Stabilization of Proteins against Aggregation by Solution Additives

by

Brian M. Baynes

Submitted to the Department of Chemical Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
[September 2005]
September 2004

Author .....
Department of Chemical Engineering
September 24, 2004

Certified by .....
Bernhardt L. Trout
Associate Professor of Chemical Engineering
Thesis Supervisor

Certified by .....
Daniel I. C. Wang
Institute Professor of Chemical Engineering
Thesis Supervisor

Accepted by .....
Daniel Blankschtein
Professor of Chemical Engineering
Chairman, Committee for Graduate Students

Stabilization of Proteins against Aggregation by Solution Additives

by

Brian M. Baynes

Submitted to the Department of Chemical Engineering
on September 24, 2004, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

In order to develop protein formulations that limit aggregation, researchers heuristically screen potential solution additives (excipients). Such screening is necessary because current understanding of mechanisms of aggregation and molecular-level effects of additives on aggregation is limited. In this study, we developed a statistical-mechanical method in order to model the thermodynamic effects of additives in molecular-level detail. This method uses no adjustable parameters and was validated by quantitative comparison with experimental data on proteins in glycerol and urea solutions. We then applied our molecular simulation technique to study the mechanism by which arginine, a common refolding buffer additive, deters protein aggregation. We find that arginine acts as a weak surfactant at the protein-solvent interface, with its guanidino group tending to face the protein. We propose that arginine is a member of a class of anti-aggregation additives, which we term "neutral crowders," characterized by their (1) negligible effect on the free energy of isolated protein molecules and (2) large size relative to water. With a simplified statistical-mechanical model, we have shown that such additives selectively increase the free energy of protein-protein encounter complexes by being preferentially-excluded from the gap between the protein molecules in such complexes. This "gap effect" will therefore slow protein association reactions. We showed experimentally that, in accordance with the gap effect model predictions, arginine slows association of model globular proteins (antibody+antigen) and of folding intermediates and aggregates of carbonic anhydrase II. We predict that neutral crowders larger than arginine will be superior anti-aggregation additives.

Thesis Supervisor: Bernhardt L. Trout
Title: Associate Professor of Chemical Engineering

Thesis Supervisor: Daniel I. C. Wang
Title: Institute Professor of Chemical Engineering

Acknowledgments

I have been very fortunate to have met and had the help of so many talented people during the course of my time at MIT.

My advisors, Profs. Bernardt Trout and Danny Wang, taught me what it means to do academic research. I am thankful for countless hours of consultation with them and for their thoughtful guidance, the mark of which is indelibly written into this thesis. I hope Prof. Wang can "find a pony in here" somewhere.

I have also benefited tremendously from the help of my thesis committee– Dr. Noubar Afeyan, Prof. Alan Hatton, Prof. Alex Klibanov, and Dr. Sharon Wang– all of whom provided challenge and a variety of useful insights from very diverse perspectives.

Dr. Jhih-Wei Chu and Dr. Jin Yin, my predecessors in protein stabilization group, helped me hit the ground running in the field. Aaron Risinger, Andre Ditsch, Brad Cicciarelli, and James and Christy Taylor provided expert assistance with equipment and experiments, and were always helpful when I came calling with new questions. Gwen Wilcox and Susan Lanza offered great advice and made sure everything ran smoothly. Compound Therapeutics, Inc. generously allowed me to use company equipment and conduct some key experiments in their laboratories.

I am also indebted to the Rosenblith Fellowship fund, the National Institutes of Health, and the National University of Singapore for financial support during my studies.

Finally, none of this would have been possible without the love and support of my parents, who have always been there for me in pursuit of my dreams.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Therapeutic proteins, such as insulin, interferon, and EPO (erythropoietin), represent an important and rapidly growing class of pharmaceuticals, presently accounting for $35B/yr in revenue worldwide. Proteins are useful as therapeutics because they have a wide range of physiological functions and are extremely potent. Natural proteins in the body, as well as man-made proteins, can often carry out their functions at extremely low concentrations, such as 10^{-9} M, 10^{-12} M, or even lower. Unfortunately, proteins are also only marginally stable, and are degraded and inactivated rapidly.

In industry, the inherent instability of proteins presents a serious problem, and a disadvantage relative to small molecule therapeutics. To optimally serve patients, it is desirable to store proteins at high purity and for long times, often for up to two years after manufacture [22]. Thus, proteins must not only be removed from their natural cellular environment, but they must also be stable against degradation for unnaturally long periods of time. This is the challenge faced by researchers and practitioners in the area of "protein stabilization."

Specific degradation routes that must be addressed include aggregation, deamidation, oxidation, and hydrolysis. Of these, the most prevalent is aggregation, the focus of this thesis.

Empirically, it has been observed that by adding low molecular weight components, such as salts, sugars, or polyols, to protein solutions, the propensity of the protein to aggregate (as well as degrade by other routes) can often be significantly affected. Unfortunately, because proteins are tremendously diverse in chemistry and structure, additives that work well for a particular protein generally do not work universally [87]. In addition, current understanding of the mechanisms by which additives confer stability on proteins is limited. Thus, there is often no theoretical guidance to aid selection of optimal additives.

This lack of understanding necessitates that protein stabilization be carried out on a case-by-case basis using heuristic experimental screens. This limits the additive search space and the possible formulation patent protection to those additive combinations which are explicitly tested. In some cases, additives that confer a useful level of stability cannot be identified.

As protein therapeutics branch out into new routes of administration, such as inhalers, implants, and stents, significant new stability challenges are presented. These new routes of administration involve protein-damaging factors such as atomization, elevated temperature, and high protein concentration, all of which can contribute in an unfavorable way to aggregation and other routes of degradation. Thus, there is an ever-increasing need to understand how to control these degradation processes to ensure that the full potential of proteins as therapeutics can be realized.

1.1 Thesis Objective

The objective of this thesis is to develop *fundamental understanding* of the *mechanisms* by which solution additives stabilize proteins against aggregation. This fundamental understanding will provide a basis for important future developments in *rational design and selection* of protein stabilizers.

1.2 Organization of This Thesis

The remaining seven chapters in this thesis give a synopsis, in roughly chronological order, of the approach taken and results obtained in pursuit of the above objective. Chapter 2 (Literature Review) summarizes the state of the art in the current literature. Chapters 3-5 (Results) cover the technical achievements of this thesis in detail. Finally, chapter 6 (Conclusions) concisely summarizes the major conclusions drawn from these results, and chapter 7 (Future Work) discusses opportunities for future work.

Chapter 2

Literature Review

2.1 The Connection between Protein Aggregation and Folding

Aggregation is a ubiquitous protein stabilization problem because aggregation is related to the natural process of protein folding. These two phenomena are united by a common driving force, the hydrophobic effect.

Protein folding is the process by which a nascent, unfolded polypeptide chain undergoes conformational transitions to ultimately arrive in its native state. One of the principal forces involved in this process is the hydrophobic force [26]. Hydrophobicity is a key driving force because in general, a significant fraction of the chemical groups in proteins are hydrophobic and prefer not to be solvated by water. When a protein is first synthesized, all of its functional groups, including the hydrophobic ones, are exposed to the solvent. Gradually, the protein orients itself spatially so these hydrophobic residues can be sequestered into a core away from the solvent. Because this hydrophobic sequestration brings together hydrophobic groups that are not proximal in the protein sequence, it imposes significant constraints on the protein's conformational entropy. In general, the result is one structure with an optimal balance of hydrophobic sequestration and conformational entropy. This state is called the "native state" and generally is the only one with biological activity.

Unfortunately, the process of hydrophobic sequestration can occur in an intermolecular fashion as well as in an intramolecular fashion as described above (Figure 2-1). When protein molecules sequester their hydrophobic residues in an intermolecular fashion, an aggregate species results. Thus, protein folding and aggregation are linked via the hydrophobic effect.

Aggregation may also proceed beyond the dimer species shown in Figure 2-1 to higher order mers. When aggregates become sufficiently large, they can phase separate and precipitate from solution.

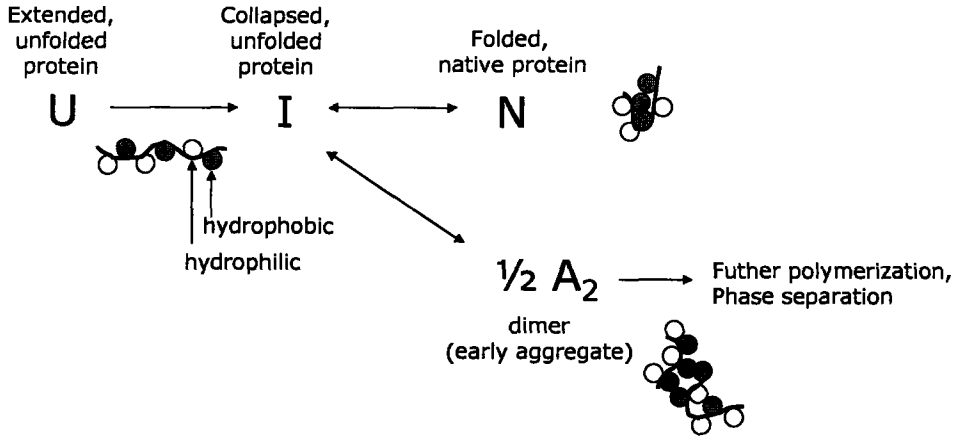

Figure 2-1: Protein folding and aggregation are driven by the hydrophobic effect. Proteins can sequester their hydrophobic side chains in either an intramolecular fashion (proper folding) or intermolecular fashion (aggregation).

2.2 Models of Protein Aggregation

2.2.1 Aggregation from the Native State

In a seminal 1954 paper, Lumry and Eyring [56] presented the first kinetic model of protein aggregation from the native state. They proposed that the native protein (N) undergoes an intramolecular transformation to an aggregation-competent intermediate (I) which associates to form the first aggregate (A_2). This framework is captured via the following two chemical reactions:

$$N \rightleftharpoons I \tag{2.1}$$

$$I + I \rightarrow A_2 \tag{2.2}$$

The most important contribution of this model is that it allows for the overall aggregation process to exhibit first or second-order kinetics, both of which have been observed experimentally [61, 46, 88, 43, 72]. Such aggregation studies are typically performed by isolating native protein, then subjecting the protein to either a thermal or denaturant stress to induce aggregation.

Either kinetic order can arise depending on which step in the reaction is rate-

limiting. If the intramolecular isomerization step is slower, the overall process will exhibit first-order kinetics. If the association step is slower, second-order kinetics will be observed.

Recently, Roberts [71] extended the Lumry-Eyring framework to consider explicitly the reaction of species larger than the dimer and the possibility of solubility limits being reached for these higher mers. This model was applied to predict the shelf life of a pharmaceutical protein (granulocyte colony simulating factor, GCSF) from data obtained in thermally-accelerated aggregation studies [72].

2.2.2 Aggregation during Folding

Kinetic studies have also been performed on aggregation during protein folding or refolding. Such experiments are typically performed by denaturing a protein in a high concentration of guanidinium chloride or urea, and then diluting or dialyzing away the denaturant to allow the protein to refold. Since when the denaturant is removed, the protein is already in an unfolded or partially-unfolded and aggregation-competent state, aggregation may proceed directly from this initial state and there is no requirement of an intermediate specie as in the Lumry-Eyring model. In refolding, there is direct, kinetic competition between proper folding and aggregation:

$$U \rightarrow N \tag{2.3}$$

$$U + U \rightarrow A_2 \tag{2.4}$$

where U, N, and A_2 are the unfolded, native, and dimer states, respectively.

Since the refolding and aggregation reactions are of different order (the refolding reaction generally being first-order and the association reaction being second-order), this kinetic competition can be observed by measuring the yield of native protein as a function of the initial protein concentration. This type of experiment has been performed by Zettlmeissl et al. on lactate dehydrogenase [94] and by Hevehan et al. on lysozyme [40].

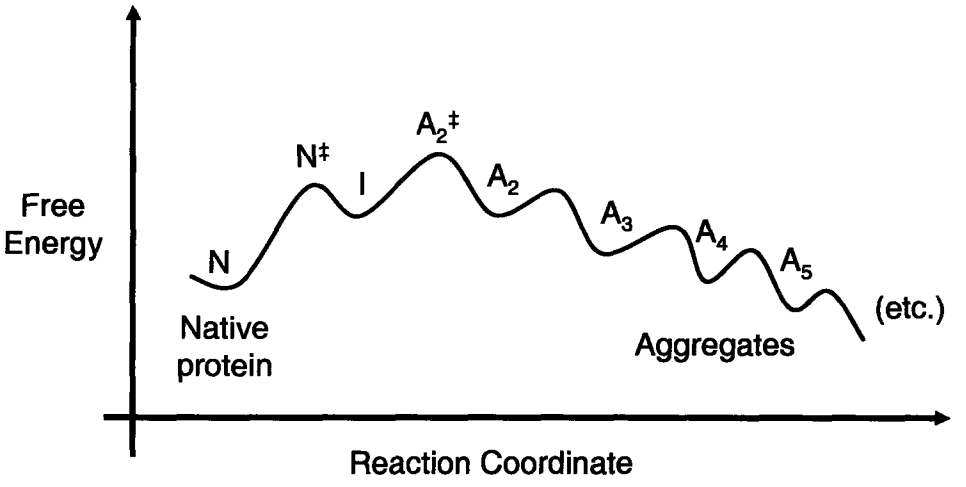

Figure 2-2: Generalized free energy-reaction coordinate diagram for protein aggregation.

