shell is considered a separate region. The autocorrelation function is

$$C_{res}(t) = \frac{\langle R(0)R(t)\rangle}{\langle R(0)\rangle^2} \tag{2.1}$$

where $R(t)$ is the residence function defined by

$$R(t) = \sum_i \sum_j R_{ij}(t) \tag{2.2}$$

where i is the region index and j is the atom index. $R_{ij}(t)$ is 1 when the jth water oxygen is within r_{sh} of the ith solute; otherwise it is 0.

First order behavior was assumed for fitting. The 95% confidence intervals were calculated by blocking multiple simulations into 100 τ_{res} blocks.

The hydrogen-bonding was defined using geometric criteria.[50] A water and a donor/acceptor are considered hydrogen-bonding if the O-H$\cdots$O or H-O$\cdots$ H (for H-O$\cdots$ H-N bonds only) angle is less than 30° from 180° and the distance between oxygen pairs or nitrogen oxygen pairs is < 3.5Å. Although nitrogen has a slightly larger Van der Waals radius than oxygen, changing the definition to account for this had a negligible effect on the hydrogen-bonding count. The hydrogen bonding lifetime was calculated using an autocorrelation function:

$$C_{hb}(t) = \frac{\langle h(0)h(t)\rangle}{\langle h(0)\rangle^2} \tag{2.3}$$

where $h(t)$ is the hydrogen-bond population function defined by

$$h_i(t) \equiv \begin{cases} 1 & i \text{ H-bond to solute} \\ 0 & i \text{ no H-bond} \end{cases}, \; h(t) = \sum_i h_i(t) \tag{2.4}$$

Note that only water is considered here, so solute-solute interactions decrease hydrogen-bonding. This has been chosen intentionally, since this research considers solute-solute interactions as a decrease in hydration. hydrogen-bonding for water to water was defined using the same correlation function and geometric definition, where the O$\cdots$O distance is < 3.5Å and the H$\cdots$O-H angle is less than 30° from 180°. First order behavior again was assumed for calculating lifetimes. Although a first order exponential fit does not completely capture the autocorrelation function, it does provide a single numerical value with which to compare the similar molecules. Using, for example, a double exponential does not change the qualitative trends.

The tetrahedral order parameter, q, was calculated for water. Here, a low q is less tetrahedral order and $q = 1$ when there is a maximum.[51] This is a common modification of the traditional s tetrahedral order parameter.[29] q is defined by:

$$q_i \equiv 1 - \frac{3}{8} \sum_{j=1}^{3} \sum_{k=j+1}^{4} \left[\cos \theta_{ijk} + \frac{1}{3} \right]^2 \tag{2.5}$$

θ_{ijk} is the angled formed between two neighbor atoms, j and k, and the central atom i. q is presented as a line plot of probability density, determined from histogramming. Negative q values are possible, though these are excluded from the plots for simplicity.

Voronoi tesselations and Voronoi polyhedra analysis in the disordered condensed phase have been described elsewhere.[52] Qualitatively, it provides a method for measuring an ensemble molecular volume and surface area from a trajectory using a geometric tessellation. We used monodisperse tesselations, which means that the tessellation is not weighted by atomic radii. The analysis is still sensitive to the force field because a molecular dynamics trajectory is used for the tessellation. An external program, Voro++, was used.[53]

Pair-pair correlation functions in multiple dimensions were calculated to overcome the radial symmetry implied in $g(r)$. These pair-pair correlation functions are based on polar spherical coordinates, with a polar angle, azimuthal angle and radial distance. The equation is:

$$g(r, \theta, \phi) = \left\langle \frac{1}{\rho_{\text{bulk}}} \frac{N(r, \theta, \phi)}{\int_0^r \int_0^{2\pi} \int_0^{\pi} r^2 \sin(\phi)\, d\phi\, d\theta\, dr} \right\rangle \tag{2.6}$$

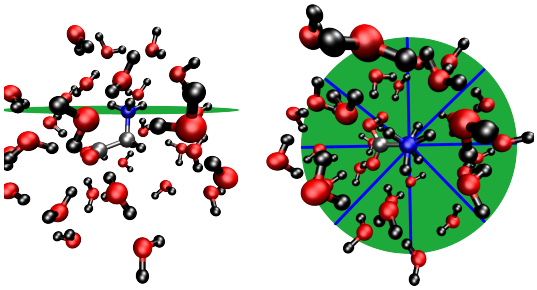


Figure 2.4: Cartoon demonstrating the $g(u,v)_{\mathrm{N,O}}$. The left picture shows the way the normal vector defines a plane and the right pictures shows the azimuthal dependence.

where ρ_{bulk} is the bulk number density. To create easily interpretable plots, the polar angle was removed and thus the dimensions reduced to two using the equation $\int g(r,\theta,\phi)\sin\phi\, d\phi$. This creates volume elements shaped like concentric orange peels. The plots are shown as heat plots of the radial and azimuthal coordinates; however they are parametrized to a u, v with $u = r\cos\phi$ and $v = r\sin\phi$ in order to make a square plot. Two reference vectors are required to define the local coordinates for each molecule. Two atom-atom vectors are used. The first defines the polar coordinate, which is the direction normal to planes plotted in the figures. The second vector defines the azimuthal origin (which is on the left).

All of the figures in the paper use the same coordinate vectors and are $g(u,v)_{\mathrm{N,O}}$ for amine nitrogen and water oxygen atom pairs. The N $\cdots\alpha-$C atom-atom vector defines the normal vector. This makes the plots show a top view of the amine group with the $\alpha-$carbon directly below the nitrogen atom. The second atom-atom vector is N-Ca, where Ca is the carboxylate anion carbon. Thus the left side of the figures represents the direction of the carboxylate anion. The radial distance corresponds to $g(r)$, so concentric circles in the plot are equivalent to r values in $g(r)$. For further clarification, the reader is directed to a cartoon in Figure 2.4 which shows this graphically.

2.2.3 Results and Discussion

The bulk dynamics were investigated where bulk refers to a system wide property. Its behavior was found to be dependent on solute species, namely noncolligative. Glycine aggregation was found to explain some of the differences between the three zwitterions. The dynamics were also shown to correlate with the tetrahedral order of water. This led us to investigate the local hydration. We used the spherically symmetric $g(r)$ but found it insufficient for a complete description of local hydration. Next, we used $g(u, v)$, local hydrogen-bond lifetimes and Voronoi tessellations to further elucidate the local hydration.

Bulk Dynamics

The self-diffusion coefficient of water (D_w) is a measure of water slowing. Our previous studies show that a nonfouling surface resists nonspecific protein adsorption due to its strong surface hydration. Proteins are often considered "slaved" to waters and adopt configurations subject to the tetrahedral geometry of water molecules.[34] D_w may be a predictor of nonfouling because it estimates the dynamic influence of a surface on water and thus a protein. Thus, we consider these solution phase simulations to be a model for the more complicated surface phenomenon. Also, some research suggests that slower solvent dynamics can inhibit protein aggregation;[54] D_w is a measure of these solvent dynamics.

D_w of the different species and concentrations is shown in Figure 2.5. The abbreviations used in figures and tables are glycine (gly), dimethylglycine (dmg), trimethylglycine (tmg), and oligo(ethylene glycol) (oeg). The pure water simulation D_w matches experiments quite well, as expected for TIP4P-Ew.[55] As expected, all solutes slow water diffusion. The hydrophobic dodecane affects water diffusion least due to its hydrophobicity, seen from its lower partial charges. Oligo(ethylene glycol) is a commonly used hydrophilic nonfouling molecule and slows water considerably more than the hydrophobic dodecane. It has a stronger effect on D_w than the zwitterions, but this oligo(ethylene glycol) has a molecular weight of 194 Da compared with trimethylglycine's weight of 118 Da. Also, oligo(ethylene glycol) is not soluble up to the impressively high concentrations of trimethylglycine: 5.5 M at 25° C.[56]

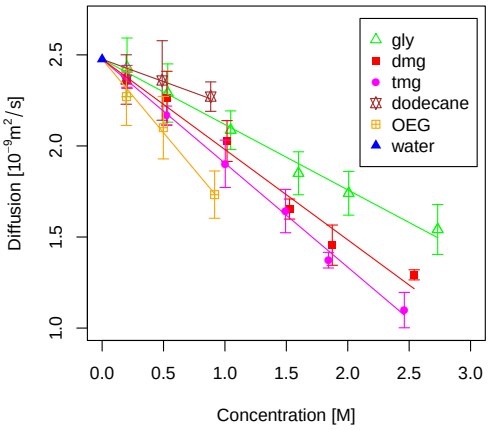


Figure 2.5: Self-diffusion coefficient of water, D_w, as a function of solute concentration. The decrease in D_w is due to an increase in bound water. Notice the steep linear decline for trimethylglycine D_w. Slopes are, in order of the legend, -0.36, -0.50, -0.57, -0.25, and -0.81. Error bars represent 95% confidence intervals from multiple simulations or blocked (split up) single simulations.

Trimethylglycine has D_w values similar to those of the other zwitterions at the lowest concentration but gradually changes to the lowest D_w at higher concentrations. Glycine and dimethylglycine are not as steeply decreasing. The slope of glycine is $-0.36 \cdot 10^{-9}\mathrm{m}^2/\mathrm{M}$ s whereas the slope for the trimethylglycine is $-0.57 \cdot 10^{-9}\mathrm{m}^2/\mathrm{M}$ s. For reference, NMR experiments have shown CH3COOK, the negative moeity on these zwitterions, to have a slope of $-0.25 \cdot 10^{-9}\mathrm{m}^2/\mathrm{M}$ s. It is perhaps unexpected that glycine influences D_w less than trimethylglycine. Glycine has five hydrogen-bonding sites, compared with trimethylglycine's two. One might expect the opposite trend: more hydrogen-bonding would have a larger affect on D_w. An explanation for this is offered below, that glycine is aggregating.

On a per-weight basis, the trimethylglycine lowers D_w more than all other simulated molecules. At an interface meant to be nonfouling, for example, trimethylglycine is likely better because of its large effect on D_w. Additionally, proteins are known to be stabilized by trimethylglycine,[23] perhaps because of trimethylglycine's large effect on water dynamics.

Glycine Aggregation Formation

These different degrees of hydration among the solutes can also be quantified using interaction energy between the solute and solvent, E_i. A feature noticed in E_i is a decrease in magnitude of the glycine E_i values: $E_i = -140$ kcal/mol at dilute to $E_i = -91$ kcal/mol at high concentration on a per solute molecule basis. A decrease in the water-glycine interaction means that there must be an increase in the glycine-glycine interactions. An increase in intermolecular glycine interaction presents an explanation for the higher D_w of glycine relative to trimethylglycine and dimethylglycine: there is glycine aggregation in the simulation. This can be observed in the trajectory and quantitatively in Figure 2.6. Figure 2.6 shows $g(r)_{\mathrm{C,N}}$ for glycine at three concentrations. The $g(r)_{\mathrm{C,N}}$ is the radial distribution function between glycines (excluding intramolecular pairs) and is a metric for how near glycines are to one another relative to the bulk glycine density. The large peak indicates pairing of glycines. Dimers, cyclic dimers, and small clusters of glycines were observed forming and dissipating in the trajectories. Trimethylglycine does not aggregate, and this leads to the

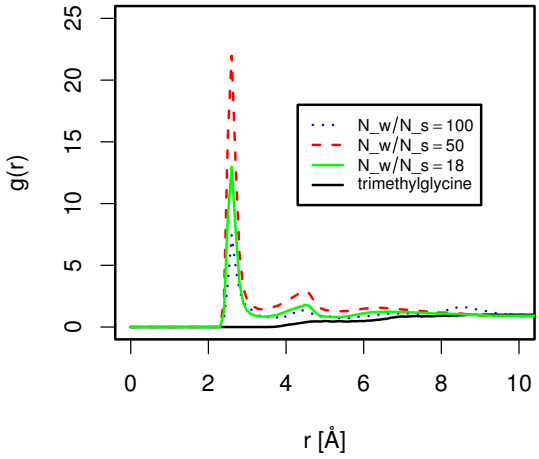


Figure 2.6: $g(r)_{\mathrm{C,N}}$ for glycine at various concentrations in number of waters per solute. The pair is the carboxylate anion carbon and the primary amine nitrogen. The peak indicates glycine aggregation. Trimethylglycine is included as a reference, at 16 waters per solute. The disperse trimethylglycine creates the smoother $g(r)_{\mathrm{C,N}}$.

larger magnitude values of E_i and lower D_w values. This further demonstrates the high hydration of trimethylglycine.

The formation of cyclic glycine dimers has been the topic of previous research. There is experimental evidence that cyclic dimers are not dominant forms, as seen in this work, although some previous research has suggested otherwise.[57] The number of cyclic dimers vs. singly bonding $\mathrm{N-H\cdots O}$ pairs was found to match the heterogeneous aggregation of glycine observed in other recent molecular dynamics studies.[58]

Connecting Bulk Structure and Bulk Dynamics

We use the tetrahedral order parameter q to describe how bulklike water is and thus if it is solute bound or unbound. Bulklike here means water unaffected by solute and structurally similar to neat water. This can be seen in Figure 2.7, which shows q as a histogram for neat water (blue line). The peak on the right, at $q = 0.8$, represents water that is tetrahedrally

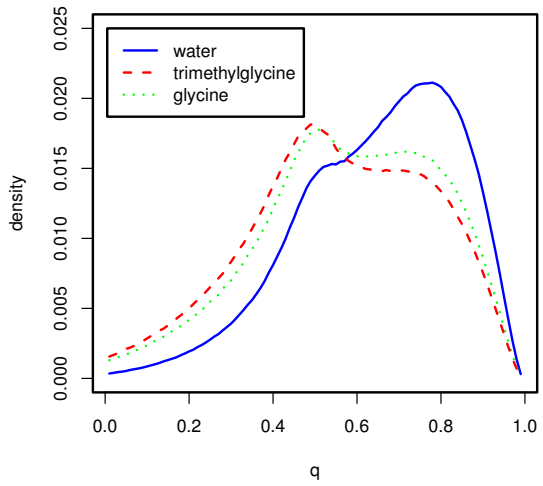


Figure 2.7: Histogram of tetrahedral order of water. The location of the tetrahedral order peak ($q = 0.8$) that corresponds to "ordered" water. The decrease of this peak corresponds to an increase in bound water.

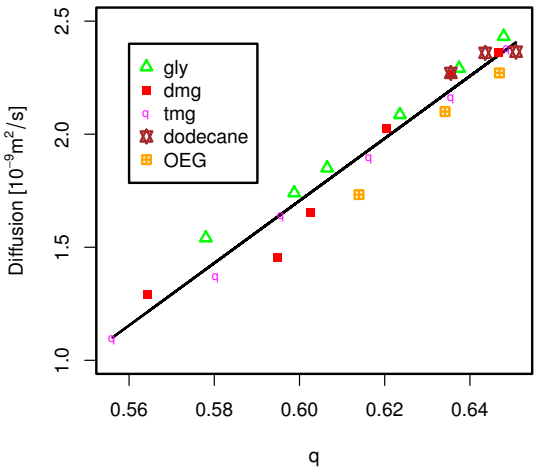


Figure 2.8: Fit of the average q (tetrahedral water order) against diffusion. $R^2 = 0.95$. The good fit shows that it is possible to connect dynamic and structural features of water, regardless of solute.

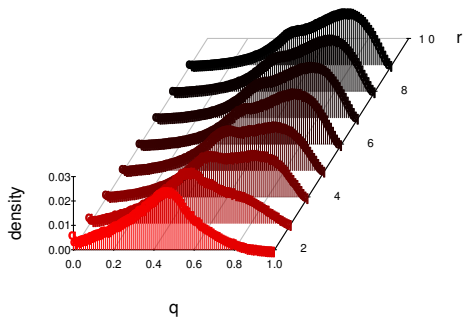


Figure 2.9: Histogram of the tetrahedral order parameter as a function of the distance from amine nitrogen of glycine at 0.5M. The density axis is the probability density, or histogram, of q. The r axis is displaced because not enough water molecules are sampled within 2 Å of the nitrogen for sufficient sampling. The right peak, $0.65 < q < 0.87$, is tetrahedrally oriented water. Water adopts the solute's shape and returns to bulk water after approximately 1.5 water coordination shells.

aligned with four other water molecules, which occurs only with bulklike water. The peak on the left, at $q = 0.47$, is a mix of other water types, including unbound water only coordinated with one other water, bound water, or other unstructured water. Thus, a decrease in the average q can be considered a decrease in bulklike water, as seen upon the addition of glycine or trimethylglycine shown in Figure 2.7. Adding either of these two solutes increases the $q = 0.47$ peak as well as decreases the proportion of bulklike water. This trend can be used to fit the D_w values (Figure 2.8). The fit shows that much of the change in D_w is correlated with the tetrahedral order of water. The quality of the fit ($R^2 = 0.95$) means there is a strong correlation between a dynamic and structural property of water: as the water structure is deformed from additional solute, the self-diffusion coefficient of water decreases. Bulk water is quite fast,[59] and an increase in bound water from solutes would be expected to slow the dynamics. Tetrahedral order allows this to be quantified.

Local Hydration through $g(r)$ and q

Additional evidence for the assumption that one peak is predominately bound, local water and the other is bulklike water can be found in Figure 2.9. This figure shows the q histogram as a function of distance from a glycine nitrogen. The r value is the distance to the nearest solute. As we look farther from the glycine, there is a shift from a single peak at $q = 0.47$ to the two peak motif found in pure water. Thus, a disappearance of the second peak corresponds to bound water (less tetrahedral). The characteristics of local water can be quantified using the tetrahedral order and plots such as Figure 2.9. Among the three zwitterions, the emergence of a bulklike q distribution is found between 1.5-2.0 water coordination shells away. Thus, the solvation shell can be interpreted as extending 1.5-2.0 water shells, which is similar to what is seen in other molecular simulations.[60] Among the different amines, trimethylglycine has bulklike water closest to its nitrogen, followed by dimethylglycine and glycine. This is interesting since the methyl groups on the quaternary amine are larger and exclude much more solvent than the hydrogen atoms on the primary amine. The above analysis breaks down at higher concentrations because the solutes are within each others hydration shells. Nevertheless, in general glycine, dimethylglycine, and trimethylglycine affect water two coordination shells away at high concentration. The tetrahedral order requires 4 neighboring water molecules and is not effective at characterizing water adjacent to a solute's surface. $g(r)$, on the other hand, can provide excellent structural data. Using $g(r)$, Table 2.2 can be generated.

All the properties in Table 2.2, with the exception of hydrogen bonding, are derived from the $g(r)$ of the amine group in the zwitterions. These properties are meant to quantify both the structure and dynamics of hydration near the different amines. Looking at the table, one can see a 30% decline in the glycine amine water coordination number between the values at the low and high concentrations. This change is not seen in the amine methyl water coordination numbers (dimethylglycine, trimethylglycine), instead only in the amine hydrogens. The reduction in amine water coordination is related to inter-solute aggregation. This matches the trends seen before which indicate glycine and to a lesser extent dimethylglycine

	N_w/N_s	r_{sh} [Å]	$N(r_{sh})$	τ_{res} [ps]	H-bond
gly	250	3.4	4.04	0.53 ± 0.02	0.98
gly	18	3.3	2.82	0.69 ± 0.07	0.67
dmg	250	3.3(4.6)[a]	1.09(10.37)[a]	$0.46 \pm 0.25(1.67 \pm 0.03)$[a]	0.94
dmg	18	3.2(4.6)[a]	0.72(10.37)[a]	$0.39 \pm 0.02(0.31 \pm 0.02)$[a]	0.66
tmg	250	4.6	9.04	1.75 ± 0.03	NA[b]
tmg	18	4.5	8.69	0.129 ± 0.004	NA[b]

[a] Values in parenthesis are based on methyl groups attached to the nitrogen, whereas the other values are based on the nitrogen. [b] There are no hydrogen-bond donors on the quaternary amine.

Table 2.2: Hydration properties for amine group. N_w/N_s is the number of water molecules per solute. It is inversely related to concentration. r_{sh} is the first water coordination shell radius. $N(r_{sh})$ is the number of water molecules in the coordination shell. τ_{res} is the residence time of water in the first coordination shell. The properties are calculated from either the nitrogen (glycine, dimethylglycine) or methyl carbon (trimethylglycine, dimethylglycine) and water oxygen radial distribution function. The hydrogen-bonds are water hydrogen-bonds per hydrogen-bond donor, in this case per amine hydrogen.

aggregate. Bulk aggregation affects local hydration structure; consequently, it changes the dynamics of the bound water. Glycine bound water has a slight increase in residence time. Considering the structure and dynamics together, glycine has few waters at higher concentration but they are bound longer. Trimethylglycine and dimethylglycine amine methyls' residence times are more sensitive to the concentration: they retain water as long as the hydrogen-bonding primary amine on glycine, but at higher concentrations the water moves rapidly between the dimethylglycine or trimethylglycine molecules. Recall that the residence time here measures how long a water resides at a given trimethylglycine molecule. Thus, the water molecules may be exchanging rapidly with bulk or other trimethylglycines. In order to determine which, the residence time of water within the coordination shell of any trimethylglycine molecule was calculated. It was found to be 150 ps at the highest concentration as opposed to 0.13 ps between a given trimethylglycine. Thus, water is near any trimethylglycine amine for a long period but changes between individual trimethylglycine molecules quickly.

A similar table for dodecane and oligo(ethylene glycol) based on $g(r)_{\mathrm{EO,O}}$ and $g(r)_{\mathrm{CH3,O}}$ is shown in Table 2.3. There is a significant difference in residence times between the zwitterions and the oligo(ethylene glycol) ($0.7-1.7$ ps versus 0.3 ps). Oligo(ethylene glycol) ethers do not retain water as well as the zwitterions. There is a reduction in dodecane water coordination similar to glycine mostly due to phase separation. The dodecane values also show there is a true difference between the methyls seen on trimethylglycine and those on an uncharged hydrophobic molecule.

The same techniques were applied to the carboxylate anion side, as seen in Table 2.4. The hydration of the carboxylate anion is in general stronger than that of the amines. There are about two hydrogen-bonds per oxygen and the water resides there for much longer. There is again a decrease in hydrogen-bonding number for the aggregating solutes. Trimethylglycine has a much larger residence time at dilute concentrations: 5.5 ± 0.7 ps compared with the 1.2 ± 0.16 and 4.21 ± 0.24 ps seen for glycine and dimethylglycine, respectively. This trend stands out from the typically similar results seen for the zwitterions. In order to better

species	N_w/N_s	r_{sh} [Å]	$N(r_{sh})$	τ_{res} [ps]
alkyl	250	5.7	14.64	0.58 ± 0.06
alkyl	50	5.7	6.25	0.378 ± 0.007
oeg	250	3.1	1.09	0.3 ± 0.2
oeg	50	3.1	1.00	0.26 ± 0.1

Table 2.3: A summary of local hydration properties for the non-zwitterion species. N_w/N_s is the number of waters per solute. It is inversely related to concentration. The dodecane properties are based on $g(r)_{\mathrm{C,O}}$, where the C is the methylene carbon and O is the water oxygen. The oligo(ethylene glycol) properties are based on the $g(r)_{\mathrm{EO,O}}$, where the EO is the ether oxygen and O is the water oxygen. r_{sh} is the first water coordination shell radius. $N(r_{sh})$ is the number of waters in the first coordination shell. τ_{res} is the residence time of waters in the first coordination shell.

	N_w/N_s	H-bond	τ_{res} in $[ps]$
gly	250	2.39	1.2 ± 0.16
gly	18	1.79	0.45 ± 0.05
dmg	250	2.35	4.21 ± 0.24
dmg	18	2.18	0.63 ± 0.03
tmg	250	2.44	5.5 ± 0.7
tmg	18	2.35	0.63 ± 0.03

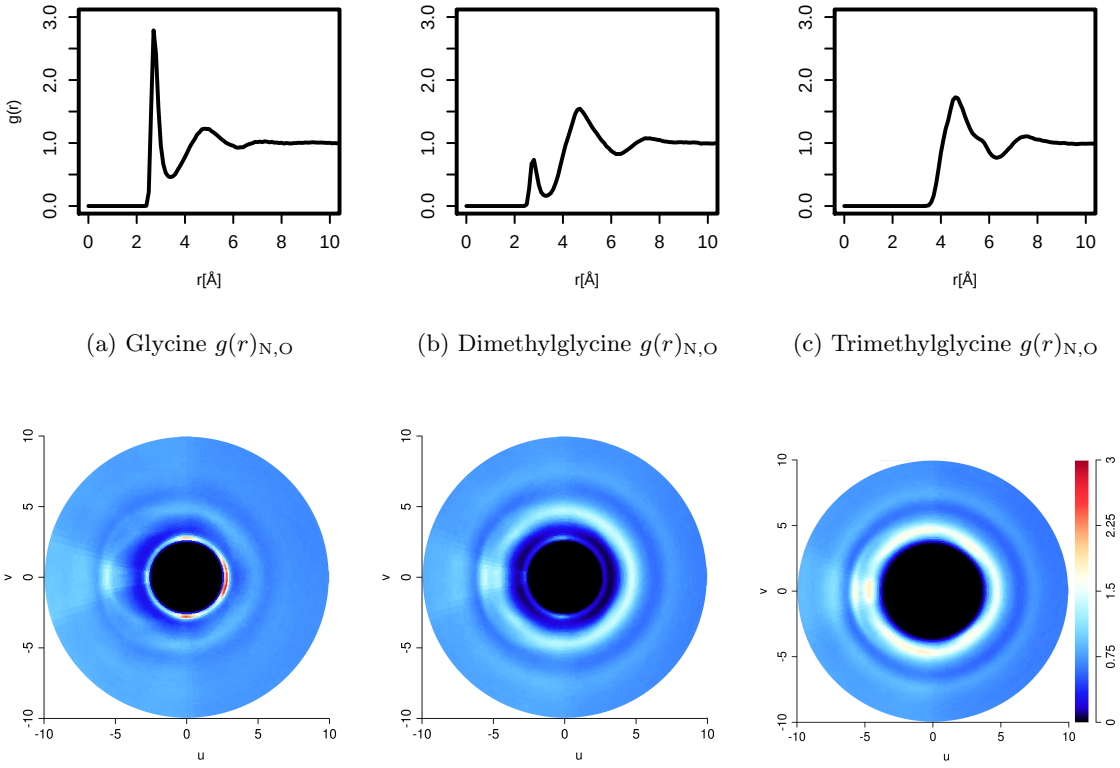
Table 2.4: Hydration properties for carboxylate anion. N_w/N_s is the number of water molecules per solute. It is inversely related to concentration. The hydrogen-bonds are water hydrogen-bonds per solute site, in this case per carboxylate anion oxygen. τ_{res} is the first water coordination shell residence time.

explain this result and the others, a two dimensional pair-pair correlation function was calculated, $g(u, v)$.

Beyond Spherical Symmetry

The two dimensional pair-pair correlation, $g(u, v)$, is based on polar spherical elements which are projected onto a plane for visualization. This creates a density distribution which is dependent on radial distance and orientation relative to the vectors that define the coordinate system.

The $g(u, v)$ results are seen in Figure 2.10. The plots show the zwitterions' $g(u, v)_{\mathrm{N,O}}$ looking down the molecule with nitrogen as the center. $g(u, v)_{\mathrm{N,O}}$ for glycine is shown in Figure 2.10d. The hydrogen-bonding sites are easily recognized as the red patches in the first shell. Also, the water hydrogen bound to the carboxylate anion is seen on the left of the plot. The dimethylglycine $g(u, v)_{\mathrm{N,O}}$ is shown in Figure 2.10e. The single hydrogen donor is seen in the tiny inner shell and the second shell is larger than the glycine second shell. The trimethylglycine $g(u, v)_{\mathrm{N,O}}$ is shown in Figure 2.10f. The second shell has a significantly larger peak than both dimethylglycine and glycine which results in larger coordination numbers. The advantage of this method is seen at the point (-5,0) where the trimethylglycine's amine second coordination shell aligns with the carboxylate anion coordination shell. This adjacency of the coordination shells gives the trimethylglycine a single coordination shell that extends all the way across the molecule. This unified coordination shell is similar to a combined hydration shell seen in ion association. The unified coordination shell of trimethylglycine is seen to a lesser extent in dimethylglycine. Near opposite behavior is observed with glycine which has unaligned coordination shells. Notice how the $(-5, 0)$ is far from the hydrogen bonding peaks in the glycine plot. The positive and negative functional groups are too near to form independent shells and the coordination shells of water for those two functional groups overlap. The lack of alignment of coordination shells between the amine and carboxylate anion causes the water near the interface of the positive and negative groups to be weakly bound.

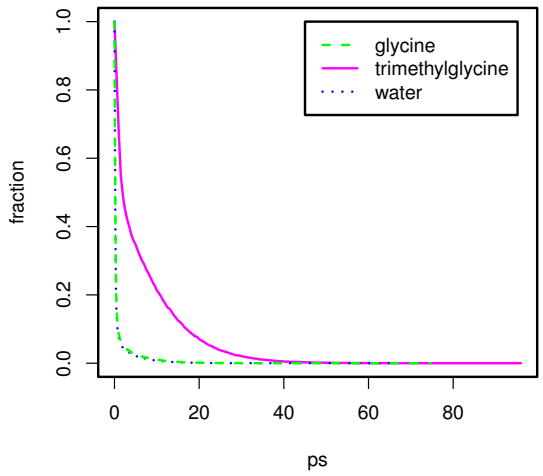


(a) Glycine $g(r)_{\mathrm{N,O}}$ (b) Dimethylglycine $g(r)_{\mathrm{N,O}}$ (c) Trimethylglycine $g(r)_{\mathrm{N,O}}$

(d) Glycine $g(u,v)_{\mathrm{N,O}}$ (e) Dimethylglycine $g(u,v)_{\mathrm{N,O}}$ (f) Trimethylglycine $g(u,v)_{\mathrm{N,O}}$

Figure 2.10: Amine water pair-pair correlation functions. $g(r)_{\mathrm{N,O}}$ is the amine nitrogen water oxygen radial distribution function. $g(u,v)_{\mathrm{N,O}}$ is a spherical two dimensional pair-pair correlation. The color gradient goes from black to white to red. Note the red spots on the glycine $g(u,v)$ which are the hydrogen bonding sites. Also, notice the trend to a more diffuse hydration shell as the number of methyl groups is increased. See the text for more details. The radial distances in $g(u,v)$ correspond to those in $g(r)$.

Figure 2.11: Hydrogen-bonding autocorrelation function around $\alpha-$carbon. Only hydrogen bonds within the methyl hydration shell of 4.7Å. The glycine water is similar to bulk whereas the trimethylglycine water is different than bulk and resides for longer times. Thus, the trimethylglycine $\alpha-$carbon water is bound.



The water between the amine and carboxylate anion is near the middle α-carbon. Figure 2.11 shows the hydrogen-bonding lifetime of this water. "Near" is defined as water within the methyl coordination shell of 4.7Å. The water near the trimethylglycine α-carbon has longer hydrogen-bonding lifetimes as result of the aligned coordination shells. Near the glycine α-carbon, water has hydrogen-bonding lifetimes similar to bulk water and thus is not considered to be hydrating. A monodisperse Voronoi tessellation was calculated on the water near the α-carbon and Voronoi volumes of water molecules were calculated. This gives an estimate of the compression of the water molecules. There is a statistical difference between the glycine and trimethylglycine, with the trimethylglycine waters being more compressed. The glycine waters approach bulk water volumes more quickly, thus showing that there is less hydration around the molecule. The Voronoi plot can be seen in Figure 2.12. Thus, the lack of alignment affects both the structure and the dynamics of the molecular hydration of glycine.

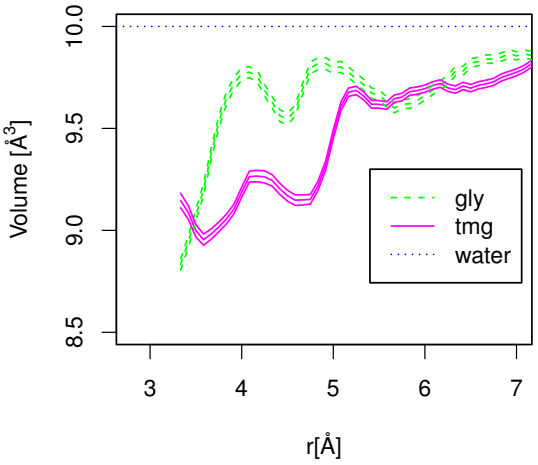


Figure 2.12: Voronoi tessellation volume of water molecules as a function of distance from the central $\alpha-$carbon's hydrogens on glycine and trimethylglycine. The lines represent the mean and confidence intervals. The water line is bulk water volume. The trimethylglycine waters are more compressed and this points towards more hydration around trimethylglycines.

2.2.4 Conclusions

Molecular dynamics simulations have been used to characterize the hydration of five important molecules relevant to protein aggregation and biointerfaces. Both structuring and dynamics of bulk and bound water were examined using a variety of properties and multiple simulations. This enables a holistic view of hydration and provides a method to predict the behavior of these molecules in more complicated systems where hydration is key, especially nonfouling.

Trimethylglycine was found to have the most bound water, slow bulk water diffusion the most, and remain well solvated even at high concentrations. Glycine was found to form dimers and aggregate more than the other two zwitterions. Each of the three zwitterions was found to only affect the tetrahedral geometry of water within 1.5-2.0 coordination shells, with trimethylglycine structurally affecting water the least. Water near the quaternary amine on trimethylglycine resides in the first coordination shell briefly (0.13 ps) but is almost always near any trimethylglycine. Trimethylglycine was found to have a contiguous hydration shell that extends across the entire molecule. This was observed using two dimensional pair-pair correlation functions, hydrogen-bond lifetimes, and Voronoi tessellations. Conversely, glycine has overlapping hydration shells between the positive and negative sides of the molecule. This results in less bound water even at dilute concentrations.

2.3 Free Energy of Solvated Salt Bridges from Simulations and Experiments

The aggregation of glycine detracted from its hydration. It is clear from Section 2.2 that minimizing self-interactions of a nonfouling group is key to hydration. Glycine contained both hydrogen bond donors and acceptors, hence the strong self-interaction. This helps us better understand as well why pCB is such a successful nonfouling material: it contains only hydrogen bond acceptors thus minimizing self-interaction.

Returning to peptides, we see that it is possible to select positively and negatively charged amino acids. However, it is necessary minimize the interactions between the two amino acids which in turn maximizes the hydration. Inter- and intra-peptide interactions

always detract from the hydration. In this section I model the interaction among the charged amino acids in order to both find the minimal self-interaction and maximal hydration.

2.3.1 Introduction

The interactions between charged amino acids, glutamate (E), lysine (K), arginine (R) and aspartate (D) are crucial for protein-protein interactions, protein stability and protein function. Charged amino acids are the most common types on the surfaces of proteins as determined from a recent analysis of many structures from the protein data bank.[61] They are especially prevalent on proteins found in the cytoplasm, an aggregation-prone crowded environment,[62] and thermophilic organisms.[61,63] Charged amino acids are also found on the interior of molecular chaperones where they are thought to be responsible for the repair of misfolded proteins.[64] It has been hypothesized that K and E are important for stabilizing proteins in aggregation-prone environments.[61] The common occurrence of R on protein surfaces is less clear; however, it has been shown that R is present in protein functional sites more than three times as often as K.[65] What difference in chemical properties makes nature prefer R over K, despite both being positively charged? A similar phenomenon occurs with E and D; D is more common in functional sites of proteins yet it has similar properties and structure to E. Motivated by these questions, we describe here the molecular interactions of salt bridges between K, R, E and D.

Simulations have been done in the past on free amino acid salt bridging using umbrella sampling,[66] a technique which can quantify the free energy of binding along a particular reaction coordinate. However, these techniques do not allow complete freedom of motion. Another simulation system which has been studied is short peptides which have a designed salt bridge that stabilizes helix formation.[67,68] Such systems are easily compared with experiment through circular dichroism, infrared spectroscopy, or nuclear magnetic resonance experiments (NMR), although the conclusions are sensitive to the peptide sequence used. The most commonly used experimental technique is the double-mutagenesis experiment.[69,70] This technique quantifies the effect of the salt bridge on the folding free energy of a protein

and has been successfully used to elucidate the role of salt bridges in specific proteins.[69,71] However, few general conclusions have been reached about salt bridging through these experiments because they study salt bridges fixed by the geometry of a protein. In some cases, buried salt bridges destabilize folding free energies, perhaps by adding conformational flexibility to the unfolded state or from the desolvation penalty of forming the salt bridge.[72,73] In other cases they stabilize structures, providing anchors between different locations on a protein.[69]

In this work, we study the interaction of E, K, D, and R in explicit solvent, similar to the environment of surface salt bridges. Utilizing well-tempered metadynamics,[74] a free energy measurement technique, it is possible to quantitatively rank the binding of the four possible salt bridges (K+E, R+E, K+D, R+D). We also make use of replica-exchange[75,76] to ensure sufficient sampling for this technique, allowing the removal of fixed reaction coordinates required for umbrella sampling simulations. Well-tempered metadynamics allows us to also provide quantitative measurements of hydration, similar to the recently developed indirect umbrella sampling technique from Patel *et al.*[77] which also quantifies the free energy as a function of the number of coordinating water molecules. Finally, we have used NMR to test predictions from simulations and found agreement in the ranking of strength of interactions. By combining these detailed atomistic simulations with experimental data, we can better address questions on salt bridging on protein surfaces, the role of charged amino acids, and the more general chemical phenomenon of ion pairing.

2.3.2 Theoretical Methods

Molecular dynamics simulations were used in order to answer questions about salt bridging and hydration of the four most commonly occurring charged amino acids: R, K, E and D. Two types of molecular dynamics simulations were conducted. In order to study hydration, free single amino acids were simulated in water. In order to study salt bridging, two free single amino acids of opposite charge were simulated in large simulation boxes. The large box size is important to allow binding and unbinding of the two amino acids. The interatomic

forces for the simulations were calculated using the OPLS-AA force field[38] and TIP4P-Ew water model.[41] The OPLS-AA force field is parametrized to replicate properties of small organic molecules and was chosen over traditional biophysical force fields which are generally parametrized for large peptides or proteins. The simulations were conducted in the NVT ensemble using the stochastic Bussi-Donadio-Parrinello NVT thermostat.[78] Replica exchange was used to enhance the sampling of many configurations.[75,76] 150 replicas were simulated with each replica simulating the system at a different temperature. This allows the replica of interest, the 300 K temperature simulation, to explore phase space not only through stochastic molecular dynamics, but also by exchanging configurations with other replicas. Finally, in order to calculate free energy we utilized the well-tempered metadynamics technique,[74] which also enhances sampling. This technique creates a "map" of where a simulation has been before and pushes the simulation towards unexplored regions. Keeping track of where the simulation must be pushed allows the algorithm to recover the free energy surface despite the biasing of the simulation.

The N- and C-termini of the amino acids were taken to be neutral (NH_2, COOH) to eliminate non-sidechain salt bridging. This convention was used in the four water/amino acid hydration simulations to make comparisons consistent. The four simulations studying the interactions between water and amino acids were conducted with single amino acids in 30 Å cubes. These will be referred to as the hydration simulations. The simulations studying salt bridging were done in 45 Å cubes containing two free amino acids: one anionic acid and one cationic base. These will be referred to as the salt bridging simulations. The systems were prepared with energy minimization, 200 ps annealing (100-500 K), and 20 ns equilibration molecular dynamics in Parrinello-Rahman NPT[79] at 300 K. For the hydration simulations, five independent well-tempered metadynamics NVT simulations (different water configurations) were conducted for 45 ns each amino acid (20 total simulations). Replica-exchange was not used for the water/amino acid simulations. The salt bridging simulations used 150 replicas with temperatures from 300K to 550K for 20 ns with exchanges attempted every

100 fs. The temperatures were distributed according to:

$$T_i = T_c \left(\frac{T_h}{T_c} \right)^{i/N} \tag{2.7}$$

where T_i is the ith replica temperature, T_c is the coldest replica temperature, i is the replica index (starting from 0), T_h is the hottest replica, and N is the number of replicas. The resulting replica exchange frequency was between 68-75% for the salt bridging simulations.

The stochastic Bussi-Donadio-Parrinello NVT thermostat (τ = 0.5 ps) was used.[78] Particle-mesh Ewald summation was used for the long range coulombic force calculations[47] and a shifted, truncated Van der Waals potential was used. The cutoff was 8 Å for interatomic forces. All covalent hydrogen bonds were constrained using the LINCS algorithm.[80] The equivalent concentration of 100 mM sodium and chloride ions was added to all systems and to ensure the system is neutral. Simulations were conducted in the GROMACS simulation engine.[81] The PLUMED plug-in was used for implementation of the well-tempered metadynamics algorithm.[82] Solvent density plots were produced using VMD.[48]

Well-tempered metadynamics is an enhanced sampling technique which also allows calculation of free energies. Details of the method are described elsewhere.[74] The difficulty in applying well-tempered metadynamics is choosing collective variables (CVs) which describe the physical phenomena of interest. The output of the well-tempered metadynamics is the free energy as a function of the CVs of the system. In the hydration simulations, the CV was the coordination number of water with the acid or base group as implemented in PLUMED.[82] The coordination number is continuous in order for derivatives of it to be taken, which are necessary for the metadynamics algorithm. See Table S1 for the parameters. The coordination of salt ions was monitored as well and when states with ion-amino acid coordination were excluded, no change was found in the free energies. In the salt bridge simulations, three CVs were biased: the coordination number between the negative and positive functional groups, the distance between the negative and positive functional groups, and the orientation of the negative and positive groups. The metadynamics parameters σ for these CVs are: 0.25, 1, and 0.1 rad. See Table S2 for the parameters. Additionally, interactions between the N-termini, C-termini, and side chains of the amino acids were

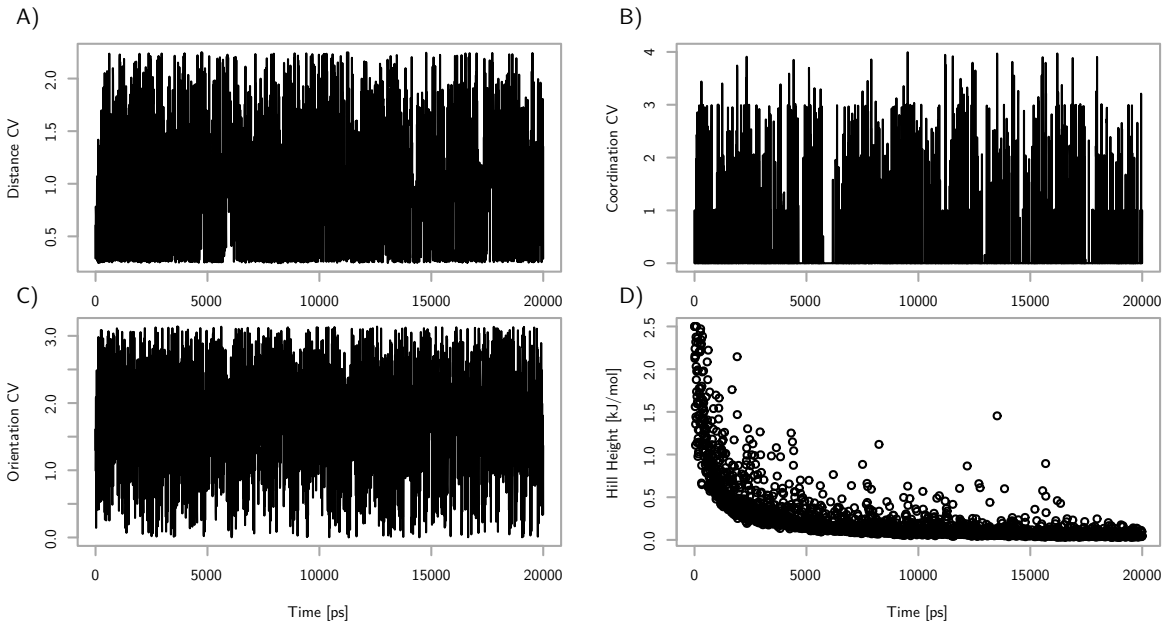


Figure 2.13: The value of the collective variables and metadynamics hill height as a function of simulation time for the R+E salt bridge simulation. Figures A-C show that the CV space is quite well explored. Figure D shows that the amount of bias added to the simulation has decayed to a small amount by the end of the simulation.

prevented through a harmonic restraint as implemented in PLUMED (κ = 50 kJ / mol, exp = 2). The PLUMED code removes the influence of that restraint on the free energy. The salt bridge simulation well-tempered metadynamics parameters were a bias factor of 5, an initial hill height of 2 kJ / mol and a hill deposition rate of 600 fs. The hydration simulations used a bias factor of 10. Convergence of the simulation may be addressed by examining the exploration of the CV space as shown in Figure 2.13. The decay in amount of bias added to the simulations further demonstrates convergence of the free energy values and is shown in Figure 2.13.

Experimental Materials and Methods

Single amino acids were synthesized to compare simulation results with NMR experiments. In simulations, the N- and C-termini may be uncharged to prevent coulombic interactions with the charged sidechains. However, such a titration state is not synthetically possible and the two termini were replaced with hydrophilic groups that may not participate in salt bridges to approximate this. These are an acetly group at the N-terminus and an amide group at the C-terminus; both uncharged hydrophilic functional groups.

N-Fluorenylmethoxycarbonyl (Fmoc)-protected amino acids with the amine and sidechain protected (Fmoc-Glu(OtBu)-OH, Fmoc-Asp(OtBu)-OH, Fmoc-Lys(Boc)-OH, Fmoc-Arg(Pbf)-OH), Rink amide AM resin, *O*-Benzotriazole-*N,N,N',N'*-tetramethyl-uronium-hexafluorophosphate (HBTU), *N*-Hydroxybenzotriazole (HOBt), and *N,N*-dimethylformamide (DMF) were bought from Aapptec (Louisville, KY). Trifluoroacetic acid (TFA), pyridine, and acetic anhydride were bought from EMD (Darrnstadt, Germany). Piperidine, dichloromethane (DCM), triisopropylsilane (TIPS), and phosphate buffered saline (PBS), were purchased from Sigma Aldrich (St. Louis, MO). Ethanol was purchased from Decon Labs, Inc. (King of Prussia, PA). Tris(2-carboxyethyl)phosphine hydrochloride (TCEP) was purchased from Pierce (Rockford, IL). AG 50W-X8 sulfonic acid cation exchange resin was purchased from Bio-Rad (Hercules, CA). Deuterated dimethyl sulfoxide (DMSO) and hydrogen dioxide were purchased from Cambridge Isotopes (Andover, MA).

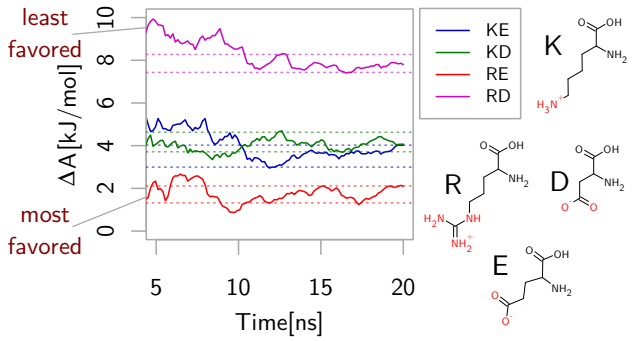


Figure 2.14: Free energy of forming a salt bridge as a function of simulation time to demonstrate convergences. The dashed lines indicate the final values as determined from a 95% confidence interval from the last 10 ns of simulation time. The structures of the four amino acids composing the salt bridges are shown on the right. The simulations were replica-exchange well-tempered metadynamics with 150 replicas.

The compounds were synthesized using the Aapptec Titan 357 automated synthesizer by a solid-phase technique, starting from Rink amide AM resin (0.58 mmol/g loading capacity). Coupling was performed using amino acid, HBTU, HOBt, and DIPEA prepared in DMF in a molar ratio of 1.1:1:1:2 in four times excess of the loading capacity of the resin. Deprotection of Fmoc groups was achieved using 20% piperidine in DMF. N-terminal acetylation was achieved with a solution of pyridine (5%), acetic anhydride (5%) and DMF (90%) (v/v/v). The cleavage of the final product was performed by a TFA (20%), DCM (70%), water (5%) and TIPS (5%) (v/v/v/v) cleavage cocktail. Am-R-Ac was purified through precipitation in cold ether and TFA. Am-K-Ac was purified via chloroform extraction. Am-D-Ac and Am-E-Ac were deprotonated on a sulfonic acid cation-exchange column in water followed by purification via cold ether precipitation in methanol. Synthesis was confirmed via NMR. All NMR was conducted on Bruker Avance AV 500 MHz NMR.

2.3.3 Results and Discussion

Figure 2.14 shows the free energies of binding for the four possible salt bridges over the course of the simulations. This plot was calculated by integrating the free energy difference between conformations where the coordination number was 0 (no salt bridge) and greater than 0.75 (salt bridge). The free energy is shown as a function of time to demonstrate convergence. The dashed lines represent 95% confidence intervals from the last 10 ns of simulation time. This free energy is not the well-depth of a PMF as in Masunov and Lazaridis[66], but the Boltzmann-averaged free energy difference beween the associated and dissociated state. This means our dissociated state includes the negative free energy of solvation that may be observed in the second free energy well of distance dependent PMFs. This is important for understanding the interplay between hydration and salt bridging. The binding free energy has a dependence on concentration because the salt bridging translational entropy change is linear in volume. That dependence vanishes, however, when looking at differences between salt bridging free energies.

All the salt bridges considered have a positive free energy, indicating that salt bridging is disfavored when the amino acids are completely solvated at this concentration (20 mM). This is consistent with the NMR results in water at a similar concentration (30mM). Additionally, the lowest free energy salt bridge is the R+E pair, which has been noted in Masunov and Lazaridis[66]. The K+E and K+D salt bridges are the next lowest in free energy and are indistinguishable. Finally, the R+D pair is the highest in free energy, which is puzzling due to the small structural differences between the R+E and R+D pair. The only difference is an additional methylene group on E between the acid side chain and amide backbone, as shown in Figure 2.14. The R+D simulation was extended for an additional 20 ns to confirm its convergence and a change of less than 0.5 kJ / mol was observed. There is no difference in the partial charges of the OPLS-AA force field between the acid groups and thus there is not a change in electrostatic interactions. As described below, the conformations that contribute most to the salt bridging involve the acid group (D / E) being in plane with the guanidinium group of R. The reduction of one methylene, and thus one torsion, reduces

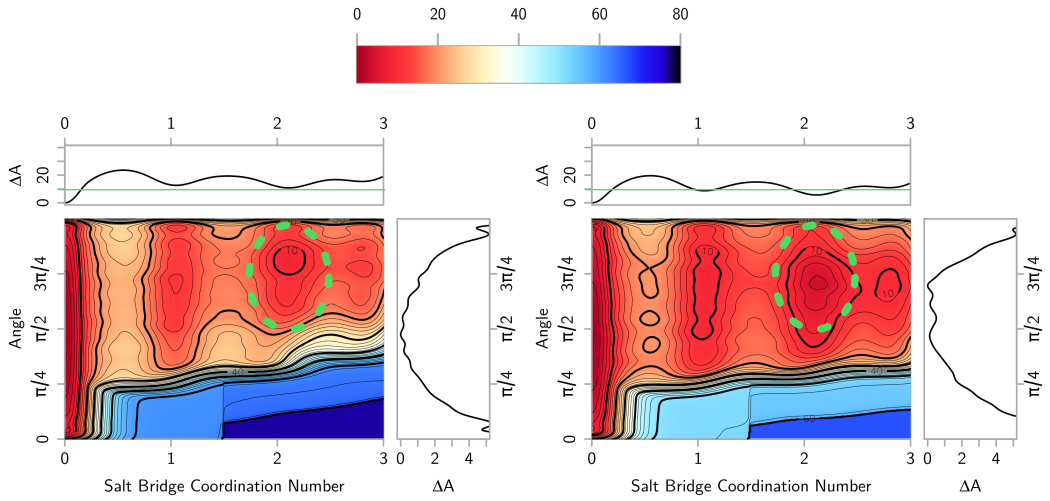


Figure 2.15: The uppermost one-dimensional plots shows free energy as a function of coordination number between the two amino acids. The coordination number is the sum of all coordinating pairs. The heatmap plot shows the free energy as a function of both. The green annotations show the decrease in free energy, and orientations (y-axis), of R+D relative to R+E in the double coordinating salt bridge.

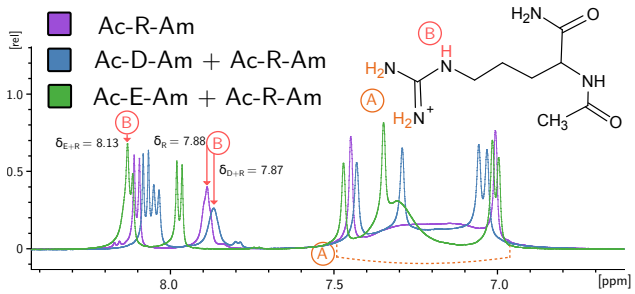


Figure 2.16: ^{1}H NMR spectra (500 MHz) for Ac-R-Am and 1:1 Ac-R-Am + Ac-E-Am or 1:1 Ac-R-Am + Ac-D-Am in DMSO. Upfield peak shifts indicate hydrogen bonding. The hydrogens labeled A see a significant change in peak shape when Ac-E-Am is added, indicating hydrogen bonding. This is not seen with Ac-D-Am, correlating with what is observed in simulations. The hydrogen labeled B shows a more discernible upfield shift, which has its peak shift annotated for the different conditions. Ac-E-Am + Ac-R-Am form stronger salt bridges than Ac-D-Am + Ac-R-Am.

the number of available salt bridge orientations of D when the acid group is in plane with the R group, explaining the significant increase in the free energy of the R+D pair relative to the R+E pair. This may be observed in the contraction of the projection of the free energy surface onto the salt bridge orientation in R+D relative to R+E seen in Figure 2.15. Such a large change from the removal of one methylene group has been previously observed in the large difference of solubilities between glutamine and asparagine.[83] Pair–pair radial distributions between the $\beta-$ and $\delta-$ carbons of D/E and R from unbiased simulations (no metadynamics) are similar, ruling out aliphatic interactions from the additional methylene being responsible for the large free energy difference.

The simulations show that R has the ability to discriminate between the two acids E and D, whereas K cannot. This is intriguing due to the significant amounts of R in the active sites of proteins as compared to K.[65] R appears to be suited for specific interactions whereas K appears to be suited for weak nonspecific interactions. The free energy difference between

K+D and R+D binding is 3.6 ± 1.2 kJ / mol, so a similar, though tenuous, argument can be made for aspartic acid which is also more common in active sites.[65] In order to confirm these results, R, K, D and E amino acids were synthesized. The goal of the experiments are to ascertain if one methylene difference (D vs E) can result in an observable experimental difference and corroborate the specific R+E salt bridge structure seen in simulations. In simulations, the N- and C-termini may be uncharged to prevent interference with the salt bridging simulations. However, such a titration state is impossible experimentally and thus the two termini were replaced with hydrophilic groups that may not participate in salt bridges. The structure of the backbone is shown in R in Figure 2.16. As discussed in the Supporting Information, the most stable salt bridges were formed on the secondary amine of R and the chemical shift of that proton can be monitored using ^{1}H NMR in DMSO. That peak is labeled as B in Figure 2.16 and the overlapping peaks were resolved via ^{1}H-^{1}H 2D-COSY90 NMR.[84] The B peak is sharp and it is possible to quantify its shift. The primary amine protons appear as a broad peak and are labeled as A. Changes in that peak are not quantitative and only serve to corroborate the presence of salt bridging. The pure 30mM Ac-R-Am spectrum is shown in purple and equimolar solutions of Ac-R-Am + Ac-D-Am and Ac-R-Am + Ac-E-Am are shown in blue and green, respectively. Upon the addition Ac-D-Am there is no change in either peak A or B, indicating little change in the interactions of the protons. Upon the addition of Ac-E-Am there is an upfield shift in both peaks A and B, indicating the presence of hydrogen bonding.[85] Peak B shows a shift of 0.25 ppm for Ac-R-Am + Ac-E-Am, compared with a shift of -0.01 ppm for Ac-R-Am + Ac-D-Am. The sharpening and upfield shift of peak A corroborates the presence of strong interactions between Ac-E-Am and Ac-R-Am. The large difference in spectra between the R+E and R+D solutions confirm the large difference in interactions between the salt bridge types.

The metadynamics free energies for the salt bridge simulations may be plotted as a function of the coordination number between the two amino acids and the orientation of the salt bridge. This is shown in the potential mean force (PMF) plot in Figure 2.17. The free energies shown are for the R+E and K+E salt bridges in Figures 2.17a and 2.17b,

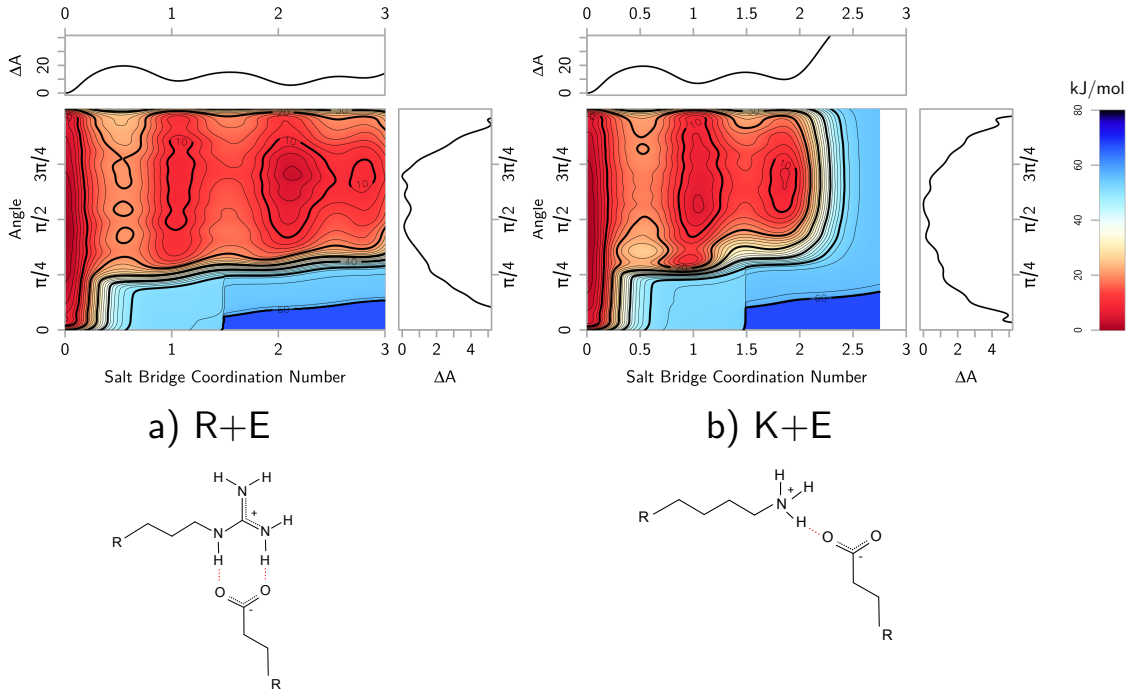


Figure 2.17: The potential mean force for the (a) R+E and (b) K+E salt bridge simulations. The uppermost one-dimensional plots shows free energy as a function of coordination number between the two amino acids. The coordination number is the sum of all coordinating pairs. The heatmap plot shows the free energy as a function of both. The bold lines are in increments of 10 kJ / mol and the thin lines are in increments of 2 kJ / mol. Both salt bridges are unfavored relative to being completely solvated. R prefers two hydrogen bonds and its lowest free energy conformation (E-NE,NH2) with a salt bridge is shown in the structure. K prefers single hydrogen bonds while salt bridging and this structure is shown on the right.

respectively. The topmost plots showing free energy as a function of coordination number show that the lowest free energy state is always non-salt bridging, which is also shown in Figure 2.14 where the free energies of salt bridging are positive. The topmost plot in Figure 2.17a shows that R has another free energy minimum at a coordination number of two (two hydrogen bonds) as opposed to K which has its salt bridge free energy minimum at one in Figure 2.17b. Comparing the minimums in the heatmaps, we see that R+E has a deep minimum (4 kJ/mol) with two hydrogen bonds whereas K+E has its minimum at one hydrogen bond (6 kJ/mol). R+E and R+D salt bridges generally occur as planar double hydrogen bonds and K+E and K+D salt bridges occur with freely rotating single hydrogen bonds.

The free energy profiles of the free amino acid hydration simulations are shown in Figure 2.18. The minimum free energy for K is at three water molecules, five for R and six for D and E. The relative rankings of K and R follow water/octanol solvation free energy experiments.[86] This may be seen by comparing the free energy of removing all water molecules from the two amino acids, which is highest for R, with the solvation free energy values from literature. However, something which is not observable in experiments is the partial desolvation of the amino acids. The partial desolvation of R is easier than K. The free energy difference to remove the first two water molecules from R is the same as removing a single water molecule from K. Thus, K may be considered to bind water more tightly than R. The two acids, D and E, are nearly identical in their hydration, although E appears to have a slightly higher free energy of solvation compared to D which is confirmed in experiments as well.[86] This is in contrast to their binding with R, where E interacts much more strongly than D.

The geometry of the interactions between the salt bridges are shown visually in Figure 2.19. The density of the side chain oxygens of E is shown in the top two panels. The preference of the secondary amine salt bridge may be seen clearly on the R+E salt bridge. The K+E is seen to have no particular orientational preference, due to the rotational symmetry of the primary amine. The water density when there is a salt bridge is shown in the

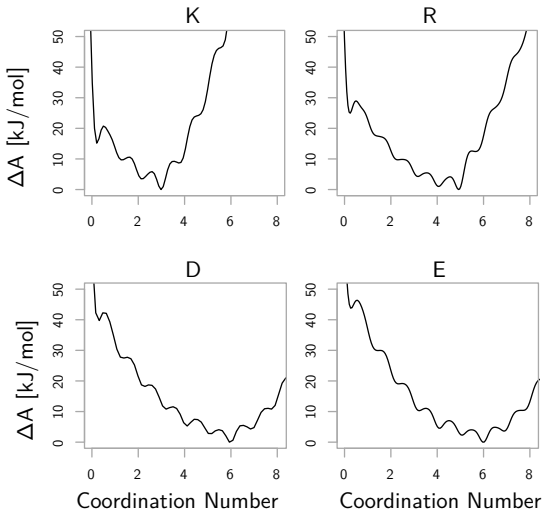


Figure 2.18: Free energy as a function of coordination number of water. K has a lower free energy of complete desolvation (coordination number of 0), as compared with R. However, removing a single water from K has a higher free energy than removing one from R. Thus, R is easier to partially desolvate as compared with K. The acids have similar free energy profiles, although D has a lower free energy of complete desolvation relative to E. D and E bind more water than R and K.

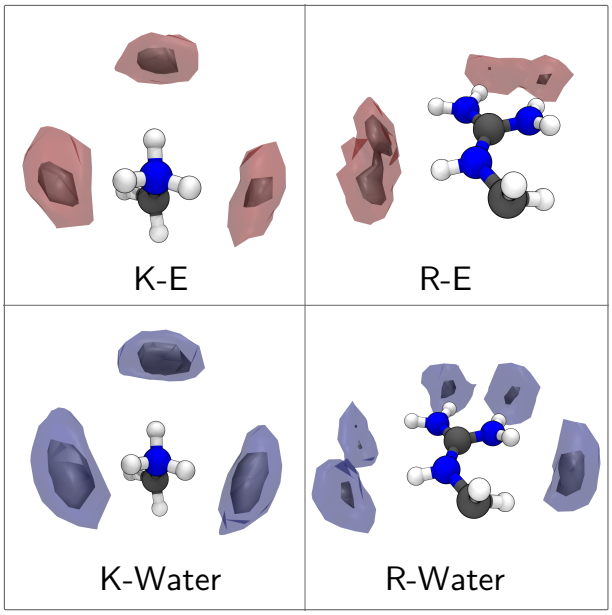


Figure 2.19: Solvent density and pair density for salt bridges. The top panel shows the density of the oxygen atoms of E when R or K is salt bridging with E. These densities are found by aligning salt bridge conformations observed in the simulation and histogramming the location of the oxygens, with metadynamics reweighting according to Bonomi *et al.*[87]. There is a preferred salt bridging orientation with R with the hydrogen of the secondary amine. K has no preference in orientation. The bottom two panels show the oxygen density of water when there is a salt bridge. The water is removed equally from the K, whereas water is removed from the secondary amine in R.

bottom two panels of Figure 2.19. There again, one may observe the difference between R and K. R-water has a higher density of water on the primary amines; the opposite preference compared to the salt bridge in R+E. This is because the water is excluded from that site while R and E are salt bridging. K-water is symmetric. The free energy of the salt bridges may be further broken down into the individual conformations through the reweighting technique described in Bonomi *et al.*[87]. If the salt bridges between R+E are divided into the five possible coordinating sites on R (E-NH1; E-NH2; E-NE; E-NE,NH2; and E-NH1,NH2) it is found that the E-NE,NH2 salt bridge is lower by 8 kJ/mol from the other conformations. This is the conformation shown in Figure 2.17. Thus, given that there is a salt bridge between R and E, 75% of the conformations will be the E-NE,NH2 conformation according to the Boltzmann distribution. Two replica exchange simulations without well-tempered metadynamics were conducted as well, to ensure the difference was not an artifact of the free energy technique.

2.3.4 Conclusions

The most common amino acids on the surface of proteins are charged.[61] Due to their prevalence, it is important to understand their interactions with one another. In this work, we have used state-of-the art molecular dynamics techniques to rank the interactions of the four most commonly occurring charged amino acids: R, D, E, and K. The free energy of salt bridging in increasing order (most favored to least), is: R+E, K+D, K+E, R+D. K is found to interact equally well with both D and E, whereas R can discriminate between the two acids. This result has been confirmed by NMR. R also has a specific orientation when salt bridging and favors two hydrogen bonds, whereas K has no preferred orientation and prefers single hydrogen bonds. D and E bind more water molecules than R and K. R liberates single water molecules more easily than K, but it requires more free energy to remove all water molecules from R. The partial desolvation, though inaccessible in experiment, is likely more relevant to understanding binding behavior; salt bridging does not require complete desolvation. The ability of R to discriminate between E and D and its strong geometric

dependence may help explain its prevalence in active sites of proteins compared with K.[65] Finally, the remarkable binding differences with R is one of the few distinguishing features of E and D.

2.4 Rational Design of Nonfouling Peptide

R+D has the weakest interactions among the charged amino acids. However, R can form specific strong salt bridges with proteins through the R+E salt bridging. This may in fact be experimentally observed, where R is found to interact more with proteins than E, K and D.[88] Further, R seems to be easier to partially desolvate than K. Thus, K appears to be a better positively charged group for self-assembling peptides. Amongst the two acids, E and D appear to pair equally well with K. D would be a better choice due to its weak interactions with any R on the surface of a protein. However, D can degrade through a deamidation reaction and due to its predicted near equal performance compared with E, E appears to be a more practical choice. Through a better understanding of nonfouling mechanism and salt bridging, we've narrowed down our design of nonfouling peptides to K+E or K+D.

A nonfouling peptide SAM consists of three ingredients: a nonfouling moiety, a covalent bond to attach the peptide and a functional group which encourages regular packing. The nonfouling moiety will be KE. Our experimental detection is done using surface plasmon resonance, which is easily accomplished on gold. Cysteine (C) is an amino acid which contains an -SH group and can form gold-thiol bonds.[89] Thus, we use C in our peptide SAM design for surface attachment.

The choice of a self-assembling group is less clear. Proline (P) was selected for three reasons. First, P is hydrophobic, which is essential for creating a strong self-assembling force between peptides. Second, P is rigid amino acid. This reduces the conformational entropy loss for adsorption of the peptide, encouraging adsorption of the peptides.[90] The rigidity also should promote a more regular structure. Finally, P inhibits α-helix formation which allows for extended chains.[91] Extended chains allows for better packing and higher density.

2.4.1 Methods

In order to test the hypothesized design, first the number of P residues must be determined. The study of addition of P to the peptides was done using the Rosetta protein structure prediction package.[92] It has had excellent performance in the past with *de novo* structure prediction.[92] The goal of the simulation was not to determine the structure of the peptides, but to instead see if the peptides have a more extended conformation. The peptides were first built to have their ideal rotamers with ideal α-helix rotamers and ideal β-sheet (non-P residues only), thus giving two starting structures for each sequence. Next, their bonds, angles, and dihedrals were optimized according to the Rosetta scoring function using the "idealize" module of Rosetta. Then, a Monte Carlo optimization of the structures was done for 400,000 steps. The Monte Carlo moves are backbone dihedral changes. The top 12 structures for the α-helix and β-sheet were chosen for continuation, resulting in 24 structures for each sequence. Next, the "relax" module of Rosetta was used to minimize side-chain clashes by changing χ-rotamers of the sidechains of the top 24 structures. This was followed by another backbone Monte Carlo optimization on each of the 24 structures for 300,000 steps and the structures were again relaxed. Finally, the 10 structures with the lowest scores among the 24 candidates were considered the result.

To test the proposed extended conformations of the P containing peptides, molecular dynamics simulations were conducted on three systems. The first sequence, Am-C-PPPP-EKEKEKE, contains P and expected to be most extended. The second sequence is Am-C-GGGG-EKEKEKE and is a flexible hydrophilic reference. The last sequence is the linker free Am-C-EKEKEKE. The peptides were built using the α-helical Dunbrack rotamers.[93] MD simulations were conducted using the GROMACS 4.5.3 simulation engine[81] and the AMBER99sb-ildn[94] force field. The peptides were energy minimized using 10,000 steps of steepest descent. The minimized systems were solvated in a box such that the peptides had 1 nm of solvent in all directions. Ions were added to charge equilibrate the system. Annealing was then done for 200 ps with a schedule from 100 K to 500 K in the NPT ensemble.[45] Finally, an equilibration for 100 ps again in the NPT ensemble was conducted.

Replica exchange with 100 replicas,[75] exchange attempts every 50 fs, and a time step of 2 fs in the NVT ensemble was used for the simulations. The temperature distribution was from 300 to 450 K according to the Equation (2.7).

Particle-mesh Ewald sums were used to treat electrostatics.[47] The van der Walls cutoff was 1 nm with an appropriate shifting function. The "v-rescale" thermostat, a stochastic thermostat not to be confused with velocity rescaling, was used.[78] The thermostat time constant τ was 0.5 ps. The simulations were run for 20 ns, and the C-α distances between the two terminal amino acids were measured.

Peptides were synthesized according to Nowinski *et al.*[95]. An SPR sensor was used to detect protein adsorption from fibrinogen and lysozyme, a negatively charged and positively model protein, respectively.[25] Peptides were self-assembled in PBS buffer for 24 hours at 0.2mg/ml. Protein adsorption was measured by flowing 1 mg/ml protein solution for 10 minutes over the peptide SAM followed by a 10 minute PBS buffer rinse. More detail on these experiments may be found in Nowinski *et al.*[95].

2.4.2 Results

The conformations of the 10 lowest energy conformations of the P containing peptides are shown in Figure 2.20. Their protein adsorption data is shown as well. There is a visible trend in both the conformational ensemble and the SPR protein adsorption data. Additional Ps increases the nonfouling performance of the peptides. The peptide with the most P, EKEKEKE-PPPP-C-Ac has nonfouling below the level for ultra-low-fouling,[96] below which blood will not clot on a material.

The results from the MD simulation are shown in Figure 2.21. It is clear that the P residues extend the conformation of the peptides. The glycine linker is also shown to be more flexible by having a wider distribution of end-to-end distances. These results match the nonfouling performance, where the P linker had 4.4 ± 2.9 ng/cm^2 fibrinogen adsorption and the G linker had 17.9 ± 11.4 ng/cm^2.