2.2.3 Unified Reaction Coordinate-Free Energy Diagram

The models presented for aggregation from the native state and from the unfolded state can been unified as shown in the reaction coordinate-free energy diagram in Figure 2-2. The figure qualitatively shows the free energies of the native protein, unfolded or partially-unfolded intermediate state, multiple aggregate states at different mer number (A_i), and the transition states between these states. Note that the relative free energy values are meant only to be illustrative at this point. As drawn, the native protein is only metastable with respect to large aggregates. It is separated from the large aggregate states via a free energy barrier that is maximal at the transition state for dimer formation, $A_2^{\ddagger}$.

Under such conditions, aggregation from the native state would exhibit second-order kinetics and would eventually proceed to completion, that is, the formation of very large aggregates, with no native protein remaining. The experimentally-observable aggregation rate (or rate of loss of native protein) can be related to the reaction coordinate-free energy diagram via transition state theory:

$$k = \nu \frac{k_B T}{h} \mathrm{e}^{-\Delta\mu_{\ddagger}/RT} \tag{2.5}$$

where k is the rate constant, ν is the transmission coefficient, k_B is Boltzmann's constant, T is the absolute temperature, h is Planck's constant, $\Delta\mu_{\ddagger}$ is the activation free energy, and R is the gas constant. In this case, the appropriate activation free energy is the free energy difference between the transition state to form the dimer and that of the native state, $\mu_{A_2^{\ddagger}} - \mu_N$. If aggregation from the native state exhibits first-order kinetics, then it can be assumed that the intramolecular isomerization ($N \rightarrow I$) is rate limiting, and the appropriate activation free energy is $\mu_I - \mu_N$ [46, 88].

The same reaction coordinate-free energy diagram can be used to visualize the refolding process. In this case, the initial state is the unfolded or partially-unfolded intermediate, I. This species can either refold to form N, or aggregate to form A_2 and subsequent larger mers. After this initial kinetic partitioning which forms native protein and aggregates, the native protein will slowly aggregate. Ultimately, the system will consist entirely of large aggregates, as above.

2.3 Theories of Additive Effects on Proteins

The presence of a solution additive potentially affects the free energy of all the states along the folding-aggregation reaction coordinate. When such free energy effects have different magnitudes at different points along the reaction coordinate, they give rise to changes in experimentally observable quantities such as equilibrium constants and reaction rates. Most importantly, the aggregation rate can change. To understand how additives affect aggregation, we must understand how they affect the free energy barrier of the rate-limiting step in the aggregation process.

For any protein state of interest along the reaction coordinate, the effect of an additive on the free energy of a state can be expressed in terms of a transfer free energy, via:

$$\mu_P = \mu_{P,0} + \Delta\mu_P^{tr} \tag{2.6}$$

where μ_P is the free energy of the state in the mixed solvent, $\mu_{P,0}$ is the free energy of the state in the reference solvent, and μ_P^{tr} is the transfer free energy from the reference

solvent to the mixed solvent.

Because of the central role of transfer free energies in describing additive effects on protein processes, a significant body of theoretical and experimental work has been done over the past forty years in this area. These contributions are summarized in the following sections.

2.3.1 Measurement via the Preferential Binding Coefficient

For any stable protein state, it is possible to measure the transfer free energy via the preferential binding coefficient. These two quantities are related in the following way [51]:

$$\Delta\mu_P^{tr} = \int_0^{m_X} \left(\frac{\partial\mu_P}{\partial m_X}\right)_{m_P} dm_X \tag{2.7}$$

$$= -\int_0^{m_X} \left(\frac{\partial\mu_X}{\partial m_X}\right)_{m_P} \left(\frac{\partial m_X}{\partial m_P}\right)_{\mu_X} dm_X \tag{2.8}$$

where $\Delta\mu_P^{tr}$ is the transfer free energy of the protein from pure water into the mixed solvent system, m is molality, and subscripts X and P identify the additive and protein respectively. Two partial derivatives appear in equation 2.8. The first captures the dependence of the additive chemical potential on additive molality and can be evaluated by experiments on a binary mixture of additive and water ($m_P \to 0$). The second partial derivative is the "preferential binding coefficient," Γ_{XP}:

$$\Gamma_{XP} \equiv \left(\frac{\partial m_X}{\partial m_P}\right)_{\mu_X} \tag{2.9}$$

The preferential binding coefficient is the thermodynamic definition of binding. It is also a measure of the excess number of additive molecules in the domain of the protein per protein molecule (Figure 2-3). The connection between the thermodynamic definition (equation 2.9) and the intuitive notion of binding (local excess number of

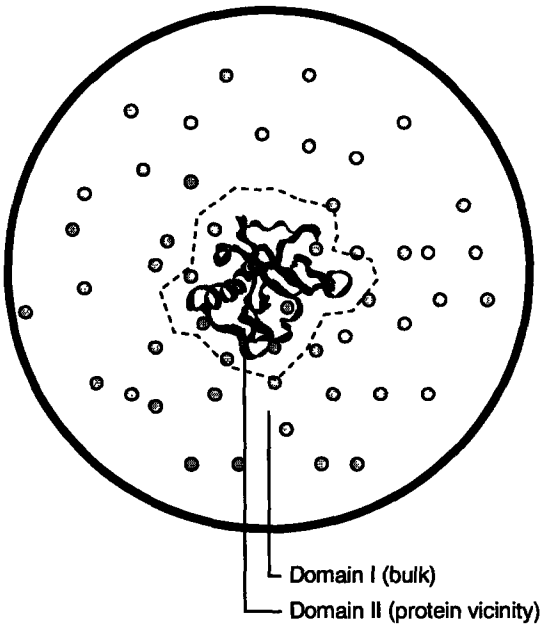

Figure 2-3: Physical interpretation of the preferential binding coefficient. Interactions of solvent molecules with the protein at the protein-solvent interface generally induce solvent concentration differences in the local (II) and bulk (I) domains. Γ_{XP} is the thermodynamic measure of the number of additive molecules bound to the protein, or in other words, the excess number of additive molecules in the vicinity of the protein versus the number of additive molecules in an equivalent volume of bulk solution.

molecules) comes from statistical mechanics, where Schellman has shown that [47, 75]:

$$\Gamma_{XP} = \left\langle n_X^{II} - n_W^{II} \left(\frac{n_X^I}{n_W^I} \right) \right\rangle \tag{2.10}$$

In the above equation, n denotes the number of a specific type of molecule (subscript X for the additive and subscript W for water) in a certain domain (superscript I for a bulk volume outside of the vicinity of the protein and superscript II for a volume in the protein vicinity), and angle brackets denote an ensemble average. Note that Γ_{XP} is independent of the choice of the boundary between the domains, as long as the boundary is far enough from the protein.

If the additive concentration is higher in the vicinity of the protein than in the bulk, Γ_{XP} is greater than zero, and μ_P is lower in the presence of the additive than in its absence. Denaturants such as urea and guanidinium chloride exhibit this type of

binding behavior [84, 37, 50, 53]. The reverse is true for sugars and polyols, such as trehalose, sucrose, and sorbitol [3, 32, 33, 4, 51, 91, 92, 93]. In one of these solutions, there is generally a deficiency of the sugar or polyol and an excess of water in the vicinity of the protein. For this "preferential hydration" case, Γ_{XP} is less than zero, and μ_P is higher in the presence of the additive.

Preferential binding coefficients have been studied extensively by high-precision densitometry over the past thirty years in the laboratory of Serge Timasheff [50]. Earlier techniques based on sedimentation [39] and isopiestic composition measurements [37] have also been employed. More recently, differential scanning calorimetry (DSC) [66] and vapor pressure osmometry (VPO) [25] have been used to a similar end. Preferential binding coefficients are rigorous thermodynamic quantities and are related to virial coefficients, activity coefficients, and free energies via standard thermodynamic relations for multi-component solutions [19].

Experimental studies by the above methods have led to some generalizations about preferential binding coefficients:

1. Γ_{XP} may be positive or negative, indicating that interactions of the protein and cosolvent are favorable or unfavorable, respectively.

2. Γ_{XP} is proportional to cosolvent molality at low concentration of cosolvent (often as high as $m_X \sim 1$ m and higher) [25, 35, 68].

3. Γ_{XP} is roughly proportional to the protein-solvent interfacial area [51].

The second generalization above, together with the fact that many binary mixtures of cosolvent and water ($m_P \rightarrow 0$) are nearly ideal at low concentration of cosolvent, leads to a useful simplification of equation 2.8:

$$\Delta\mu_P^{tr} = -\int_0^{m_X} \left(\frac{\partial RT \ln m_X}{\partial m_X}\right)_{m_P} \left(\frac{\Gamma_{XP}}{m_X}\right) m_X \, dm_X \tag{2.11}$$

$$= -RT \left(\frac{\Gamma_{XP}}{m_X}\right) \int_0^{m_X} dm_X \tag{2.12}$$

$$= -RT \, \Gamma_{XP} \tag{2.13}$$

Equation 2.13 provides a simple and convenient link between preferential binding coefficients and free energies. This relation leads to the useful rule that when Γ_{XP} is proportional to m_X, for each cosolvent molecule that preferentially interacts with the protein, the protein's free energy is reduced by approximately 0.6 kcal/mol at 25°C. The simplicity of this relation is a natural result of the close relationship between Γ_{XP} and a second virial coefficient.

When approximation 2 above also applies, the relationship between preferential binding and transfer free energies can be expanded to

$$\Delta\mu_P^{tr} = -RT\ \gamma_{XP} a_P m_X \tag{2.14}$$

where γ_{XP} is the preferential binding coefficient per surface area of the protein per concentration of additive and a_P is the protein-solvent interfacial area. This formalism is the most convenient method for evaluating the transfer free energy effects of additives such as denaturants, polyols, and sugars, for which all of the above approximations are valid.

2.3.2 Relation to Mechanistic Models

To be able to predict preferential binding coefficients and understand their origins, the above thermodynamic framework and general observations must be augmented by a mechanistic model. Several such models have been presented in the literature, including models based on the binding polynomial or statistical mechanical partition function, solvent-cosolvent exchange at defined sites, cosolvent partitioning between the local and bulk domains, group contribution methods for estimating transfer free energies. Some specific models are discussed below.

Binding Polynomial

The most general model of how an additive affects the free energy of a state comes from considering an equilibrium of all possible protein-cosolvent complexes, from

which it can be shown that [90]:

$$\Delta\mu_P^{tr} = -RT\ln(1 + \sum_i \sum_j K_{ij} c_W^i c_X^j) \tag{2.15}$$

where K_{ij} is the association equilibrium constant for a reaction of a protein molecule, i molecules of water, and j molecules of cosolvent into a complex; c_W is the concentration of water; and c_X is the concentration of the additive. This formalism is related to the so-called "binding polynomial" due to Wyman. While this model is completely general, its utility is limited because it is not possible to determine experimentally the many K_{ij} parameters present in equation 2.15.

Weak Binding at Multiple Equivalent Sites

If the additive binds at a large number of equivalent sites, the binding polynomial representation (equation 2.15) reduces to:

$$\Delta\mu_P^{tr} = -nRT\,\langle K\rangle\, c_X \tag{2.16}$$

where n is the number of sites and $\langle K\rangle$ is the average association equilibrium constant at a site. The single parameter $\langle K\rangle$ can then be determined from an experimental measurement of the transfer free energy or Γ_{XP}. When equation 2.13 holds, the relation between $\langle K\rangle$ and Γ_{XP} is simply:

$$\langle K\rangle = \Gamma_{XP}/n\; m_X \tag{2.17}$$

Values of $\langle K\rangle$ for the same additive with different proteins in the linear binding regime are roughly equal [76]. $\langle K\rangle$ cannot, however, be determined without knowledge of Γ_{XP} or other free energy data on the particular cosolvent system of interest. In fact, one can say that $\langle K\rangle$ is defined by Γ_{XP}.

Single Site Binding

If the additive binds only at a single site, the binding polynomial representation (equation 2.15) reduces to:

$$\Delta\mu_P^{tr} = -RT\ln(1 + Kc_X) \tag{2.18}$$

where K is the association equilibrium constant for the additive and the protein.

This formalism is the most convenient method for evaluating the transfer free energy effects of strong binding additives such as surfactants, antibodies, and folding chaperones.