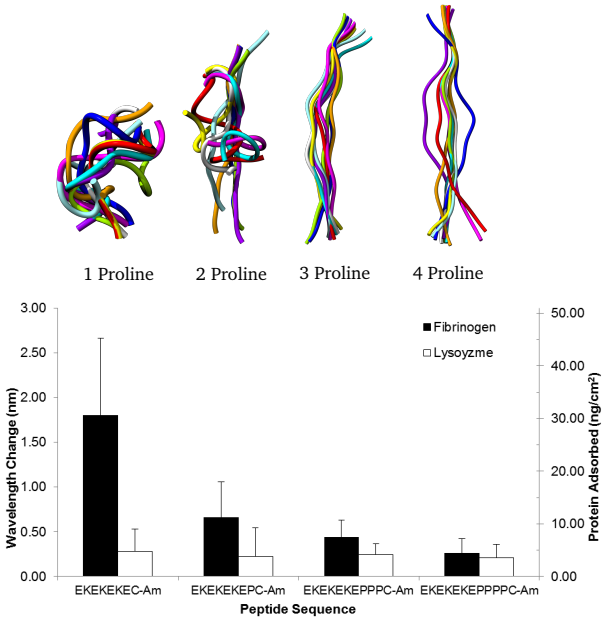


Figure 2.20: Effect of peptide linker length on protein adsorption. (Top) Ten lowest energy (score) conformations from 24 different starting structures as predicted using the Rosetta structure prediction package. They are aligned and oriented with the cysteine (C-terminus) at the bottom. The important feature to note is the increasing extended conformations as proline is added, due to the change from random coil to polyproline helix. The proline also makes the conformations more rigid, which reduces the entropic penalty of peptide adsorption onto a gold surface. Sequences from left to right are EKEKEKEPC-Am, EKEKEKEPPC-Am, EKEKEKEPPPC-Am, and EKEKEKEPPPPC-Am. (Bottom) Adsorption of protein on peptide SAMs composed of proline linker series EKEKEKEC-Am, EKEKEKEPC-Am, EKEKEKEPPPC-Am and EKEKEKEPPPPC- Am determined from a surface plasmon resonance sensor (SPR) in the unit of wavelength shift (nm) or converted surface concentration (ng/cm^2). (Black fibrinogen, White lysozyme). Each data point represents an average value $\pm$ standard deviation (SD) from at least three independent measurements.

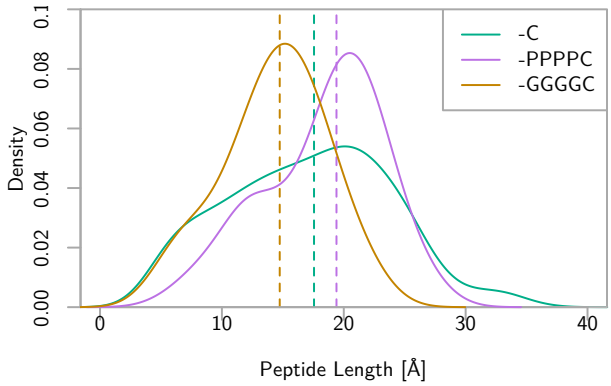


Figure 2.21: NVT MD simulation histogram of end-to-end distance of self-assembling peptides. -C is EKEKEKE-C-Ac, -PPPPC is EKEKEKE-PPPP-C-Ac, and -GGGGC is EKEKEKE-GGGG-C-Ac. The vertical lines indicate that the P sequence has the most extended conformation median.

2.4.3 Conclusions

Based on atomistic simulations, I've designed self-assembling nonfouling peptides. E and K can be mixed to create net neutral, zwitterionic, highly hydrated peptides which have minimal self-interactions. Adding P induces a more rigid and extended conformation in the peptides, creating strong self-assembly forces. If gold-attachment is desired, C may be added. Based on these principles, the sequence EKEKEKE-PPPP-C-Ac was designed and synthesized. Its nonfouling properties are excellent, reaching the ultra-low fouling standard necessary to prevent coagulation of blood.

2.5 Chapter Summary

Nonspecific interactions are difficult to model directly because they are, by definition, interactions between one compound with many others. In this chapter, we exploited the hydration of a material as a technique to indirectly study nonspecific interactions. If a material strongly binds to water, then it will resist interactions with other molecules. In

Section 2.2, I studied the effect of self-interactions and methylating amines on nonspecific interactions through the hydration of the molecules. In Section 2.3, I maximized the hydration of charged molecules by minimizing their interactions with one another. Additionally, by studying these four amino acids we saw a molecular example of specific binding where R is able to discriminate between the acids E and D. K has no such ability. These simulation results were accompanied by experiments wherever possible and the principles learned resulted in a self-assembling peptide that resists nonspecific interactions.

Chapter 3

NONSPECIFIC INTERACTIONS IN PROTEINS AND CHAPERONES

3.1 Nonspecific Interactions in Nature

In Chapter 2, I indirectly studied nonspecific interactions in simulations by studying hydration. In this chapter, nonspecific interactions are directly studied using bioinformatics. The protein data bank contains a wealth of structural data and this may be used to study nonspecific interactions among proteins. In Section 3.2, the adaption of proteins to resisting nonspecific interactions is characterized. This does not involve any modeling, but the conclusions are tested by experiments. In Section 3.3, I model a nonspecific biophysical process directly, protein stabilization, using a combination of bioinformatics and phenomenological modeling.

3.2 Decoding Nonspecific Interactions from Nature

Proteins resist nonspecific adsorption in order to be stable in complex environments such as the cytoplasm of a cell, which contains thousands of protein types.[62] The cytoplasm is a crowded environment and provides many spurious binding targets for proteins; yet proteins have evolved high selectivity and stability through resisting nonspecific interactions. The non-interacting property of proteins is often put into practice in biomaterials and biosensors research where proteins, bovine serum albumin for example, is used as a blocking agent to block nonspecific adsorption of non-target proteins onto surfaces.[97] Another example of nature resisting nonspecific interactions is found in molecular chaperones, which guide proteins from a misfolded or unfolded conformation back into a native conformation.[98] The defective (substrate) proteins fold while enclosed inside a cavity of the molecular chaperone. The chemistry of this cavity is unique among biological surfaces in that it contacts not only thousands of proteins, but many conformations of each protein.[99] Yet chaperone

proteins do not irreversibly bind with proteins. The cavity is sometimes described as a "non-stick" surface.[100] Thus molecular chaperones provide a second system which has this non-interacting property.

By examining many proteins from both these systems, it is possible to separate nonspecific effects from the many specific functions of proteins. We use two types of bioinformatics methods for this. The first studies the sequence and abundance of amino acids in the proteins, similar to the molecular formula of a molecule. The second set examines the structure and interactions among the amino acids in a protein and solvent, similar to the 3D structure of a molecule. Through these two methods and two systems, it is possible understand the way these proteins avoid nonspecific interactions. We analyzed a database of protein surfaces and molecular chaperone cavity surfaces using these two techniques. The questions to be answered are which amino acids are most common, how often do they interact, do they interact with water more than other amino acids, and do they prefer to interact with protein surfaces or interiors. Next, the modeling conclusions were used to design peptide based materials which should resist nonspecific interactions. Finally, these peptides were synthesized to test their resistance to nonspecific interactions with proteins. These peptides do create surfaces which resist nonspecific interactions and compare well to others low-fouling peptides which have been reported in the literature.[95,101–103]

3.2.1 Methods

Protein Dataset Construction

Three datasets of proteins with Protein Data Bank (PDB) structures were constructed. The first dataset consists of 1,162 unique human proteins. The second dataset is a subset of the first with 34 proteins and the additional criterion that the proteins are located in the extracellular space (GO ID:5615).[104] The last is again a subset of the human dataset with the additional criterion of being located in the cytoplasm (GO ID: 5737). It has 221 proteins. We ensure a diversity of structures by using a 40% homologue cut-off and all structures containing residue gaps are excluded. Further constraints are that the X-

ray resolution is $\leq$ 2.5Å, 'mutant' must not appear in title, no large ligands (e.g., DNA, RNA), and the only macromolecule in the structure is a protein. The molecular chaperones considered are all crystal structures in the *cis* or closed conformation. This is the conformation during which substrate proteins refold. The molecular chaperons considered are: GroEL-GroES isolated from *E. coli* ('*E. coli* GroEL'),[105] a GroEL-GroES complex from *Thermus thermophiles* ('Thermo GroEL-GroES'),[106] a group II chaperonin protein isolated from *Methanococcus maripaludis* ('Group II'),[107] a eukaryotic molecular chaperone isolated from yeast ('HSP90')[108] and a cytosolic chaperonin isolated from yeast ('CCT').[109]

Statistical analysis was performed using the R statistical package.[110] SQLShare was used for managing data.[111] The PDBs and data used for each dataset are available from: http://sqlshare.escience.washington.edu The human, cytoplasm, and extracellular datasets are available as SQL data tables under the 'h2', 'cph2', and 'eh2' tags, respectively. X_1, X_2, and X_3 (where X is the dataset) contain the protein information, residue information, and atomic information, respectively. X_c contains the surface contacts (only available for 'h2').

Surface Identification

All water was removed from the PDB structures before calculating the accessible surface area. Surface area was calculated using Accessible Surface Area.[112] A residue is classified as 'surface' when it is 30% or above its maximum surface area. Maximum surface area for a given residue is defined as the surface area occupied by the side chain atoms of a Gly-X-Gly peptide, with X being the residue of interest. The rotamers were taken to be α-helical and the lowest energy $\chi-$rotamers were used.

The identification of interior residues in molecular chaperones consists of three steps: (1) identify surface residues, (2) tabulate heavy atoms from the surface residues which are occluded, (3) identify which residues have more than h heavy atoms occluded. Once a residue is identified as a surface residue, it may be either an interior or an exterior surface residue. To be an interior surface residue, h heavy atoms (non-hydrogen atoms), or more,

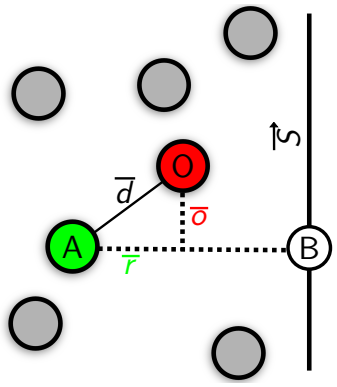


Figure 3.1: A graphic showing how to test if atom A is occluded by atom O. Atom A's point is projected onto S to obtain line $\overline{r}$. By the Pythagorean theorem, $\overline{o}$ may be found knowing $\overline{r}$ and $\overline{d}$, the distance from Atom A to atom O. If $\overline{o}$ is shorter than the occlusion margin, M, atom O is occluding atom A.

in the residue must not be occluded by atoms from other residues.

Occluded atoms are atoms which have another atom occluding its orthogonal vector to the principal axis of the protein ($\vec{S}$). The orthogonal vector is occluded when another atom, O, has an occlusion line segment $\overline{o}$ which is less than the occlusion margin, M, away from $\overline{r}$ (see Figure 3.1). $\overline{r}$ is given by the formula for the projection of a point onto a line:

$$P_B = \vec{S}\,\frac{P_A \cdot \vec{S}}{|S|}$$

$$\overline{r} = P_B - P_A,\ \overline{d} = P_A - P_O$$

Using the Pythagorean theorem, $\overline{o}$, can be found:

$$|\overline{o}|^2 = \left|\overline{d}\right|^2 - \frac{\left(\overline{d} \cdot \overline{r}\right)^2}{|\overline{r}|^2}$$

where the last term is from the projection of atom O onto $\overline{r}$. If $|\overline{o}| < M$, the atom is occluded. Only heavy atoms are tested and they are tested against only heavy atoms. After tabulating all occluded atoms, if a residue contains greater than or equal to h occluded heavy atoms, it is considered an exterior residue.

Short Name	PDB ID	Cutoff	h	M[Å]	$\vec{S}$[Å]	ϵ	r_{max} [Å]
GroEL Open	1SX4	0.3	2	2	$(0, 0, 1)$	$(-25, 35)$	40
GroEL-GroES	1SX4	0.3	2	2	$(0, 0, 1)$	$(-70, 35)$	60
Thermo GroEL Open	1WE3	0.3	2	1.8	$(-0.01, -0.68, 0.73)$	$(-15, 23)$	45
Thermo GroEL-GroES	1WE3	0.3	2	2.2	$(-0.01, -0.68, 0.73)$	$(-75, 30)$	60
Group II	3KFB	0.3	2	1.8	$(0.71, 0, 0.71)$	$(-35, 20)$	70
HSP90	2CG9	0.3	2	1.3	$(1, 0, 0)$	$(-30, 45)$	22
Yeast CCT	3P9D	0.3	3	1.8	$(0.83, 0.33, 0.45)$	$(-61, 61)$	∞

Table 3.1: The parameters used for identifying interior residues for the various chaperone proteins.

A few additional details are necessary as well. First, the occluding atom O must not lie on the other side of $\vec{S}$ from the atom being considered A. O must not be behind A ($\overline{d} \cdot \overline{r} > 0$). Residues whose projection onto $\vec{S}$ exceeds an extent, $\epsilon\vec{S}$, are considered occluded. Finally, residues which are farther away than r_{max} are considered occluded. A table containing the specific parameters used for each chaperone protein are shown in Table 3.1. The free parameters, M and h were chosen by visual inspection of the resulting interior. This may be seen in Figure 3.2.

Expected Number of Amino Acid Pairs

To estimate the number of residues which should be next to each other in sequence if the sequence is random, a background model was calculated. It is a multinomial model. The number of pairs can be calculated as:

$$E\left[N_{ij}\right] = N\hat{p}_i\hat{p}_j,\ i \neq j \tag{3.1}$$

$$E\left[N_{ii}\right] = \frac{1}{2}N\hat{p}_i^2 \tag{3.2}$$

$$\hat{p}_i = \frac{N_i}{N} \tag{3.3}$$

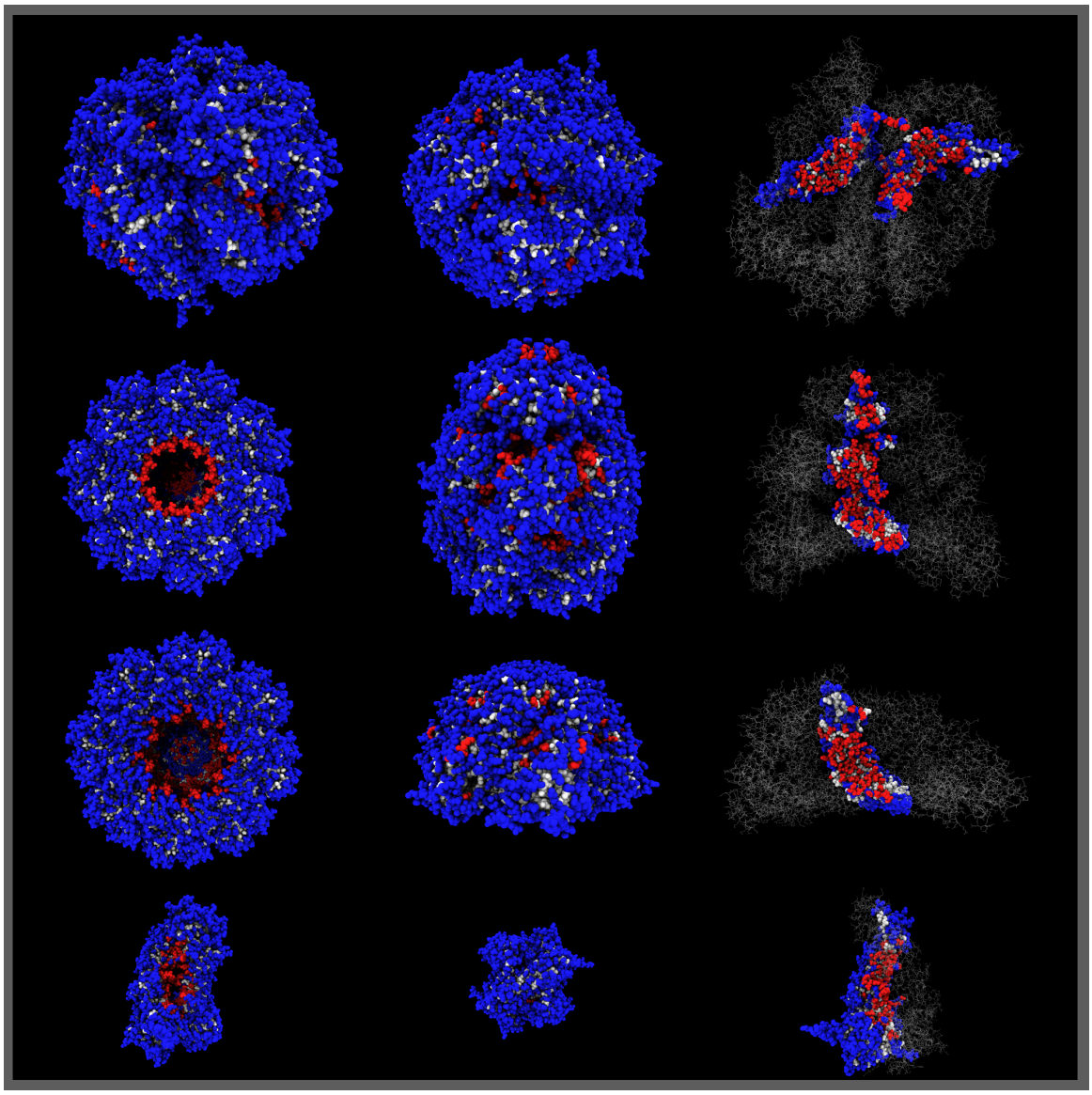


Figure 3.2: Visualization of the interiors of the molecular chaperones considered. Interior residues are in red and exterior residues are shown in blue. The top row is Yeast CCT, the second row is GroEL-GroES (same as Thermo GroEL-GroES), the third row is Group II, and the final row is HSP90. The resulting interiors demonstrate the effectiveness of the techniques described here for identification of the interior of molecular chaperones

N_{ij} is the number of sequence residue pairs between types i and j on the surface, N_i is the number of residues of type i on the surface, and N is the total number of residues on the surfaces. The uncertainty in this estimate, relying on a truncated Taylor expansion, is:

$$\sigma^2_{N_{ij}} = N^2\hat{p}_i^2\sigma^2_{\hat{p}_j} + N\hat{p}_j^2\sigma^2_{\hat{p}_i}\; i \neq j \tag{3.4}$$

$$\sigma^2_{N_{ii}} = N^2\hat{p}_i^2\sigma^2_{\hat{p}_i} \tag{3.5}$$

where $\sigma^2_{\hat{p}_i}$ is the sample variance with each protein treated as an observation. This estimate is reasonably valid if the number of categories is high, which in this case is true with 20 amino acids.

Structure and Interaction Equations

The proportion of amino acids which are interacting may also be calculated as:

$$\hat{p}_i = 1 - \frac{N_i^C}{N_i} \tag{3.6}$$

where N_i^C is the number of residues of type i that are in contact with at least one side-chain and N_i is the number of residues of type i observed on the surface. The preference for amino acids to interact with water is calculated as:

$$\hat{p}_i^w = \frac{N_{iw}}{N_{iw} + \sum_j N_{ij}} \tag{3.7}$$

where N_{iw} is the number of contacts between surface residues of type i and water and N_{ij} is the number of contacts between surfaces residues of type i and side-chains of residue type j. This statistic is slightly different from the propensity of a residue to be in contact with water, because the normalization is relative to all contacts. The energy of interaction between a single amino acid and a protein interior (buried) or surface is calculated according to:

$$E_i = \sum_j^{20} p_j\chi_{ij} \tag{3.8}$$

where p_j is the fraction of each residue present either on the protein surface (dark bar) or buried (light bar). The interaction energies, χ_{ij}, are described below.

Interaction Energy Definition

The interaction energies are derived from counts of residue contacts and total number of residues. The interaction energy is defined as:

$$\chi_{AB} \equiv U_{AB} - U_A - U_B$$

where U indicates energy and A and B are residue types. Now, substituting the Boltzmann distribution:

$$\chi_{AB} = -\frac{1}{\beta}\ln\left(e^{-\beta U_{AB}}\right) + \frac{1}{\beta}\ln\left(e^{-\beta U_A}\right) + \frac{1}{\beta}\ln\left(e^{-\beta U_B}\right), \quad \beta = kT$$

where k is Boltzmann's constant and T is the temperature. Replacing the Boltzmann distribution with the probabilities:

$$\chi_{AB} = -\frac{1}{\beta}\ln P_{AB} + \frac{1}{\beta}\ln P_A + \frac{1}{\beta}\ln P_B = -\frac{1}{\beta}\ln\frac{P_{AB}}{P_A P_B} \tag{3.9}$$

The probabilities may be found using maximum likelihood estimators[1] and shown to be:

$$\hat{P}_A = \frac{N_A}{\sum_i N_i}, \quad \hat{P}_{AB} = \frac{N_{AB}}{N_{\text{Free A}} + \sum_y N_{Ay}} \tag{3.10}$$

where N_A indicates the number of residues of type A, N_{AB} indicates the number of A, B pairs and 'Free' indicates unpaired A. The summation in the denominator is across all pairs where A is part of the pair.

Residue contacts were calculated by finding the pair-wise distance between each side-chain heavy atom on each residue. Neighboring residues, as determined by residue indices in PDB files, were excluded from being in contact. If any of the heavy-atom pairs were below the cutoff distance, the side-chain pair is said to be in contact. The heavy atom pair cutoffs were taken to be Van der Waals energy minimum radii. The following Van der Waal radii were used: nitrogen: 3.25Å, oxygen 2.96Å, and sulfur 3.55Å. The following mixing rule was applied:

$$r_{ij} = 2^{1/6}\sqrt{\sigma_i \sigma_j} \tag{3.11}$$

where σ is the Van der Waals radius.

Error Analysis

Figure 3.5 uses standard error for the error bars and the amino acid surface fractions errors reported in text were standard errors. Standard error is calculated as:

$$\sigma_i = \frac{1}{\sqrt{N}} \sqrt{\frac{\sum_j (p_{ij} - \hat{p}_j)^2}{N-1}} \tag{3.12}$$

where i is the residue type, j is the protein, p_{ij} is the fraction of residue type i on protein j, $\hat{p}_i$ is the average residue fraction of type i, and N is the number of protein. Notice that the standard error calculations were done by considering each residue fraction on each protein as a single observation. The choice of standard error means the error considered is uncertainty in the mean. This is different from standard deviation. This is because our design principles should be based on our materials being in contact with a population of proteins, not a single protein. Thus, uncertainty in the mean is most important. The use of these statistics assumes normality, which can be seen visually in Figure 3.3. As one can see, the higher residue fractions are normal and the assumption is valid. For the more rarely observed residues, for example cysteine, the assumption of normality is not correct and the error bars may not be exact in Figures 3.5 and 3.6. Note that this does not affect the values themselves.

A sensitivity analysis similar to bootstrap resampling was used for the error analysis called "bootstrap error," specifically in Figure 3.7, except for the error bars in Figure 3.7a. The sensitivity analysis was done by treating each protein as an independent observation. Pseudo-replicates of the data were created by sampling from the protein dataset with replacement. Each pseudo-replicate has the same number of proteins as the original dataset; there are repeats and omissions in the pseudo-replicates. The sampling was done 500 times for each statistic. Quantiles were calculated on the pseudo-replicates to obtain errors. 95% confidence intervals are shown as error bars in Figures 3.7b, 3.7c, and 3.7d.

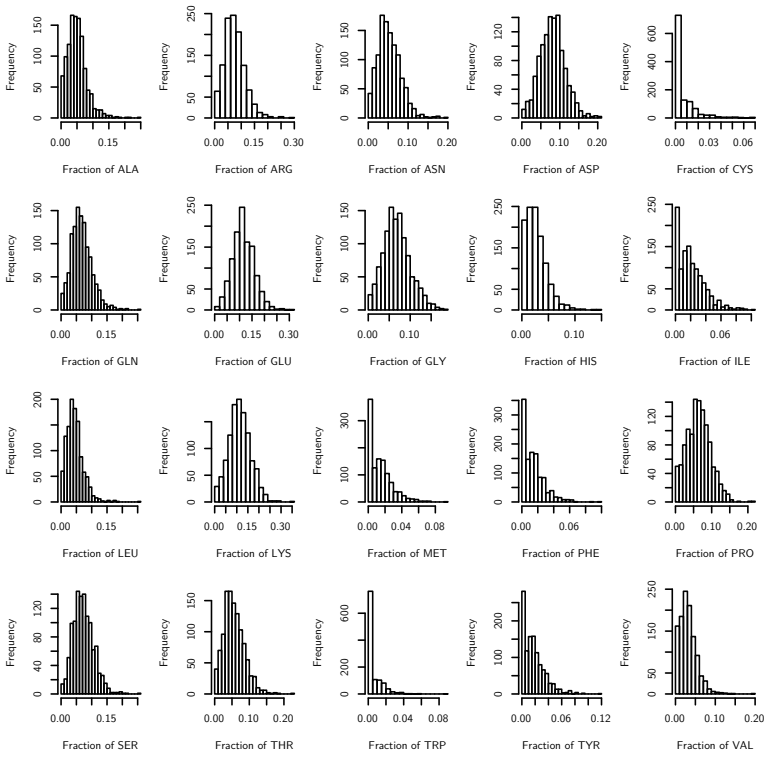


Figure 3.3: A histogram of the 20 amino acid surface fractions on protein surfaces. This plot shows that the assumption of normality is true for the more commonly observed residues.

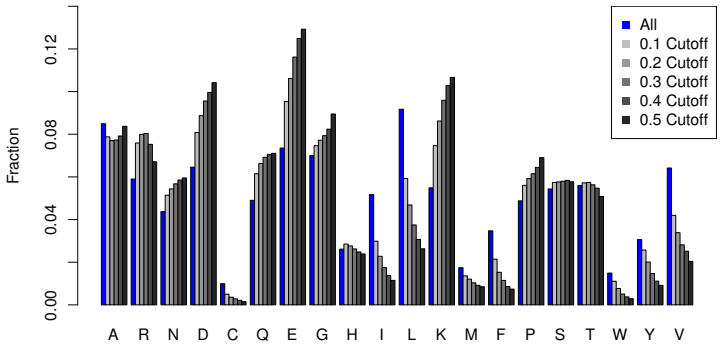


Figure 3.4: The *E. Coli* residue distribution as a function of surface cutoff. The total residue distribution (no cutoff) is shown in blue. As the cutoff changes, the residue fractions change slightly. There is a tend towards more hydrophilic residues as the cutoff increases, which is expected since hydrophilic residues are generally more surface exposed than hydrophobic residues. A cutoff of 0.3 was chosen for all calculations.

Sensitivity to Surface Cutoff

The effect of changing the surface cutoff is shown in Figure 3.4. The residue fractions are relatively stable to cutoffs, with lysine (K) and glutamic acid (E) being the most sensitive. The total change across the cutoffs considered here is 2% for the E and K fractions. The cutoff chosen in text was 0.3. Hydrophilic residues tend to increase as the cutoff is increased, which is expected since hydrophilic residues are generally more solvent exposed than hydrophobic residues.

Peptide Synthesis and Characterization

N-Fluorenylmethoxycarbonyl (Fmoc)-protected amino acids with the amine and sidechain protected (Fmoc-Glu(OtBu)-OH, Fmoc-Lys(Boc)-OH, Fmoc-Cys(Trt)-OH, Fmoc-Pro-OH,

Fmoc-Gly-OH), Rink amide AM resin, O-Benzotriazole-N,N,N',N'-tetramethyl-uronium-hexafluoro-phosphate (HBTU), N-Hydroxybenzotriazole (HOBt), and N,N-dimethylformamide (DMF) were bought from Aapptec (Louisville, KY). N,N-diisopropylethylamine (DIPEA) was bought from TCI America (Portland, OR). Trifluoroacetic acid (TFA), pyridine, and acetic anhydride were bought from EMD (Darrnstadt, Germany). Piperidine, dichloromethane (DCM), triisopropylsilane (TIS), 1,3-dimethoxybenzene(DMB), 1,2-ethanedithiol (EDT), 1,1,1,3,3,3-hexafluoro-2-propanol (HFIP), phosphate buffered saline (PBS), fibrinogen from bovine plasma, and lysozyme from chicken egg white were purchased from Sigma Aldrich (St. Louis, MO). Ethanol was purchased from Decon Labs, Inc. (King of Prussia, PA). Tris(2-carboxyethyl)phosphine hydrochloride (TCEP) was purchased from Pierce (Rockford, IL).

The peptides were synthesized using the Aapptec Titan 357 automated synthesizer by a solid-phase technique, starting from a polystyrene Rink amide AM resin (0.58 mmol/g loading capacity). Coupling was performed using amino acid monomer, HBTU, HOBt, and DIPEA prepared in DMF in a molar ratio of 1.1:1:1:2 in four times excess of the loading capacity of the resin. Deprotection of Fmoc groups was achieved using 20% piperidine in DMF. N-terminal acetylation was achieved with a solution of pyridine (5%), acetic anhydride (5%) and DMF (90%) (v/v/v). Random peptide sequences were created using the mix and split capability of the Aapptec Titan 357. The cleavage of the final product was performed by a TFA (75%), DCM (15%), DMB (4%), water (2%), TIS (2%), and EDT (2%) (v/v/v/v) cleavage cocktail. The peptide purity was evaluated by preparative reverse phase high pressure liquid chromatography (RP-HPLC) for known sequences and purified as needed. The purity of of the glycine peptide sequences was 92% and the asparagine peptide was 97%. Peptide were analyzed by matrix-assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF-MS).

Peptide Self-assembly and Protein Adsorption

Peptide self-assembled monolayers were prepared as previously described.[101] Gold coated surface plasmon resonance (SPR) chips were cleaned by rinsing with Millipore water, ethanol, and then drying with filtered air. They were placed in the UV cleaner for 20 minutes. Once removed, the gold chips were rinsed again with Millipore water, ethanol, and dried by filtered air. The clean chips were incubated with a phosphate buffered saline (PBS) aqueous solution (pH 7.4 and Ionic Strength of 150 mM) of 0.2 mg/ml peptide for 24 hours. After incubation the chips were removed from solution, rinsed with Millipore water, and evaluated by SPR.

A laboratory SPR sensor developed at the Institute of Photonics and Electronics, Prague, Czech Republic was used[113] as described previously[25] to evaluate protein adsorption. Gold chips covered with peptide SAMs were rinsed with Millipore water, dried by filtered air, and mounted to the device. The temperature controller was set to 25 $\pm$ 0.01°C. Protein adsorption was measured by flowing PBS buffer at 40 μL/min for 10 minutes, 1 mg/mL protein solutions of fibrinogen and lysozyme for 10 minutes, and PBS buffer again for 10 minutes. The wavelength shift between baselines before protein injection and after buffer rinse was used to quantify the total amount of protein adsorbed. A reference channel containing solely PBS buffer was flown for each chip and its baseline drift was subtracted from the final wavelength change. A 1 nm wavelength shift from 750 nm corresponds to 17 ng/cm^2 adsorbed proteins.[114] The detection limit for the SPR sensor is 0.2 ng/cm^2.[25] For statistics reported in the paper, each chip corresponds to one data point for calculating standard deviations.

3.2.2 Results

Abundance and Sequence

After constructing a dataset of proteins found in humans, it is possible to tabulate the most abundant amino acids on protein surfaces. The average fractions of amino acids on

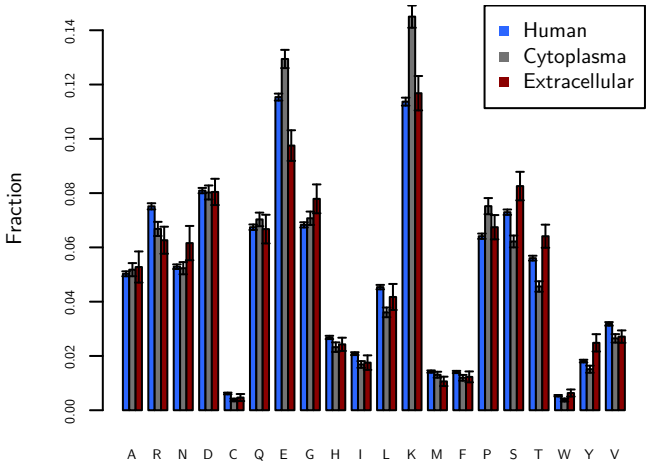


Figure 3.5: Fraction of amino acids on protein surfaces. The fraction of amino acids on the surface of proteins found in three different locations: human proteins ($N = 1,162$), human cytoplasmic proteins ($N = 221$), and human extracellular ($N = 34$) proteins. The y–axis is the median of the fractions of each amino acid over the entire dataset. The figure shows the large fraction of charged residues on protein surfaces, in particular E and K. The error bars are standard error.

protein surfaces found in humans are shown in Figure 3.5. The most striking observation is the large fractions of lysine (K) and glutamic acid (E). In fact, charged amino acids, including E, K, arginine (R), aspartic acid (D), and histidine (H), comprise $41\% \pm 0.3\%$ of the surface of proteins. Those amino acids comprise only 27% when considering both surface and non-surface amino acids. It is generally assumed that proteins have the most hydrophilic amino acids on their exteriors and the hydrophobic on the interior in order to maintain their native conformation. Thus the ordering of most common to least common amino acids on the surface may follow how hydrophilic the amino acids are. However, K and E do not have the lowest free energy of solvation;[86] hydrophilicity is not the determining property for surface fraction.

The protein dataset is further broken into the cytoplasm and extracellular. The extracellular environment, the space between cells, generally is not as crowded as the cytoplasm and we expect that nonspecific binding is not as interfering compared to the cytoplasm. As seen from our results, the largest difference between cytoplasmic and extracellular proteins is the larger fraction of charged amino acids: $43.8\% \pm 0.3\%$ for the cytoplasm and $37.5\% \pm 0.7\%$ for the extracellular. The extracellular dataset is lower in charged amino acids, but higher in polar hydrophilic amino acids ($27.3\% \pm 0.2\%$ vs. $23.7\% \pm 0.6\%$), among which specifically serine (S) and threonine (T) are more abundant than those in the cytoplasm. Polar hydrophilic amino acids are S, T, asparagine (N), and glutamine (Q). These results generally follow the trends seen by Andrade, *et al.*,[115] who examined a similar though smaller dataset. In the cytoplasm, a crowded environment prone to protein aggregation,[62] the proteins have more E and K. Thus, the K and E play an important role in these nonspecific interactions. Evidence that E and K are located in nonspecific regions of protein surfaces (unrelated to function) can be found in work from Jimenez as well, who analyzed protein surfaces broken into regions related to protein function and regions unrelated to function.[116] He showed an increase of 36% of charged amino acids in regions unrelated to function relative to regions related to function.

A similar abundance analysis on the interior of molecular chaperones abundance is shown

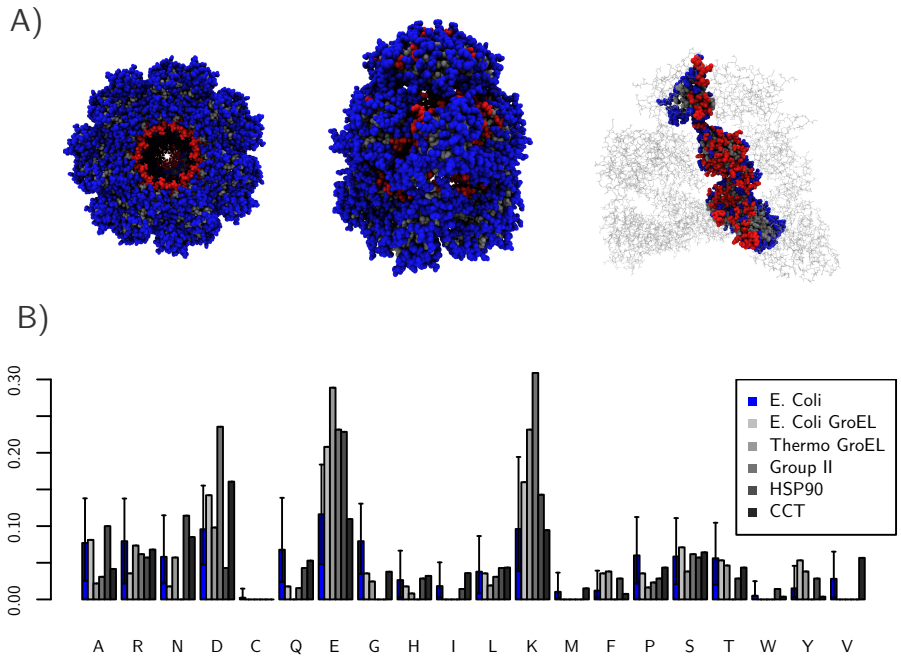


Figure 3.6: Fraction of amino acids on interior of chaperones. A) shows three views of location of interior residues (red) for the E. Coli GroEL. B) The fraction of each amino acid on the interior of 5 molecular chaperones. The median fraction of each amino acid type on the surface of a collection of 528 *E. coli* proteins is shown in blue for reference. All the chaperone structures used for these calculations were the *cis* or "closed" forms. The error bars come from a 95% confidence interval from quantiling. The fractions of E and K are the most different relative to *E. coli* proteins. The thermophilic GroEL-GroES mutant has a very high fraction of charged residues, 70%.

in Figure 3.6. The analysis here though is for a the interior of a single chaperone protein complex. All the chaperones are in the closed (*cis*) conformation, during which encapsulated substrate proteins are folding. Again E and K are the most common amino acids in the interior cavities of this collection of chaperone proteins. The fraction of E and K is much larger than the previous dataset and the fraction of each amino acid is over 20% on these large proteins. The fraction of charged amino acids, in general, is much higher as well for the molecular chaperones. A dataset containing 528 *Escherichia coli* proteins was constructed to calculate if the fraction of charged amino acids is significant. The fraction of charged amino acids on the surface of *E. coli* GroEL-GroES is 56%, which is a higher fraction than 98% of all of the *E. coli* proteins considered (shown in blue for reference). This large fraction of charged amino acids has been noted before.[98,99] However, here we can see exactly which amino acids are more common than expected (E) and how significant it is. Figure 3.6 also shows how the interior cavity surface changes between the mesophilic GroEL-GroES and a thermophilic GroEL-GroES (optimal growth temperature of 65°C). Protein folding is typically more difficult at higher temperatures due to the increasing importance of entropy as temperature increases, and thus the thermophilic GroEL-GroES represents a more challenged chaperone. The fraction of charged amino acids is increased to 70% for the thermophilic GroEL-GroES, with most of the increase coming from E and K. Molecular chaperones, where nature requires strong nonfouling against many protein types, appear to use charged amino acids to accomplish this. E and K are the most utilized charged amino acids in both systems considered.