2.3.3 Other Mechanistic Models

Local-bulk Domain Model

Another model that recasts preferential binding coefficient data in terms of a single model parameter is the local-bulk domain model developed by Courtenay et al [25]. The parameter in this model is the partition coefficient K_P, relating the number of water molecules and cosolvent molecules in the local and bulk domains via:

$$K_P = \frac{n_X^{II}/n_W^{II}}{n_X^{I}/n_W^{I}} \tag{2.19}$$

Similar to the Schellman site exchange model, the convention used in this model is that the local domain consists of a monolayer of water and enough cosolvent to obtain the experimentally observed Γ_{XP}. Note that because the absolute occupancy of water and cosolvent in the local domain cannot be easily determined by experiment, the local-bulk domain model effectively defines n_W^{II}. Like $\langle K \rangle$, values of K_P can be used to predict Γ_{XP} at other cosolvent concentrations or for other proteins in the same cosolvent, but predictions cannot be made in the absence of Γ_{XP} or free energy data on the same cosolvent system.

Group Contribution Methods

Lastly, transfer free energy group contribution models, pioneered by Bolen's group [55], take a different approach. These models conceptually divide whole proteins into groups [84] such as the amino acid side chains and the protein backbone and model the transfer free energy of the whole protein as a sum of the transfer free energy of the groups it comprises, via:

$$\Delta\mu_P^{tr} = \sum_i \alpha_i \Delta g_i^{tr} \tag{2.20}$$

where Δg_i^{tr} is the transfer free energy of the model group and α_i is the solvent accessible area of the group in the whole protein, normalized to the solvent accessible area of the model compound. The overall $\Delta\mu_P^{tr}$ can then be predicted for any system of known structure. In the context of the previously described models, the transfer free energy model can be thought of as a linearized binding model where each surface group or amino acid in the protein represents a different type of independent binding site, and the binding constants for those sites are determined by experiments on model compounds, such as free amino acids or cyclic di-amino acid compounds. Predictions made by transfer free energy models have met with mixed success. A linear group contribution model (equation 2.20) may be too simple to capture all of the important contributions to $\Delta\mu_P^{tr}$ [14].

Ionic (Debye-Huckel) Binding

If the additive is ionic and the charge of the protein mers is nonzero, an additive can also affect aggregation rates via a Debye-Huckel ionic screening effect. The transfer free energy of a charged molecule into a solvent with an ionic additive is then [59]:

$$\Delta\mu_P^{tr} = -\frac{\kappa}{\epsilon} q^2 \tag{2.21}$$

where ϵ is the permittivity of the solvent medium, q is the charge of the molecule, and κ is the inverse Debye length:

$$\kappa = \sqrt{\frac{4\pi}{k_B T \epsilon} \sum_i q_i^2 c_i} \tag{2.22}$$

where the sum is over all ionic species in the medium, q_i is the charge on a species, and c_i it its concentration.

At infinite separation, two protein molecules act independently and have a total transfer free energy:

$$\Delta\mu_P^{tr} = -\frac{2\kappa}{\epsilon} q^2 \tag{2.23}$$

Once in an encounter complex or associated state, the transfer free energy of the state is:

$$\Delta\mu_P^{tr} = -\frac{\kappa}{\epsilon}(2q)^2 = -\frac{4\kappa}{\epsilon} q^2 \tag{2.24}$$

Thus, increasing ionic strength favors aggregate formation by a factor of $2\kappa q^2/\epsilon$.

2.4 Experimental Observations of Different Additive Effects

In a few specific cases, experimentally observed effects of additives on aggregation processes have been related to a mechanism. These observations and conclusions are summarized below.

2.4.1 Single Site Binding

In studies of refolding of carbonic anhydrase, interferon-γ, tissue plasminogen activator, and deoxyribonuclease, Cleland et al. found that polyethylene glycol (PEG) favored refolding over aggregation (Figure 2-4) [20, 21]. They showed that the mechanism behind this change was that PEG bound selectively to a folding intermediate, rendering it unable to associate. Thus, binding decreases the free energy of the

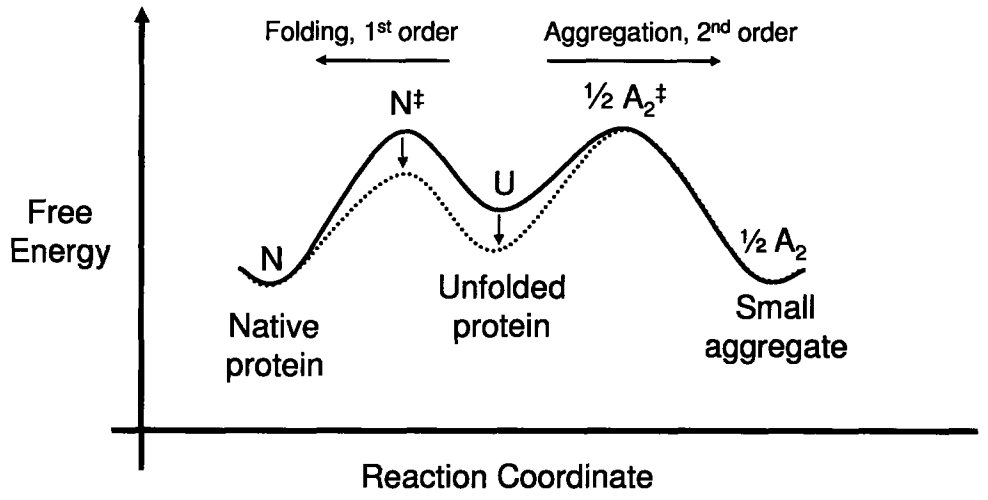

Figure 2-4: In refolding of carbonic anhydrase, interferon gamma, and DNase from the unfolded state (U), Cleland [20] showed that polyethylene glycol (PEG) binds selectively to the unfolded protein and folding intermediates. This slows aggregation and increases the yield of native protein. The free energy in the absence of additive is shown as a solid line and, in the presence of PEG, as a dotted line.

unfolded protein and refolding transition state, increases the activation energy for aggregation, slows the rate of aggregation, and increases the final yield of active protein.

In vivo it is known that folding chaperones such as the GroEL/GroES system sequester unfolded proteins in their core, allow the nascent proteins to fold in isolation, and then release them into the cytoplasm [38]. This can be modeled as a selective binding of the unfolded state and folding intermediates by the chaperone system which increases the free energy barrier to aggregation. Thus, the chaperone system is analogous to the PEG case above in a mechanistic sense.

Similarly, Dumoulin et al [29] found that a cameloid antibody (cAb) bound native D67H lysozyme and prevented it from forming amyloid fibrils. The mechanism by which the cAb acted was by specific binding of the native protein but not the transition state to form the aggregation transition state. The binding free energy of the cAb to lysozyme thus constituted an increase in the activation free energy for association, analogously to the way PEG deterred aggregation of CA from its molten state (Figure 2-5).

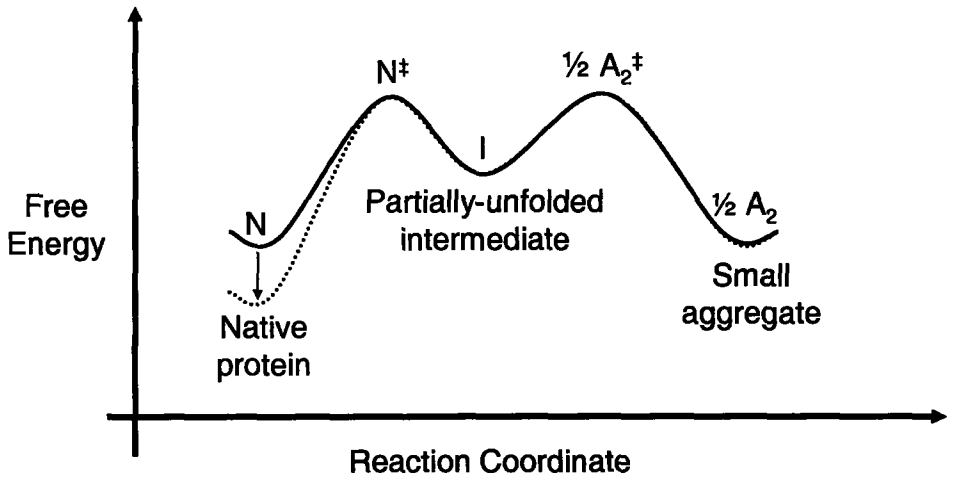

Figure 2-5: Dumoulin et al. [29] found that an antibody selectively bound native D67H lysozyme with high affinity and prevented it from aggregating into amyloid fibrils. The free energy in the absence of additive is shown as a solid line and, in the presence of the antibody, as a dotted line.

2.4.2 Nonspecific Preferential Binding

Sucrose and other molecules that are preferentially-excluded from the protein-solvent interface have been shown to stabilize native protein molecules against unfolding [51, 4, 3]. Analogously, in studies of interferon-γ aggregation from the native state, Kendrick et al. [46] found that aggregation in 0.9M guanidinium chloride (GuHCl) exhibited first-order kinetics, and that sucrose deterred aggregation linearly with concentration (Figure 2-6). They modeled this observation using the Lumry-Eyring framework with the first-order conformational change being rate limiting for the protein aggregation process, and a preferential binding model for the additive effect on this process. Webb et al. [88] made a similar observation and used the same mechanism to explain experimental data at 0.45M GuHCl for the same model system.

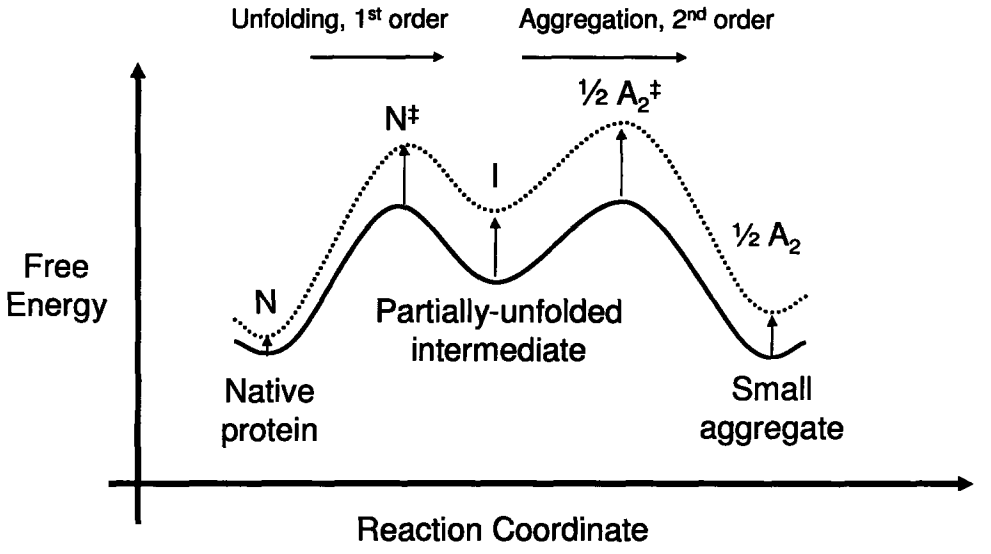

Figure 2-6: In studies of interferon-gamma aggregation from the native state, Kendrick *et al.* [46] showed that sucrose increases the activation free energy for formation of aggregates (A_2) from the native state (N). The free energy in the absence of additive is shown as a solid line and, in the presence of sucrose, as a dotted line.

Chapter 3

Computation of Preferential Binding Coefficients with no Adjustable Parameters

While the mechanistic models discussed in sections 2.3.2 and 2.3.3 have helped in the understanding of the phenomenon of preferential binding, they generally incorporate strong assumptions, and they necessitate the use of experimental data on highly analogous systems in order to determine model parameters and make predictions. Thus, their uses as predictive tools and as tools to gain insight into specific systems are limited.

In this chapter, we develop a predictive, molecular-level approach for the study of preferential binding based on all-atom, statistical mechanical models that use no adjustable parameters. To date, statistical mechanical models of preferential binding have only been developed for interactions of ions with charged cylinders [2, 60] and for interactions of two-dimensional, "hard circles" with a linear interface [85], both far too simple to be generally applied to protein-additive systems. Other explicit mixed solvent simulations of proteins and amino acids have been performed [95, 12, 86, 1, 18], but these studies did not compute thermodynamic quantities related to preferential binding. In our approach, we define the number of "bound" molecules in a thermodynamically consistent way and do not *a priori* incorporate any information about "binding sites." The use of our approach for the computation of preferential binding coefficients was validated in two systems by comparison with experimental data from the literature. Additionally, the molecular-level detail of the approach provides new insights into the following issues:

1. The changes in solvent and additive concentration as a function of distance from the protein surface.

2. A precise definition of the "local domain" (Figure 2-3).

3. The differences in preferential binding or apparent binding equilibrium constant at different locations on the protein-solvent interface.

The success of this method in modeling preferential binding indicates that it captures the important underlying physics of protein-additive-water systems and that the difficulty in quantitative prediction to date can be surmounted by explicitly incorporating the complex protein-solvent and solvent-solvent interactions.

3.1 Computational Approach

Our strategy is to use molecular dynamics to generate an equilibrium ensemble of protein, water, and additive. From this ensemble, the average concentration of additive as a function of position relative to the protein can be determined. This enables the preferential binding coefficient to be computed directly from a single trajectory.

To accomplish this, we utilize explicit atomic interaction potentials (force fields), such as Lennard-Jones, Coulombic, spring, and torsion interactions, with pre-fit coefficients [15, 36]. Thermodynamic properties (such as preferential binding coefficients) are computed by averaging in the time domain via molecular dynamics (MD). A snapshot from a dynamic simulation of RNase T1 in a urea solution is shown in Figure 3-1.

Molecular dynamics uses Newton's second law of motion, that acceleration is the quotient of force and mass, to compute the positions of each atom in the system as a function of time. To do this, an energy model, sometimes called a "force field," that can be used to compute the net force on any atom in any configuration is employed.

During the MD run, the positions of each atom are recorded at fixed intervals in time. These "snapshots" form an ensemble of configurations which can then be used to compute thermodynamic properties, such as Γ_{XP}.

Importantly, this method of computing Γ_{XP} does not introduce any adjustable parameters to model preferential binding or any other aspect of a system containing a protein and two solvent components. All of parameters required by the MD method for energy computations are determined independently of this particular modeling objective, and in fact have been shown to be generally applicable to biological systems [45]. Thus, the method developed here could be used to estimate Γ_{XP} and $\Delta\mu_P^{tr}$ in systems where no experimental data is available. It therefore facilitates the study of preferential binding when direct experimental study is difficult, such as at transition state configurations or at marginally stable states of proteins. Furthermore, it yields detailed, local, molecular-level insight into the system studied.

Another benefit of this approach is that when equation 2.13 holds (such as for urea

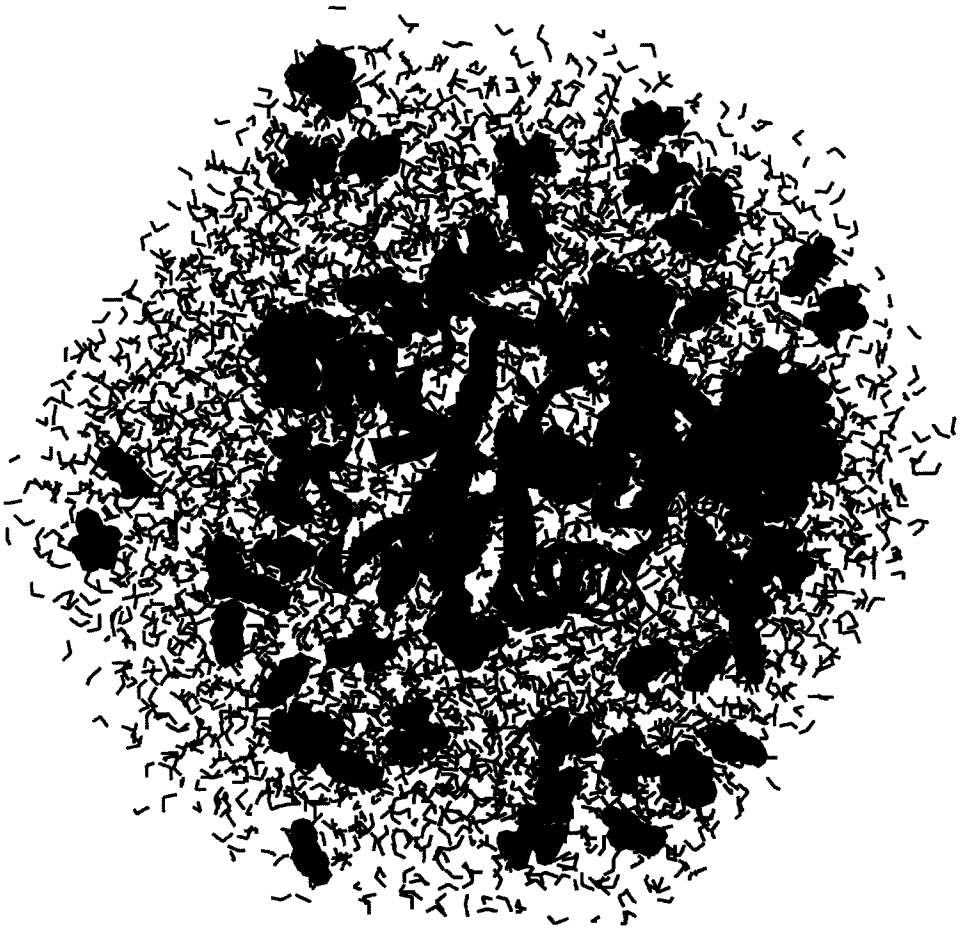

Figure 3-1: A simulation cell containing RNase T1 (ribbon) solvated by water (thin lines) and urea (spheres). Figure generated with VMD [42].

and glycerol), the protein transfer free energy ($\Delta\mu_P^{tr}$) can be calculated from a single Γ_{XP} simulation. Traditional free energy calculation methods such as thermodynamic integration [9, 48] require 15-20 trajectories, which is computationally difficult for protein systems of this size.

3.1.1 Preferential Binding Coefficients of Constituent Groups

Because proteins have a range of different functional groups in different orientations on their surfaces, the concentrations of solvent and additives near different patches on the protein's surface may be different. For example, the vicinity of a hydrophobic patch on the protein may have a lower concentration of water and a higher concentration of additive than in the vicinity of a hydrophilic patch. Preferential binding experiments capture only the average effect arising from all of the interactions over the entire protein-solvent interface; however, molecular simulations allow more detailed analyses of the local contributions to preferential binding coefficients.

A protein can be thought of as a set of non-overlapping constituent groups [84], each of which has its own preferential binding coefficient defined by the composition of the solvent in its immediate vicinity. Similar to group contribution methods for computing transfer free energies (see Introduction), one possible group definition is that each type of amino acid side chain (up to 20) and the amino acid backbone are distinct groups. To compute a preferential binding coefficient for a constituent group, the solvent molecules in the local domain are assigned only to the nearest group (i), and the "group preferential binding coefficients" ($\Gamma_{XP,i}$) can be defined as:

$$\Gamma_{XP,i} \equiv \left\langle n_{X,i}^{II} - n_{W,i}^{II} \left(\frac{n_X^I}{n_W^I} \right) \right\rangle \tag{3.1}$$

where $n_{X,i}^{II}$ and $n_{W,i}^{II}$ are the number of additive and water molecules in the local domain that are nearest to group i. If each solvent molecule in the local domain is assigned to a group, the overall preferential binding coefficient is simply the sum of

all of the group preferential binding coefficients:

$$\Gamma_{XP} = \Sigma\Gamma_{XP,i} \tag{3.2}$$

The group preferential binding coefficients decompose the effect of each small subset of the protein on the overall preferential binding coefficient. This is analogous to the group contribution models for transfer free energy except that the parameters are extracted from a simulation of an entire protein instead of experiments on model compounds.

3.1.2 Minimum Simulation Time

Sufficient sampling of position-space configurations in time is required for the accurate calculation of Γ_{XP} via equation 2.9. Assuming that the average protein solution structure is close to that of the initial (crystal) structure and that water molecules sample position space rapidly because of their high density, the most important time scale to be captured is that of the additives sampling position space. One way to estimate this time is that it must be much larger than the average time between additive-additive contacts.

An estimate of the time between contacts can be obtained as:

$$t_{contact} \approx \frac{1}{12\mathcal{D}} \left(\frac{V_{solv}}{n_X} \right)^{\frac{2}{3}} \tag{3.3}$$

where $\mathcal{D}$ is the additive diffusivity, V_{solv} is the solvent volume, and n_X is the number of additive molecules. For the simulations performed here, the solvent is mostly water, so equation 3.3 can be further simplified to yield:

$$t_{contact} \approx \frac{1}{12\mathcal{D}} \left(\frac{1}{N_A \rho_W m_X} \right)^{\frac{2}{3}} \tag{3.4}$$

where N_A is Avogadro's number and ρ_W is the density of water in kg/m^3. For a 1m additive in water system with an additive diffusivity of 2x10^{-9} m^2/s (a lower

bound on the diffusivities of the additives studied here), $t_{contact}$ is about 30ps. Thus, nanosecond trajectories will be required for good sampling of additive position space. Importantly, this time increases as the additive concentration decreases, implying that there is a minimum concentration that can be studied with any given amount of computational resources.

3.2 Methodology

3.2.1 Molecular Simulations

Molecular dynamics was used to sample the phase space of proteins solvated by water and an additive. Version 28 of the CHARMM [15] molecular dynamics package was used for all simulations. The CHARMM force-field was used for the protein, and the TIP3P model [44] was used for water. A force-field was constructed for glycerol using the standard CHARMM geometries and partial charges for the atoms in a -CHOH- unit [15, 36]. Urea was assumed to be planar with bond lengths equal to the CHARMM standards and partial charges recomputed as done previously [28] but using the CHARMM van der Waals mixing rules in the objective function.

The structures of RNase A (PDB code: 1fs3) and RNase T1 (PDB code: 1ygw) were obtained from the Protein Data Bank [13]. In total, three simulations were performed: RNase A in 1m glycerol (pH 3), RNase T1 in 1m glycerol (pH 7), and RNase T1 in 1m urea (pH 7). Details of each simulation are shown in Table 3.1. Each protein was solvated in a trucated octahedral box extending a minimum of 9Å from the protein. The pH of each simulation was fixed by setting the protonation states of each ionizable side chain to the dominant form expected for each amino acid at the pH of interest. Arginine, cysteine, lysine, and tyrosine were protonated in all of the simulations. Aspartate, glutamate, and histidine were assumed to have pKa values of 3.4, 4.1, and 6.6 [31, 30], respectively, and were therefore protonated in the simulation at pH 3 and deprotonated at pH 7. Initial placement of water and additive molecules were random. Protein counterions were placed using SOLVATE 1.0. The system was

Additive	Protein	T (°C)	pH	n_X	n_W	$<l>$ (Å)
Urea	RNase T1	25	7	90	4274	57.48
Glycerol	RNase T1	25	7	87	4582	59.24
Glycerol	RNase A	25	3	90	5480	62.86

Table 3.1: Details of four molecular dynamics (MD) simulations performed. n_X is the number of additive molecules; n_W is the number of water molecules; and $<l>$ is the average dimension of the primary unit cell (which varies during the run at constant pressure).

first energy minimized at 0K, next heated to 298.15K, and then equilibrated for 1ns in the NTP ensemble at one atmosphere. For the computation of the properties of interest, two nanoseconds of dynamics were then run, during which statistics were computed from snapshots of the trajectory every picosecond.

3.2.2 Calculation of Preferential Binding Coefficients

The trajectories were then used to define the local and bulk regions and compute Γ_{XP} in the following manner. For the purpose of computing Γ_{XP} and other thermodynamic and structural parameters, each water and additive molecule was treated as a point at its center of mass. The distance of each of these points to the protein's van der Waals surface was computed, and then $\rho_W(r)$ and $\rho_X(r)$, defined as the number densities of these points at a distance r from the protein, were computed. In all cases, the $\rho(r)$ functions exhibited peaks and valleys characteristic of solvation shells in the range $0 < r < 6$Å. At distances in the range of 6-8Å and higher, such variations are no longer seen, and the local number density is defined as bulk number density, $\rho(\infty)$. Such a region far from the protein containing a spatially uniform concentration of water and additive must be present in the simulation cell in order to define the local and bulk regions and calculate Γ_{XP}.

The position of the boundary between the local and bulk domains, a distance of r_* away from the surface of the protein, was then determined by choosing the minimum distance at which no significant difference between $\rho(r_*)$ and $\rho(\infty)$ was apparent for either water or additive. All solvent molecules whose centers of mass fell inside a distance of r_* from the protein's van der Waals surface were defined as belonging to

the local domain (II), and all other solvent molecules were defined as belonging to the bulk domain (I). With these definitions of the domains, the instantaneous preferential binding coefficient, $\Gamma_{XP}(t)$, was computed as

$$\Gamma_{XP}(t) \equiv n_X^{II} - n_X^I \left(\frac{n_W^{II}}{n_W^I} \right) \tag{3.5}$$

for each time point in each trajectory. The preferential binding coefficient, Γ_{XP}, was then computed for each trajectory as the time average of these instantaneous values:

$$\Gamma_{XP} = \frac{1}{t} \int_0^t \Gamma_{XP}(t')dt' \tag{3.6}$$

The radial distribution functions $g_X(r)$ and $g_W(r)$ are defined as:

$$g_i(r) \equiv c_i(r)/c_i(\infty) \tag{3.7}$$

where i represents water (W) or an additive (X) species and c is concentration. These functions provide another route to compute Γ_{XP}:

$$\begin{aligned} \Gamma_{XP} &= \langle n_X^{II} \rangle - \left\langle \left(\frac{n_X^I}{n_W^I} \right) n_W^{II} \right\rangle & (3.8) \\ &= c_X(\infty) \int g_X \, dV - \left(\frac{c_X(\infty)}{c_W(\infty)} \right) c_W(\infty) \int g_W \, dV & (3.9) \\ &= c_X(\infty) \int (g_X - g_W) \, dV & (3.10) \end{aligned}$$

where each integral is over the local domain or the entire system (since $g_X - g_W = 0$ in the bulk domain).