The large fraction of K and E suggests that they may be part of a general sequence pattern on protein surfaces; this was found to not be the case. The data considered here are the amino acid pair frequencies in sequence space. The most frequently occurring pairs for the human protein dataset are shown in Figure 3.7a. The plot of the pair frequency in white and the black shows what the predicted distribution would be if the pairs followed random chance, a multinomial distribution as a background model. The multinomial model is the number of pairs that would occur by chance if we knew the amino acid surface

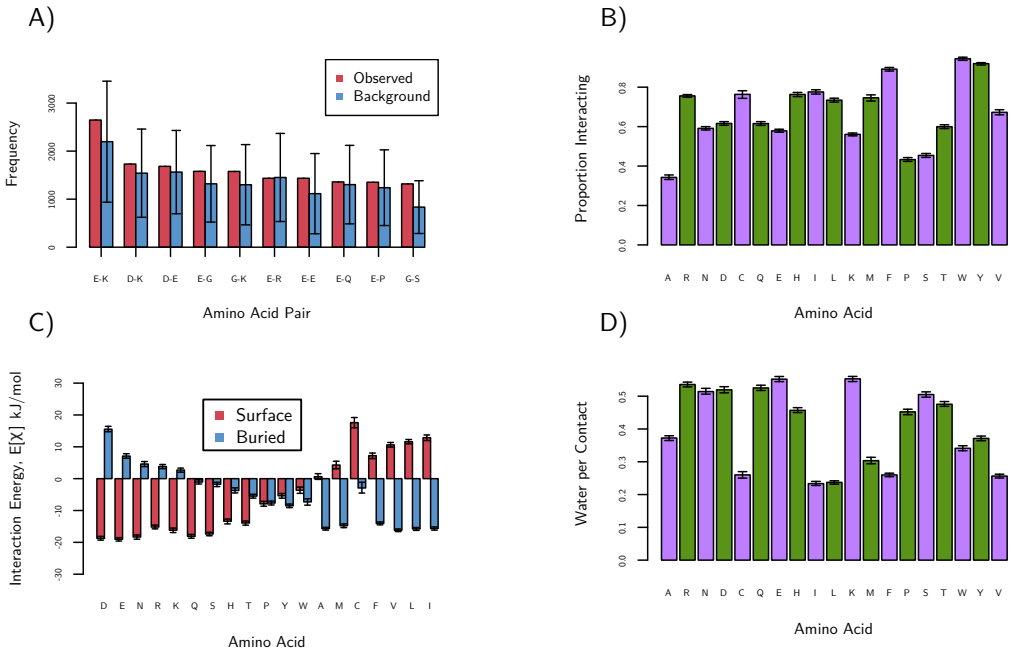


Figure 3.7: Interactions of amino acids. These plots show statistics based on protein sequences (a) and structure (b,c,d). Panel a shows the observed number of pairs of amino acids on the surface of proteins. The white bar is the expected number of pairs if the sequence were random. The order of left to right is from most frequently observed to least. Only the G–S pair is considered to be significantly more common than expected. Panel b shows the proportion of amino acids interacting from human dataset. An amino acid is considered 'interacting' if it is in contact with any other amino acid. We see small chains (S, P, A), E and K have the lowest proportions of interactions. Panel c shows the preference of each amino acid for protein surfaces or protein interiors. See text for details. D, E, and N have the highest preference for protein surfaces relative to protein interiors. Panel d shows the preference of amino acids for water relative to interacting with another amino acids. E and K have the highest water per contact. The green and purple colors are to guide the eyes.

fractions. Interestingly, few pairs occur next to each other more or less often than the multinomial model suggested within the error. Results show that on the surface at least, there are no global sequence patterns. There is one exception to this trend in Figure 3.7a: the glycine/serine pair. Based on analysis of the G-S ramachandrin plots, that pair is most commonly found as a type–II turn, resulting in its increased frequency over the multinomial model. This demonstrates that the methodology can discover surface motifs; in this case solvated type–II turns. The large fractions of K and E, however, are not correlated with a frequently occurring sequence motif and are instead nearly randomly distributed.

Structure and Interaction

In order to understand why nature chooses E and K, a series of statistical measures based on 3–dimensional structures were calculated and are plotted in Figures 3.7b, 3.7c and 3.7d. These results are only for the protein surfaces dataset, not the molecular chaperones. Briefly, Figure 3.7b shows that E, K and S have very few interactions with other amino acids, Figure 3.7c shows that charged (E, K, R, D, H) and amide amino acids (N and Q) have the most disfavored interactions with protein cores, and Figure 3.7d shows that charged amino acids, amide (Q, N), and hydroxyl (S) are the most hydrated. These results are discussed in detail below.

Figure 3.7b shows how often an amino acid is interacting with any other amino acid, which characterizes their nonspecific interactions. These data are normalized by the number of the amino acids, so that each bar is comparable. This data is only for amino acids observed on the surface. S, K and E have the lowest proportion interacting among the charged and polar hydrophilic amino acids. Alanine and proline are lower due in part to the small size of their side-chains. K and E have large side-chains but still rarely interact. R has the highest proportion of interactions among the hydrophilic amino acids, perhaps explaining why it is so much less often observed on protein surfaces.

In addition to the amount of interactions described above, it is important to discover with what amino acids interact. For nonspecific interactions, reversible interactions are

preferred to irreversible. For example, aggregation is often an irreversible process which we expect proteins to disfavor. Therefore we plot the strength of interactions of each amino acid with surface and buried amino acids on an average protein. Generally, more favorable interactions with the buried residues of a protein destabilize the protein, possibly leading to unfolding and aggregation. The average protein is a hypothetical protein which has a surface and buried residue distribution given by Figure 3.5. The strength of interactions was calculated using quasi-chemical theory.[117] The results are plotted in Figure 3.7c where the dark colored bar indicates strength of interaction between an amino acid and the average protein's surface. The light colored bar shows the strength of interaction between an amino acid and the average interior of proteins. Amide, charged, and alcohol amino acids have the highest preference for protein surfaces. Combined with our previous results, we see that E and K have fewer interactions (Figure 3.7b) and also favor interactions with protein surface amino acids, not interior amino acids (Figure 3.7c). Nature seems to prefer these amino acids as well in locations where resisting nonspecific interactions is necessary for function (i.e., the cytoplasm and molecular chaperones).

It is well established that hydration is the key to resisting nonspecific interactions.[118,119] Thus, we also analyze crystallographic proximity of waters to amino acids. The numbers of water per contact are shown in Figure 3.7d. The choice of using the number of contacts as the normalization was done to eliminate the size effects of amino acids. Those amino acids with more atoms tend have more waters near them. Therefore, Figure 3.7d should be thought of as the preference of an amino acid to interact with water relative to interacting with another amino acid. Again, we see E and K resisting other amino acids and preferring water. The amides and alcohols follow the trend as well. Overall, Figure 3.7 shows E and K are randomly distributed and rarely interact. If E and K do interact, they prefer water to amino acids. If E and K do interact with amino acids, they prefer to interact with those found on the surface of proteins instead of the interior. Regardless of the abundances of E and K found, these structural results strongly indicate E and K resist nonspecific interactions.

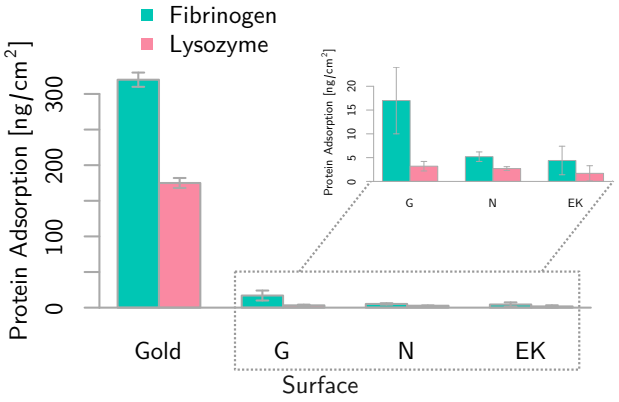


Figure 3.8: Protein adsorption of peptide SAMs. Protein adsorption results as determined by SPR. Ac-[EK]$_7$PPPPC-Am (EK), Ac-GGGGGGGPPPPC-Am (G), and Ac-CPPPPNNNNNNN-Am (N) sequences were self-assembled onto gold and protein solution was flowed over. Bound protein after buffer wash is shown in the bars. Untreated gold and poly-glycine are shown for reference. EK and N show similar nonfouling performance.

Implications for Design of Nonfouling Materials

There are three separate results indicating E and K resist nonspecific interactions. First, E and K are the most common amino acids on protein surfaces and even more so on protein surfaces in the cytoplasm, a crowded environment. Second, E and K are the most common amino acids found on the interior of molecular chaperones, a location which must resist nonspecific binding with thousands of protein types and conformations. Lastly, the structure and interaction results indicates that K and E rarely interact with other amino acids, favorably interact with water, and lack observed sequence patterns. We tested the ability of E and K to resist nonspecific interactions through the design of a nonfouling material. Nonfouling materials are an ideal experimental system and provide a direct measure of nonspecific interactions. When a nonfouling material is coated onto surfaces, the degree of its resistance to nonspecific adsorption can be quantitatively measured using highly-sensitive

surface plasmon resonance (SPR) sensors. The choice of using both K and E is to balance charges and avoid a cationic or anionic surface. A random mix of E and K was used to test the conclusion that nature distributes E and K without a sequence pattern (Figure 3.7a). Among the uncharged residues, N had the most similar properties to E and K, especially in its preference for protein surfaces (Figure 3.7c).

Experimental studies of protein adsorption on a peptide surface containing random K and E motifs were performed. The chosen sequence is Ac-$[EK]_7$PPPPC-Am, where the square brackets indicate seven random E and K. Ac and Am indicate acetylation and amidation, respectively. The proline repeat and cysteine provides a stable anchor for self-assembling peptides on gold surfaces.[95] The peptides were self-assembled onto a gold surface and adsorbed fibrinogen and lysozyme measured via SPR. The results are shown in Figure 3.8. Untreated gold was and poly-glycine were used as controls. The EK surface performance is comparable to the ultra-low protein fouling standard of <5 ng / cm^2 of fibrinogen.[96] Additionally a AC-$[N]_7$PPPPC-Am peptide was synthesized because N had best modeling results among the uncharged amino acids. It also had results below the ultra-low fouling threshold.

3.2.3 Conclusions

A fundamental understanding of nonspecific interactions has been gained by examining protein surfaces via bioinformatics. It was observed that (a) K and E are the most abundant on the surface of proteins and molecular chaperones, two disparate systems; (b) the KE content increases from extracellular proteins (21%) to cytoplasm proteins (27.5%) to molecular chaperones (38%) and to thermophilic molecular chaperones (52%); (c) K and E are distributed randomly; (d) K and E have strong water-binding capabilities, but weak binding with surrounding amino acids; (e) The uncharged amino acids which preferred water to other interactions were also abundant, N and S. These amino acids are used by proteins and molecular chaperones to resist nonspecific interactions, as supported by evidence from experiments with the random K and E peptide SAMs from this work, zwitterionic poly–

EK alternating peptide SAMs from our previous studies[95,101] and poly–serine SAMs by others.[102,103] Understanding and mimicking nature's resistance to nonspecific interactions is key to addressing emerging challenges in chemistry, especially in practical applications where complex environments can degrade materials and surface coatings. The techniques and conclusions here provide new insights and directions into the understanding, characterization and design of nonspecific interactions.

3.3 Role of Nonspecific Interactions in Molecular Chaperones through Model-based Bioinformatics

Section 3.2 demonstrated that bioinformatics may be used to better understand nonspecific interactions. Interestingly, nearly the same conclusions were reached when mimicking nature as when designing rationally designing nonfouling peptides in Chapter 2. Nonfouling has been considered only as a technique to resist nonspecific interactions thus far. However, it may be viewed more as resisting irreversible protein adsorption. Protein adsorption is inevitable for a biomaterial. If a surface, however, causes a protein to remain in its native conformation, the adsorption is more likely reversible. In this section, I model and study how nature stabilizes native conformations through nonspecific interactions.

3.3.1 Introduction

Molecular chaperones stabilize proteins is through a nonspecific process. These molecular complexes are able to nonspecifically help proteins fold or correct their misfolding. They contact many protein types and reversibly bind with them while stabilizing their folding.[98] This property of nonspecifically resiting irreversible binding with the proteins to be folded is sometimes called "non-stick."[100] The most well characterized chaperone protein is the GroEL-GroES complex found in *E. coli* from the chaperonin family. This large complex has seven-fold symmetry and may be described as a macromolecular machine.[98] Briefly, misfolded substrate proteins bind to the interior cavity of GroEL followed by the binding of GroES, a protein 'cap' which encases the substrate protein inside a chamber. Seven ATP

molecules bind to the complex and it undergoes a conformational change which affects the interior cavity conformation. The substrate protein then attempts to refold in the cavity until ATP hydrolysis, after which GroES unbinds, the substrate protein unbinds and the cycle begins anew, either with the same protein because it failed to refold correctly or another protein.[120] This description is a simplification of a relatively complicated process and there are open questions about this process. For example, it is unknown how deterministic the process is.[121] Also, it is unknown if the protein refolding is actively encouraged by changing the folding free energy, or if GroEL-GroES simply gives proteins another chance to fold free of other macromolecules interfering. A few recent reviews have been written on the topic.[98,99]

Recent research has quantified the interior cavity surface of GroEL-GroES along with four other molecular chaperones.[61] It was found that these interior surfaces have a high number of lysine and glutamic acid. Furthermore, when these two particular amino acids are used to create a surface, proteins are unable to irreversibly adsorb onto the surface, providing direct experimental evidence for the "non-stick" property of molecular chaperones. However, there remains several outstanding questions. Is the resistance of protein adsorption a passive effect which allows proteins to fold free of the interference of other macromolecules or are proteins actively stabilized? Can a molecular chaperone actively interact *nonspecifically* to stabilize protein folding? In this work we show that a simple model of interactions between proteins and the GroEL-GroES interior captures the trends observed in experiment. The model demonstrates that it is possible for GroEL-GroES to nonspecifically assist the folding of a large number of proteins through changes to the free energy of protein folding in its interior cavity. The mechanism of this stabilization is the enhancement of the hydrophobic effect inside the cavity; GroEL-GroES strengths the energetic difference between having hydrophobic groups exposed on the surface and having polar groups exposed.

In addition to enlightening the role of nonspecific interactions in biological systems, there is an interesting connection between surfaces which resist nonspecific protein adsorption and molecular chaperones. Nonspecific protein adsorption impacts biosensors, biomedical

implant coatings, and even marine coatings of commercial ships,[4] where the prevention of nonspecific protein adsorption is essential. As mentioned above, molecular chaperones provide a naturally occurring example of resisting protein adsorption. They contact many misfolded protein types, yet the molecular chaperones are able to unbind them. Thus a better understanding of GroEL-GroES can directly apply to the creation of new materials which resist nonspecific protein adsorption.

There has been previous research into molecular simulations of specific proteins in GroEL-GroES models.[122,123] Such research can be used to test scaling arguments or validate other models which generalize to many proteins. In this work, we exploit the lack of specificity in GroEL-GroES to make a model simple enough to be independent of the folding details of substrate proteins enclosed by GroEL-GroES yet accurate enough to match current experimental understanding. This avoids the nearly impossible molecular simulations necessary to consider fully atomistic protein folding inside the GroEL-GroES complex. The model developed in this work treats proteins at a simple level, which enables analysis of the folding of hundreds of proteins in GroEL-GroES. Confinement effects are included in order to determine which effects are most important and because confinement has been proposed as the mechanism of action for the folding assistance of proteins.[124]

This work can be divided into three parts. In part 1, we develop our model of protein folding inside the GroEL-GroES system. In part 2, we describe our model results on proteins found in *E. coli*. Finally, in part 3 we optimize our model to predict a hypothetical "best" GroEL-GroES interior surface and compare this to both the molecular chaperones proteins found in other organisms and the surface chemistry seen in materials which resist protein adsorption.

3.3.2 Methods

The PDBs and data used are available at http://sqlshare.escience.washington.edu.[111] The *E. coli* data are split into three tables: "ecoli_nogaps_1.csv" which contains the per protein data, "ecoli_nogaps_2.csv" which contains the per residue data, and "ecoli_backbone_contacts.csv"

which contains the data used to calculate the interactions energies. The *E. coli* proteins were selected from the protein data bank using a homologue cut-off of 40%, X-ray resolution < 2.5Å, "mutant" must not appear in the title, no large ligands (e.g., DNA, RNA), the only macromolecular in the structure is a protein, the number of residues is greater than 100, and no gaps within a chain appear in the structure. The statistical analysis was done using the R statistics language and program.[110] A residue was classified as on the surface when its surface area is 30% of its maximum surface area. The choice of 30% is commonly used and, in general, residues that are buried have no surface area so the results are relatively insensitive to cutoffs between 10-50%.[61,115] Surface area was calculated using accessible surface area.[112] The maximum surface area for a given residue was defined as the surface area occupied by the side chain atoms of a free Gly-X-Gly peptide, with X being the residue of interest. The backbones were taken to be $\alpha-$helical and the lowest energy χ-rotamers were used[93] for these tripeptides. Once the surface residues are identified, the surface residue fractions, p_i^f, can be calculated for each protein, which are used in the model.

The algorithm describing the identification of interior surface residues of the GroEL-GroES complex can be found in 3.2. The radii of gyration for the GroEL *trans* and GroEL-GroES *cis* conformation were calculated from the C-α carbons from residues which were identified as interior. The characteristic lengths for these two conformations are 29.96 and 46.4 Å for *cis* and *trans* conformations, respectively. The number of residues used in the model calculations was taken from the ATOM lines in the PDB files. This number was used for the calculation of the random coil radius of gyration.

3.3.3 Results

Model Description

A model is developed here which describes the influence of GroEL-GroES on protein folding. There have been previous efforts to quantify the influence of molecular chaperones on protein folding.[125,126] Here, we extend these entropy based arguments to include an enthalpy term derived from the distribution of residues from White *et al.*[61]. In that work,

the fractions of each amino acid found on the interior cavities of molecular chaperones was calculated. This quantitative model can provide folding free energy perturbations ($\Delta\Delta A$) which describe the effect of encapsulation in GroEL-GroES on protein folding using the knowledge of this residue distribution and the radii of the protein and GroEL-GroES. The folding free energy perturbation may also be interpreted as a type of free energy excess function, which quantifies the difference between folding free energy in the ideal case (no chaperone) and with chaperone. This term includes both the entropic confinement effects and energetic interactions between amino acids. There are some features which are lacking from this model, among which the most important are considering clusters of residues (e.g., a hydrophobic patch), the flexibility of GroEL, and the kinetics of this process. This model is only parametrized for *E. coli* GroEL-GroES, which will subsequently referred to as GroEL-GroES. We begin with the expression of free energy of protein folding:

$$\Delta A^{\circ} = U^{\circ} - T\Delta S^{\circ} \tag{3.13}$$

where $\circ$ indicates without the influence of GroEL-GroES. We will use a simple two state model for protein folding, where the first state is all unfolded conformations, represented by a random coil state, and the second state is the folded conformation. The entropic effect of confinement on the folded state is considered to be negligible due to that state's collapsed conformation. The entropic effect of confinement on the unfolded state inside GroEL-GroES is described by a scaling exponent from Takagi et al. who used a Gō-like model to evaluate scaling arguments for spherical confinement.[125] The entropic effect on folding is given by:

$$\Delta S = (S^{\circ}_{fold} + S_{fold}) - (S^{\circ}_{coil} + S_{coil}) = \tag{3.14}$$
$$\Delta S^{\circ} + \left(0 - \left[-k\left(\frac{R_g}{L}\right)^{3.25}\right]\right)$$

where S_{fold} is the folded state entropy with GroEL-GroES, S°_{fold} is the folded state entropy wihtout GroEL-GroES, likewise for the unfolded state (coil), L is the characteristic size of the confinement, which is the radius of the GroEL-GroES cavity, k is the Boltzmann's constant, R_g is the random coil radius of gyration, and 3.25 comes from Takagi *et al.*[125].

The confinement effect is positive because the unfolded protein state cannot occupy conformations that are larger than the cavity. For GroEL, the confinement of the substrate protein is modeled as a random coil in cylindrical confinement, which changes the exponent to 5/3 from 3.25.[127] R_g for random coil state proteins is given by:

$$R_g = N^{\nu} l \tag{3.15}$$

where ν was shown to be 0.6 by Kohn et al.[128] N is the number of amino acids and l is the Kuhn length, which was estimated to 1.93Åfor a large collection of proteins.[128]

The internal energy perturbation comes from interactions between the proteins and the interior surface of the GroEL-GroES complex. It is given by:

$$\Delta U = \Delta U^{\circ} + E_{fold} - E_{coil} \tag{3.16}$$

where E is the interaction energy between the protein and chaperone, given by:

$$E = N_s \sum_{i}^{20} \sum_{j}^{20} p_i \chi_{ij} p_j^g \tag{3.17}$$

where N_s is the number of residues on the surface, p_i is the fraction of residue type i on the surface of the protein, χ_{ij} is the energy of the interaction between residues of type i and type j, and p_j^g is the fraction of residues of type j on the interior surface of the GroEL-GroES complex. The p_j^g were calculated as described in White et al.[61] In that work, surface residues were identified by measuring their accessible surface area[112] and the interior surface residues of GroEL-GroES were identified using a geometric algorithm. N_s is used because we assume the number of interactions to be equal to the number of surface residues. This assumes there are not multiple interactions and that the number of chaperone residues is large enough that they do not limit the interactions.

The interaction energies, χ_{ij}, are the only energy terms in the model. Thus, all of our energy values depend on the accuracy of these interaction energy terms. These values were derived using the 'knowledge based' or 'quasi-chemical' approach which is employed in protein structure prediction, among other fields.[129] The process uses experimental crystallography data and decsribed below. The last term is p_i, the fraction of each amino acid.

This was calculated for the folded state by tabulating the residues present on the surface of a protein according to White *et al.*[61]. The unfolded state's residue distribution is assumed to be the same as the whole protein residue fractions. A more sophisticated model would consider the influence of the chaperone on the unfolded residue distribution, which would likely lower the magnitude of the change in interaction energy between folded and unfolded proteins. However, such a model would inevitably involve simulating protein geometry and greatly complicate the model. The collection of p_i's will be called the residue distribution.

Equations 2-5 can be combined into Equation 1 and rearranged to give:

$$\Delta A = \Delta A^{\circ} + \Delta\Delta A$$
$$\Delta\Delta A = \sum_{i}^{20} N_s \left(p_i^f - \frac{R_g^f}{R_g^u} p_i^u \right) \sum_{j}^{20} \chi_{ij} p_j^g - kT \left(\frac{R_g}{L} \right)^{3.25} \tag{3.18}$$

where the f indicates the folded state, u indicates the coil state, and the $\frac{R_g^f}{R_g^u}$ term is due to the change in the number of surface residues in the unfolded state. The number of residues on the surface of the random coil unfolded state is different than the folded state. We can assume that the surface density is proportional to the density:

$$\frac{N_s}{A} \propto \frac{N}{V}$$

where N is the total number of residues. This leads to:

$$N_s \propto \frac{A}{V} \propto \frac{1}{R}$$

The increase in the radius of gyration upon the transition from folded to random coil states can be used to predict the number of surface residues:

$$\frac{N_s^u}{N_s^f} = \frac{R_g^f}{R_g^u}$$

Thus leading the term appearing in Equation 7.

Model Results

The model results from a sample of *E. coli* proteins are shown in Fig. 3.9. The $x-$axis shows the free energy of folding perturbation ($\Delta\Delta A$) due to confinement in the closed, or *cis*,

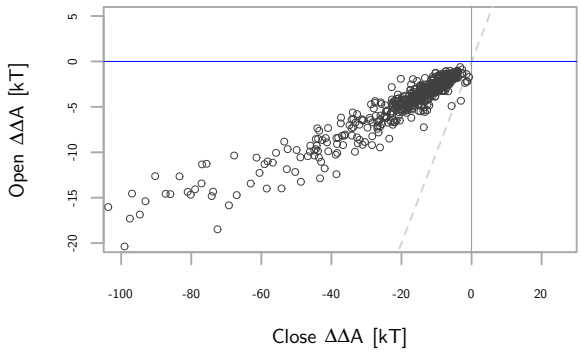


Figure 3.9: The $y-$axis is the predicted folding free energy perturbation from the open or *trans* GroEL complex. A negative number indicates stabilization. The $x-$axis indicates the perturbation from the closed or *cis* GroEL-GroES complex. Most hypotheses of the GroEL-GroES complex action predict that the closed form should be most stabilizing, which is indeed observed for the majority (98%) of the proteins. This is indicated by being above the dashed line.

GroEL-GroES complex and the $y-$axis shows $\Delta\Delta A$ for the open, or *trans*, GroEL protein. The two deltas indicate free energy perturbations to free energy differences, specifically perturbations to the free energy difference upon protein folding. As expected, all the points are on the left of the green line, meaning all proteins examined are stabilized by the closed form of the GroEL-GroES complex. This is essential for a model of GroEL-GroES. The median $\Delta\Delta A$ for the closed GroEL-GroES complex is -14.8 kT, or -36.9 kJ/mol at 300 K (0.10 kJ/mol·residue). The dashed gray line separates those proteins which are stabilized more by the open form from those which are stabilized more by the closed form. As seen in the plot, most proteins are stabilized more by the closed form because they lie above that line. A few small proteins near the origin lie below this line. This matches the mechanism for GroEL-GroES[130] which shows that the model performs well. The median change in $\Delta\Delta A$ between the closed and open forms of GroEL is -10.5 kT or -26.2 kJ/mol at 300 K. The blue line indicates which proteins have a stabilized fold when encapsulated by the open form. The proteins are weakly stabilized by the open form; all the points lie below the blue line. This does not follow the expected mechanism for GroEL, which should show positive $\Delta\Delta A$ values. GroEL should destabilize proteins; it interacts preferentially with misfolded proteins. This discrepancy is described in detail below. Overall, the model can assign a quantitative free energy to the conformational changes observed between the open and closed forms of the GroEL-GroES complex, it matches what is understood from experiments, and it shows that the closed form of GroEL-GroES stabilizes protein folding.

Arguably the most well-established hypothesis on the mechanism of action for GroEL-GroES includes a preferential binding of misfolded proteins to GroEL.[99,131–133] Our model does not predict that the open form generally stabilizes the unfolded state. The cause of this can be determined by breaking down the model into entropy and internal energy terms as shown in Fig. 3.10. The open form is indeed energetically destabilizing, as indicated by the positive values for the internal energy. However, the entropy is stabilizing in the open form. This is due to the small size of the open form, especially relative to the closed form of GroEL-GroES. The characteristic open length is 29.96 Å, compared with 46.4 Å for the closed form.

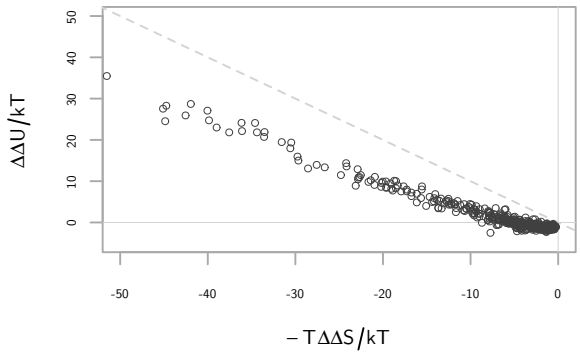


Figure 3.10: The contribution to the folding free energy perturbation from the change in internal energy is plotted on the $y-$axis. The internal energy is nearly always positive, indicating all proteins are energetically destabilized in the closed form. The entropy or confinement contribution to the folding free energy perturbation is shown in the $x-$axis. It is negative for most of the proteins. Points below the dashed lines are stabilized more by confinement than the internal energy.

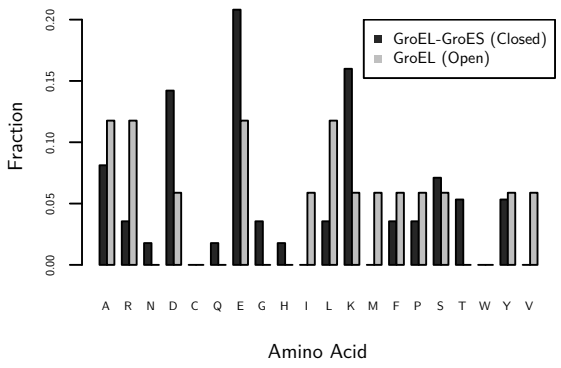


Figure 3.11: The key feature here is which residues change. There is a higher fraction of charged residues in the closed form and the open form does have more hydrophobic residues, specifically leucine, isoleucine, valine, and methionine.

The preferential binding of a misfolded protein may still occur because the hydrophobic regions exposed while unfolding may be the only part of a substrate protein sequestered in the open form, whereas the remaining residues are outside. Or, more likely, the two state model is an oversimplification. Many of the misfolded protein conformations may be smaller than the random coil state yet still have hydrophobic residues exposed. Such conformations would indeed bind more favorably to GroEL than the native conformations. Their smaller size would decrease the magnitude of confinement yet the misfolded conformations would still interact favorably with the GroEL residues which favorably interact with unfolded residue distributions.

It is important that the model does predict that proteins will be energetically perturbed towards unfolding in the open form from the interaction energy. The cause for the difference in interaction energies between the open and closed form is from the residue distribution, which is shown in Fig. 3.11. As one can see, there is a large change in the number of hydrophobic residues and charged residues. The leucine and isoleucine perturb proteins

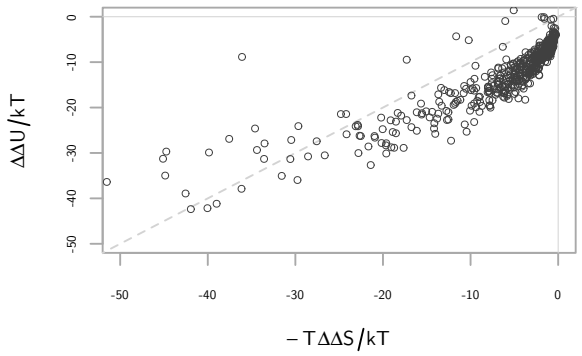


Figure 3.12: The contribution to the folding free energy perturbation from the change in internal energy is plotted on the $y-$axis. The internal energy is nearly always negative, indicating nearly all proteins are energetically stabilized in the closed form. The entropy or confinement contribution to the folding free energy perturbation is shown in the $x-$axis. It is negative for all of the protein. Points above the dashed lines are stabilized more by confinement than the internal energy, which is rare except for the larger proteins. Note that proteins which cannot fit into the GroEL-GroES complex while folded are not included in this analysis.

towards unfolding, whereas aspartic acid, glutamic acid, and lysine strongly perturb proteins towards folding (see Table 3.2). Another interesting feature to note is the exchange of arginine with lysine between the two distributions. Lysine is more stabilizing than arginine (see Table 3.2).

The effects of confinement and the cavity surface in the closed form can be compared by again breaking the model equations into entropy and internal energy terms, as shown in Fig. 3.12. The effect of confinement, or entropy, on the folding free energy perturbation is plotted on the $x-$axis and the effect of the cavity surface, or internal energy, is plotted on the $y-$axis. This plot shows that the internal energy is more important at the smaller

protein sizes, where confinement has a negligible effect. The exponent on the entropy term, however, causes the term to grow quickly as the radius of gyration of the random coil state increases. This causes the trend to curve at the bottom of the plot. The dashed line indicates the separation between proteins which are stabilized more by entropy and those which are stabilized by internal energy. In general, the proteins are more stabilized by the internal energy due to interactions between the proteins and the surface of the cavity. The reason for this strong effect is the high number of charged residues on the interior cavity. Charged residues enhance the hydrophobic effect through their unfavorable interactions with hydrophobic groups. Aspartic acid in particular has the most unfavorable interactions with hydrophobic groups followed by asparagine and glutamic acid (see Table 3.2). Equally important is that the charged residues have favorable interactions with charged and polar residues, which are most common on the surface of proteins.[61] In particular, lysine is the most common residue found on the surface of *E. coli* proteins, and thus aspartic and glutamic acid, which interact very favorably with lysine, have a very low interaction energy with the surface of proteins (Table 3.2). This effect of internal energy and the role of the interior cavity of the GroEL-GroES complex is still an open question with much of the recent discussion in literature focused on the role of water.[134–137] Water is included implicitly in our model through the residue-residue interaction terms, which come from experimental crystal structures containing water. The internal energy effect on protein folding is a nonspecific effect and independent of the geometry of interactions.

Detailed Description of Interaction Energy

The interaction energies are derived from counts of residue contacts and total number of residues. The interaction energy is defined as:

$$\chi_{AB} \equiv U_{AB} - U_A - U_B$$

where U indicates energy and A and B are residue types. Now, substituting the Boltzmann distribution:

$$\chi_{AB} = -\frac{1}{\beta}\ln\left(e^{-\beta U_{AB}}\right) + \frac{1}{\beta}\ln\left(e^{-\beta U_A}\right) + \frac{1}{\beta}\ln\left(e^{-\beta U_B}\right), \quad \beta = kT$$

where k is Boltzmann's constant and T is the temperature. Replacing the Boltzmann distribution with the probabilities:

$$\chi_{AB} = -\frac{1}{\beta}\ln P_{AB} + \frac{1}{\beta}\ln P_A + \frac{1}{\beta}\ln P_B = -\frac{1}{\beta}\ln\frac{P_{AB}}{P_A P_B} \tag{3.19}$$

The probabilities may be found using maximum likelihood estimators[138] and shown to be:

$$\hat{P}_A = \frac{N_A}{\sum_i N_i}, \quad \hat{P}_{AB} = \frac{N_{AB}}{N_{\text{Free A}} + \sum_y N_{Ay}} \tag{3.20}$$

where N_A indicates the number of residues of type A, N_{AB} indicates the number of A, B pairs and 'Free' indicates unpaired A. The summation in the denominator is across all pairs where A is the first residue. The interaction energies for each residue is shown in Table 3.2.

Residue contacts were calculated by finding the pair-wise distance between each side-chain heavy atom on each residue. The PDB files described in the main text were used for these calculations. Neighboring residues, as determined by residue indices in PDB files, were excluded from being in contact. If any of the heavy-atom pairs were below the cutoff distance, the side-chain pair is said to be in contact. The heavy atom pair cutoffs were taken to be Van der Waals energy minimum radii. The following Van der Waal radii were used: nitrogen: 3.25Å, oxygen 2.96Å, and sulfur 3.55Å. The following mixing rule was applied:

$$r_{ij} = 2^{1/6}\sqrt{\sigma_i \sigma_j} \tag{3.21}$$

where σ is the Van der Waals radius.