The boundary between domains I and II must be placed far enough from the protein to ensure that it is in the bulk, yet at the smallest such distance so that statistical fluctuations in the number of molecules in the domains can be minimized. We can use the values of $g_X(r)$ and $g_W(r)$ to determine the optimal boundary. Defining Γ_{XP}^* as the apparent preferential binding coefficient resulting from defining the local domain

as those molecules whose centers of mass lie inside a distance r_* from the protein:

$$\Gamma^*_{XP}(r_*) = c_X(\infty) \int_0^{r_*} (g_X - g_W) \frac{dV}{dr}\, dr \tag{3.11}$$

The error in Γ_{XP}, E_Γ, introduced by selecting a particular value of r_* is then

$$\begin{aligned} E_\Gamma &= \Gamma^*_{XP}(r_*) - \Gamma_{XP} && (3.12) \\ &= -c_X(\infty) \int_{r_*}^{\infty} (g_X - g_W) \frac{dV}{dr}\, dr && (3.13) \end{aligned}$$

When r_* is selected properly, the surface defined by $r = r_*$ is entirely in the bulk solution, $g_X(r_*) = g_W(r_*) = 1$, and $E_\Gamma = 0$. Thus, selecting r^* as the minimum distance for which all $r \geq r^*$ satisfy $g_X(r) = g_W(r) = 1$ (within the error of the simulation) is optimal.

3.2.3 Calculation of Constituent Group Preferential Binding Coefficients

For each simulation, up to 21 constituent group preferential binding coefficients were calculated. The 21 groups were each type of amino acid side chain present in the protein (up to 20) and the protein backbone. The "protein backbone" was defined as the -NH-CH-COO- unit, as well as the two extra protons at the N-terminus and extra oxygen atom at the C-terminus of the protein. The glycine side chain was defined as the proton bound to the alpha carbon that would be replaced by a substituent to form a different L-amino acid.

For the simulation of RNase T1 in glycerol solution, the constituent group preferential binding coefficients for the 15 individual serine residues in the protein were also calculated. For this calculation, solvent and additive molecules that were nearest to an atom in the protein that was not part of a serine side chain were not considered.

Water and additive molecules were associated with a specific constituent group by computing the distance from the center of mass of each solvent molecule to the van der Waals surface of every atom in the protein, selecting the protein atom that

was nearest to the solvent molecule, and then determining to what constituent group this nearest protein atom belonged.

3.2.4 Estimation of Statistical Error

The statistical error arising from computing averaged properties from a finite trajectory was estimated in the following fashion:

1. The dynamic trajectory of interest was divided into n pieces.

2. The mean of the property of interest was computed in each piece. These means were designated $\overline{x}_i$ where $i = 1..n$.

3. The standard deviation of the $\overline{x}_i$ values was computed.

4. This standard deviation was divided by $\sqrt{n}$ and the quotient was designated σ_m, an estimate of the error in the mean determined by time averaging the full trajectory.

The number of pieces n into which the trajectory is divided must be small enough to ensure that the means of each piece (the $\overline{x}_i$) are statistically independent. An autocorrelation analysis (not shown) of several trajectories of $\Gamma_{XP}(t)$ data and the underlying molecular counts (n_i^I and n_i^{II}) indicates that a window of about 0.2ns is sufficiently large for this to be true. Therefore, for a 2ns dynamics trajectory, a value of $n = 2/0.2 = 10$ was used.

For long trajectories, the statistical error σ_m is roughly proportional to the inverse square root of the trajectory length. This property can be used to estimate the trajectory length required to achieve a given level of statistical accuracy after a small trajectory has been generated and analyzed.

3.3 Results and Discussion

3.3.1 Radial Distribution Functions of Water and Additives

The radial distribution functions of water, urea, and glycerol were computed for all three simulations as described in Methodology and are shown in Figure 3-2.

At very short distances, $r < 0.6$Å for water and $r < 1.0$Å for glycerol and urea, regions of total solvent and additive exclusion due to very strong van der Waals repulsion can be seen. The size of these "totally excluded" regions is much smaller than one would expect based on the apparent van der Waals radii of the solvent and additive molecules alone (for example, $r \approx 1.5$Å for water and 2.2Å for urea [77]), indicating that electrostatic attractive forces play an important role in solvation even at these distances. After the regions of total exclusion, strong first coordination shells of these three molecules can be clearly seen. The peaks of the first coordination shells become more distant from the protein as the size of the molecules they correspond to increases. Significantly smaller second coordination shell peaks are also visible for urea solvating RNase T1 and glycerol solvating RNase A. At distances greater than 6-7Å from the protein, solvation shells cannot be discerned, and the number densities of water, urea, and glycerol reach their bulk values.

In the simulations of RNase T1 in glycerol and urea solutions, the radial distribution functions for glycerol and urea are quite different. The maximum value of $g_X(r)$ for urea is over 4.5, while that for glycerol is about 2.5. The difference in these maximum values, while significant, is not sufficient to say that the number of urea molecules coordinated to the protein (n_X^{II}) is higher than the number of glycerol molecules coordinated; this can only be done by integrating each $g_X(r)$ function appropriately via equation 3.10.

The radial distribution functions for both water and glycerol are similar in the simulations of RNase A and RNase T1 in glycerol solution, despite the fact that the proteins and the pHs of the solutions are different. Given that the proteins are of similar size, this observation is consistent with the fact that the values of Γ_{XP} for the two solutions are close.

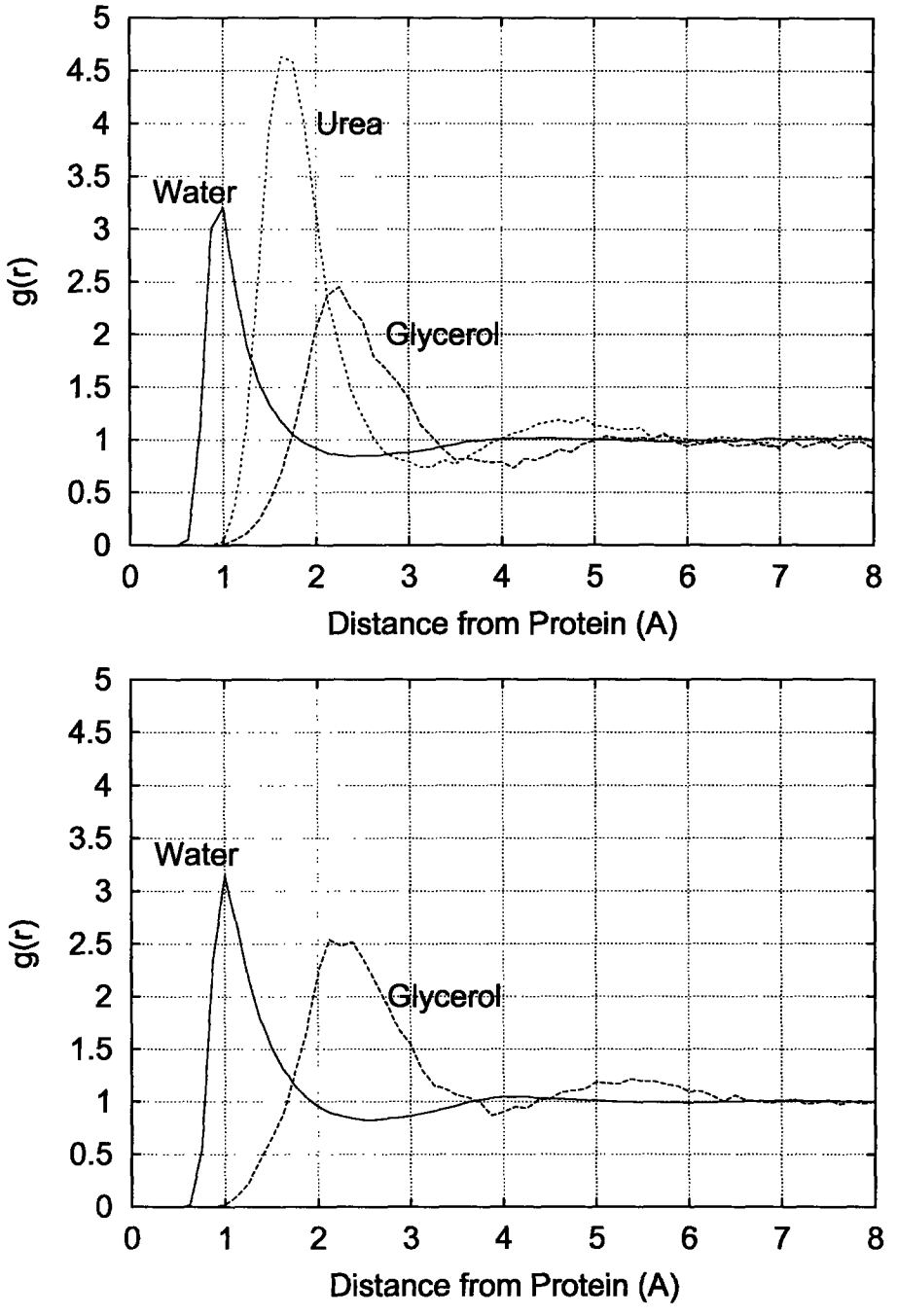

Figure 3-2: Radial distribution functions of water, urea, and glycerol are shown for simulations of RNase T1 in glycerol and urea solutions (left) and RNase A in a glycerol solution (right). In the left-hand figure, the difference between the two $g_W(r)$ functions is not visible at this scale.

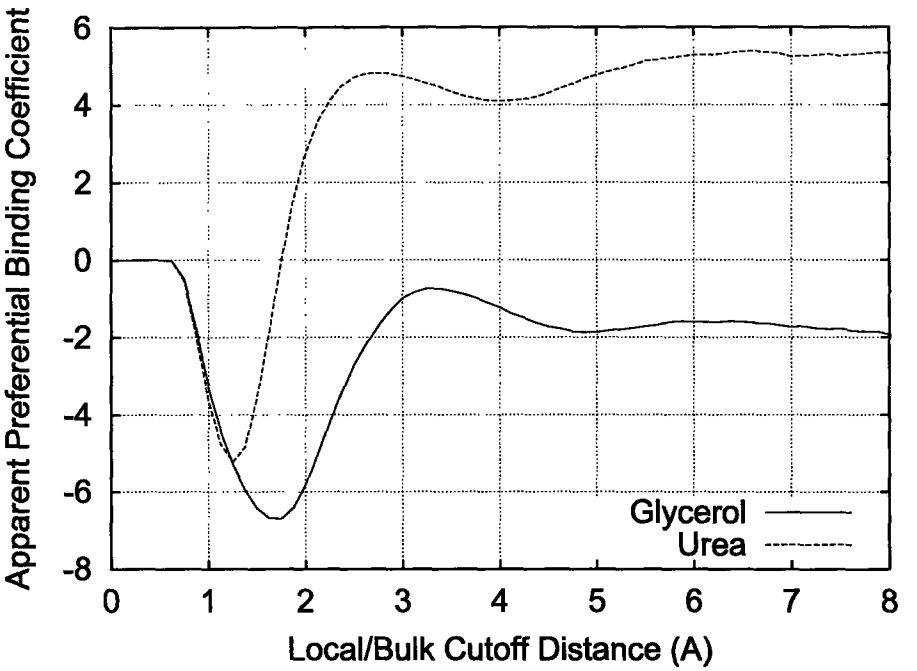

Figure 3-3: Apparent preferential binding coefficient as a function of the cutoff distance between the local and bulk domains for simulations of RNase T1 in glycerol and urea solution.

3.3.2 Preferential Binding Coefficients

The radial distribution functions in Figure 3-2 suggest that r_* in the range of 6-8Å is an appropriate choice of boundary between the local and bulk domains. The error in Γ_{XP} introduced by a particular choice of the boundary distance, r_*, can be estimated by plotting the apparent preferential binding coefficient (Γ^*_{XP}) versus r_* (Figure 3-3). Γ^*_{XP} depends very strongly on r_* in the first solvation shell ($r = 0 - 4$Å) and weakly on r_* in the second solvation shell ($r = 4 - 6$Å). In the range $r = 6 - 8$Å, the dependence of Γ^*_{XP} on r_* is small (± 0.5), and is less than the statistical error in Γ_{XP} (shown in Table 3.2, explained below). Therefore, a cutoff distance of 6Å, or about two solvation shells, is sufficiently large to minimize systematic error in Γ_{XP} caused by the choice of r_*. If only a single solvation shell were considered ($r_* \sim 3.5 - 4$Å), a systematic error in Γ_{XP} of approximately 0.5 - 1 molecules would be introduced as a result of neglect of the second solvation shell.