	ALA	ARG	ASN	ASP	CYS	GLN	GLU	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
ALA	-0.34	0.08	0.04	0.15	-0.13	0.01	0.03	0.12	-0.2	-0.2	0.02	-0.18	-0.16	-0.01	0.03	-0.02	0.03	-0.02	-0.2
ARG	0.08	0.08	-0.08	-0.74	0.26	-0.2	-0.78	0.08	0.23	0.14	0.3	0.17	0.15	-0.1	-0.08	-0.03	-0.11	0.01	0.26
ASN	0.04	-0.08	-0.54	-0.47	0.12	-0.32	-0.26	-0.08	0.42	0.34	-0.25	0.23	0.14	-0.1	-0.32	-0.24	-0.01	0	0.25
ASP	0.15	-0.74	-0.47	-0.03	0.43	-0.3	0.22	-0.46	0.51	0.52	-0.78	0.23	0.36	-0.01	-0.5	-0.28	0.23	0.06	0.49
CYS	-0.13	0.26	0.12	0.43	-1.93	0.25	0.2	-0.25	-0.19	-0.1	0.35	-0.29	-0.24	-0.17	0.05	0.14	-0.11	0.22	-0.19
GLN	0.01	-0.2	-0.32	-0.3	0.25	-0.38	-0.21	-0.17	0.23	0.16	-0.13	0.16	0.16	-0.17	-0.23	-0.24	-0.18	0.06	0.18
GLU	0.03	-0.78	-0.26	0.22	0.2	-0.21	0.07	-0.34	0.33	0.3	-0.81	0.28	0.25	-0.1	-0.25	-0.17	0.01	-0.03	0.32
HIS	0.12	0.08	-0.08	-0.46	-0.25	-0.17	-0.34	-0.6	0.18	0.18	0.13	-0.03	0.17	-0.13	-0.29	-0.19	-0.01	-0.05	0.13
ILE	-0.2	0.23	0.42	0.51	-0.19	0.23	0.33	0.18	-0.49	-0.41	0.25	-0.19	-0.22	0.05	0.18	-0.01	0.04	-0.09	-0.4
LEU	-0.2	0.14	0.34	0.52	-0.1	0.16	0.3	0.18	-0.41	-0.46	0.28	-0.19	-0.23	0.05	0.2	0.08	-0.1	-0.05	-0.39
LYS	0.02	0.3	-0.25	-0.78	0.35	-0.13	-0.81	0.13	0.25	0.28	0.07	0.11	0.12	0.03	-0.11	-0.05	0.04	-0.18	0.26
MET	-0.18	0.17	0.23	0.23	-0.29	0.16	0.28	-0.03	-0.19	-0.19	0.11	-0.62	-0.26	-0.06	0.05	-0.05	-0.1	-0.11	-0.23
PHE	-0.16	0.15	0.14	0.36	-0.24	0.16	0.25	0.17	-0.22	-0.23	0.12	-0.26	-0.27	-0.16	-0.01	0.08	-0.04	-0.1	-0.27
PRO	-0.01	-0.1	-0.1	-0.01	-0.17	-0.17	-0.1	-0.13	0.05	0.05	0.03	-0.06	-0.16	-0.18	0.07	-0.06	-0.49	-0.38	0.03
SER	0.03	-0.08	-0.32	-0.5	0.05	-0.23	-0.25	-0.29	0.18	0.2	-0.11	0.05	-0.01	0.07	-0.23	-0.17	-0.08	0	0.16
THR	-0.02	-0.03	-0.24	-0.28	0.14	-0.24	-0.17	-0.19	-0.01	0.08	-0.05	-0.05	0.08	-0.06	-0.17	-0.17	0.1	0.08	0.01
TRP	0.03	-0.11	-0.01	0.23	-0.11	-0.18	0.01	-0.01	0.04	-0.1	0.04	-0.1	-0.04	-0.49	-0.08	0.1	-0.21	0.01	-0.05
TYR	-0.02	0.01	0	0.06	0.22	0.06	-0.03	-0.05	-0.09	-0.05	-0.18	-0.11	-0.1	-0.38	0	0.08	0.01	-0.15	-0.09
VAL	-0.2	0.26	0.25	0.49	-0.19	0.18	0.32	0.13	-0.4	-0.39	0.26	-0.23	-0.27	0.03	0.16	0.01	-0.05	-0.09	-0.39
Aromatic	-0.08	0.05	0.06	0.22	-0.07	0.05	0.1	0.05	-0.13	-0.15	-0.02	-0.18	-0.17	-0.31	-0.02	0.08	-0.05	-0.1	-0.17
Polar	0.03	-0.11	-0.31	-0.31	0.03	-0.25	-0.17	-0.2	0.14	0.15	-0.1	0	-0.04	-0.11	-0.18	-0.19	-0.22	-0.13	0.1
Charged	0.08	-0.38	-0.25	-0.4	0.22	-0.21	-0.4	-0.25	0.31	0.29	-0.39	0.17	0.21	-0.06	-0.26	-0.15	0.03	-0.03	0.31
Hydrophobic	-0.23	0.17	0.25	0.39	-0.16	0.14	0.24	0.14	-0.37	-0.36	0.19	-0.23	-0.22	0.03	0.14	0.01	-0.04	-0.06	-0.34
Proteins	-23.68	-11.49	-15.57	2.38	1.59	-22.78	-10.59	-16.61	-4.6	-5.68	-11.05	-11.49	-9.23	-18.39	-21.45	-23.88	-12.24	-16.39	-6.31
Surface	-1.92	-18.46	-24.27	-22.8	15.14	-21.83	-21.78	-16.13	19.1	17.97	-19.78	8.21	9.08	-8.47	-19.93	-15.77	-3.89	-5.92	17.29
Interior	-21.61	6.32	7.74	23.85	-13.11	-0.73	10.49	-0.92	-23.16	-23.6	8.91	-19.65	-18.09	-10.28	-1.65	-8.03	-8.17	-9.95	-23.73
Surface - Proteins	21.61	-6.32	-7.74	-23.85	13.11	0.73	-10.49	0.94	23.16	23.6	-8.91	19.65	18.09	10.28	1.65	8.03	8.17	9.95	23.73

Table 3.2: Interaction energies, χ, for the 19 amino acid paris (glycine is excluded) in units of kT. The Aromatic group is PHE, TYR, and TRP. The polar group is SER, THE, ASP, GLU, and PRO. The charged group is LYS, ASP, GLU, HIS, and ARG. The hydrophobic group is ALA, VAL, LEU, ILE, and MET. Proteins indicates the interaction between one residue type and all the amino acids weighted by their frequency in proteins. The surface is the same, except weighted by the amino acids' frequency on the surface of proteins. Interior is the complement of the surface amino acids. Surface - proteins is the difference between those two interaction energies and represents the difference in interaction energies between a protein surface and random coil. These distribution level interaction energies treat each protein equally, whereas in the model in the main text each protein is weighted by its size and number of residues.

Ideal Cavity Surface

Now that it is possible to quantify folding free energy perturbations, we can treat the interior surface of the GroEL-GroES complex as a design variable in order to maximize the magnitude of the folding free energy perturbation. That is, we may consider an idealized GroEL-GroES where all residues are alchemically transformed to give an ideal residue distribution. This will demonstrate which residues are most important in stabilizing proteins in GroEL-GroES. The geometry of the cavity will not change in this analysis, only the residue identities. The model equations are linear in the residue fractions of the GroEL-GroES surface (Equation 6) and the fractions are bound between 0 and 1. The extreme values of the model thus occur when any residue fraction is 1 and the others are 0. This is plotted in Fig. 3.13, where the $x-$axis labels indicate which residue fraction is 1 (with all others being 0) and the $y-$axis is the median $\Delta\Delta A$. For example, the bar labeled "D" shows what the median free energy of the 528 *E. coli* proteins would be if every residue on the interior surface GroEL-GroES were replaced with aspartic acid. Aspartic acid is the most stabilizing residue, followed by asparagine. Their median $\Delta\Delta A$s are -27.6 and -23.1 kT respectively. The actual GroEL-GroES residue fraction has a median $\Delta\Delta A$ value of -14.8 kT which is shown in the horizontal dashed line. After aspartic acid and asparagine comes glutamic acid, lysine, and arginine, meaning four of the five most stabilizing residues are charged. The charged residues are the most important for stabilizing protein folding. Therefore, the high number of charged residues seen in chaperone proteins is not so unexpected. The reason for charged residues being so stabilizing is that they have the largest difference in interaction energies between a folded and unfolded protein (see Table 3.2). Among the uncharged residues, asparagine is the most stabilizing. Asparagine is similar in geometry to aspartic acid but contains an amide functional group. We can gain some insight into the effect of the negative charge by examining the difference between aspartic acid and asparagine. Specifically, asparagine has more favorable interactions with hydrophobic residues (0.25 vs. 0.49 kT) and much more favorable interactions with aromatic residues (0.06 vs. 0.22 kT) (Table 3.2). Thus the negative charge seems to repel aromatic groups and

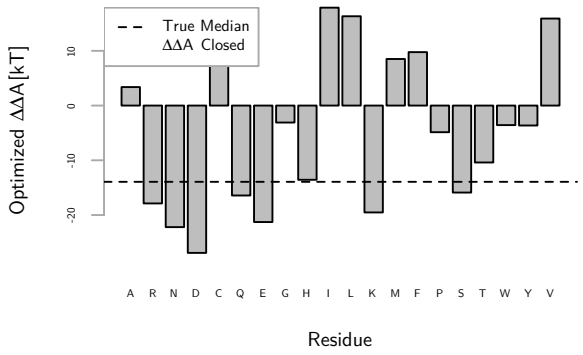


Figure 3.13: The median folding free energy perturbation as predicted by the model from GroEL-GroES on 528 *E. Coli* proteins is shown on the $y-$axis. The $x-$axis is the single residue type which is maximal. For example, the A indicates that only alanine is present on the surface of the GroEL-GroES and the bar height is the median folding free energy for *E. coli* proteins if only alanine were present in GroEL-GroES. The plot shows that the isoleucine destabilizes proteins the most, along with other hydrophobic residues as expected. Aspartic acid is the most stabilizing residue, followed by asparagine, glutamic acid, and lysine. Cysteine is not expected to be stabilizing or destabilizing due to its unique disulfide bonding.

hydrophobic residues slightly more, which are more common on the interior of proteins as opposed to their surface. As expected, the residues which most perturb folding free energies away from folding are the hydrophobic residues: valine, leucine, and isoleucine.

The mutations between the *Thermus thermophiles* and *E. coli* GroEL-GroES complex[61] can be better understood from the results described above and those shown in Fig. 3.13. The large increase in lysine and glutamic acid relative to *E. coli* GroEL-GroES stabilizes protein folding more inside the *Thermus thermophiles* GroEL-GroES complex. There is a corresponding decrease in alanine, glycine, proline, serine, threonine, leucine and isoleucine on the surface. All those residues stabilize protein folding less than glutamic acid and lysine or destabilize protein folding. There is a decrease in aspartic acid as well, though it is small in comparison to the increase of the other charged residues.

3.3.4 Discussion and Experimental Comparisons

There is debate in the study of molecular chaperone proteins about whether they directly perturb the free energy of folding for a protein.[99] Some researchers have argued that the mechanism of GroEL-GroES may be explained without GroEL-GroES stabilizing protein folding.[139] It is clear that it is possible to stabilize a protein fold through nonspecific interactions. For example, osmolytes can accomplish this. Further, there exists a general residue distribution on the surface of proteins and on the interior of proteins.[61] It is possible for a surface to create favorable interactions with those residues and thus nonspecifically make a folded state more favored than an unfolded state. Alternatively, a surface may have strongly unfavorable interactions with those residues which are not seen on the surface of proteins (hydrophobic residues). Therefore, it is possible for the cavity of GroEL-GroES to stabilize protein folding via its surface chemistry. It would be surprising if nature did not make use of this. The model presented here shows that the effect is significant: the folding free energy perturbation is 10 kJ/mol per 100 residues. Additional evidence for this can be seen from the fact that the interior cavity of GroEL-GroES is unlike what is typically seen on the surface of *E. coli* proteins. There is a folding free energy perturbation from GroEL-GroES.

There is a remaining question of why aspartic acid is less common than glutamic acid inside the chaperone cavities. It is not always the case, for example a group II chaperonin protein isolated from *Methanococcus maripaludis*[107] has slightly more interior aspartic acid than glutamic acid.[61] However, generally glutamic acid is present in more than double the amount of aspartic acid. The two acids have similar hydration free energies,[140] similar interaction energies, and size. A difference which may explain the preference of GroEL-GroES for glutamic acid is that glutamic acid interacts less than aspartic acid.[61] Essentially, aspartic acid has more protein stabilizing interactions given that it is interacting but glutamic acid interacts less while still having stabilizing interactions.

The effects of confinement have been studied experimentally through tail-multiplication studies.[139,141] A Gly-Gly-Met 'tail' found at the C-terminal of WT-GroEL-GroES may be extended to decrease the volume of the closed conformation by approximately 4% per tail multiplication. At four tails a drastic decrease in activity is observed for large and small substrate proteins, though the effects at shorter tail lengths are imperceptible.[139] The change in entropy predicted from the model for these tail multiplication studies is approximately the same as the change in volume ($V \propto R^3$, $\Delta\Delta S \propto R^{-3.25}$, $\Delta\Delta S \propto V^{-1.08}$). Thus, even at four tails the effect on entropy is modest, with a 13% decrease. This is supported from the results of Farr et al., which showed most of the change in GroEL-GroES activity for the four tail GroEL-GroES mutants is due to changes in ATPase activity.[139] Thus, the magnitude of confinement in our model is consistent with experiments.

There have been experiments exploring the hydrophilic character of the interior cavity of GroEL-GroES through mutations.[139,142] Most mutations which affect only the closed GroEL-GroES conformation have shown negligible effects on the activity of GroEL-GroES,[100] with the exception of two mutations which created a net neutral charge for the interior cavity.[142] Our model is unable to account for the loss of activity when the interior cavity becomes neutral; those particular results may be due to a change in the structure of GroEL-GroES or some other large scale effect. Neglecting these two mutation results, the other 11 mutation sets referenced were all without effect in experiments. According to the

model, two of the most mutated variants have median folding free energy perturbations of −13.3 and −14.7 kT for the E252A/D253A/E255A mutant and D359N/D361N/E363Q respectively. These mutations are too slight to overcome the effect of the other charged residues and confinement. According to the model, only drastic mutations may produce an effect without removing the negative charge inside the cavity. Eight Lysine to leucine mutations and eight aspartic acid to leucine mutations would keep the negative charge but strongly increase the folding free energy perturbations (destabilizing proteins) enough to cause positive folding free energies in 25% of the proteins. Interestingly, mutation counts below that number are still not significant enough to overcome both the other hydrophilic residues stabilization and the confinement entropy. Alanine to aspartic acid mutations may increase the folding free perturbations (stabilizing proteins), although it is generally difficult to increase wild-type activity.

There are experimental results showing which proteins strictly require GroEL-GroES, so-called 'Class III' proteins.[143] Model calculations on Class III proteins which have crystallography structures produced a higher median free energy perturbation value of -22.0 kT compared with -14.8 kT for the *E. coli* dataset. The perturbation is stronger, as expected. Compared to proteins of similar size, though, it is not significantly different. Thus, it appears that their GroEL-GroES dependence is a combination of their higher free energies of folding and stronger effect from GroEL-GroES.

More direct evidence for the role of charged residues and asparagine may be found from experiments studying nonspecific protein adsorption on self-assembled monolayers (SAMs) of oligo-peptides. Glutamic acid and lysine combinations and poly-aspargine have been shown previously to resist irreversible protein adsorption, demonstrating that these residues strongly prefer interacting with protein surfaces (reversible binding) and not all protein residues (irreversible binding).[61,101,144] In fact, these SAMs function quite similarly to GroEL-GroES in their ability to bind proteins reversibly. Other research has shown that asparagine and aspartic acid have the lowest nonspecific protein adsorption, following the trend seen in the model.[145]

3.3.5 Conclusions

Molecular chaperones have a unique distribution of residues in their interior cavities which have large fractions of charged residues.[61,99] In the GroEL-GroES chaperonin found in *Thermus thermophiles*, the interior cavity has 70% charged residues.[61] The role of these charged residues is to stabilize protein folding inside the chamber by increasing the hydrophobic effect, as demonstrated through the simple model of protein folding inside GroEL-GroES presented here. Thus GroEL-GroES stabilizes a large number of proteins through interactions between protein surfaces and the interior cavity of GroEL-GroES. The median free energy perturbation on folding free energy in GroEL-GroEL isolated from *E. coli* is -10 kJ/mol per 100 residues for a diverse sample of 528 *E. coli* proteins. This model provides predictions which are qualitatively consistent with the hypothesized mechanism of GroEL-GroES, experiments, and captures the behavior of both confinement entropy and energetic effects from the surface chemistry of GroEL-GroES. The residues which provide the most protein stabilization are aspartic acid, glutamic acid, asparagine, and lysine of which lysine, aspartic acid, and glutamic acid are present in high amounts on the interior surface of many chaperone proteins.[61] This research brings a better understanding of how nature is able to interact nonspecifically with proteins.

3.4 Chapter Summary

Section 3.2 demonstrated that proteins have general patterns on their surfaces and that these may relate to resisting nonspecific interactions. Mimicking the patterns leads to molecules that resist nonspecific interactions, strengthening this hypothesis. Interestingly, the same two amino acids, E and K, as seen in Chapter 2 were found on protein surfaces in harsh biological environments and make good nonfouling peptides. New designs, based on asparagine, were found as well. Section 3.3 showed how to model nonspecific interactions directly through model based bioinformatics and addressed the relationship between resisting nonspecific interactions and stabilizing protein conformations.

Chapter 4

ANALYZING PEPTIDE LIBRARIES

The previous chapters have utilized nonfouling as motivation throughout and many of the results have been directly related to the design of nonfouling materials. In this Chapter, the knowledge of nonspecific interactions is applied to high throughput peptide library screening. Currently, experimental collaborators in my research group are developing techniques to create libraries for screening nonfouling peptides. The analysis of this data is limited by the general lack of techniques for analyzing peptide library data. In Section 4.1 I address this by describing a set of state-of-the-art statistical tools to analyze peptide libraries and apply them to known peptide libraries in the literature. These tools can analyze peptides either by examining their motifs or quantitative structure-activity relationships (QSARs). Known peptide libraries in literature are generally concerned with specific interactions such as protease activity on peptides, peptide-protein binding, or antimicrobial peptide activity. Thus, they are concerned with motifs and to a lesser extent with the chemical properties of the peptides, such as solubility. In contrast, the peptides which our group searches for are more impacted by the chemical properties, but motifs are of interest to a lesser degree. In Section 4.2, I develop new statistical tools based on current research in the machine learning of speech recognition which can integrate information about both motifs and QSARs. These new models are shown to be flexible, interpretable, quick to train and perform as well as the best classifiers in literature at predicting antimicrobial activity, a difficult benchmark task.

4.1 Standardizing and Simplifying Peptide Library Analysis

4.1.1 Introduction

Combinatorial peptide libraries are powerful tools for quickly screening millions of peptides for activity. With an appropriate assay, it is possible to obtain the individual sequences of active peptides. This may be used to discover protein ligands,[146] antimicrobial peptides,[147] and even molecules for protein separation.[148] Peptide libraries are a large collection of peptides each with different sequences. A process is applied to the library that separates active peptides from inactive peptides. For example, running the library over a column with an immobilized target molecule will elute the non-binding peptides away from those which bind. Then the peptides which are bound may be examined to identify active sequences. The remarkable aspect of peptide libraries is that millions of sequences can be tested in parallel, enabling high-throughput experiments.

One of the most accurate types of peptide libraries are solid-phase combinatorial libraries.[149,150] Solid-phase libraries are unique in their ability to eliminate biases in amino acid frequency while still providing individual active sequences. Some peptide library methods have confounding factors; for example, FLITRX libraries, which display peptides using *E. coli*, are estimated to lose around 10% of peptides due to expression problems.[151] Bacteria or phage libraries require multiple iterations and tuning to ensure that multiple active sequences are discovered.[152] Other methods can have convoluted results. For example, using affinity columns with the target bound to the column and peptides in the mobile phase screens for both peptide abundance and affinity,[153] whereas peptide affinity is the only variable with which we are concerned. Solid-phase libraries are well suited to analysis because they provide individual sequences based only on affinity.[149] Utilizing solid-phase libraries consists of three basic steps: synthesis, activity determination, and sequencing. During synthesis, the library of peptides is constructed on solid particles, ranging from 90 μm up to 200 μm depending on the chemistry and application[150,154]. Each particle, or bead, contains tethered peptides, all with the same sequence. Between beads, however, there are different

sequences. During the activity determination step, each bead is tested and beads which are active are isolated. This is done, for example, using a colorimetric or fluorescence assay and the beads may be separated via an automated sorter.[155] Finally, during sequencing the peptides are cleaved one amino acid at a time and sequenced using MALDI-TOF in a technique called partial edman degradation.[149] The result of these steps is a list of sequences which are active.

There are two techniques used for analyzing the peptide libraries in this work. The first is quantitative structure-activity relationships (QSAR), which excel at describing small molecules. A descriptor is a quantitative metric based on chemical structure, for example number of double bonds, which may then be correlated with activity. The correlation of a descriptor with activity is called a QSAR. The second technique is motif discovery. Motifs are frequently occurring short strings of amino acids. For example, 'RGD' is a motif, the one letter amino acid abbreviations are used. Typically, motif lengths are between 3-10 amino acids long.

Analysis of these solid-phase peptide libraries is still relatively unexplored, excepting traditional consensus sequence analysis. There are three main challenges for analysis of such experiments. The first is that the variable regions of the peptides are typically too short (3-10 amino acids) to be analyzed using existing techniques from proteomics or genomics. For example, the popular Multiplied EM for Motif Elicitation algorithm, which discovers sequence motifs, is suggested to work on at least 8 amino acid length peptides and typically used for searching whole proteins or long gene sequences for motifs[156,157]. The second challenge is that, when viewed from a traditional QSAR descriptor based perspective, the peptides have molecular weights far beyond what most descriptors were designed for. This limits the applicability of QSAR techniques, and even analysis of two amino acid dipeptides is challenging.[158] Finally, the results of peptide libraries are a list of active sequences. This makes it difficult to utilize the large number of QSAR classification techniques which require examples of both active and inactive structures.[159] Experimentally, it is possible to sequence inactive sequences but the results will be quite close to random peptide sequences, providing

little information. These challenges limit the use of the large amount of data generated from such experiments.

In this work, we describe a collection of algorithms meant to solve these challenges and simplify the analysis of peptide library data. Most of the algorithms operate on the sequence view of the peptides; each peptide is represented as a string of letters. This is the most relevant perspective in biology, where a consensus motif or sequence is the desired output from a peptide library. It is often the case, however, that certain active sequences do not contain a consensus motif and thus we also describe algorithms which examine such peptides from a molecular perspective using traditional QSAR descriptors.

The algorithms produce four important results. The first is the number of motifs present in a peptide library. The second is the grouping of the sequences based on the number of motifs. Although not discussed in this work, the grouping is enough to produce a substitution matrix for PSI-BLAST to find examples of the grouped sequences in protein databases.[160] The third result is the motifs of the grouped sequences, as determined from a model fitting procedure. These motifs are ultimately the output of a peptide library. The motifs indicate the preferred substrate for the peptide library assay. Finally, if the motif fit is unsatisfactory, QSARs may be calculated to test if other structure-property relationships fit the sequences better than motifs.

4.1.2 Methods

An overview of the techniques of this work is given in Figure 4.1. Sequence-sequence distances are calculated, a distance matrix may be derived from these distances, the matrix is used to separate the sequences into clusters and motifs are fitted to each cluster. Separately, QSARs are calculated on all sequences and we test if the distribution is different than that of the entire peptide library.

K-means clustering was used to cluster the sequences.[161] K-means is a clustering technique which finds the local optimum grouping of sequences for a given number of clusters. It is a "hard" clustering technique, meaning that each sequence may belong to only one

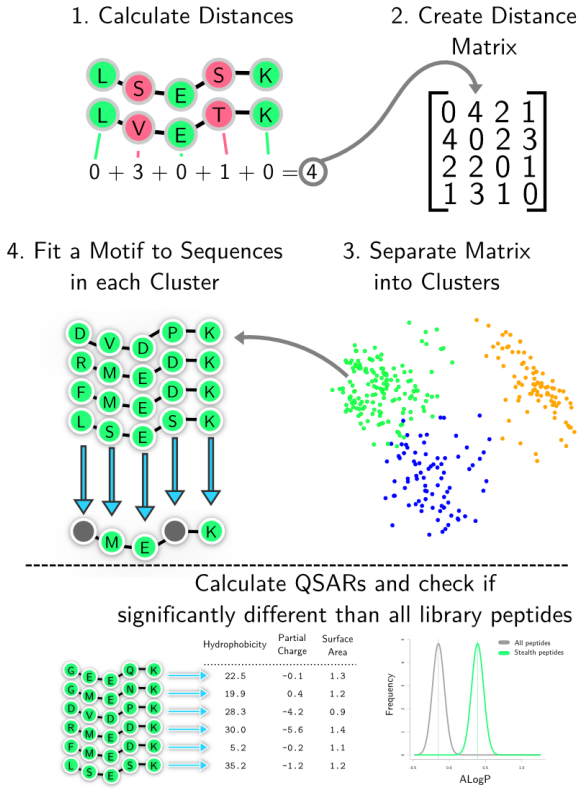


Figure 4.1: An overview of the techniques presented in this work for analyzing peptide libraries. Sequence-sequence distances are calculated (1), a distance matrix is derived from these distances (2), the matrix is used to separate the sequences into clusters (3) and motifs are fitted to each cluster (4). Separately, QSARs are calculated on all sequences and we test if the distribution is different than that of the entire peptide library.

cluster. The Hartigan and Wong algorithm[162] was used as implemented in the statistical computer program R[110] with 20 random starts to account for the local optimization.

Principal component plots were calculated by first computing the eigenvector matrix of the sequence-sequence distance matrix using singular value decomposition and then multiplying the data matrix by the eigenvector matrix. Only the first two columns are retained, which is the data projected onto the principal two components. The implementation of this algorithm in R was used.[110]

The motif model used is derived from the MEME model presented in Bailey[156]. The expectation-maximization (EM) algorithm is used to optimize the motif model on the sequences, where the "hidden" data is the motif starting position.[163] In the implementation of the MEME algorithm, initial guesses based on statistics derived from genomic analysis were used; the frequency of amino acids is non-uniform in naturally occurring proteins. However, in our implementation we use uniform distributions throughout for initial guesses due to the uniformity of solid-phase peptide libraries. The EM algorithm is considered converged here once the sum of the squares of the model parameters of the model changes by less than 0.1%.

The equation to update a motif (M-step) is:

$$m_{kja} = \frac{\sum_i^N z_{ik} \mathbf{1}_{\{a=s_{i(j+k)}\}} + \frac{Q}{\mathcal{A}}}{\sum_i^N \sum_a^{\mathcal{A}} z_{ik} \mathbf{1}_{\{a=s_{i(j+k)}\}} + Q} \tag{4.1}$$

where m_{kja} is the estimated probability of amino acid a occurring at position j in the kth motif. $j \in [1, w]$. w is the motif width. The number of starting positions is $L-w$, where L is the sequence lengths. $k \in [1, L-w]$. z_{ik} is the estimated probability for the ith sequence to start the motif in the kth position. $\mathbf{1}_{\{x\}}$ is the indicator function, which is 1 if the condition x is true. $s_{i(j+k)}$ is amino acid at the $(j+k)$th position in the ith sequence. The other unknown parameter, z_{ik}, is updated (E-step) according to:

$$z_{ik} = \frac{\sum_{j=k}^{k+w} m_{k(j-k+1)(s_{ij})}}{\sum_{k=1}^{L-w} \sum_{j=k}^{k+w} m_{k(j-k+1)(s_{ij})}} \tag{4.2}$$

where $m_{k(j-k+1)(s_{ij})}$ is the estimated probability of the amino acid belonging to the ith

sequence at the jth position occurring at the $(j-k+1)$th position in the kth motif. The initial guesses for z_{ij} and m_{kja} are uniform. The background distribution is not updated as described in Bailey[156]. Instead, it is known to be uniform for solid-phase peptide libraries and is constant $\frac{1}{\mathcal{A}}$. The equation to calculate the pseudocount, a sort of "guess" for the motifs which becomes less important as the amount of data grows, is given below:

$$Q = \min(N, \mathcal{A}) \tag{4.3}$$

where Q is the pseudocount, N is the number of sequences, and $\mathcal{A}$ is the size of the alphabet. Here the alphabet size is the number of amino acids (20) plus a gap character and unknown residue character.

The width of the motif model is chosen *a priori* and is ultimately a decision of the expert using the algorithms. The procedure used in this paper which recovered motifs seen by experts analyzing the datasets is to start with a motif width of three and increase by one amino acid as long as the additional motif positions have one amino acid with greater than 15% probability mass. Another method is to create elbow plots of the log-likelihood as a function of motif width.[164] The starting position, z, of the motifs may be set as to be the same for all sequences or different for each motif. If it is the same for all sequences, the EM algorithm operates on z as a vector and if it's different for each sequence a matrix is used for z. Throughout this work, it is not assumed that the motif starts at the same position for each sequence.

The Wilcoxon paired signed rank test (Wilcoxon T-test) was used as implemented in R.[165] It is used to determine if two populations are significantly different. In order to estimate the QSARs on inactive sequences, 500 sequences were randomly generated assuming a uniform distribution of amino acids. QSARs were calculated on the randomly generated sequences for the Wilcoxon T-test.

The QSARs reported in this research are intentionally simple and meant to illustrate the methods. The counts of groups were calculated as follows: the basic groups are H, R, and K. The acid groups are E and D. The aromatic groups are W, Y, and F. The polar groups are S, T, C, P, N, Q, K, R, H, E and D. The charged groups are E, D, H, K, and R.

Matrix	Agglomerative	K-means
Hamming	193	264
BLOSUM50	276	283
BLOUSM62	279	282
BLOSUM85	280	**283**
BLOSUM90	266	274

Table 4.1: Comparison of different clustering and substitution matrix types for clustering the SHP2 Dataset. The table entries are the number of peptides which match the clustering done by experts in[149], which contains 331 peptide sequences. The version used in the main text is bold.

ALogP is the average AlogP of the amino acids in the sequence, as calculated according to Ghose and Crippen[166] and implemented in the Chemistry Development Kit.[167]

Comparison of methods

A comparison of the choice of substitution matrix and clustering methods are given in the Tables 4.1 and 4.1. A hamming distance is a substitution matrix where all off-diagonal elements are 1 and the diagonal is 0. This provides none of the chemical similarity information encoded into a BLOSUM substitution matrix. Based on these results, the K-means clustering method was selected and the BLOSUM85 substitution matrix was selected.

4.1.3 Results

Five previously published datasets are used to test the analysis techniques presented. The first three come from Sweeney *et al.*[149] and the second two come from Chen *et al.*[168]. All five peptide libraries target different phosphatase enzymes and are solid-phase peptide

Matrix	Agglomerative	K-means
Hamming	86	151
BLOSUM50	108	101
BLOSUM62	109	102
BLOSUM85	86	**102**
BLOSUM90	93	107

Table 4.2: Comparison of different clustering and substitution matrix types for clustering the TULA-Pre Dataset. The table entries are the number of peptides which match the clustering done by experts in[168], which contains 151 peptide sequences. The version used in the main text is bold.

libraries. The algorithms are general and do not require that the data come from solid-phase peptide library techniques[149] or phosphatase enzymes.

Generally multiple sequences will be active in a peptide library. If most of these sequences display a similar sequence, they may be considered to contain a common sequence or motif. The motif may be determined by an expert examining the collection of active sequences. In some cases, there may even be multiple unrelated motifs. However, it is difficult to effectively and objectively categorize the sequences into a small number of motifs and also choose the number of motifs. This is also complicated due to the large number of sequences generated from solid-phase peptide libraries (100-500). One technique which is particularly suited to alleviating this is clustering. If the number of motifs is known in advance, it is possible to collect the sequences into groups based on how chemically similar they are.

K-means is a clustering technique that groups points, or sequences, into groups based on their similarity. K-means clustering performs operations on a symmetric $N \times N$ distance matrix, where element i, j encodes the distance from sequence i to j. Although it is possible to consider variable length peptide libraries using multiple sequence alignment tools,[169] we'll

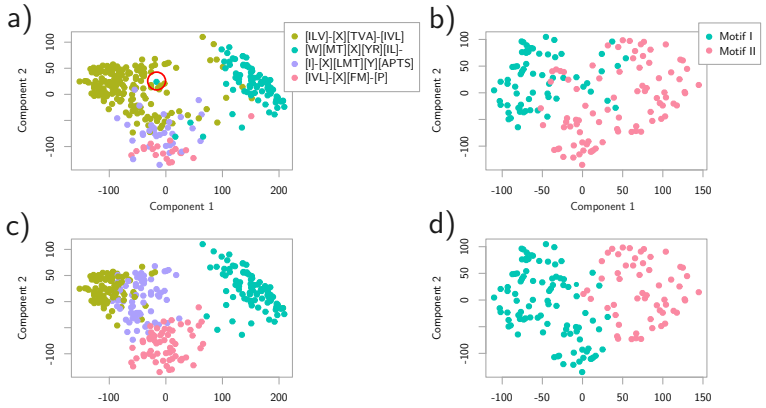


Figure 4.2: A plot of the two principal components of the distance matrix of the peptide library data from two previously published datasets[149,168]. The colors in a and b represent the segregation of sequences according to Sweeney *et al.*[149] and Chen *et al.*[168], respectively. The consensus sequences developed by Sweeney *et al.*[149] are in panel a in the legend. No consensus sequences were reported for panel b but they were separated by Chen *et al.*[168] into two motifs. Square brackets denote the motif positions and "-" indicates a non-motif, or background, position. The amino acids in brackets are ordered by most frequently occurring to least in the motif positions. The results from K-means clustering are shown in panels c and d. The K-means clustering produces similar results, capturing the key features of the sequences. The point circled in red in panel a is a possible misclassification.

assume that each sequence has the same length within each peptide library. The distance between two sequences may be calculated using substitution scores according to:

$$d_{ij} = \sum_{k=1}^{l} f(s_{ik}, s_{jk}) \tag{4.4}$$

Where l is the length of the peptide, $f()$ is the substitution function (for example, the BLOSUM62 matrix[170]) which measures how chemically similar two amino acids are, and s_{ik} is the kth position of sequence i. If a BLOSUM matrix is utilized for the substitution function, the distance matrix must be shifted. Positive numbers indicate similarity in BLO-

SUM and low negative numbers indicate dissimilarity. Thus the shift should make the most positive distance be the zero element and the most negative element become as large as the largest difference in the un-shifted distance matrix. BLOSUM62 is used in this work. The distance matrix may be visualized using a principal component analysis. This is shown for two previously published peptide libraries in Figure 4.2. The colors in Figure 4.2a and 4.2b represent the clustering of the sequences as done by the authors (experts), and not an algorithm. It can be seen that the distance matrix does a good job of separating the clusters which were chosen by experts. Some possible mistakes in the classification also become visible as well. For example, sequence 133 which is circled in Figure 4.2a, begins with isoleucine yet was classified into a cluster where each other sequence begins with tryptophan. The results using the K-means clustering are shown in bottom two panels in Figure 4.2c and 4.2d. The clustering finds similar patterns but is automatic, reducing the risk of accidentally misclassifying a sequence. The K-means algorithm operates on the entire distance matrix, yet only the principal two components are shown in Figure 4.2, thus overlapping clusters there do not necessarily mean they are overlapping in the other dimensions which are not shown.

Choosing the number of clusters, and thereby the number of motifs, is an ongoing research problem. One way to choose the number is the "elbow" technique, which is utilized here.[164] In the elbow technique, some measure of the goodness-of-fit as a function of the number of clusters is plotted. The goodness-of-fit generally increases as a function of the number of clusters; more model parameters create a better fit. The number of clusters just before the goodness-of-fit flattens may be chosen as the number of clusters. This involves searching for an elbow in a plot, hence the name "elbow" technique. This is shown in Figure 4.3 for a collection of previously published peptide libraries. The solid vertical line indicates the number of clusters as determined from experts. The dashed vertical line indicates the number of clusters as determined from the elbow technique. In Figure 4.3a the technique agrees well with the choice of experts, again differing by one. In Figures 4.3b and 4.3c the authors did not consider multiple motifs. In Figure 4.3d the technique disagrees

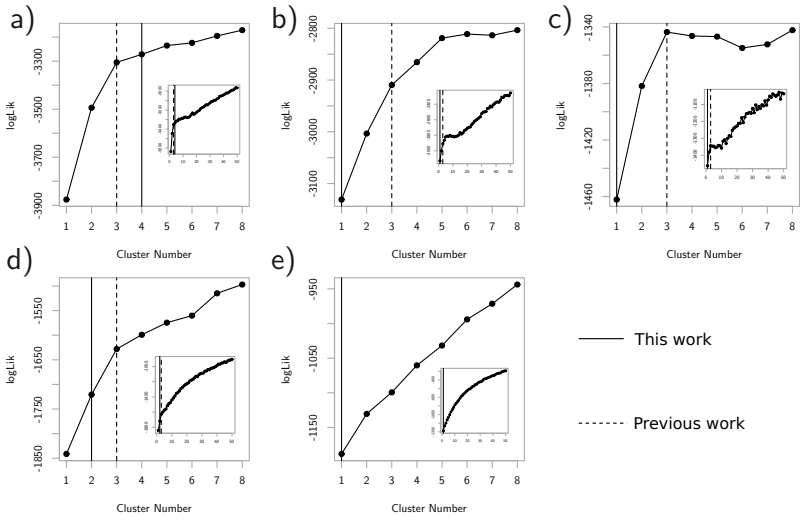


Figure 4.3: The elbow plots of the number of clusters or motifs in the five previously published datasets[149,168]. The y-axis is the log-likelihood of the motif models over the dataset and the x-axis is the number of clusters. The dashed vertical line indicates where we choose the number of clusters, based on the elbow technique. The insets show the same plot but to a larger number of clusters. The elbows are more visible in these plots. The solid vertical lines show the choice of the experts who originally analyzed the datasets. Panels b, c, and e were not considered to have multiple motifs by experts. Panel e is not considered to have multiple motifs based on information described in text. These plots provide a justifiable and straightforward technique for choosing the number of motifs.

with the choice of the experts by one motif. Only one cluster was chosen for Figure 4.3e because increasing the cluster number did not significantly change the motifs, which is another test to determine the number of motifs. Although the process is in some ways still subjective with elbow plots, they do provide an easily communicated and justifiable method to choose the number of motifs.

The clustering algorithms described above assign each sequence to a particular cluster. The next step in the analysis is to extract the motif or consensus sequence. Here, we use a modified version of MEME.[156,157,171] MEME is an expectation-maximization algorithm that identifies the motif in a set of sequences and the location of the motif in the sequences. The algorithm optimizes the likelihood of a particular motif, which is a measure of how well the proposed motif fits the data. The likelihood for a collection of n sequences, each of length l, given model M is given by:

$$\mathrm{L}(S;M) = \prod_{i=1}^{N} \prod_{z=1}^{l-w} \left[\prod_{j=0}^{z-1} \Pr\left(s_{ij} \mid b\right) \right] \left[\prod_{j=z}^{z+w} \Pr\left(s_{ij} \mid m_j\right) \right] \left[\prod_{j=z+1}^{w} \Pr\left(s_{ij} \mid b\right) \right] \tag{4.5}$$

Where s_{ij} is the jth position of the ith sequence, b is the background model, which models the non-motif positions, m_j is the jth position of the motif, N is the number of sequences, l is the length of the sequences, w is the length of the motifs, and z is the starting location of the motifs. Each probability, $\Pr\left(\cdot \mid \cdot\right)$, is a normalized vector of probabilities, with one probability for each possible amino acid occurring. Following the general EM algorithm, this equation is not maximized directly, but instead the expectation of its log over the possible starting positions (z) is maximized. Details of general EM algorithms may be found in Dempster *et al.*[163]. The motif model equations produce a $w \times A$ matrix, where w is the width of the motif and A is the number of possible amino acids (typically 20). Each column in the matrix represents the probabilities of each amino acid at that motif position. If the motif position is not variable, then the model reduces to the proportion of each amino acid at the motif positions. This is the usual quantity analyzed for determining motifs from a collection of sequences. See Sweeney *et al.*[149] or Chen *et al.*[168] for examples. The expected

log-likelihood, the maximized quantity in fitting the model, is the goodness-of-fit used above in the elbow technique.

The results of the motif models on a collection of 5 datasets are shown in Figure 4.4. Figure 4.4a may be compared with the clustering as accomplished by experts and their motif choices in Figure 4.2a. Although there are some minor differences, the motifs and clusters are quite similar. The most significant partition is between the blue and other clusters. This is captured in both Figures 4.4a and 4.2a. The advantage of the technique presented here is that the process takes a few seconds and the effect of changing the number of motifs and motif width may be tested just as quickly. Figure 4.4b shows the effect of segregating the sequences into clusters in contrast to what is presented in Sweeney *et al.*[149], where there is no separation. The motif model determined the last three residues to be the most important, with the blue points being the most conserved motif. The tyrosine in position 5 is unique in that cluster. In Figure 4.4c, it is possible to see a significant difference between the three motifs in positions two and three, where the green cluster shows a hydrophilic serine in position two and the blue cluster shows an aromatic tyrosine. The red cluster shows a threonine (similar to the serine in the green cluster) in position two but no aromatics in the third position.

A complementary approach to motif searching is to examine quantitative structure-activity relationships QSARs.[172] QSARs are functions which take a peptide sequence as an input and output a number representing some property of the peptide. For example, the length of a peptide sequence is a QSAR. QSARs may be more appropriate for analyzing peptide libraries when the chemical properties of the peptides seem more important than the sequence. Searching for a motif is relatively simple, because if a pattern appears multiple times in active sequences then it is likely significant (assuming a uniform background distribution of amino acids). However, a QSAR may contain the same value for each active sequence simply because all peptides have that property. A trivial example would be the number of chlorine atoms being zero for all active sequences. There are no amino acids with chlorine atoms and thus all peptides in the library would have the same QSAR value. A

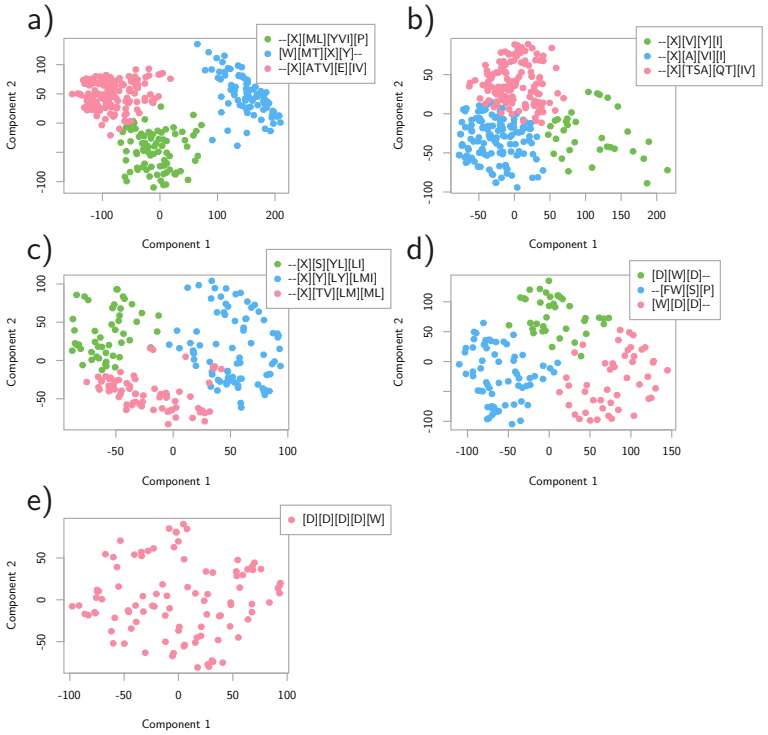


Figure 4.4: A plot of the two principal components of the distance matrix of the peptide library data from five previously published datasets. The colors correspond to the motifs shown in the legend. Square brackets denote the motif positions and "-" indicates a non-motif, or background, position. The amino acids in brackets are ordered by most frequently occurring to least in the motif positions. The numbers of motifs were chosen according to Figure 4.3. Panels a and c may be compared directly to experts' clustering and motif choices shown in Figure 4.2.

QSAR	p-value TULA-2 Pre	p-value TULA-2 Post
Basic Group Count	$6.71 \cdot 10^{-4}$	$2.88 \cdot 10^{-7}$
Acid Group Count	$8.71 \cdot 10^{-2}$	$7.77 \cdot 10^{-22}$
Polar Group Count	$2.77 \cdot 10^{-15}$	$4.57 \cdot 10^{-3}$
Aromatic Group Count	$2.65 \cdot 10^{-48}$	$4.60 \cdot 10^{-29}$
Charged Group Count	$1.11 \cdot 10^{-1}$	$2.22 \cdot 10^{-5}$
ALogP	$7.38 \cdot 10^{-9}$	$1.83 \cdot 10^{-10}$

Table 4.3: QSARs from two peptide library datasets. A lower p-value indicates a significant QSAR, as determined from a Wilcoxon T-test. A significant QSAR means that the sequences which were active in the library have a QSAR value significantly different than what was seen in the inactive sequences.

more subtle example is the surface area of the molecules, which is similar because it is mostly a function of the length of the peptides and all peptides are generally the same length in a peptide library. Thus, it is important to calculate QSARs on both the active sequences and the inactive sequences. If the distributions of QSARs are significantly different, then it may be said to be a relevant QSAR. This significance may be calculated by using a Wilcoxon test,[165] a non-parametric version of the Student's t-test. This hypothesis test is used to compare the median of two distributions and results in a p-value. A lower p-value means the distributions are different and that the QSAR is significant. The only complication is calculating a QSAR on the distribution of inactive sequences, which is unknown. One approximation that may be easily checked during the assaying phase of peptide libraries is that the active sequences are a small fraction of the peptide library. If that is the case, and due to the lack of bias in solid-phase peptide library synthesis, then we may estimate the inactive sequences as uniformly random sequences. These may be randomly generated computationally to construct a distribution of QSARs on the inactive sequences for the

Wilcoxon test.

Figure 4.4e, which shows the TULA-2 Pre dataset, is a good example of when QSARs are more appropriate than motifs. The motif in Figure 4.4e is four aspartic acids and a tryptophan. However, that particular motif actually does not appear in the list of active sequences. In fact, most of the active sequences contain two acids and one aromatic group. A small set of QSARs on the active sequences were calculated along with their p-values, which are shown in Table 4.3 under the TULA-2 Pre column. It is clear that the number of aromatic groups and acid groups is a significant QSAR. For comparison, the p-values are shown for the same analysis on the TULA-2 Post peptide library from the same publication. That dataset is shown in Figure 4.4d as well. The number of acidic groups is no longer a significant QSAR, but the number of aromatic groups still is. This is corroborated by the motif shown in Figure 4.4d, where each of the motifs contains an aromatic amino acid but not necessarily an acidic amino acid. QSAR analysis provide a complementary technique for finding patterns in peptide libraries when there is no clear motif.

The algorithms presented here, along with additional analysis techniques for analyzing peptide library data, have been packaged into a plug-in for R called "peplib." It is available on the CRAN repository (http://cran.r-project.org) along with a manual describing its us and further information may be found at http://peplib.org.

To summarize the techniques presented, analysis of peptide libraries should begin with elbow plots to estimate the number of motifs present in a dataset. Next, the sequences should be clustered and a motif model be fit to each cluster to describe the motifs. If the motifs have many variable positions, then QSARs may be fit instead to the sequences.

4.1.4 Conclusions

Solid-phase peptide libraries are powerful experimental techniques for quickly screening millions of peptides for activity. Analysis of such libraries is generally accomplished by experts analyzing hundreds of sequences by hand and using intuition for the number of motifs and consensus sequences. Reliable and freely available algorithms have been described here to

analyze the data generated from such experiments. We have described how to choose the number of motifs in a peptide library, how to group sequences together based on similarity, how to extract the motifs from similar sequences, and how to analyze the chemical properties of the peptides with QSARs. The algorithms compare well with the work of experts in the field on five previously published datasets and excel in their speed and consistency compared with the current techniques. Implementations of these algorithms, documentation and a tutorial for them may be obtained at http://cran.r-project.org/web/packages/peplib/. An online application is also available at http://peplib.org. It allows researchers to use a basic version of the algorithms presented here on their data.

4.2 Modeling QSARs and Motifs Simultaneously with Graphical Models

Previously the analysis of peptide libraries has either been only motifs or only QSARs. In the next section, I develop models that combine both and provide an introduction to the power of graphical models, which are quickly permeating the fields of speech recognition, computer vision, bioinformatics, and other fields. One of the most noticeable advantages of graphical models is that the large and difficult to derive equations (Equation 4.2 and 4.1) are computed algorithmically. Thus, models may be quickly tested without the need to re-derive equations.

4.2.1 Introduction

Graphical models are finding incredible success in speech recognition,[173] computer vision,[174] bioinformatics[175,176] and other areas of machine learning.[177] The graphical modeling framework allows common models (e.g., PCA, LDA, LSA), dynamic models (e.g., genomics models, speech recognition), and models with complex structure to all be specified, trained, and applied using the same algorithms. The power of graphical models lies in their ability to express sophisticated model structure with completely general training techniques. Thus little to no time is spent on training algorithms for a given model and instead time may be spent on searching for the best models. Furthermore, due to the generality of the training

algorithms, their efficiency has been tuned to the point that they can be applied to study entire genomes.[178] A recent review of them may be found in Bilmes[177].

Graphical models are ideally suited for QSAR modeling due to their ability to encode chemical knowledge, their speed, and ability to design interpretable models. Graphical models can be constructed to have a large number of constraints that allow one to incorporate chemical knowledge into the model. For example, when applying a graphical model to small drug-like molecules, one could specify that the molecule must have a molecular weight below a cut-off and that at least two QSARs must be in a certain range. Such constraints are difficult to embed into linear discriminant analysis, for example, and require a change to the model fitting procedure. There is no such requirement in graphical models due to the generality of their training procedures. That generality also means that very fast algorithms have been developed that make use of such constraints to reduce the training space. This has made graphical models one of the techniques used on the massive ENCODE database,[178] a deep sequencing project of the entire human genome. Finally, the combination of the constraints and speed allow models to be constructed that are easily interpretable.[175] For example, it is possible to embed a dimensionality reduction into a graphical model so that we may simultaneously find a reduced dimension for interpreting a QSAR model and fit the model.

Another benefit of the use of graphical models is their ability to do integrative modeling.[179] Integrative modeling is is the combination of heterogeneous data into a unified model. For example, combining the sequence data of a peptide with chemical descriptors of a peptide is a challenging task. In a graphical models, two parts of the model may deal with the different data types and be connected through probability distributions. This has an advantage over other model combination techniques, such as consensus modeling,[180] in that both models may be simultaneously trained. Applying graphical models to chemistry problems will open new ways of combining data such as bioavailability descriptors, sequence models, and perhaps even simulation results.

Graphical models require the data to be viewed with probability distributions. QSAR

modeling is traditionally unsuited to this because the space of all chemical compounds is both unbounded and difficult to enumerate. Even when restricting a model to a chemical domain of applicability, it is difficult to find all member compounds of that domain so that a normalized probability distribution may be constructed. When modeling combinatorial libraries, the space of chemical compounds is both bounded and known. This allows QSAR descriptors to be constructed with normalized probability distributions using knowledge of the complete chemical space. One particular example of this is peptide libraries and the topic of this work.

As a case study, graphical models are constructed on two datasets. The first is the antimicrobial peptide database (APD).[181] The second is a collection of fragments of sequences from the surface of human proteins. Using graphical models, we ask the following questions regarding these datasets: (a) what molecular descriptors are abnormal for these datasets, (b) how different are the descriptors between the datasets, (c) are there antimicrobial peptides that are similar to human proteins and might that be a predictor of low cytotoxicity, (d) what are the motifs in the antimicrobial peptide database, and (e) what are the components of the best classifier that can be built on the antimicrobial dataset: motifs, descriptors or both. These questions are important from a biological standpoint and by answering them, we also demonstrate the flexibility of graphical models.

In the Methods section, we describe the construction of the datasets and the training procedure used for the graphical models. In the Results section we discuss how to prepare descriptors for use in a probability model, introduce the models to be examined, discuss parameter selection, and present the results for the models.

4.2.2 Methods

Two datasets were used. The first, 'APD', is the complete antimicrobial database (APD) as of April 2013[181] and contains 1,783 sequences. The second dataset, 'Human', is built upon the protein Section 3.2 dataset. All contiguous amino acid sequences of length greater than 4 amino acids present on the surface of proteins from that dataset were tabulated as

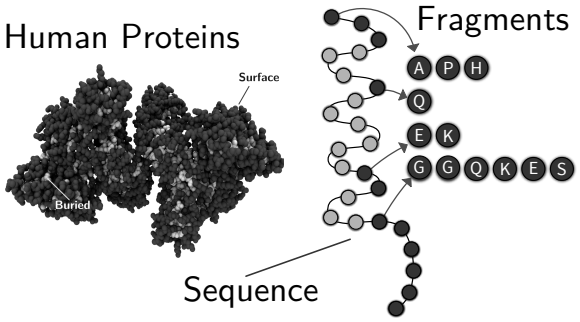


Figure 4.5: Diverse proteins isolated from humans with structure in the protein data bank created a database of 1,162 proteins. The surface was found as described in Section 3.2. Contiguous surface sequences were found (darky gray) and converted into sequence fragments. All with length greater than 4 were used.

independent sequences, as depicted in Figure 4.5. This yields 4,840 unique sequences.

Descriptors, except ALogP, were calculated using the peplib R plugin.[182] ALogP was calculated according to Ghose and Crippen[166] as implemented in the Chemistry Development Kit.[167] To estimate the distribution of ALogP values, the calculation was performed on all peptides from lengths 1 to 3 and on a random sample of 5,000 peptides for each length up to 10. The results of that calculation are available on SQLShare[111] under dataset 'alogp.csv' and username 'whitead'.

Graphical models were trained using the expectation-maximization algorithm[163] as implemented in the graphical model tool kit (GMTK).[177,183] The models were first triangulated and then junction tree inference was used for EM training.[177] All prior distributions were set to uniform, except classification priors which were set to their true values. This ensures that if training on only one class of data, the active class for example, the prior reflects this. EM training was terminated when the expected log-likelihood changed by less than 0.1% per iteration. The training parameters were initialized randomly and EM training was done 5 times for each model. If the minimized expected log-likelihoods differed by more than

1%, 15 more EM trainings were done to ensure the algorithm was not trapped in a local optimum. After EM training, classification prior distributions were replaced with uniform distributions. Prediction was done using the Viterbi algorithm with the trained parameters as implemented in GMTK.[183]

4.2.3 Results and Discussion

Normalization of Descriptors

In order to use a structural descriptor in a graphical model, it must be converted into a form that may be described by a probability distribution. An additional challenge for modeling combinatorial libraries is that the descriptors can only adopt the values of the chemical space of the library. This is difficult for a descriptor such as the number of chlorine atoms in a library with no chlorines or a library where every compound has a single chlorine atom. Thus, the first step towards converting descriptors into graphical models is converting them to a form that corrects for the bias in the chemical space and allows them to be described by a probability distribution.

The probability that a descriptor $f(\cdot)$ equals a value x in a molecule c is given by:

$$\Pr(f(c) = x) = \frac{1}{Z} \sum_i w_i \mathbf{1}_{\{f(c)=x\}},\ Z = \sum_i w_i \tag{4.6}$$

where Z is the partition coefficient, w_i is the unnormalized probability (weights) of observing the ith compound in the chemical space, and $\mathbf{1}_{\{\}}$ is the indicator function. Knowledge of the complete probability distribution means one may calculate the significance of a given descriptor value in the active set using a student's t-test or a non-parametric (do not assume normal distribution) Wilcoxon significance test.[165] No knowledge of all inactive compounds or quantitative activities is required; a significant benefit of this representation. Further, the compounds may be weighted by their synthetic difficulty, recognizing the fact that the experimentally active compounds are likely chosen with a bias. However, one must be able to to enumerate all synthetically feasible structures, which is generally unbounded and unintegrable.

When modeling combinatorial libraries, the space of chemical compounds is both bounded and known. This allows QSAR models to be constructed with normalized probability distributions using knowledge of the complete chemical space. The combinatorial library type for this work is peptide libraries and the same techniques that apply here apply to combinatorial small molecule libraries. The weights of each compound are unity since peptide libraries have little to no synthetic bias. The partition coefficient simply becomes A^l, where A is the size of the alphabet (generally 20 for amino acids) and l is the length of the amino acid sequence.

Constructing probability distributions for group-wise additive descriptors follows two approaches. When $l < 3$, all possibilities (20^3) may be enumerated to create a probability distribution. When $l > 3$, the probability distribution may be approximated as a sum of l identical normal distributions. The approximation is accurate, provided the number of 0's is low (e.g., the number of sulfur atoms in the peptides will not fit into this approximation). The mean of the normal distributions is the mean (μ) of the descriptor calculated on the amino acids and the variance (σ) is calculated likewise. The sum of the l normal distributions will have a mean of $l\mu$ and variance $l\sigma^2$. For non-groupwise distributions, the probability distribution may be estimated by sampling from the combinatorial library where the sampling is done according to the weights w_i.

Once the probability distributions are known for each descriptor across the entire combinatorial library, descriptors for the active compounds may be converted into a quantile score between 0 and 100. This is done by first quantiling the probability distributions for each descriptor across the whole library and then descriptors calculated on the active compounds are ranked based on those quantiles. For example, if the number of double bonds has a score of 95 for a compound, that means it is in the 95th percentile relative to the entire peptide library. This approach has three benefits. First, it is now immediately obvious if a descriptor is at an extreme value. Second, now when examining multiple descriptors, their range corresponds exactly to their span of the entire combinatorial library. Thus, if a descriptor range is 5–95, it is not significant. If it is within the range of 20–25, then

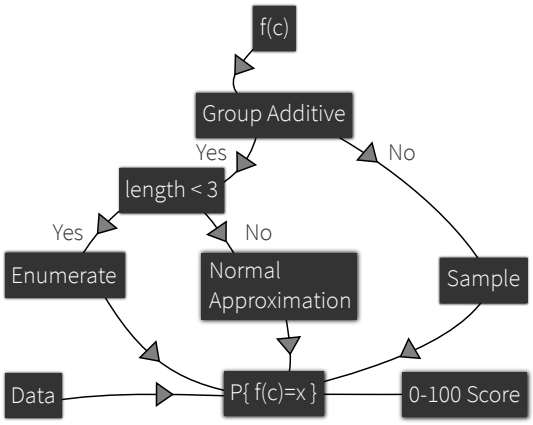


Figure 4.6: A flowchart for converting a descriptor, $f(c)$, into a score from 0–100 that both removes biases from the chemical space of the library and normalizes it for use in a graphical models.

the descriptors occupy a range that only 5% of the chemical space of the library occupies. Third, the effect length on the descriptors may be removed by only comparing descriptors against uniform length probability distributions. For example, if there are sequences from lengths 3–10 in a library, the descriptors may be calculated relative only to sequences of the same length. Then a score of 5 is interpreted as in the bottom 5% relative to sequences of the same length. If this is not desired, only the probability distribution on the longest 2 lengths need to be calculated since that corresponds to 99.75% $(1 - 1/20^2)$ of the possible values. This process is depicted in Figure 4.6.

A set of group-wised additive descriptors were calculated on two datasets and are shown in Table 4.4. The left values are the 2.5% quantile, the middle are the median, and the right are the 97.5% quantiles. As corroborated in Section 3.2, the number of charged residues is indeed high on human protein surfaces. This is also reflected in the high water solubility (low ALogP values). It is also interesting that there is no dominant net charge in one direction or another for human protein surfaces. Finally, the number of aromatic residues

Dataset	ALogP	HB Acceptors	BH Donors	Charged Groups
Human	(0, 31, 92)[a]	(7, 65, 99)	(5, 57, 99)	(8, 77, 99)
APD	NA[b]	(0, 8, 94)	(0, 14, 99)	(1, 33, 99)

Dataset	Polar Groups	NonPolar Groups	Aromatic Groups	Net Charge
Human	(10, 66 99)	(1, 33, 90)	(11, 18, 83)	(0, 50, 99)
APD	(0, 24, 97)	(3, 76, 100)	(1, 13, 81)	(29, 90, 100)

Table 4.4: [a] 95% confidence interval from quantiling. Middle value is median. [b] ALogP is not calculated due to its poor correlation at long peptide lengths found in the APD datasets.

is low which is expected since the number of aromatic residues is low across all proteins. The APD dataset shows an abnormally low number of charged residues and a high number of non-polar groups. Despite the lower number of charged residues though, the charges are extremely skewed towards positively charged. This is consistent with past analysis of antimicrobial peptides.[184–186]

Graphical QSAR Classifier

A graphical 2–state mixture model is used to classify sequences using the descriptors in Table 4.4 and is shown as a graphical model in Figure 4.7a. When $c = 0$, all the descriptors are drawn from whole chemical space descriptor probability distributions. Due to the descriptors scores used, this is a uniform distribution from 0–100. When $c = 1$, the active state, the descriptors are drawn from the descriptor distributions from a distribution trained to match data. With simple graphical model techniques and normalized descriptors, a powerful model-based classifier has been constructed which is similar to state-of-the art model-based classification techniques.[187]

Three descriptors were chosen based on inspecting Table 4.4: net charge, number of non-polar groups, and number of charged groups. The classifier was fit to the Human dataset

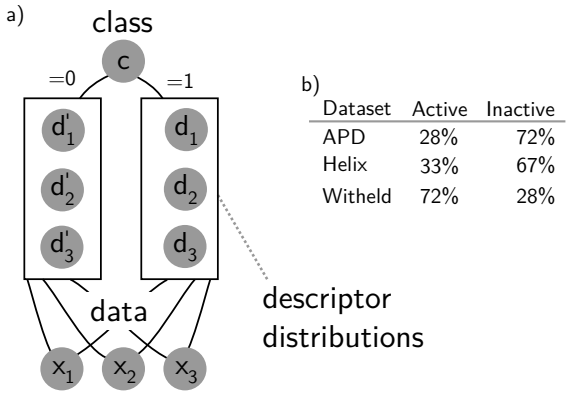


Figure 4.7: Panel a is the graph of a 2-state classifier that fits 3 observed descriptors to either distribution 0 or 1. Panel b is the results of the graphical model trained to the Human dataset. Perfect fit would be 100% active on the withheld row. Withheld is 20% of the Human dataset which was not used in training.

using EM training on 80% of the dataset. Classification was done with the Viterbi algorithm. The confusion matrix, which shows the classifier performance, is shown in Figure 4.7b. Only 28% of the APD dataset matches the descriptors seen in Human dataset. This shows antimicrobial peptides are different from human proteins surfaces, which are thought to be optimized for minimal nonspecific interactions.[61,64] The classifier's performance on the withheld 20% of the data is 72% correct classification; relatively poor. This is likely due to the large heterogeneity in the Human dataset.

There were 8 antimicrobial peptide sequence tied for most similar to human proteins by the classifier. The shortest two (typically easiest to synthesize) are "GWMSKIASGIGT-FLSGMQQ" (ADP ID:AP00167) and "FLPILGNLLSGLL" (ADP ID:AP01725). The first sequence, called Phylloxin, was isolated from hylid frogs of South America and has activity against both gram-positive (*Bacillus megaterium*, *Corynebacterium glutamicum* , *Micrococcus luteus*) and gram-negative bacteria (*Rhizobium meliloti* , *Escherichia coli*).[188] The

second sequence was isolated from *Rana boylii* frogs from Oregon, USA and also has activity against both gram-positive (*Staphylococcus aureus*) and gram-negative (*Escherichia coli*) bacteria.[189] They are similar to human proteins due to their low net charge and low number of non-polar residues. Due to the connection between low protein adsorption and human protein surfaces,[61] these two sequences may be good candidates for antimicrobial surface coatings.

Classifying Antimicrobial Peptides with Motifs

The most important tasks in QSAR modeling is predicting activity. In this section we develop graphical models that predict if a given sequence should be an antimicrobial peptide or not. One challenge of approaching this task is that there is no database of non-antimicrobial peptides and thus no data to negative examples with which to train a classifier. Torrent *et al.*[190] approached this problem by using sequences not reported to have activity, which may be a good assumption since antimicrobial peptides are rare. We use a similar approach to ensure our models are not overfit, but we do not use them in training. A decoy dataset is generated by replacing each residue in the APD dataset with a randomly selected amino acid. The models are then tested to see if they correctly reject the decoys.

The first model is the 2-state mixture model as described above and shown in Figure 4.7. It was fit to 80% of the APD dataset with the EM training procedure and classification was done via the Viterbi algorithm on the withheld 20% of the data. The same descriptors were used as above (net charge, count of non-polar groups, and count of charged groups). The median prediction success of 20 training/classification runs is 86.2%. This simple 3 descriptor classifier has performance near as good as the best performing examples in the literature (88–94%).[190–194] The classifiers performance at rejecting decoys was 73.0%, which indicates the model is overfit to some extent.

Next, a graphical motif model is constructed and tested as a classifier on the APD dataset. An emphasis is placed on keeping the model interpretable. A motif model with the following attributes was chosen:

1. There are 0 to k motif types
2. Sequences may contain 0 to ∞ motifs
3. Motifs may not be partially expressed
4. Non-motif residues in a sequence are drawn from the same distribution (background)
5. Motifs are of fixed length w and regularized to be sparse
6. The probability of a motif starting at a sequence position is iid

The changes that are unique to this description relative to other motif models[157,182] are the regularization of motifs, tied background distribution, uniform motif start probability, and the ability to deal with variable length sequences. The regularization forces the motifs to be sparse so that each motif position only has one or two possible residues. This makes motif interpretation more intuitive. The tied background distribution reduces the number of model parameters by $(k-1)(A-1)$, where A is the number of amino acids. Such a change greatly complicates traditional algebraic analysis of the model and is only practical with graphical models. The uniform motif start probability reduces the number of parameters by $k(l-1)$, where l is the length of the sequences. The ability to deal with variable length sequences without pre-alignment is a significant feature and is what allows modeling of the highly heterogeneous APD.

Classification is done by adding a membership variable that represents the probability a sequence is antimicrobial or not. Now the model identifies motifs, classifies sequences based on which motif they contain, and also predicts if a sequence will have antimicrobial activity or not. The complete dynamic graphical representation of this model is shown in Figure 4.8a and the inset Figure 4.8b. The complexity of the model is due to the large amount of structure embedded into the model, as seen from the high number of structure variables (nodes). For example, in order to enforce no partial expression of motifs, the

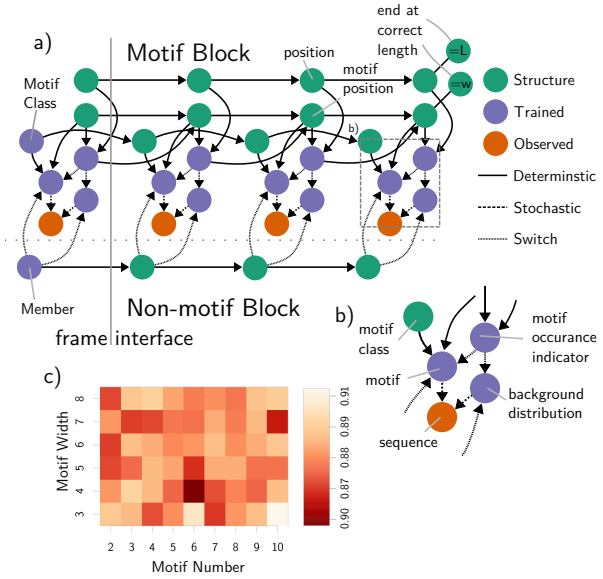


Figure 4.8: Panel a is the graph of the motif-model used. The middle frames of the model may be repeated as many times as necessary to fit the length of a sequence. Panel b is an inset showing how the motif indicator can switch the model to generate data from a motif or from the background distribution. Panel c is the prediction accuracy from Viterbi decoding on withheld data as a function of the motif width and motif number.

motif position must be tracked and the position along the sequence. The structure does not complicate interpretation of the model. The trained parameters will be a list of motifs, the number of motif classes, the motif width, which motif class each peptide belongs to, and the distribution of amino acids not described by motifs. This model is an example of a dynamic Bayesian multi-net.[195]

Understanding Figure 4.8a is most easily done by interpreting it as a generative model; pretending it is generating the data as opposed to being fit to data. First, it is randomly decided if motif sequence data or non-motif sequence data will be generated at the member node. If non-motif data is to be generated, the member node switches the data to be drawn from the uniform background distribution for the first sequence position, also known as frame. The member node is copied to the next frame of the model and again the data is drawn from the uniform background distribution. This repeats until enough frames are generated. If instead the member is set to generate motif data, the sequence data in the first frame is generated either from the trained background distribution or a motif. The sequence data is drawn from the motif if the motif occurrence indicator has been switched on. Now we must traverse more of the graph to see where how the motif is generated. The motif random variable is a kxwxA dimension distribution. The motif class is the first index and chosen randomly in the first frame. The motif position is at 0 in the first frame, but increments each frame when a motif is being expressed. The last index is the number of amino acids. A similar walk along the graph can be used to understand the other parts of the model. The complete model specification is given in a dynamic graphical model grammar[183] in Appendix A.

The models were trained using EM training on 80% of the APD data as described in the Methods section. The figure of merit for the models is classification error on withheld 20% data. The adjustable parameters are the motif width (w) and the motif number (k). 40 combinations were trained ($w = [3, 8]$, $k = [2, 10]$) 5 times each and the median classification error on these 40 combinations was 10.9%, with $k = 10$, $w = 3$ being the best at 9.0% classification error and $k = 6$, $w = 7$ being the worst at 11.9%. Those results are

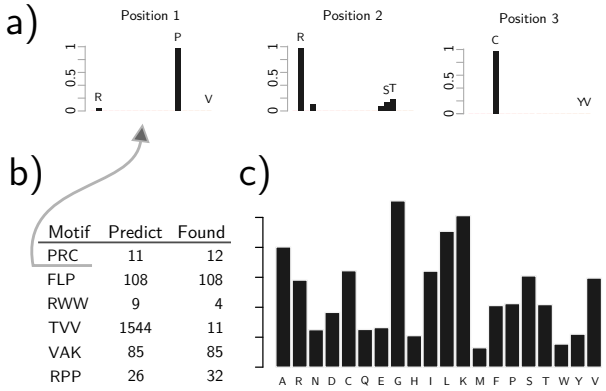


Figure 4.9: Panel a is the probability of each amino acid in the 3 positions of a motif. Notice the sparsity from regularization. Panel b is the list of motifs found by the model. The predict column is the number of sequences which are predicted to be members of that motif class by the motif model. The found column is the number of sequences that contain the motif. Panel c is the background distribution of amino acids from the motif model. The y-axis is probability.

shown in Figure 4.8. In order to determine if the motif model portion is most important for performance or the background distribution (amino acid composition), training of the background distribution was disabled. The same 40 combinations had a significantly higher median classification error of 94.0%. There was also a stronger trend in the motif width and length. The final parameters chosen were $k = 6$ and $w = 3$ since they were have the second best prediction error (9.4%) and a smaller number of motifs than the best ($k = 10$). The fraction rejected decoy sequences set was 91.3%, indicating much lass overfitting than the QSAR classifier. This ability to reject decoys is inherit in the model due to the incorporation of the "no-motif" block in Figure 4.8 which models sequences as uniformly random.

The ability to interpret the model may be seen in Figure 4.9. Figure 4.9a shows the probability distribution from the first motif. The regularization operates as expected and the motifs are sparse; only one or two amino acids have non-zero probability. Figure 4.9b shows

all the motifs. The predict column shows the number of sequences which were assigned to each motif by the model. The next column contains the number sequences which actually contain each motif. We see the model has done well in assigning each motif to the correct sequence based on the close match between the predict and found columns, but the sequences which don't have one of the motifs are simply placed into the TVV motif. Thus, we may see that there are relatively few examples of the motifs discovered by the model. Only two of the motifs found are likely significant, since the number of 3 amino acid combinations is 8,000 and the dataset contains 53,000 3 amino acid positions. Only the FLP, VAK, and possibly RPP motifs are frequent enough to be considered significant. Finally, the background distribution is shown in Figure 4.9c. This may be considered the amino acid composition of the library with the motifs observed in sequences removed. It is different than uniform or the Human database[61] and as mentioned above, contributes significantly to the performance of the classifier. This is not unexpected since amino acid composition is a well-used descriptor for analyzing peptides and proteins.

A classifier has been built which uses motifs and, implicitly, the amino acid composition which works as well as the best previously described classifiers in the literature.[190–194] The significant advantage of the model described here is that it is simple to interpret, simple to modify, and can be trained on the 1,500 sequence APD dataset in minutes.