The preferential binding coefficient, Γ_{XP}, was computed via equation 2.9 using

System	m_{bulk}	Simulation Γ_{XP}	Experiment Γ_{XP}
Urea / RNase T1	1.10m	5.2 ± 1.0	6.4 [53]
Glycerol / RNase T1	1.07m	-1.6 ± 0.8	
Glycerol / RNase A	0.91m	-0.9 ± 1.0	-1.7 ± 0.8 [32]

Table 3.2: Preferential binding coefficients computed from MD simulations and compared with available experimental data at similar additive concentrations. A wide range of behavior (positive and negative preferential binding coefficients) can be modeled without the use of adjustable parameters. The confidence intervals on Γ_{XP}(MD) are an estimate of the statistical error resulting from the use of a finite trajectory. For easier comparison, the experimental values of Γ_{XP} reported above were interpolated to m_{bulk} from data sets spanning the molality of interest.

$r_* = 6$Å as the boundary between the local and bulk domains. A confidence interval for this ensemble average was computed as described in Methodology. The binding coefficients and their statistical uncertainties are shown in Table 3.2. Experimental values from the literature were available for two out of three of these protein-additive systems, and our computed values of Γ_{XP} agree quite favorably with these. The fact that this occurs for both positive and negative values of Γ_{XP} without the use of any adjustable parameters is very encouraging. For an additive that obeys equation 2.13, the confidence intervals of ± 1.0 in Γ_{XP} represents a confidence limit in the transfer free energy of about 0.6 kcal/mol, which is a typical value for free energies calculated via this type of molecular simulation. Achievement of this level of accuracy despite the fact that structural fluctuations in the native state ensemble of proteins have been observed on much longer time scales [27] than the time scale of the simulations performed here suggests that solvent dynamics are more important than protein structural dynamics in determining Γ_{XP}.

$\Gamma_{XP}(t)$ probability density functions for the simulations of RNase T1 in urea and glycerol solution are shown in Figure 3-4. The range of instantaneous values of the preferential binding coefficient, $\Gamma_{XP}(t)$, is quite large relative to the absolute values of Γ_{XP}. $\Gamma_{XP}(t)$ values in excess of $\Gamma_{XP} \pm 15$ are observed. The breadths of these distributions are related to the size of the interface between the local and bulk domains and indicate the importance of sampling a large number of solvent configurations to obtain the macroscopic, averaged Γ_{XP} (equation 3.6).

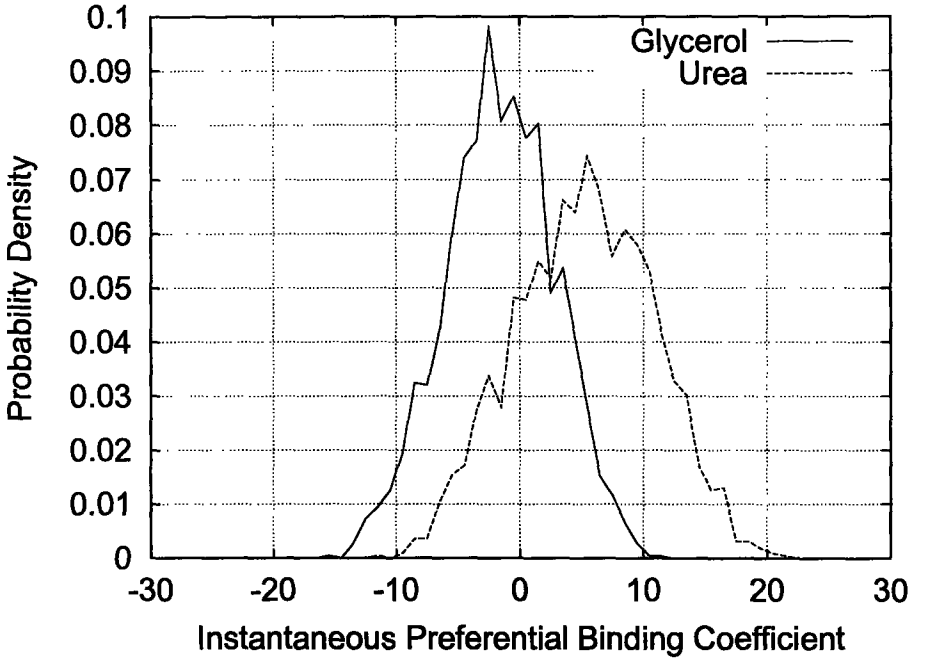

Figure 3-4: $\Gamma_{XP}(t)$ probability density function. A wide range of values of $\Gamma_{XP}(t)$ are sampled as water and additive molecules diffuse between the local and bulk domains.

3.3.3 The Relation between Solvent Accessible Area and the Number of Molecules in the Local Domain

The solvent accessible areas of whole proteins (SAA) and constituent groups (SAA_i) in crystal structures have been used extensively in analyzing proteins. SAA and SAA_i are essentially simple ways of measuring water coordination numbers. In models developed to date, SAA or SAA_i has been used to estimate n_W^{II} or $n_{W,i}^{II}$ by assuming that the local domain is a monolayer of water and each water molecule occupies approximately 10Å^2 of the solvent accessible area. Since we have introduced a new notion of the local domain, it is worthwhile to see what relationships exist between SAA_i and the coordination numbers $n_{W,i}^{II}$ and $n_{X,i}^{II}$ that utilize this definition.

A scatter plot of the solvent accessible area of a set of constituent groups (amino acid side chains and the protein backbone) versus the number of water molecules in the local domain for three different simulations is shown in Figure 3-5. Solvent accessible area was calculated analytically in CHARMM (based on Richmond's method [69]) using a 1.4Å probe. There is a strong, linear correlation of these variables with slope

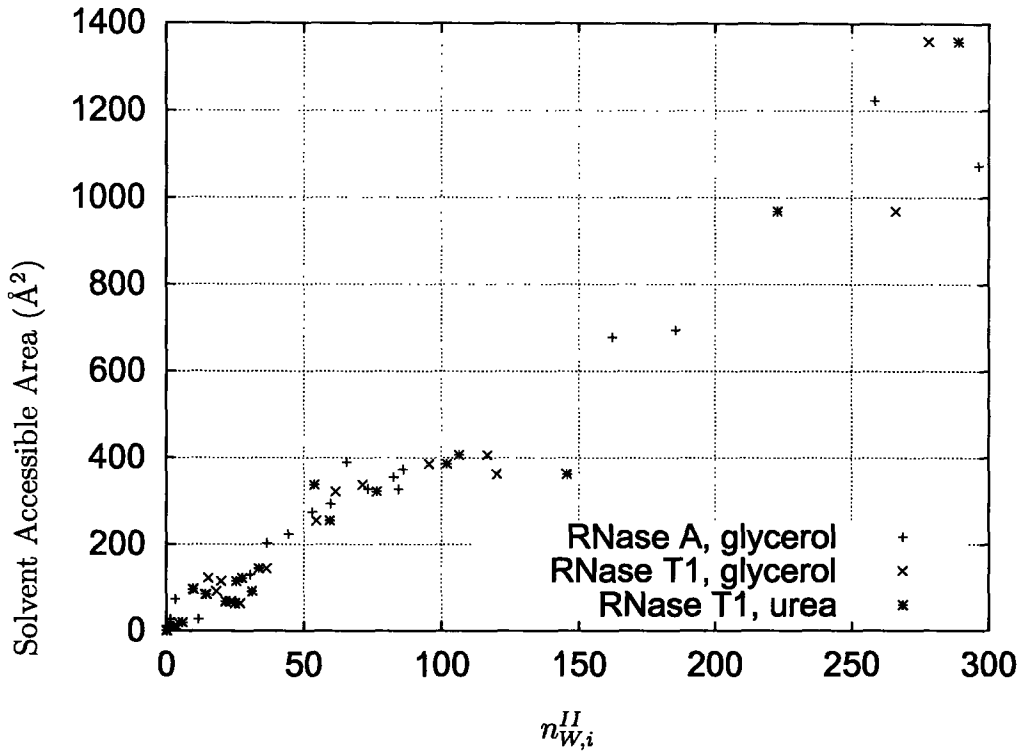

Figure 3-5: Correlation of solvent accessible area and the number of water molecules in the local domain of constituent groups. Each point represents a constituent group of either a type of amino acid side chain or the protein backbone in one of the three simulations shown in Table 3.2. The solvent accessible area of a constituent group and the number of water molecules in the local domain of the solvent near the group ($n_{W,i}^{II}$) are highly correlated.

4.2 Å^2/molecule and correlation coefficient 0.96. Similarly strong correlations are seen for SAA_i with $n_{X,i}^{II}$ in individual simulations. A summary of proportionality constants and correlation coefficients for these relationships is shown in Table 3.3. If the time average SAA_i from each dynamics simulation is used instead of the crystal structure SAA_i values, the correlation coefficients increase slightly. Because the time average solvent accessible areas are higher than those in the crystal structure, the proportionality constants shown in Table 3.3 also increase.

3.3.4 Constituent Group Preferential Binding Coefficients

The constituent group preferential binding coefficients were calculated for each simulation as described in Methodology and are shown in Figures 3-6 - 3-9 as the number

Species (i)	Protein	Avg Protein SAA/n_i^{II} (Å^2/molecule)	r^2
Water	RNase A/T1	4.2	0.96
0.91m Glycerol	RNase A	290	0.96
1.07m Glycerol	RNase T1	230	0.93
1.10m Urea	RNase T1	170	0.98

Table 3.3: Relationships between solvent accessible area in each protein crystal structure and number of solvent molecules in the local domain for different protein-additive systems. r^2 symbolizes the correlation coefficient.

of water and additive molecules coordinated to each constituent group. In each figure, a line at the bulk solution composition is also plotted, enabling a quick determination of the composition of the solvent in the vicinity of a constituent group compared to the bulk solvent. The statistical uncertainties in the values of $n_{W,i}^{II}$ and $n_{X,i}^{II}$ (and consequently $\Gamma_{XP,i}$) are high. Because of these uncertainties, we will not report specific values of the group preferential binding coefficients, but rather classify them into broad categories based on their statistical likelihood of being either positive, negative, or zero/indeterminate.

The average number of water and glycerol molecules coordinated to each of the 15 serine residues in RNase T1 are shown in Figure 3-6. A wide range of binding behavior can be seen among the serine residues, all of which have a good degree of solvent exposure. Ser 17, 35, and 72 fall above the bulk concentration line and have positive preferential binding coefficients, Ser 63 falls below the line and has a negative preferential binding coefficient, and the preferential binding coefficients of the remaining 11 serine residues are not statistically different from zero. The wide range of local concentrations in the vicinities of these serine residues indicates that developing a group contribution method to estimate Γ_{XP} or $\Delta\mu_P^{tr}$ based on primary sequence information and solvent accessibility ($n_{W,i}^{II}$) alone may be difficult. In addition to the type of amino acids present at the protein-solvent interface, other effects such as specific combinations of residues and secondary or tertiary structure must be important in determining water and additive binding behavior. These factors probably contribute to the range of local concentrations seen in Figure 3-6. For example, Ser35 and Ser72 are proximal to each other and several Gly and Tyr side