The last classifier considered combines features from the previous two. It is shown in Figure 4.10. The "omitted motif model" is shown in Figure 4.8. This classifier incorporates the descriptors shown in Figure 4.7 into a new "QSAR" block in the model. The incorporation of descriptors demonstrates the integrative model of graphical models. The complete model specification is given in Appendix A. The model takes in both sequence information and descriptors. A lower motif number ($k = 3$), the same motif width ($w = 3$), and the same descriptors (net charge, number of charged groups, and number of non-polar groups) were used. The lower motif number was based on the marginal difference in misclassifications seen in Figure 4.8c and the results in Figure 4.9b showing that likely only three of the motifs are significant. The model was trained again on 80% of the APD data and its prediction

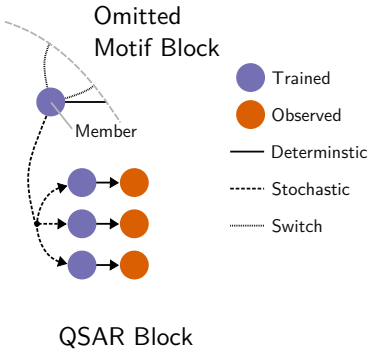


Figure 4.10: Graph of combined QSAR/Motif model. The omitted motif block refers to Figure 4.8.

was tested on the withheld data. The motifs and descriptor distributions were similar to those described in the previous two separate models. Its classification performance on withheld data is similar to the motif model at a median 9.6% vs the 9.4% of the motif model. The QSAR-motif model sees a significant improvement in decoy rejection, predicting 97.8% of randomly generated sequences will not be active vs. 91.3% of the motif model. This demonstrates adding the extra descriptor distributions did not result in over-fitting; in fact it improved the ability of the classifier to filter randomly generated sequences.

4.2.4 Conclusions

The application of graphical models to QSAR modeling has been introduced. Graphical models are flexible and may encode sophisticated chemical knowledge, as seen from the motif models presented. Their flexibility also allows models to be constructed with easy to interpret parameters. This has been demonstrated with the motif models, where regularization forced each motif position to only contain one amino acid, as opposed to previous models where motif positions have non-zero probability on each of the 20 amino acids.[157,182] Finally, the graphical models show excellent classification performance with 90.4% accuracy

on withheld data and 2.2% misclassification error on decoy sequences. This is as good as the more opaque and complex strategies such as multilayer artificial neural networks.[193] As a case-study they were applied to compare human proteins and antimicrobial peptides. It was found that human proteins and antimicrobial peptides differ significantly. Human proteins have more charged groups, less non polar groups, and a lower absolute net charge. Two antimicrobial peptides were found that appear like human proteins and may show potential as a nonfouling and antimicrobial surface. Overall, graphical models provide a significant leap in the type of modeling that can be done in QSAR and the ease in which models can be constructed.

4.3 Chapter Summary

In Section 4.1 I reviewed the state of existing peptide library analysis tools and found there were few in existence, especially for the case of synthetic peptide libraries. A new standard set of techniques was developed by combining existing tools from bioinformatics,[157] machine learning,[161] and statistics.[165] These techniques are well tested and perform well at clustering peptide libraries and finding motifs as judged from their performance on peptide library data from literature. Then, in Section 4.2, I introduced the newly developed graphical models which come from machine speech recognition. These new techniques are capable of both motif discovery and clustering. Additionally, they make excellent classifiers, can combine QSAR and motif techniques, be quickly trained, and are flexible. These models will transform the way QSAR modeling is done. This research has laid the foundation for the modeling of the peptide library experiments being performed by collaborators in my research group and provide a complementary "experiment first, analyze second" approach to the previous Chapters. This research has studied the opposite side of nonspecific interactions. Previously, a particular surface would interact with many unknown molecules. Here, there are many peptides interacting in a particular assay. The common theme is the ambiguity in the structure and molecules involved in the interactions.

Chapter 5

CONCLUSIONS

Maximizing activity of a molecule in biology should always involve minimizing nonspecific activity. It is easy to design a peptide which binds to a target protein. It is difficult to make a peptide which binds *only* to a target protein. Through analyzing the different systems in this research, the following principles for minimizing nonspecific interactions may be derived:

1. Maximize hydration by using zwitterionic or strongly hydrophilic groups and avoiding self-interactions.

2. Create an entropic penalty for binding. For example, by making a polymer more flexible.

3. Minimize known unwanted specific binding. For example, be net neutral to avoid electrostatic interactions with charged proteins. As another example, prevent strong binding with amide groups which are present on protein backbones.

In Chapter 2, I designed self-assembling nonfouling peptides by applying these principles. The resulting sequence matches the best performance of existing protein resistant peptide SAMs reported in the literature. The change in hydration between a primary amine and quaternary amine on zwitterions was studied as well to understand how self-interactions affect hydration. Free energy simulations of solvated salt bridges were conducted to better understand the association between charged amino acids. These simulations are especially important when considering the results from Section 3.2 showed that these amino acids are the most common on the surfaces of proteins and interior of molecular chaperones.

Chapter 3 described research on the surfaces of proteins, how they interact with one another, and how protein folding is affected nonspecifically by molecular chaperones. These different methods led to the design of the same KE nonfouling peptides, as well as new ideas on how proteins remain stable for so much longer than synthesized protein mimics[5,6]. This will allow a more direct study of nonspecific regions and a better understanding of nonspecific protein interactions.

New peptide library analysis techniques were developed in Chapter 4. These new techniques solve the difficulties in utilizing peptide libraries for more complicated systems, where many binding motifs are present. By utilizing new advances in machine speech recognition, we may also combine QSAR and motif analysis into single models. This is critical when using peptide libraries to study nonspecific effects. The introduction of these new graphical models is also a significant breakthrough in the field of QSAR modeling.

The work presented demonstrates that it is possible to model nonspecific interactions despite their ambiguity. They can be modeled indirectly by simulating hydration, directly through bioinformatics, or experimentally with high-throughput screening.

BIBLIOGRAPHY

[1] T. Ben-Yedidia. 1997. Design of peptide and polypeptide vaccines. Curr Opin Biotechnol 8:442–448.

[2] Heidi K. Privett, Gert Kiss, Toni M. Lee, Rebecca Blomberg, Roberto A. Chica, Leonard M. Thomas, Donald Hilvert, Kendall N. Houk, and Stephen L. Mayo. 2012. Iterative approach to computational enzyme design. Proc Natl Acad Sci USA 109:3790–3795.

[3] David Farrusseng, Sonia Aguado, and Catherine Pinel. 2009. MetalOrganic frameworks: Opportunities for catalysis. Angewandte Chemie International Edition 48:7502–7513.

[4] Shaoyi Jiang and Zhiqiang Cao. 2010. Ultralow-fouling, functionalizable, and hydrolyzable zwitterionic materials and their derivatives for biological applications. Adv Mater 22:920–932.

[5] T. Peters. 1985. Serum albumin. Advances in protein chemistry 37:161–245.

[6] Francesco M. Veronese and Anna Mero. 2008. The impact of PEGylation on biological therapies. BioDrugs : clinical immunotherapeutics, biopharmaceuticals and gene therapy 22:315–329.

[7] Douglas S. Clark and Harvey W. Blanch. 1997. Biochemical Engineering (Chemical Industries). CRC Press, 1 edition.

[8] Angus Hucknall, Dong-Hwan Kim, Srinath Rangarajan, Ryan T. Hill, William M. Reichert, and Ashutosh Chilkoti. 2009. Simple fabrication of antibody microarrays on nonfouling polymer brushes with femtomolar sensitivity for protein analytes in serum and blood. Adv Mater 21:1968–1971.

[9] Hana Vaisocherová, Zheng Zhang, Wei Yang, Zhiqiang Cao, Gang Cheng, Allen D. Taylor, Marek Piliarik, Jiří Homola, and Shaoyi Jiang. 2009. Functionalizable surface platform with reduced nonspecific protein adsorption from full blood plasmamaterial selection and protein immobilization optimization. Biosens Bioelectron 24:1924–1930.

[10] R. Erik Holmlin, Xiaoxi Chen, Robert G. Chapman, Shuichi Takayama, and George M. Whitesides. 2001. Zwitterionic sams that resist nonspecific adsorption of protein from aqueous buffer. Langmuir 17:2841–2850.

[11] J. Zheng. 2005. Strong repulsive forces between protein and oligo (ethylene glycol) self-assembled monolayers: A molecular simulation study. Biophys J 89:158–166.

[12] Jason C. Hower, Matthew T. Bernards, Shengfu Chen, Heng-Kwong Tsao, Yu-Jane Sheng, and Shaoyi Jiang. 2009. Hydration of nonfouling functional groups. J Phys Chem B 113:197–201.

[13] Igal Szleifer. 1997. Polymers and proteins: interactions at interfaces. Current Opinion in Solid State and Materials Science 2:337–344.

[14] Hillary J. Taunton, Chris Toprakcioglu, Lewis J. Fetters, and Jacob Klein. 1990. Interactions between surfaces bearing end-adsorbed chains in a good solvent. Macromolecules 23:571–580.

[15] Lingyan Li, Shengfu Chen, and Shaoyi Jiang. 2007. Protein interactions with oligo(ethylene glycol) (OEG) self-assembled monolayers: OEG stability, surface packing density and protein adsorption. Journal of biomaterials science. Polymer edition 18:1415–1427.

[16] Emanuele Ostuni, Robert G. Chapman, R. Erik Holmlin, Shuichi Takayama, and George M. Whitesides. 2001. A survey of Structure-Property relationships of surfaces that resist the adsorption of protein. Langmuir 17:5605–5620.

[17] Qing Shao, Yi He, Andrew D. White, and Shaoyi Jiang. 2010. Difference in hydration between carboxybetaine and sulfobetaine. J Phys Chem B 114:16625–16631.

[18] Lingyan Li, Shengfu Chen, Jie Zheng, Buddy D. Ratner, and Shaoyi Jiang. 2005. Protein adsorption on oligo(ethylene glycol)-terminated alkanethiolate self-assembled monolayers: The molecular basis for nonfouling behavior. J Phys Chem B 109:2934–2941.

[19] Thomas Wyttenbach, Matthias Witt, and Michael T. Bowers. 2000. On the stability of amino acid zwitterions in the gas phase: The influence of derivatization, proton affinity, and alkali ion addition. J Am Chem Soc 122:3458–3464.

[20] D. O. Bayles and B. J. Wilkinson. 2000. Osmoprotectants and cryoprotectants for listeria monocytogenes. Lett Appl Microbiol 30:23–27.

[21] J. O'Callaghan. 2000. Growth of lactococcus lactis strains at low water activity: correlation with the ability to accumulate glycine betaine. Int J Food Microbiol 55:127–131.

[22] C. Rajashekar. 1999. Glycine betaine accumulation and induction of cold tolerance in strawberry (fragaria x ananassa duch.) plants. Plant Science 148:175–183.

[23] Teresa Caldas, Nathalie Demont-Caulet, Alexandre Ghazi, and Gilbert Richarme. 1999. Thermoprotection by glycine betaine and choline. Microbiology 145:2543–2548.

[24] M. Civera. 2003. Molecular dynamics simulation of aqueous solutions of glycine betaine. Chem Phys Lett 367:238–244.

[25] Hana Vaisocherova, Wei Yang, Zheng Zhang, Zhiqiang Cao, Gang Cheng, Marek Piliarik, Jiri Homola, and Shaoyi Jiang. 2008. Ultralow Fouling and Functionalizable Surface Chemistry Based on a Zwitterionic Polymer Enabling Sensitive and Specific Protein Detection in Undiluted Blood Plasma. Anal Chem 80:7894–7901.

[26] T. Tokushima, Y. Harada, O. Takahashi, Y. Senba, H. Ohashi, L. G. M. Pettersson, A. Nilsson, and S. Shin. 2008. High resolution x-ray emission spectroscopy of liquid water: The observation of two structural motifs. Chem Phys Lett 460:387–400.

[27] Haigang Lu, Yuekui Wang, Yanbo Wu, Pin Yang, Lemin Li, and Sidian Li. 2008. Hydrogen-bond network and local structure of liquid water: An atoms-in-molecules perspective. J Chem Phys 129:124512–124516.

[28] Marco Paolantoni, Paola Sassi, Assunta Morresi, and Sergio Santini. 2007. Hydrogen bond dynamics and water structure in glucose-water solutions by depolarized rayleigh scattering and low-frequency raman spectroscopy. J Chem Phys 127:024504–024512.

[29] Chau and A. J. Hardwick. 1998. A new order parameter for tetrahedral configurations. Mol Phys 93:511–518.

[30] Jeffrey R. Errington and Pablo G. Debenedetti. 2001. Relationship between structural order and the anomalies of liquid water. Nature 409:318–321.

[31] Matias H. H. Pomata, Milton T. Sonoda, Munir S. Skaf, and M. Dolores Elola. 2009. Anomalous dynamics of hydration water in carbohydrate solutions. J Phys Chem B 113:12999–13006.

[32] C. Branca, S. Maccarrone, S. Magazù, G. Maisano, S. M. Bennington, and J. Taylor. 2005. Tetrahedral order in homologous disaccharide-water mixtures. J Chem Phys 122:174513–174519.

[33] David W. McCall and Dean C. Douglass. 1965. The effect of ions on the self-diffusion of water. i. concentration dependence. J Phys Chem 69:2001–2011.

[34] Daisuke Matsuoka and Masayoshi Nakasako. 2009. Probability distributions of hydration water molecules around polar protein atoms obtained by a database analysis. J Phys Chem B 113:11274–11292.

[35] C. J. Camacho, Z. Weng, S. Vajda, and C. DeLisi. 1999. Free energy landscapes of encounter complexes in protein-protein association. Biophys J 76:1166–1178.

[36] Thomas Kleinert, Wolfgang Doster, Harald Leyser, Winfried Petry, Veronika Schwarz, and Marcus Settles. 1998. Solvent composition and viscosity effects on the kinetics of co binding to horse myoglobin†. Biochem 37:717–733.

[37] I-Feng W. Kuo, Christopher J. Mundy, Matthew J. McGrath, J. Ilja Siepmann, Joost VandeVondele, Michiel Sprik, Jurg Hutter, Bin Chen, Michael L. Klein, Fawzi Mohamed, Matthias Krack, and Michele Parrinello. 2004. Liquid water from first principles: Investigation of different sampling approaches. J Phys Chem B 108:12990–12998.

[38] Wolfgang Damm, Antonio Frontera, Julian Tirado-Rives, and William L. Jorgensen. 1997. Opls all-atom force field for carbohydrates. J Comput Chem 18:1955–1970.

[39] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. Al A. Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople. 2004. Gaussian 03, revision c.02.

[40] U. Chandra Singh and Peter A. Kollman. 1984. An approach to computing electrostatic charges for molecules. J Comput Chem 5:129–145.

[41] Hans W. Horn, William C. Swope, Jed W. Pitera, Jeffry D. Madura, Thomas J. Dick, Greg L. Hura, and Teresa H. Gordon. 2004. Development of an improved four-site water model for biomolecular simulations: Tip4p-ew. J Chem Phys 120:9665–9678.

[42] Jan Zielkiewicz. 2005. Structural properties of water: Comparison of the spc, spce, tip4p, and tip5p models of water. J Chem Phys 123:104501–104508.

[43] S. Plimpton. 1995. Fast parallel algorithms for short-range molecular dynamics. J Comput Phys 117:1–19.

[44] L. Martínez, R. Andrade, E. G. Birgin, and J. M. Martínez. 2009. Packmol: a package for building initial configurations for molecular dynamics simulations. J Comput Chem 30:2157–2164.

[45] Shuichi Nosé. 1984. A unified formulation of the constant temperature molecular dynamics methods. J Chem Phys 81:511–519.

[46] Glenn J. Martyna, Douglas J. Tobias, and Michael L. Klein. 1994. Constant pressure molecular dynamics algorithms. J Chem Phys 101:4177–4189.

[47] Ulrich Essmann, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. 1995. A smooth particle mesh ewald method. J Chem Phys 103:8577–8593.

[48] W. Humphrey. 1996. Vmd: Visual molecular dynamics. J Mol Graphics 14:33–38.

[49] David Chandler. 1987. Introduction to Modern Statistical Mechanics. Oxford University Press, Inc., 198 Madison Avenue, New York, New York 10016-4314.

[50] A. Luzar and D. Chandler. 1996. Effect of environment on hydrogen bond dynamics in liquid water. Phys Rev Lett 76:928–931.

[51] Z. Yan, S. V. Buldyrev, P. Kumar, N. Giovambattista, P. G. Debenedetti, and H. E. Stanley. 2007. Structure of the first- and second-neighbor shells of simulated water: quantitative relation to translational and orientational order. Phys Rev E 76:051201–051206.

[52] G. Ruocco, M. Sampoli, and R. Vallauri. 1992. Analysis of the network topology in liquid water and hydrogen sulphide by computer simulation. J Chem Phys 96:6167–6176.

[53] Chris H. Rycroft, Gary S. Grest, James W. Landry, and Martin Z. Bazant. 2006. Analysis of granular flow in a pebble-bed nuclear reactor. Phys Rev E 74:021306–021321.

[54] Brian M. Baynes and Bernhardt L. Trout. 2004. Rational design of solution additives for the prevention of protein aggregation. Biophys J 87:1631–1639.

[55] William S. Price, Hiroyuki Ide, and Yoji Arata. 1999. Self-diffusion of supercooled water to 238 k using pgse nmr diffusion measurements. J Phys Chem A 103:448–450.

[56] Jarmo Huuskonen. 2000. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. J Chem Inf Comput Sci 40:773–777.

[57] Jun Huang, Thomas C. Stringfellow, and Lian Yu. 2008. Glycine exists mainly as monomers, not dimers, in supersaturated aqueous solutions: Implications for understanding its crystallization and polymorphism. J Am Chem Soc 130:13973–13980.

[58] Said Hamad, Colan E. Hughes, Catlow, and Kenneth D. M. Harris. 2008. Clustering of glycine molecules in aqueous solution studied by molecular dynamics simulation. J Phys Chem B 112:7280–7288.

[59] R. Jimenez, G. R. Fleming, P. V. Kumar, and M. Maroncelli. 1994. Femtosecond solvation dynamics of water. Nature 369:471–473.

[60] Nicolas Giovambattista, Peter J. Rossky, and Pablo G. Debenedetti. 2006. Effect of pressure on the phase behavior and structure of water confined between nanoscale hydrophobic and hydrophilic plates. Phys Rev E 73:041604–041627.

[61] Andrew D White, Ann K Nowinski, Wenjun Huang, Andrew J Keefe, Fang Sun, and Shaoyi Jiang. 2012. Decoding nonspecific interactions from nature. Chem Sci 3:3488–3494.

[62] Bert van den Berg, Ellis, and Christopher M. Dobson. 1999. Effects of macromolecular crowding on protein folding and aggregation. The EMBO Journal 18:6927–6933.

[63] G. Vogt, S. Woell, and P. Argos. 1997. Protein thermal stability, hydrogen bonds, and ion pairs. J Mol Biol 269:631–643.

[64] Andrew D White, Wenjun Huang, and Shaoyi Jiang. 2012. Role of Nonspecific Interactions in Molecular Chaperones through Model-based Bioinformatics. Biophs J 103:2485–2491.

[65] Buyong Ma, Tal Elkayam, Haim Wolfson, and Ruth Nussinov. 2003. Proteinprotein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc Natl Acad Sci USA 100:5772–5777.

[66] Artëm Masunov and Themis Lazaridis. 2003. Potentials of Mean Force between Ionizable Amino Acid Side Chains in Water. J Am Chem Soc 125:1722–1730.

[67] C. A. Olson, E. J. Spek, Z. Shi, A. Vologodskii, and N. R. Kallenbach. 2001. Cooperative helix stabilization by complex Arg-Glu salt bridges. Proteins 44:123–132.

[68] Marcela P. Aliste, Justin L. MacCallum, and D. Peter Tieleman. 2003. Molecular dynamics simulations of pentapeptides at interfaces: salt bridge and cation-pi interactions. Biochem 42:8976–8987.

[69] Hans Rudolf R. Bosshard, Daniel N. Marti, and Ilian Jelesarov. 2004. Protein stabilization by salt bridges: concepts, experimental approaches and clarification of some misunderstandings. J Mol Recognit 17:1–16.

[70] George I. Makhatadze, Vakhtang V. Loladze, Dmitri N. Ermolenko, XiaoFen Chen, and Susan T. Thomas. 2003. Contribution of surface salt bridges to protein stability: Guidelines for protein engineering. J Mol Biol 327:1135–1148.

[71] Donna L. Luisi, Christopher D. Snow, Jo-Jin Lin, Zachary S. Hendsch, Bruce Tidor, and Daniel P. Raleigh. 2003. Surface salt bridges, Double-Mutant cycles, and protein stability: an experimental and computational analysis of the interaction of the asp 23 side chain with the N-Terminus of the N-Terminal domain of the ribosomal protein l9†. Biochemistry 42:7050–7060.

[72] D. Sali, M. Bycroft, and A. R. Fersht. 1991. Surface electrostatic interactions contribute little of stability of barnase. J Mol Biol 220:779–788.

[73] A. H. Elcock. 1998. The stability of salt bridges at high temperatures: implications for hyperthermophilic proteins. Journal of molecular biology 284:489–502.

[74] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. 2008. Well-Tempered Metadynamics: A Smoothly Converging and Tunable Free-Energy Method. Phys Rev Lett 100:020603+.

[75] Koji Hukushima and Koji Nemoto. 1996. Exchange Monte Carlo method and application to spin glass simulations. J Phys Soc Jpn 65:1604–1608.

[76] Giovanni Bussi, Francesco Luigi L. Gervasio, Alessandro Laio, and Michele Parrinello. 2006. Free-energy landscape for beta hairpin folding from combined parallel tempering and metadynamics. J Am Chem Soc 128:13435–13441.

[77] Amish J. Patel, Patrick Varilly, David Chandler, and Shekhar Garde. 2011. Quantifying density fluctuations in volumes of all shapes and sizes using indirect umbrella sampling. J Stat Phys 145:265–275.

[78] Giovanni Bussi, Davide Donadio, and Michele Parrinello. 2007. Canonical sampling through velocity rescaling. J Chem Phys 126:014101+.

[79] M. Parrinello and A. Rahman. 1981. Polymorphic transitions in single crystals: A new molecular dynamics method. J Appl Phys 52:7182–7190.

[80] Berk Hess, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. 1997. LINCS: A linear constraint solver for molecular simulations. J Comput Chem 18:1463–1472.

[81] David Van Der Spoel, Erik Lindahl, Berk Hess, Gerrit Groenhof, Alan E. Mark, and Herman J. Berendsen. 2005. GROMACS: fast, flexible, and free. J Comput Chem 26:1701–1718.

[82] Massimiliano Bonomi, Davide Branduardi, Giovanni Bussi, Carlo Camilloni, Davide Provasi, Paolo Raiteri, Davide Donadio, Fabrizio Marinelli, Fabio Pietrucci, Ricardo A. Broglia, and Michele Parrinello. 2009. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. Comput Phys Commun 180:1961–1972.

[83] Yasuhiko Nozaki and Charles Tanford. 1971. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. J Biol Chem 246:2211–2217.

[84] Jeremy N. S. Evans. 1995. Biomolecular NMR Spectroscopy. Oxford University Press, USA, 1 edition.

[85] A. Nikolić, B. Jović, S. Csanady, and S. Petrović. 2007. NHo hydrogen bonding: FT IR, NIR and 1H NMR study of n-methylpropionamide cyclic ether systems. J Mol Struct 834-836:249–252.

[86] R. Wolfenden, L. Andersson, P. M. Cullis, and C. C. B. Southgate. 1981. Affinities of amino acid side chains for solvent water. Biochem 20:849–855.

[87] M. Bonomi, A. Barducci, and M. Parrinello. 2009. Reconstructing the equilibrium Boltzmann distribution from well-tempered metadynamics. J Comput Chem 30:1615–1621.

[88] Y. Kita, T. Arakawa, T. Y. Lin, and S. N. Timasheff. 1994. Contribution of the surface free energy perturbation to protein-solvent interactions. Biochem 33:15178–15189.

[89] M. Boncheva and H. Vogel. 1997. Formation of stable polypeptide monolayers at interfaces: controlling molecular conformation and orientation. Biophys J 73:1056–1072.

[90] G. D. Frasman. 1989. Prediction of Protein Structure and the Principles of Protein Conformation. Springer, 1 edition.

[91] M. W. MacArthur and J. M. Thornton. 1991. Influence of proline residues on protein conformation. J Mol Biol 218:397–412.

[92] Carol A. Rohl, Charlie E. Strauss, Kira M. Misura, and David Baker. 2004. Protein structure prediction using Rosetta. Methods Enzymol 383:66–93.

[93] Roland L. Dunbrack and Fred E. Cohen. 1997. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 6:1661–1681.

[94] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David E. Shaw. 2010. Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins 78:1950–1958.

[95] Ann K. Nowinski, Fang Sun, Andrew D. White, Andrew J. Keefe, and Shaoyi Jiang. 2012. Sequence, Structure, and Function of Peptide Self-Assembled Monolayers. J Am Chem Soc 134:6000–6005.

[96] W. B. Tsai, J. M. Grunkemeier, and T. A. Horbett. 1999. Human plasma fibrinogen adsorption and platelet adhesion to polystyrene. J Biomed Mater Res 44:130–139.

[97] Ed Harlow and David Lane. 1998. Using Antibodies: A Laboratory Manual. Cold Spring Harbor Laboratory Press.

[98] Arthur L. Horwich, Wayne A. Fenton, Eli Chapman, and George W. Farr. 2007. Two Families of Chaperonin: Physiology and Mechanism. Annu Rev Cell Dev Biol 23:115–145.

[99] Paul B. Sigler, Zhaohui Xu, Hays S. Rye, Steven G. Burston, Wayne A. Fenton, and Arthur L. Horwich. 1998. STRUCTURE AND FUNCTION IN GroEL-MEDIATED PROTEIN FOLDING. Annu Rev Biochem 67:581–608.

[100] Arthur L. Horwich, Adrian C. Apetri, and Wayne A. Fenton. 2009. The GroEL/GroES cis cavity as a passive anti-aggregation device. FEBS Lett 583:2654–2662.

[101] Shengfu Chen, Zhiqiang Cao, and Shaoyi Jiang. 2009. Ultra-low fouling peptide surfaces derived from natural amino acids. Biomaterials 30:5892–5896.

[102] Olivier R. Bolduc, Christopher M. Clouthier, Joelle N. Pelletier, and Jean-François Masson. 2009. Peptide Self-Assembled Monolayers for Label-Free and Unamplified Surface Plasmon Resonance Biosensing in Crude Cell Lysate. Anal Chem 81:6779–6788.

[103] Rolf Chelmowski, Stephan D. Koster, Andreas Kerstan, Andreas Prekelt, Christian Grunwald, Tobias Winkler, Nils Metzler-Nolte, Andreas Terfort, and Christof Woll. 2008. Peptide-Based SAMs that Resist the Adsorption of Proteins. J Am Chem Soc 130:14952–14953.

[104] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29.

[105] Charu Chaudhry, Arthur L. Horwich, Axel T. Brunger, and Paul D. Adams. 2004. Exploring the Structural Dynamics of the E.coli Chaperonin GroEL Using Translation-libration-screw Crystallographic Refinement of Intermediate States. J Mol Biol 342:229–245.

[106] Tatsuro Shimamura, Ayumi Koike-Takeshita, Ken Yokoyama, Ryoji Masui, Noriyuki Murai, Masasuke Yoshida, Hideki Taguchi, and So Iwata. 2004. Crystal structure of the native chaperonin complex from Thermus thermophilus revealed unexpected asymmetry at the cis-cavity. Structure (Camb) 12:1471–1480.

[107] Jose H. Pereira, Corie Y. Ralston, Nicholai R. Douglas, Daniel Meyer, Kelly M. Knee, Daniel R. Goulet, Jonathan A. King, Judith Frydman, and Paul D. Adams. 2010. Crystal structures of a group II chaperonin reveal the open and closed states associated with the protein folding cycle. J Bio Chem 285:27958–27966.

[108] Maruf M. Ali, Mark M. Roe, Cara K. Vaughan, Phillipe Meyer, Barry Panaretou, Peter W. Piper, Chrisostomos Prodromou, and Laurence H. Pearl. 2006. Crystal structure of an Hsp90-nucleotide-p23/Sba1 closed chaperone complex. Nature 440:1013–1017.

[109] Carien Dekker, Mark M. Roe, Elizabeth A. McCormack, Fabienne Beuron, Laurence H. Pearl, and Keith R. Willison. 2011. The crystal structure of yeast CCT reveals intrinsic asymmetry of eukaryotic cytosolic chaperonins. The EMBO journal 30:3078–3090.

[110] R Development Core Team. 2011. R: A Language and Environment for Statistical Computing.

[111] Bill Howe, Garret Cole, Emad Souroush, Paraschos Koutris, Alicia Key, Nodira Khoussainova, and Leilani Battle. 2011. Database-as-a-Service for Long Tail Science. In In Proceedings of the 23rd international conference on Scientific and statistical database management (SSDBM'11), pages 480–489. Springer-Verlag.

[112] N. R. Voss and M. Gerstein. 2005. Calculation of standard atomic volumes for RNA and comparison with proteins: RNA is packed more tightly. J Mol Biol 346:477–492.

[113] Jiri Homola. 2008. Surface Plasmon Resonance Sensors for Detection of Chemical and Biological Species. Chemical Reviews 108:462–493.

[114] J. Homola. 2006. Surface plasmon resonance based sensors. Springer-Verlag.

[115] M. A. Andrade, S. I. O'Donoghue, and B. Rost. 1998. Adaptation of protein surfaces to subcellular location. J Mol Biol 276:517–525.

[116] José L. Jiménez. 2005. Does structural and chemical divergence play a role in precluding undesirable protein interactions? Proteins 59:757–764.

[117] F. Glaser, D. M. Steinberg, I. A. Vakser, and N. Ben-Tal. 2001. Residue frequencies and pairing preferences at protein-protein interfaces. Proteins 43:89–102.

[118] P. Harder, M. Grunze, R. Dahint, G. M. Whitesides, and P. E. Laibinis. 1998. Molecular Conformation in Oligo(ethylene glycol)-Terminated Self-Assembled Monolayers on Gold and Silver Surfaces Determines Their Ability To Resist Protein Adsorption. J Phys Chem B 102:426–436.

[119] Shengfu Chen, Jie Zheng, Lingyan Li, and Shaoyi Jiang. 2005. Strong Resistance of Phosphorylcholine Self-Assembled Monolayers to Protein Adsorption: Insights into Nonfouling Properties of Zwitterionic Materials. J Am Chem Soc 127:14473–14478.

[120] Mark Shtilerman, George H. Lorimer, and S. Walter Englander. 1999. Chaperonin Function: Folding by Forced Unfolding. Science 284:822–825.

[121] Hagen Hofmann, Frank Hillger, Shawn H. Pfeil, Armin Hoffmann, Daniel Streich, Dominik Haenni, Daniel Nettels, Everett A. Lipman, and Benjamin Schuler. 2010. Single-molecule spectroscopy of protein folding in a chaperonin cage. Proc Natl Acad Sci USA 107:11793–11798.

[122] Hao Fan and Alan E. Mark. 2006. Mimicking the action of GroEL in molecular dynamics simulations: Application to the refinement of protein structures. Protein Sci 15:441–448.

[123] Jianhui Tian and Angel E. Garcia. 2011. Simulation Studies of Protein Folding/Unfolding Equilibrium under Polar and Nonpolar Confinement. J Am Chem Soc 133:15157–15164.

[124] Hue S. Chan and Ken A. Dill. 1996. A simple model of chaperonin-mediated protein folding. Proteins 24:345–351.

[125] Fumiko Takagi, Nobuyasu Koga, and Shoji Takada. 2003. How protein thermodynamics and folding mechanisms are altered by the chaperonin cage: molecular simulations. Proc Natl Acad Sci USA 100:11367–11372.

[126] Jeetain Mittal and Robert B. Best. 2008. Thermodynamics and kinetics of protein folding under confinement. Proc Natl Acad Sci USA 105:20233–20238.

[127] Pierre-Giles de Gennes. 1979. Scaling concepts in polymer physics. Cornel University Press.

[128] Jonathan E. Kohn, Ian S. Millett, Jaby Jacob, Bojan Zagrovic, Thomas M. Dillon, Nikolina Cingel, Robin S. Dothager, Soenke Seifert, P. Thiyagarajan, Tobin R. Sosnick, M. Zahid Hasan, Vijay S. Pande, Ingo Ruczinski, Sebastian Doniach, and Kevin W. Plaxco. 2004. Random-coil behavior and the dimensions of chemically unfolded proteins. Proc Natl Acad Sci USA 101:12491–12496.

[129] Sanzo Miyazawa and Robert L. Jernigan. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18:534–552.

[130] Judith Frydman. 2001. FOLDING OF NEWLY TRANSLATED PROTEINS IN VIVO: The Role of Molecular Chaperones. Annu Rev Biochem 70:603–647.

[131] Zong Lin and Hays S. Rye. 2004. Expansion and Compression of a Protein Folding Intermediate by GroEL. Mol Cell 16:23–34.

[132] R. Zahn. 1994. Thermodynamic Partitioning Model for Hydrophobic Binding of Polypeptides by GroEL II. GroEL Recognizes Thermally Unfolded Mature -lactamase. J Mol Biol 242:165–174.

[133] S. Walter, G. H. Lorimer, and F. X. Schmid. 1996. A thermodynamic coupling mechanism for GroEL-mediated unfolding. Proc Natl Acad Sci USA 93:9425–9430.

[134] Martin C. Stumpe, Nikolay Blinov, David Wishart, Andriy Kovalenko, and Vijay S. Pande. 2010. Calculation of Local Water Densities in Biological Systems: A Comparison of Molecular Dynamics Simulations and the 3D-RISM-KH Molecular Theory of Solvation. J Phys Chem B 115:319–328.

[135] Ariel Fernández and Harold A. Scheraga. 2003. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. Proc Natl Acad Sci USA 100:113–118.

[136] Tetsuji Okada, Yoshinori Fujiyoshi, Maria Silow, Javier Navarro, Ehud M. Landau, and Yoshinori Shichida. 2002. Functional role of internal water molecules in rhodopsin revealed by x-ray crystallography. Proc Natl Acad Sci USA 99:5982–5987.

[137] Young M. Rhee, Eric J. Sorin, Guha Jayachandran, Erik Lindahl, and Vijay S. Pande. 2004. Simulations of the role of water in the protein-folding mechanism. Proc Natl Acad Sci USA 101:6456–6461.

[138] Peter Guttorp. 1995. Stochastic Modeling of Scientific Data. Chapman and Hall/CRC, second edition.

[139] George W. Farr, Wayne A. Fenton, and Arthur L. Horwich. 2007. Perturbed ATPase activity and not close confinement of substrate in the cis cavity affects rates of folding by tail-multiplied GroEL. Proc Natl Acad Sci USA 104:5342–5347.

[140] R. Wolfenden, L. Andersson, P. M. Cullis, and C. C. B. Southgate. 1981. Affinities of amino acid side chains for solvent water. Biochem 20:849–855.

[141] Yun-Chi Tang, Hung-Chun Chang, Annette Roeben, Dirk Wischnewski, Nadine Wischnewski, Michael J. Kerner, F. Ulrich Hartl, and Manajit Hayer-Hartl. 2006. Structural Features of the GroEL-GroES Nano-Cage Required for Rapid Folding of Encapsulated Protein. Cell 125:903–914.

[142] Yun-Chi Tang, Hung-Chun Chang, Kausik Chakraborty, F. Ulrich Hartl, and Manajit Hayer-Hartl. 2008. Essential role of the chaperonin folding compartment in vivo. The EMBO Journal 27:1458–1468.

[143] Michael J. Kerner, Dean J. Naylor, Yasushi Ishihama, Tobias Maier, Hung-Chun Chang, Anna P. Stines, Costa Georgopoulos, Dmitrij Frishman, Manajit Hayer-Hartl, Matthias Mann, and F. Ulrich Hartl. 2005. Proteome-wide Analysis of Chaperonin-Dependent Protein Folding in Escherichia coli. Cell 122:209–220.

[144] Ann K. Nowinski, Fang Sun, Andrew D. White, Andrew J. Keefe, and Shaoyi Jiang. 2012. Sequence, structure, and function of peptide Self-Assembled monolayers. J Am Chem Soc 134:6000–6005.

[145] Olivier R. Bolduc and Jean-François Masson. 2008. Monolayers of 3-Mercaptopropyl-amino acid to reduce the nonspecific adsorption of serum proteins on the surface of biosensors. Langmuir 24:12085–12091.

[146] T. Obata, M. B. Yaffe, G. G. Leparc, E. T. Piro, H. Maegawa, A. Kashiwagi, R. Kikkawa, and L. C. Cantley. 2000. Peptide and protein library screening defines optimal substrate motifs for AKT/PKB. The Journal of biological chemistry 275:36108–36115.

[147] S. E. Blondelle and K. Lohner. 2000. Combinatorial libraries: a tool to design antimicrobial and antifungal peptide analogues having lytic specificities for structure-activity relationship studies. Biopolymers 55:74–87.

[148] Lau Sennels, Mogjiborahman Salek, Lee Lomas, Egisto Boschetti, Pier G. Righetti, and Juri Rappsilber. 2007. Proteomic Analysis of Human Blood Serum Using Peptide Library Beads. J Proteome Res 6:4055–4062.