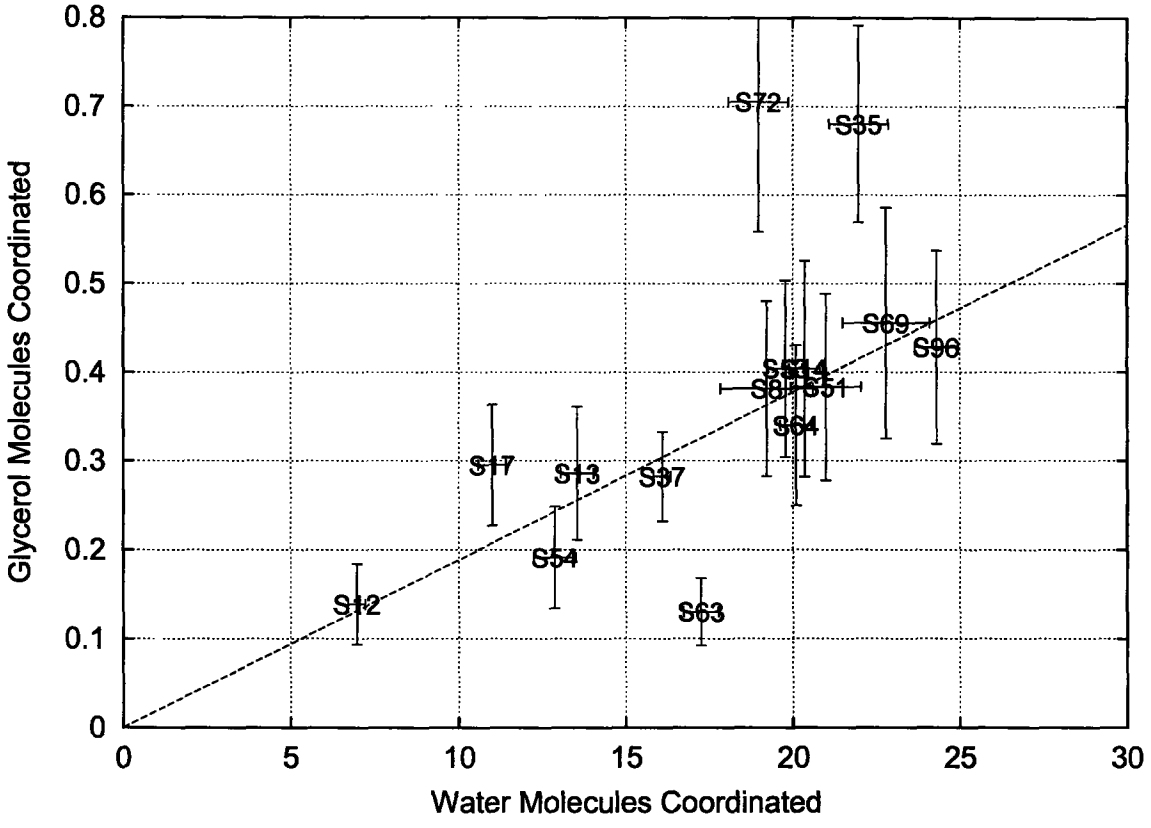

Figure 3-6: Binding behavior of glycerol and water with the 15 serine residues in RNase T1 is shown as a plot of the number of glycerol molecules in the local domain of each serine residue versus the number of water molecules in the same volume. The labels are the one-letter code for each amino acid side chain, and "B" is the protein backbone. The line represents the bulk glycerol composition. Ser 17, 35, and 72 have positive preferential binding coefficients, Ser 63 has a negative preferential binding coefficient, and the remaining 11 serine residues have essentially zero values for their preferential binding coefficients.

chains (Gly34, 70, 71, and Tyr68), which tend to have positive preferential binding coefficients in glycerol (Figure 3-8). This may be the reason that the group preferential binding coefficients for these residues are higher than those of the other serine residues.

The preferential binding behavior of urea and glycerol with each type of amino acid in RNase T1 and the protein backbone are shown in Figures 3-7 and 3-8. In urea solution, the protein backbone and Ser as well as the hydrophobic amino acid side chains of Cys, Gly, Leu, Phe, Pro, Tyr, and Val all preferentially bind urea, while the hydrophilic Asp preferentially binds water. In glycerol solution, only Tyr and Gly preferentially bind glycerol, and Asp and Glu preferentially bind water. Qualitatively, the binding behavior of the amino acid side chains of RNase T1 follow a hydrophobic series, with the hydrophobic side chains tending to bind more additive and the hydrophilic ones tending to bind more water.

The binding behavior of glycerol and water with the amino acid side chains and backbone in RNase A, shown in Figure 3-9, is significantly different than the binding behavior of these solvent components with the same constituent groups in RNase T1. (Note that the protonation states of Asp, Glu, and His are different in the two simulations.) The amino acid backbone, which occupies a large fraction of the protein-solvent interface as indicated by its high value of $n_{W,i}^{II}$, has a binding coefficient near zero in RNase T1 and a significant negative binding coefficient in RNase A. More strikingly, Tyr in RNase T1 preferentially binds glycerol whereas Tyr in RNase A preferentially binds water. This is likely because the six Tyr residues in RNase A are at or near the solvent interface (a more hydrophilic region) whereas the nine in RNase T1 are mostly buried (a more hydrophobic region). This difference in solvent exposure is evident from the crystal structures of the proteins but also can be discerned by comparing the water coordination numbers for Tyr in the two proteins: $n_{W,i}^{II}$ for Tyr in RNase A is higher than in RNase T1, even though there are 50% more Tyr residues in RNase T1.

Based on the above observations, some generalizations about the effects that these additives have on protein folding equilibria can be postulated, the validity of which must be confirmed via future studies. In urea solution, most of the constituent groups

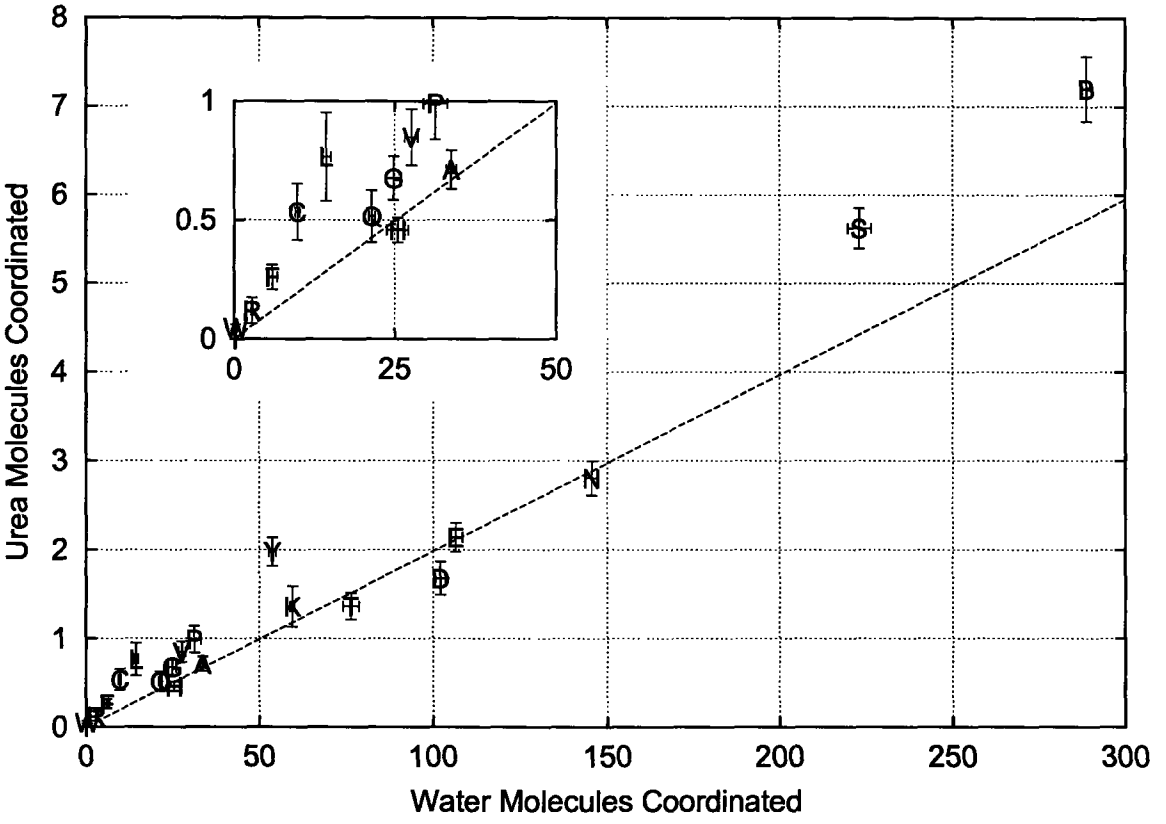

Figure 3-7: Local binding behavior of urea and water with the amino acid backbone and side chains in RNase T1. The labels are the one-letter code for each amino acid side chain, and "B" is the protein backbone. The line denotes the bulk urea concentration. In addition to the protein backbone and Ser, the hydrophobic amino acids Cys, Gly, Leu, Phe, Pro, Tyr, and Val all preferentially bind urea, while the hydrophilic Asp preferentially binds water.

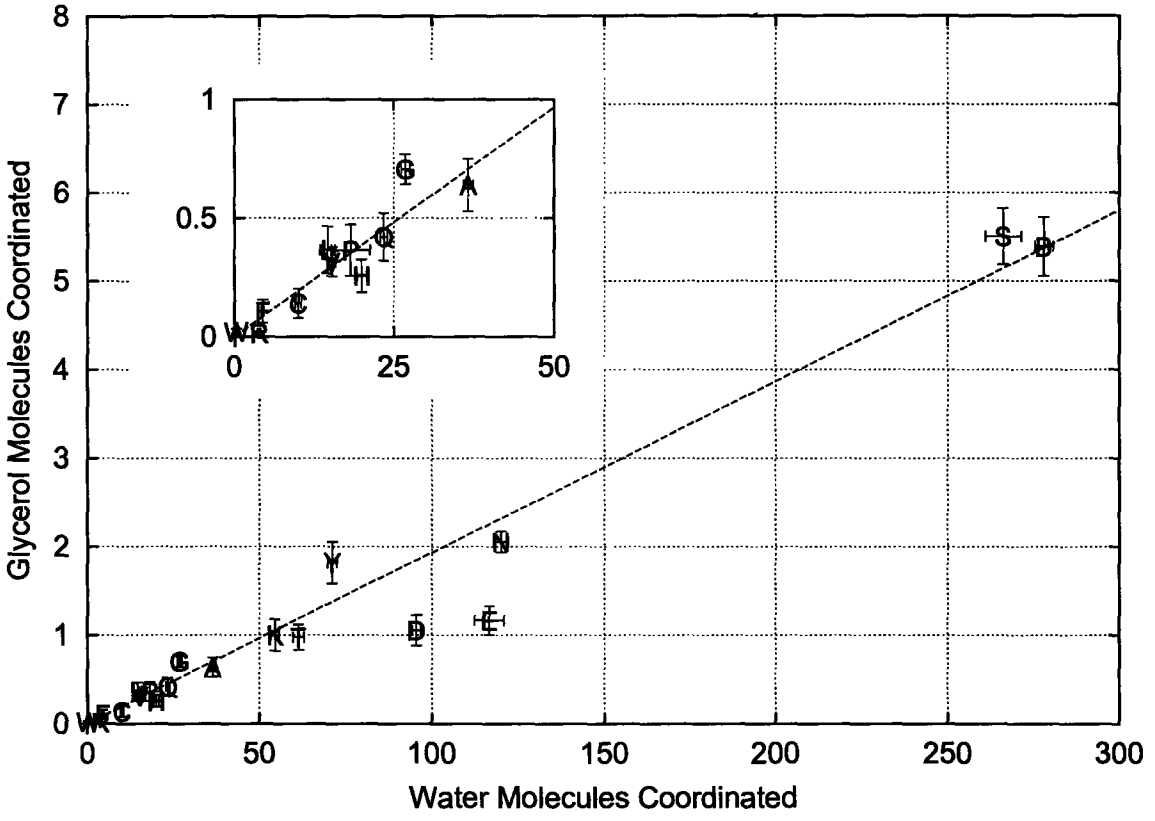

Figure 3-8: Group preferential binding coefficients for glycerol with the amino acid backbone and side chains in RNase T1. The labels are the one-letter code for each amino acid side chain, and "B" is the protein backbone. The line denotes the bulk glycerol concentration. Tyr and Gly preferentially bind glycerol; Asp and Glu preferentially bind water; and the binding coefficients of the other groups are not statistically different from zero.