[149] Michael C. Sweeney, Anne-Sophie S. Wavreille, Junguk Park, Jonathan P. Butchar, Susheela Tridandapani, and Dehua Pei. 2005. Decoding protein-protein interactions through combinatorial chemistry: sequence specificity of SHP-1, SHP-2, and SHIP SH2 domains. Biochem 44:14932–14947.

[150] Andrew J Keefe, Kyle Caldwell, Ann K Nowinski, Andrew D White, Amit Thakkar, and Shaoyi Jiang. 2013. Screening Nonspecific Interactions of Peptides with Elminiated Background Interference. Biomaterials 34:1871–1877.

[151] Zhijian Lu, Kristin S. Murray, Victor V. Cleave, Edward R. LaVallie, Mark L. Stahl, and John M. McCoy. 1995. Expression of thioredoxin random peptide libraries on the escherichia coli cell surface as functional fusions to flagellin: A system designed for exploring Protein-Protein interactions. Nat Biotechnol 13:366–372.

[152] George P. Smith and Valery A. Petrenko. 1997. Phage Display. Chem Rev 97:391–410.

[153] Z. Songyang, S. E. Shoelson, M. Chaudhuri, G. Gish, T. Pawson, W. G. Haser, F. King, T. Roberts, S. Ratnofsky, and R. J. Lechleider. 1993. SH2 domains recognize specific phosphopeptide sequences. Cell 72:767–778.

[154] Nianhuan Yao, Wenwu Xiao, Xiaobing Wang, Jan Marik, See Hyoung H. Park, Yoshikazu Takada, and Kit S. Lam. 2009. Discovery of targeting ligands for breast cancer cells using the one-bead one-compound combinatorial method. J Med Chem 52:126–133.

[155] Martin Hintersteiner, Thierry Kimmerlin, Frank Kalthoff, Markus Stoeckli, Geraldine Garavel, Jan-Marcus M. Seifert, Nicole-Claudia C. Meisner, Volker Uhl, Christof Buehler, Thomas Weidemann, and Manfred Auer. 2009. Single bead labeling method for combining confocal fluorescence on-bead screening and solution validation of tagged one-bead one-compound libraries. Chem Biol 16:724–735.

[156] Timothy L. Bailey. 1995. Discovering motifs in DNA and protein sequences: The approximate common substring problem. Ph.D. thesis, University of California at San Diego.

[157] T. L. Bailey and C. Elkan. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology 2:28–36.

[158] Jianping Wu and Rotimi E. Aluko. 2007. Quantitative structure-activity relationship study of bitter di- and tri-peptides including relationship with angiotensin I-converting enzyme inhibitory activity. J Peptide Sci 13:63–69.

[159] Douglas M. Hawkins, Subhash C. Basak, and Denise Mills. 2003. Assessing Model Fit by Cross-Validation. J Chem Inf Comput Sci 43:579–586.

[160] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

[161] J. Macqueen. 1967. Some methods for classification and analysis of multivariate observations. In Proc. 5th Berkeley Symp. Mathematical Statist. Probability, pages 281–297.

[162] J. A. Hartigan and M. A. Wong. 1979. Algorithm AS 136: A K-Means Clustering Algorithm. Applied Statistics 28:100–108.

[163] A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. In JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B, volume 39, pages 1–38.

[164] A. D. Gordon. 1999. Classification, 2nd Edition (Chapman & Hall/CRC Monographs on Statistics & Applied Probability). Chapman and Hall/CRC, 2 edition.

[165] David F. Bauer. 1972. Constructing Confidence Sets Using Rank Statistics. J Am Stat Assoc 67:687–690.

[166] A. K. Ghose and G. M. Crippen. 1987. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. J Chem Inf Comput Sci 27:21–35.

[167] Christoph Steinbeck, Yongquan Han, Stefan Kuhn, Oliver Horlacher, Edgar Luttmann, and Egon Willighagen. 2003. The chemistry development kit (CDK): an Open-Source java library for chemo- and bioinformatics. J Chem Inf Comput Sci 43:493–500.

[168] Xianwen Chen, Lige Ren, Soochong Kim, Nicholas Carpino, James L. Daniel, Satya P. Kunapuli, Alexander Y. Tsygankov, and Dehua Pei. 2010. Determination of the substrate specificity of protein-tyrosine phosphatase TULA-2 and identification of Syk as a TULA-2 substrate. The Journal of biological chemistry 285:31268–31276.

[169] Mickael Goujon, Hamish McWilliam, Weizhong Li, Franck Valentin, Silvano Squizzato, Juri Paern, and Rodrigo Lopez. 2010. A new bioinformatics analysis tools framework at EMBLEBI. Nucleic Acids Res 38:W695–W699.

[170] S. Henikoff and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89:10915–10919.

[171] Timothy L. Bailey and Charles Elkan. 1995. Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization. Machine Learning 21:51–80.

[172] Arthur Doweyko. 2008. QSAR: dead or alive? J Comput -Aided Mol Des 22:81–89.

[173] JeffreyA Bilmes. 2004. Graphical Models and Automatic Speech Recognition. In Mark Johnson, SanjeevP Khudanpur, Mari Ostendorf, and Roni Rosenfeld, editors, Mathematical Foundations of Speech and Language Processing, volume 138 of *The IMA Volumes in Mathematics and its Applications*, pages 191–245. Springer New York.

[174] Michael Isard and Andrew Blake. 1998. CONDENSATIONConditional Density Propagation for Visual Tracking, volume 29. Kluwer Academic Publishers.

[175] Michael M. Hoffman, Jason Ernst, Steven P. Wilder, Anshul Kundaje, Robert S. Harris, Max Libbrecht, Belinda Giardine, Paul M. Ellenbogen, Jeffrey A. Bilmes, Ewan Birney, Ross C. Hardison, Ian Dunham, Manolis Kellis, and William S. Noble. 2013. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res 41:827–841.

[176] Hetunandan Kamisetty, Arvind Ramanathan, Chris Bailey-Kellogg, and Christopher J. Langmead. 2011. Accounting for conformational entropy in predicting binding free energies of protein-protein interactions. Proteins 79:444–462.

[177] J. Bilmes. 2010. Dynamic Graphical Models. Signal Processing Magazine, IEEE 27:29–42.

[178] The ENCODE Project Consortium. 2012. An Integrated Encyclopedia of DNA Elements in the Human Genome. Nature 489:57—74.

[179] Matthias Kormaksson, James G Booth, Maria E Figueroa, and Ari Melnick. 2012. Integrative model-based clustering of microarray methylation and expression data. The Annals of Applied Statistics 6:1327–1347.

[180] Mark Hewitt, Mark T. D. Cronin, Judith C. Madden, Philip H. Rowe, Clara Johnson, Anrdrea Obi, and Steven J. Enoch. 2007. Consensus QSAR Models: Do the Benefits Outweigh the Complexity? J Chem Inf Model 47:1460–1468.

[181] Guangshun Wang, Xia Li, and Zhe Wang. 2009. APD2: the updated antimicrobial peptide database and its application in peptide design. Nucleic Acids Res 37:D933–D937.

[182] Andrew D. White, Andrew J. Keefe, Ann K. Nowinski, Qing Shao, Kyle Caldwell, and Shaoyi Jiang. 2013. Standardizing and simplifying analysis of peptide library data. J Chem Inf Model 53:493–499.

[183] Ozgur Cetin, Harriet Nock, Katrin Kirchhoff, Jeff Bilmes, and Mari Ostendorf. 2002. The 2001 GMTK-based SPINE ASR system. In Proc. Int. Conf. on Spoken Language Processing. Denver, Colorado.

[184] Christopher D. Fjell, Jan A. Hiss, Robert E. W. Hancock, and Gisbert Schneider. 2012. Designing antimicrobial peptides: form follows function. Nat Rev Drug Discov 11:37–51.

[185] Sara Bobone, Alessandro Piazzon, Barbara Orioni, Jens Z. Pedersen, Yong H. Nan, Kyung-Soo Hahm, Song Y. Shin, and Lorenzo Stella. 2011. The thin line between cell-penetrating and antimicrobial peptides: the case of Pep-1 and Pep-1-K. J Peptide Sci 17:335–341.

[186] V. Frecer, B. Ho, and J. L. Ding. 2004. De novo design of potent antimicrobial peptides. Antimicrobial agents and chemotherapy 48:3349–3357.

[187] Chris Fraley and Adrian E. Raftery. 2002. Model-Based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97:611–631.

[188] T. N. Pierre, A. A. Seon, M. Amiche, and P. Nicolas. 2000. Phylloxin, a novel peptide antibiotic of the dermaseptin family of antimicrobial/opioid peptide precursors. Eur J Biochem 267:370–378.

[189] J. Michael Conlon, Milena Mechkarska, Eman Ahmed, Laurent Coquet, Thierry Jouenne, Jérôme Leprince, Hubert Vaudry, Marc P. Hayes, and Gretchen Padgett-Flohr. 2011. Host defense peptides in skin secretions of the Oregon spotted frog Rana pretiosa: implications for species resistance to chytridiomycosis. Developmental and comparative immunology 35:644–649.

[190] Marc Torrent, David Andreu, Victòria M. Nogués, and Ester Boix. 2011. Connecting Peptide Physicochemical and Antimicrobial Properties by a Rational Prediction Model. PLoS ONE 6:e16968+.

[191] HÅvard Jenssen, Christopher D. Fjell, Artem Cherkasov, and Robert E. W. Hancock. 2008. QSAR modeling and computer-aided design of antimicrobial peptides. J Peptide Sci 14:110–114.

[192] T. Lejon, M. B. Strøm, and J. S. Svendsen. 2001. Antibiotic activity of pentadecapeptides modelled from amino acid descriptors. J Peptide Sci 7:74–81.

[193] Christopher D. Fjell, Håvard Jenssen, Kai Hilpert, Warren A. Cheung, Nelly Pante, Robert E. W. Hancock, and Artem Cherkasov. 2009. Identification of Novel Antibacterial Peptides by Chemoinformatics and Machine Learning. J Med Chem 52:2006–2015.

[194] Håvard Jenssen, Tore J. Gutteberg, and Tore Lejon. 2005. Modelling of anti-HSV activity of lactoferricin analogues using amino acid descriptors. J Peptide Sci 11:97–103.

[195] Jeff Bilmes. 2000. Dynamic Bayesian Multinets. In Uncertainty in Artificial Intelligence (UAI), 16th. Morgan Kaufmann Publishers.

VITA

Andrew was born in Stockton, CA and grew up near Vancouver, WA. Andrew attended Rose-Hulman Institute of Technology in Terre Haute, IN and studied abroad for a year at Otto-Von-Guericke Universität in Magdeburg, Germany. Andrew received his Bachelor of Science in Chemical Engineering in 2008. He is the great-great-great-grandson of Walter White, the assistant secretary of the Royal Society from 1844-1885.

Bachelor of Science, Chemical Engineering (2008) *Magna Cum Laude*
Minor in German Language and Culture
Rose-Hulman Institute of Technology, Terre Haute, IN

Honors:

Best Poster/Outstanding Graduate Student AICHE COMSEF division, 2012

Best Speaker, UW Chemical Engineering Symposium, 2012

Fellow, Found. of Mol. Modeling and Sim., 2012

Best Poster, Int'l. Congress on Marine Fouling and Corrosion, 2012

Best Poster, Gordon Research Conf.: Water and Aq. Soln., 2010

Best Poster, UW Chemical Engineering Symposium, 2009

Runstad Fellow, 2008−2009

Publications:

1. **White AD**, Keefe A, Ella-Menye J-R, Nowinski AK, Shao Q, Jiang S (2013) Free Energy of Solvated Salt Bridges from Simulations and Experiments. *Submitted*.
2. **White AD**, Keefe AJ, Nowinski AK, Shao Q, Caldwell K, Jiang S (2013) Standardizing and Simplifying Analysis of Peptide Library Data. *J. Chem. Inf. Model.* 53 493–499.
3. Brault ND, **White AD**, Taylor AD, Yu Q, and Jiang S (2013) A Directly Functionalizable Surface Platform for Protein Arrays in Undiluted Human Blood Plasma. *Anal. Chem.* 85 1447–1453.
4. Keefe AJ, Caldwell K, Nowinski AK, **White AD**, Thakkar A, Jiang S (2012) Screening Nonspecific Interactions of Peptides with Eliminated Background Interference. *Biomaterials* 34 1871–1877.
5. **White AD**, Huang W, Jiang S (2012) Role of Non-specific Interactions in Molecular Chaperones through Model-based Bioinformatics. *Biophys. J.* 103 2484–2491.
6. **White AD**, Nowinski AK, Huang W, Keefe AJ, Sun F, Jiang S (2012) Decoding Non-specific Interactions from Nature. *Chem. Sci.* 3 3488–3494.
7. Shao Q, He Y, **White AD**, Jiang S (2012) The Different Effects of Zwitterion and Oligo(Ethylene glycol) Solutes on Proteins. *J. Chem. Phys.* 136 225101.
8. Nowinski A, Fang S, **White AD**, Keefe A, Jiang S (2012) Sequence, Structure, and Function of Peptide Self-assembled Monolayers. *JACS* 134 6000-6005.
9. **White AD**, Jiang S (2011) Local and Bulk Hydration of Zwitterionic Glycine and its Analogues through Molecular Simulations. *J. Phys. Chem. B* 115 660-667. Cover Feat.
10. Shao Q, He Y, **White AD**, Jiang S (2010) Difference in Hydration between Carboxybetaine and Sulfobetaine. *J. Phys. Chem. B* 114 16625-16631.
11. Yang W, Zhang L, Wang S, **White AD**, Jiang S (2009) Functionalizable and Ultra Stable, Nanoparticles Coated with Zwitterionic Poly(carboxybetaine) in Undiluted Blood Serum. *Biomaterials* 30 5617–5621.

Invited Presentations:

1. **White AD**, Jiang S (2012) Using Computational Approaches to Mimic Nature. *UW eScience Institute Seminar.*
2. **White AD**, Jiang S (2010) Glycine and Its Zwitterionic Analogues' Local and Bulk Hydration through Molecular Simulation *Gordon Research Conference: Water and Aqueous Solutions.*

Contributed Presentations:

1. **White AD**, Nowinski AK, Huang W, Keefe AJ, Sun F, Jiang S (2012) Modeling Nonspecific Interactions in Biological Systems. *2012 AIChE Annual Meeting.*
2. **White AD**, Nowinski AK, Huang W, Sun F, Jiang S (2012) Designing Peptide Biomaterials Using Experiments and Modeling. *2012 AIChE Annual Meeting.*
3. **White AD** (2012) Resisting Nonspecific Interactions in Biology through Experiments and Modeling *UW Chemical Engineering Graduate Symposium.* Best Speaker 1^{st} .
4. **White AD**, Jiang S (2011) Model-Based Clustering Applied to Combinatorial Libraries. *2011 AIChE Annual Meeting.*
5. **White AD**, Jiang S (2011) Data-Mining Nature to Design New Peptide Based Biomaterials. *2011 AIChE Annual Meeting.*
6. **White AD**, Jiang S (2011) The Free Energy of a Salt Bridge from Simulations. *2011 AIChE Annual Meeting.*
7. **White AD** (2010) Hydration of Zwitterionic Glycine and Analogues *UW Chemical Engineering Graduate Symposium.*

Contributed Posters:

1. **White AD**, Huang W, Nowinski AK, Jiang S (2012) Modeling Nonspecific Interactions in Biology. *2012 AIChE Annual Meeting.* Best Poster (2 awarded) .
2. **White AD**, Jiang S (2012) Modeling and Experiments of Nonspecific Interactions. *2012 AIChE Annual Meeting.*
3. **White AD**, Nowinski AK, Huang W, Keefe AJ, Sun F, Jiang S (2012) Mimicking Non-specific Processes in Biology through Experiments, Modeling, and Simulations. *Foundations of Molecular Simulation Modeling.*
4. **White AD**, Nowinski AK, Huang W, Keefe AJ, Sun F, Jiang S (2012) Designing Anti-Fouling Materials with Modeling, Biomimetics and QSPR. *International Congress on Marine Fouling and Corrosion.* Best Poster 1^{st} .
5. **White AD**, Jiang S (2011) Hydration of Zwitterionic Glycine and Analogues Through Molecular Simulation. *Biophysical Society 55th Annual Meeting.*
6. **White AD**, Jiang S (2010) Glycine and Its Zwitterionic Analogues' Local and Bulk Hydration through Molecular Simulation. *UW Chemical Engineering Graduate Symposium.*
7. **White AD**, Jiang S (2010) Glycine and Its Zwitterionic Analogues' Local and Bulk Hydration through Molecular Simulation. *Gordon Research Conference: Water and Aqueous Solutions.* Best Poster (8 awarded)
8. **White AD**, Hower J, He Y, Shao Q, Jiang S (2009) Molecular Understanding, Mechanism, and Design of Nonfouling Materials. *UW Chemical Engineering Graduate Symposium.* Best Poster 1^{st}
9. **White AD**, Hower J, He Y, Shao Q, Jiang S (2009) Molecular Understanding, Mechanism, and Design of Nonfouling Materials. *Foundations of Molecular Simulation Modeling.*

Appendix A

GRAPHICAL MODEL SPECIFICATIONS

A.1 Motif Model

```
GRAPHICAL_MODEL MOTIF

#define MAX_LENGTH 150
#define ALPHABET 20

#define CLASS_NUMBER 6
#define DATA_NUMBER 1445
#define MOTIF_LENGTH 4
#define MAX_MOTIF_LENGTH 3

%prologue frame
frame: 0 {

  %does the sequence fit into the motif model?
   variable : membership {
     type : discrete hidden cardinality 2;
     conditionalparents : nil using DenseCPT("membership");
   }

  %Motif class, the classification variable.
  variable : motifClass {
    type : discrete hidden cardinality CLASS_NUMBER;
    conditionalparents : nil using DenseCPT("class");
  }

  %Sequence position. Starts at 0
  variable : position {
    type : discrete observed value 0 cardinality MAX_LENGTH;
    conditionalparents : nil using
              DeterministicCPT("zeroPosition");
```

```
  }

 %Whether or not a motif is occuring.
 variable : motifOccur {
   type : discrete hidden cardinality 2;
   conditionalparents : motifClass(0) using DenseCPT("motifStart");
 }

 %Position in Motif, should be 0 and then may increment or not
 variable : motifPosition {
   type : discrete observed value 0 cardinality MOTIF_LENGTH;
   conditionalparents : nil using DeterministicCPT("zero");
 }

 %The actual motif.
 variable : motif {
   type : discrete hidden cardinality ALPHABET;
   conditionalparents : motifClass(0), motifPosition(0) using DenseCPT("motifDefs");
   weight: scale 2;
 }

 %the observed sequence.
 variable : sequence {
  type : discrete observed 0:0 cardinality ALPHABET;
  switchingparents : membership(0), motifOccur(0) using mapping("tripleSwitchDT");
  conditionalparents :
    nil using DenseCPT("uniformBackground") |
    nil using DenseCPT("background") |
    motif(0) using DeterministicCPT("copyMotif");
  }

}

%Normal frames
frame: 1 {

  %Sequence position, counts up to observed length
  variable : position {
    type : discrete hidden cardinality MAX_LENGTH;
    conditionalparents : position(-1)
```

```
     using DeterministicCPT("increment");
 }

%Whether or not a motif is occuring
variable : motifOccur {
  type : discrete hidden cardinality 2;
  switchingparents : motifOccur(-1), motifPosition(-1)  using mapping("motifOccurMapDT");
  conditionalparents :
    motifOccur(-1), motifPosition(-1) using DeterministicCPT("copyMotifOccur")
    |  motifClass(0) using DenseCPT("motifStart");
}

%Position in Motif, should be 0 and then may increment or not
variable : motifPosition {
  type : discrete hidden cardinality MOTIF_LENGTH;
  conditionalparents : motifPosition(-1), motifOccur(0), motifOccur(-1) using DeterministicCPT("motifCounter");
}

%Motif class, which is just copied
variable : motifClass {
   type : discrete hidden cardinality CLASS_NUMBER;
   conditionalparents : motifClass(-1) using DeterministicCPT("copyClass");
}

%membership class, copied over
variable : membership {
  type : discrete hidden cardinality 2;
  conditionalparents : membership(-1) using DeterministicCPT("copyBinary");
}

%The actual motif.
variable : motif {
  type : discrete hidden cardinality ALPHABET;
  conditionalparents : motifClass(0), motifPosition(0) using DenseCPT("motifDefs");
  weight : scale 2;
}

%the observed sequence.
variable : sequence {
 type : discrete observed 0:0 cardinality ALPHABET;
```

```
  switchingparents : membership(0), motifOccur(0) using mapping("tripleSwitchDT");
  conditionalparents :
  nil using DenseCPT("uniformBackground") |
    nil using DenseCPT("background") |
    motif(0) using DeterministicCPT("copyMotif");
  }
}

%epilogue. All that changes is the position is observed
frame : 2 {

  %Sequence position, counts up to observed length
  variable : position {
    type : discrete observed 1:1 cardinality MAX_LENGTH;
    conditionalparents : position(-1)
      using DeterministicCPT("increment");
  }

 %Whether or not a motif is occuring
 variable : motifOccur {
   type : discrete hidden cardinality 2;
   conditionalparents :  motifOccur(-1), motifPosition(-1) using DeterministicCPT("copyMotifOccur");
 }

 %Position in Motif, should be 0 and then may increment or not
 variable : motifPosition {
   type : discrete hidden cardinality MOTIF_LENGTH;
   conditionalparents : motifPosition(-1), motifOccur(0), motifOccur(-1) using DeterministicCPT("motifCounter");
 }

 %make sure we don't end in the middle of a motif
 variable : motifPositionCheck {
   type : discrete observed value 1 cardinality 2;
   conditionalparents : motifPosition(0) using DeterministicCPT("motifPositionCheck");
 }

 %Motif class, which is just copied
 variable : motifClass {
    type : discrete hidden cardinality CLASS_NUMBER;
    conditionalparents : motifClass(-1) using DeterministicCPT("copyClass");
```

```
}


%membership class, copied over
variable : membership {
  type : discrete hidden cardinality 2;
  conditionalparents : membership(-1) using DeterministicCPT("copyBinary");
}


%The actual motif.
variable : motif {
  type : discrete hidden cardinality ALPHABET;
  conditionalparents : motifClass(0), motifPosition(0) using DenseCPT("motifDefs");
  weight : scale 2;
}

%the observed sequence.
variable : sequence {
 type : discrete observed 0:0 cardinality ALPHABET;
 switchingparents : membership(0), motifOccur(0) using mapping("tripleSwitchDT");
 conditionalparents :
   nil using DenseCPT("uniformBackground") |
   nil using DenseCPT("background") |
   motif(0) using DeterministicCPT("copyMotif");
 }
}

chunk 1:1

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%% Decisions Trees
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

DT_IN_FILE inline

9 % number of decision trees

% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
0 % a DT that is always 0
```

```
zeroDT
0 % no parents
-1 0

% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
1 % a DT that copies the parent
copyDT
1 % 1 parent
-1 {p0}

% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
2 % a DT counts downwards
incrementDT
1 % 1 parent
-1 {p0 + 1}

% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
3 % a DT that counts upwards if a motif is occuring
motifCounterDT
3 % 2 parents. Parent 0 is motif position(-1) and parent 1 is the motif occurance(0)
1 2 1 default % are we in a motif?
  0 2 MAX_MOTIF_LENGTH default % are we at the end of a motif and beginning a new one?
    -1 0 % yes
    2 2 1 default % should we increment because we already expressed a position, or are we in the default 0
      -1 {p0 + 1} % no, increment
      -1 0 % yes, we still shouldn't increment
  -1 0 %not occuring, so don't increment

% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
4 % a DT that switches between two parents into three outcomes
tripleSwitchDT
2 % 2 parents
0 2 1 default %are you a member?
 -1 {p1 + 1} %yes, then use second parent offset by one (1 or 2)
 -1 0 % no, then use the non-member rvs (0)

% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
5 % Motif occurance decision tree.
motifOccurMapDT % A motif may occur if either we're not in a motif or we were and the motif ended.
2 % motifOccur(-1), motifPosition(-1)
```

```
0 2 1 default %Is there a motif occuring?
  1 2 MAX_MOTIF_LENGTH default %is the length at its highest value?
    -1 1 % yes it is. then a new motif may start
    -1 0 % it is not at the highest length, so a motif is occuring and we'll copy
  -1 1 %there is not a motif occuring, so allow one

% % % % % % % % % % % % % % % % % % % % % % % % %
6 % Motif occurance copy decision tree.
copyMotifOccurDT % decision to copy motif occurance and end it if necessary
2 % motifOccur(-1), motifPosition(-1)
0 2 1 default %are we in a motif?
  1 2 MAX_MOTIF_LENGTH default % are we at the end of the motif length?
    -1 0 %Yes we were, thus the motif is over
    -1 1 %We are not, thus continue motif occurance
  -1 0 % no we are not in a motif and we didn't want to start one.

7 % checks that we are not in the middle of a motif. Emits 1 if we aren't
motifPositionCheckDT
1 % motifPosition(0)
0 3 0 MAX_MOTIF_LENGTH default
  -1 1;
  -1 1;
  -1 0;

8 % repeat motif
copyMotifDT
1 % motif(0)
-1 {p0}

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%% Non-training Deterministic CPT
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
DETERMINISTIC_CPT_IN_FILE inline 10
0 % first one
zero
0% no parents
MOTIF_LENGTH
zeroDT

1
```

```
copyPosition
1 % parent
MAX_LENGTH MAX_LENGTH
copyDT

2
copyMotifOccur
2 % parent
2 MOTIF_LENGTH 2
copyMotifOccurDT

3
copyClass
1 % parent
CLASS_NUMBER CLASS_NUMBER
copyDT

4
increment
1 % parent
MAX_LENGTH MAX_LENGTH
incrementDT

5
motifCounter
3 % parent
MOTIF_LENGTH 2 2 MOTIF_LENGTH
motifCounterDT

6
zeroPosition
0 % no parents
MAX_LENGTH
zeroDT

7
motifPositionCheck
1 % motif position
MOTIF_LENGTH 2
motifPositionCheckDT
```

```
8
copyMotif
1 % 1 parent
ALPHABET ALPHABET
copyMotifDT

9
copyBinary
1 % 1 parent
2 2
copyDT
```

A.2 QSAR-Motif Model Specification

```
GRAPHICAL_MODEL QSAR-MOTIF

#define MAX_LENGTH 150
#define ALPHABET 20

#define QSAR_SPACE 5
#define CLASS_NUMBER 3
#define DATA_NUMBER 1440
#define MOTIF_LENGTH 3
#define MAX_MOTIF_LENGTH 2

%prologue frame
frame: 0 {

  %does the sequence fit into the motif model?
   variable : membership {
     type : discrete hidden cardinality 2;
     conditionalparents : nil using DenseCPT("membership");
   }

  %Motif class, the classification variable.
  variable : motifClass {
    type : discrete hidden cardinality CLASS_NUMBER;
    conditionalparents : nil using DenseCPT("class");
```

```
  }

  %Sequence position. Starts at 0
  variable : position {
    type : discrete observed value 0 cardinality MAX_LENGTH;
    conditionalparents : nil using
              DeterministicCPT("zeroPosition");
  }


 %QSARs are read during in the prologue frame.

 %INSERT_QSARS_HERE
%QSAR 1

  variable : qsar1 {
    type : discrete observed 2:2 cardinality 5;
    switchingparents : membership(0) using mapping("simpleSwitchDT");
    conditionalparents :
      nil using DenseCPT("inactive") |
      nil using DenseCPT("active_1");
  }
%QSAR 2

  variable : qsar2 {
    type : discrete observed 3:3 cardinality 5;
    switchingparents : membership(0) using mapping("simpleSwitchDT");
    conditionalparents :
      nil using DenseCPT("inactive") |
      nil using DenseCPT("active_2");
  }
%QSAR 3

  variable : qsar3 {
    type : discrete observed 4:4 cardinality 5;
    switchingparents : membership(0) using mapping("simpleSwitchDT");
    conditionalparents :
      nil using DenseCPT("inactive") |
      nil using DenseCPT("active_3");
  }
```

```
 %INSERT_QSARS_HERE

 %Whether or not a motif is occuring.
 variable : motifOccur {
   type : discrete hidden cardinality 2;
   conditionalparents : motifClass(0) using DenseCPT("motifStart");
 }

 %Position in Motif, should be 0 and then may increment or not
 variable : motifPosition {
   type : discrete observed value 0 cardinality MOTIF_LENGTH;
   conditionalparents : nil using DeterministicCPT("zero");
 }

 %The actual motif.
 variable : motif {
   type : discrete hidden cardinality ALPHABET;
   conditionalparents : motifClass(0), motifPosition(0) using DenseCPT("motifDefs");
   weight: scale 2;
 }

 %the observed sequence.
 variable : sequence {
  type : discrete observed 0:0 cardinality ALPHABET;
  switchingparents : membership(0), motifOccur(0) using mapping("tripleSwitchDT");
  conditionalparents :
    nil using DenseCPT("uniformBackground") |
    nil using DenseCPT("background") |
    motif(0) using DeterministicCPT("copyMotif");
  }

}

%Normal frames
frame: 1 {

  %Sequence position, counts up to observed length
  variable : position {
    type : discrete hidden cardinality MAX_LENGTH;
    conditionalparents : position(-1)
```

```
     using DeterministicCPT("increment");
 }

%Whether or not a motif is occuring
variable : motifOccur {
  type : discrete hidden cardinality 2;
  switchingparents : motifOccur(-1), motifPosition(-1)  using mapping("motifOccurMapDT");
  conditionalparents :
    motifOccur(-1), motifPosition(-1) using DeterministicCPT("copyMotifOccur")
    |  motifClass(0) using DenseCPT("motifStart");
}

%Position in Motif, should be 0 and then may increment or not
variable : motifPosition {
  type : discrete hidden cardinality MOTIF_LENGTH;
  conditionalparents : motifPosition(-1), motifOccur(0), motifOccur(-1) using DeterministicCPT("motifCounter");
}

%Motif class, which is just copied
variable : motifClass {
   type : discrete hidden cardinality CLASS_NUMBER;
   conditionalparents : motifClass(-1) using DeterministicCPT("copyClass");
}

%membership class, copied over
variable : membership {
  type : discrete hidden cardinality 2;
  conditionalparents : membership(-1) using DeterministicCPT("copyBinary");
}

%The actual motif.
variable : motif {
  type : discrete hidden cardinality ALPHABET;
  conditionalparents : motifClass(0), motifPosition(0) using DenseCPT("motifDefs");
  weight : scale 2;
}

%the observed sequence.
variable : sequence {
 type : discrete observed 0:0 cardinality ALPHABET;
```

```
  switchingparents : membership(0), motifOccur(0) using mapping("tripleSwitchDT");
  conditionalparents :
  nil using DenseCPT("uniformBackground") |
    nil using DenseCPT("background") |
    motif(0) using DeterministicCPT("copyMotif");
  }
}

%epilogue. All that changes is the position is observed
frame : 2 {

  %Sequence position, counts up to observed length
  variable : position {
    type : discrete observed 1:1 cardinality MAX_LENGTH;
    conditionalparents : position(-1)
      using DeterministicCPT("increment");
  }

 %Whether or not a motif is occuring
 variable : motifOccur {
   type : discrete hidden cardinality 2;
   conditionalparents :  motifOccur(-1), motifPosition(-1) using DeterministicCPT("copyMotifOccur");
 }

 %Position in Motif, should be 0 and then may increment or not
 variable : motifPosition {
   type : discrete hidden cardinality MOTIF_LENGTH;
   conditionalparents : motifPosition(-1), motifOccur(0), motifOccur(-1) using DeterministicCPT("motifCounter");
 }

 %make sure we don't end in the middle of a motif
 variable : motifPositionCheck {
   type : discrete observed value 1 cardinality 2;
   conditionalparents : motifPosition(0) using DeterministicCPT("motifPositionCheck");
 }

 %Motif class, which is just copied
 variable : motifClass {
    type : discrete hidden cardinality CLASS_NUMBER;
    conditionalparents : motifClass(-1) using DeterministicCPT("copyClass");
```

```
 }


 %membership class, copied over
 variable : membership {
   type : discrete hidden cardinality 2;
   conditionalparents : membership(-1) using DeterministicCPT("copyBinary");
 }


 %The actual motif.
 variable : motif {
   type : discrete hidden cardinality ALPHABET;
   conditionalparents : motifClass(0), motifPosition(0) using DenseCPT("motifDefs");
   weight : scale 2;
 }

 %the observed sequence.
 variable : sequence {
  type : discrete observed 0:0 cardinality ALPHABET;
  switchingparents : membership(0), motifOccur(0) using mapping("tripleSwitchDT");
  conditionalparents :
    nil using DenseCPT("uniformBackground") |
    nil using DenseCPT("background") |
    motif(0) using DeterministicCPT("copyMotif");
  }
}

chunk 1:1


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%% Decisions Trees
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

DT_IN_FILE inline

10 % number of decision trees

% % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
```

```
0 % a DT that is always 0
zeroDT
0 % no parents
-1 0

% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
1 % a DT that copies the parent
copyDT
1 % 1 parent
-1 {p0}

% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
2 % a DT counts downwards
incrementDT
1 % 1 parent
-1 {p0 + 1}

% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
3 % a DT that counts upwards if a motif is occuring
motifCounterDT
3 % 2 parents. Parent 0 is motif position(-1) and parent 1 is the motif occurance(0)
1 2 1 default % are we in a motif?
  0 2 MAX_MOTIF_LENGTH default % are we at the end of a motif and beginning a new one?
    -1 0 % yes
    2 2 1 default % should we increment because we already expressed a position, or are we in the default 0
      -1 {p0 + 1} % no, increment
      -1 0 % yes, we still shouldn't increment
  -1 0 %not occuring, so don't increment

% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
4 % a DT that switches between two parents into three outcomes
tripleSwitchDT
2 % 2 parents
0 2 1 default %are you a member?
 -1 {p1 + 1} %yes, then use second parent offset by one (1 or 2)
 -1 0 % no, then use the non-member rvs (0)

% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
5 % Motif occurance decision tree.
motifOccurMapDT % A motif may occur if either we're not in a motif or we were and the motif ended.
```

```
2 % motifOccur(-1), motifPosition(-1)
0 2 1 default %Is there a motif occuring?
  1 2 MAX_MOTIF_LENGTH default %is the length at its highest value?
    -1 1 % yes it is. then a new motif may start
    -1 0 % it is not at the highest length, so a motif is occuring and we'll copy
  -1 1 %there is not a motif occuring, so allow one

% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
6 % Motif occurance copy decision tree.
copyMotifOccurDT % decision to copy motif occurance and end it if necessary
2 % motifOccur(-1), motifPosition(-1)
0 2 1 default %are we in a motif?
  1 2 MAX_MOTIF_LENGTH default % are we at the end of the motif length?
    -1 0 %Yes we were, thus the motif is over
    -1 1 %We are not, thus continue motif occurance
  -1 0 % no we are not in a motif and we didn't want to start one.

7 % checks that we are not in the middle of a motif. Emits 1 if we aren't
motifPositionCheckDT
1 % motifPosition(0)
0 3 0 MAX_MOTIF_LENGTH default
  -1 1;
  -1 1;
  -1 0;

8 % repeat motif
copyMotifDT
1 % motif(0)
-1 {p0}

% % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % % %
9 % a DT that just switches parents as expected
simpleSwitchDT
1 % 1 parent
-1 {p0}


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%% Non-training Deterministic CPT
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
DETERMINISTIC_CPT_IN_FILE inline 10
0 % first one
zero
0% no parents
MOTIF_LENGTH
zeroDT

1
copyPosition
1 % parent
MAX_LENGTH MAX_LENGTH
copyDT

2
copyMotifOccur
2 % parent
2 MOTIF_LENGTH 2
copyMotifOccurDT

3
copyClass
1 % parent
CLASS_NUMBER CLASS_NUMBER
copyDT

4
increment
1 % parent
MAX_LENGTH MAX_LENGTH
incrementDT

5
motifCounter
3 % parent
MOTIF_LENGTH 2 2 MOTIF_LENGTH
motifCounterDT

6
zeroPosition
0 % no parents
```

```
MAX_LENGTH
zeroDT

7
motifPositionCheck
1 % motif position
MOTIF_LENGTH 2
motifPositionCheckDT

8
copyMotif
1 % 1 parent
ALPHABET ALPHABET
copyMotifDT

9
copyBinary
1 % 1 parent
2 2
copyDT
```