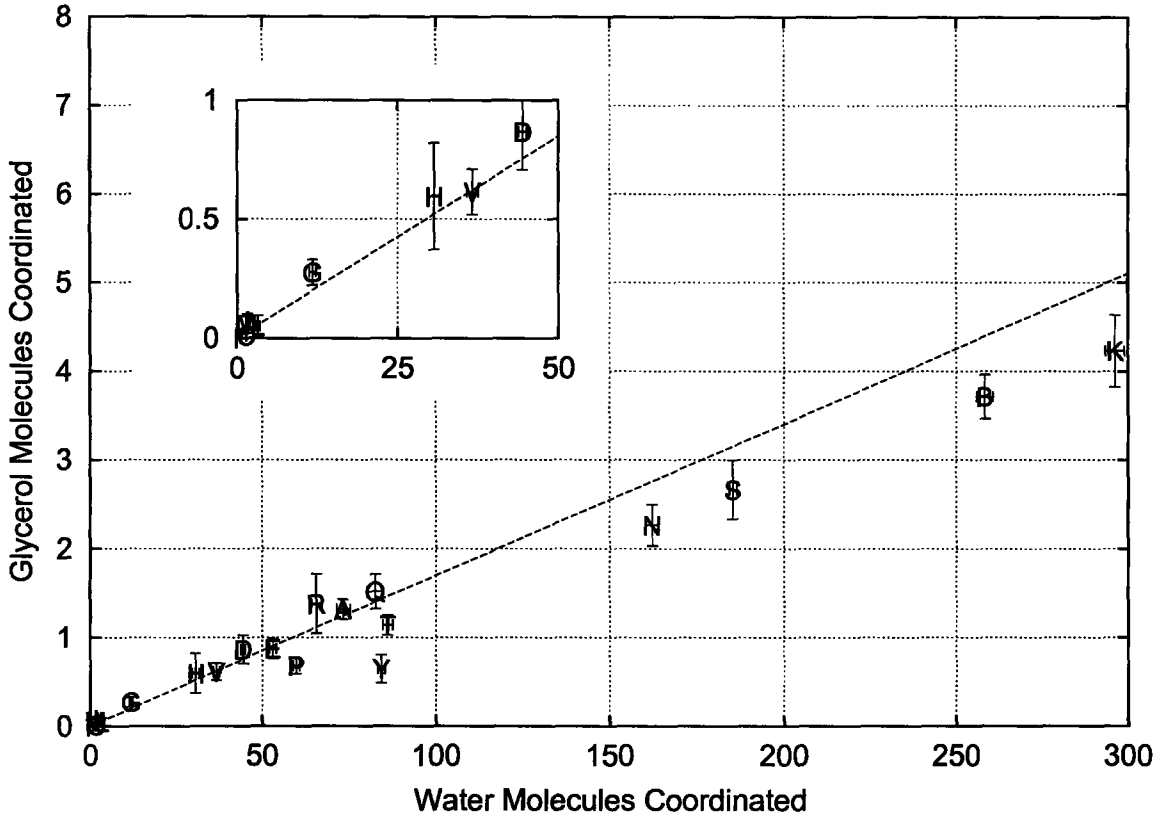

Figure 3-9: Local binding behavior of glycerol with the amino acid backbone and side chains in RNase A. The labels are the one-letter code for each amino acid side chain, and "B" is the protein backbone. The line denotes the bulk glycerol concentration. All of the constituent groups in RNase A either preferentially bind water or are neutral.

in RNase T1 either preferentially bind urea or are indifferent to urea and water. Asp, which is found on the surface of RNase T1, is the only constituent group that is significantly below the bulk concentration line in Figure 3-7 and therefore preferentially binds water over urea. Since the amino acids that compose the core of RNase T1 and are exposed upon unfolding preferentially bind urea, this pattern suggests that the preferential binding coefficient or urea with unfolded RNase T1 is higher than that with native RNase T1. This is thermodynamically consistent with urea's well-known ability as a denaturant. Inversely, in glycerol solution, almost all of the constituent groups in RNase A and T1 are neutral or preferentially bind water. This is consistent with the fact that glycerol binds less to the unfolded protein than the native state, and therefore is a protein stabilizer. Both of these generalizations are consistent with earlier work on model compounds [14].

3.4 Conclusions

A quantitative method based on molecular dynamics simulations using all atom potential models has been developed and validated for calculating preferential binding coefficients. Our method is not a derivative of thermodynamic integration or thermodynamic perturbation methods and requires only a single trajectory to compute the transfer free energy of a protein into a weak-binding additive system. Our results match experimental data well for glycerol and urea solutions, covering a range of positive and negative binding behavior. This work also augments experimentally-observable, macroscopic thermodynamics with the mechanistic insight provided by a molecular-level, statistical mechanical model. Variations in the radial distribution functions with distance for each additive are evident up to about 6Å, or two solvation shells of water, away from the protein. Glycerol is not totally excluded from close contact with the protein, but glycerol is less likely than urea to be found in such a position. The radial distribution functions of water and additives are sufficient to calculate preferential binding coefficients by integrating over a suitable solvent volume.

The binding behavior of the amino acid side chains in RNase T1 qualitatively follow a hydrophilic series, with more hydrophilic amino acids in the protein tending to have a higher concentration of water in their vicinity. The constituent group binding behavior differs between the groups in RNase A those in RNase T1. Development of a group contribution method at the amino acid level for estimating binding coefficients or transfer free energies of whole proteins is complicated by the wide range of coordination behaviors observed for single types of amino acids in different environments on the protein surface.

Chapter 4

The Gap Effect

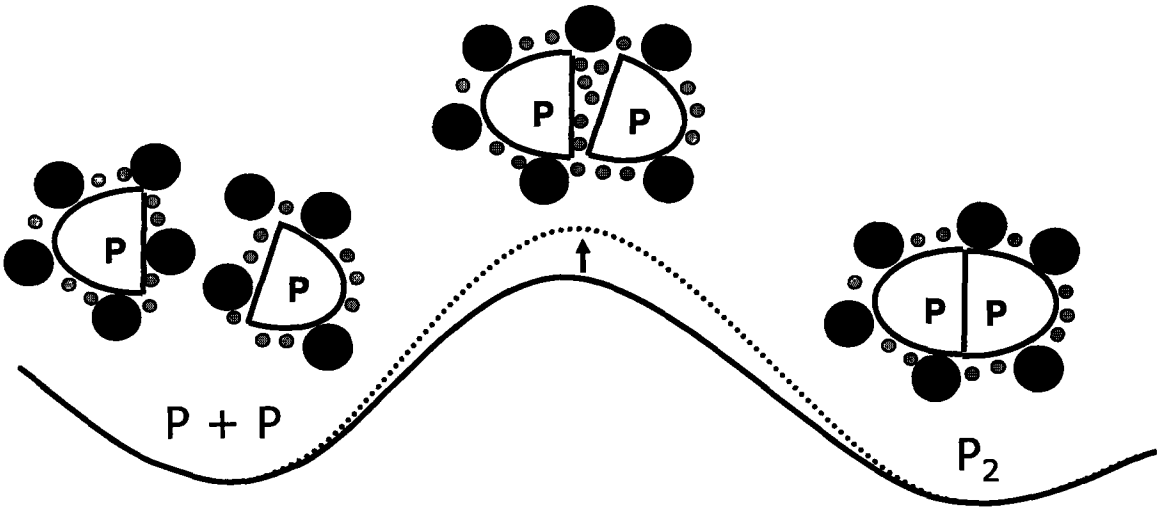

Figure 4-1: If a protein molecule (P) or complex contains narrow channels too small for a large additive (black) to enter, the cosolvent exerts an osmotic stress effect that favors the collapse of these channels and the release of the water (grey) they contain. In the case of the above protein-protein association reaction coordinate, the "gap effect" caused by the large additive selectively increases the free energy of encounter complexes that contain such narrow gaps. The gap effect therefore slows isomerization between the associated and dissociated protein states.

In order to affect the kinetics of protein association reactions selectively without affecting protein folding and solution phase structure, there must be a unique feature of the association transition state that can be exploited by a binding interaction with an additive. The emerging picture of protein association/dissociation transition states indicates that each protein in the encounter complex is still mostly solvated but near the orientation in the final complex [79, 78]. Because the complex is still mostly solvated but the two protein molecules are in close proximity to one another, there is the potential for a "gap effect" to arise in a mixed solvent if the additive is significantly larger than the primary solvent (Figure 4-1). This gap effect is analogous to osmotic stress [67]. In such a situation, the large additive will be excluded from solvating the gap between the protein molecules for steric reasons. This, in turn, results in an increase in the free energy of the protein-protein encounter complex.

In order for an additive to reduce the rate of aggregation without affecting the folding rate and equilibrium, in addition to creating a "gap effect," it is necessary

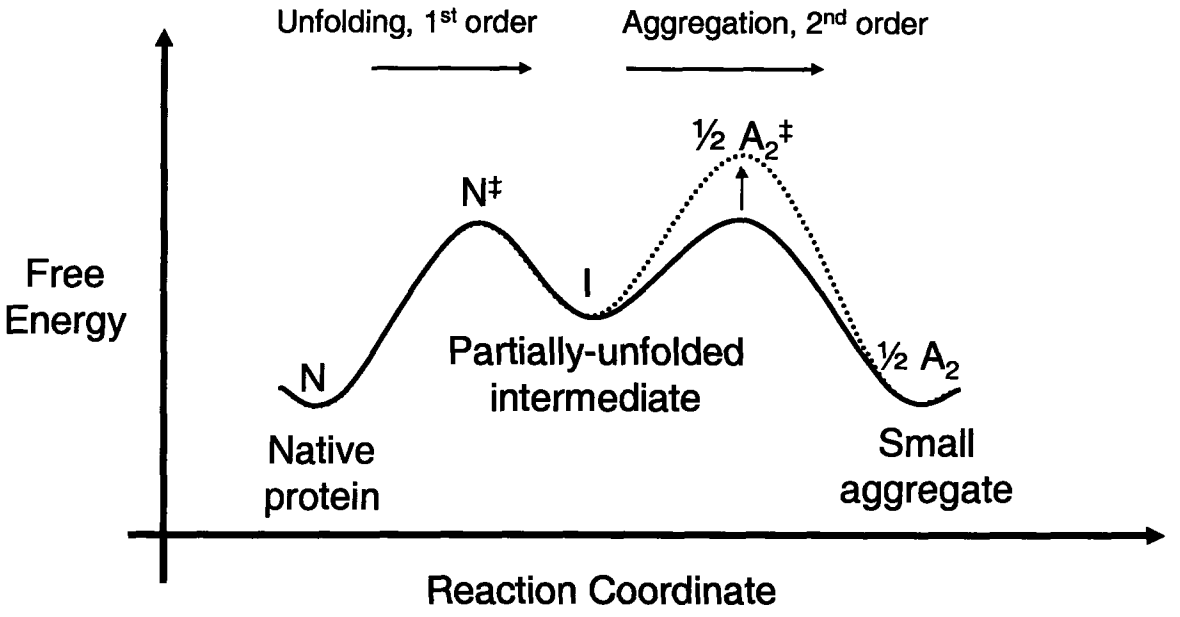

Figure 4-2: The hypothesized effect of a neutral crowder on the free energy of protein states along the refolding/aggregation reaction coordinate is shown. The free energy in the absence of additive is shown as a solid line and, in the presence of a neutral crowder, as a dotted line. The neutral crowder is preferentially-excluded from the gap between the protein molecules in the association transition state ($A_2^{\ddagger}$), increasing the free energy of this state.

for the additive not to interact with isolated protein molecules differently than water does. We call additives that exhibit both of these properties "neutral crowders." We hypothesize that a neutral crowder would affect a refolding/aggregation reaction coordinate as shown in Figure 4-2.

With this gap effect hypothesis in mind, the specific objectives of this chapter are to:

1. Develop and justify the use of a simple model for protein-protein association reactions in the presence of additives. This model should take into account protein-additive binding interactions over the entire association reaction coordinate.

2. Use this model to test the gap effect hypothesis and evaluate the potential magnitude of any observed gap effect.

3. Evaluate the potential of neutral crowders as anti-aggregation additives.

4.1 Theoretical Approach

In order to test our gap effect hypothesis, we compute the effects of various solution additives on the association and dissociation rate constants for a suitable protein-protein model system. If the only effect of an additive is to alter the free energy barrier to the transition state, the change in a rate constant can be expressed as:

$$k/k_0 = \mathrm{e}^{-\Delta\Delta\mu^{\ddagger}/k_bT} \tag{4.1}$$

where k is the rate constant in the presence of the additive, k_0 is the rate constant in the absence of the additive, $\Delta\Delta\mu^{\ddagger}$ is the change in the activation free energy induced by the additive, k_b is Boltzmann's constant, and T is the absolute temperature. Thus, in order to compute the relative rate constants, we must compute the shift in activation free energy for association and dissociation induced by the additives. Our approach to this is to:

1. Define two simple, limiting models of a protein-protein association reaction, (i) the association of two parallel planes, and (ii) the association of two spheres. We also propose with our models suitable reaction coordinates and compute the free energy as a function of reaction coordinate in each case.

2. Compute the perturbations to each free energy diagram induced by an additive. These perturbations are obtained by using data determined from explicit all-atom molecular dynamics simulations and incorporating them into our simple models.

The details of each of these steps are described in the following sections.

4.1.1 Two Model Association Reactions

As two limiting models of proteins undergoing an association and dissociation reaction, we choose:

1. The reaction of two spheres, each 20Å in radius. This is roughly the size of a 25kD protein.

2. The reaction of two parallel, planar plates, each with 400πÅ^2 of area on a face. This area was selected to make the change in protein solvent-accessible area of reaction the same for the cases of the two spheres and two planes (the reaction coordinates are explained below). The thermodynamics properties of these plates are obtained by calculating the property per unit of surface area of a pair of infinite plates and then multiplying by the area above. Thus, edge effects are ignored.

These two geometries can be considered as extreme cases by which associating proteins approach one another. Because of the symmetry of these simple model proteins, the reaction coordinates, x, can be simply defined as the shortest distance between the planes and the distance between the centers of the spheres. We are then free to choose any representative free energy of the complex as a function of this reaction coordinate, $\mu_{P,0}(x)$. For convenience, we set the reference states as the monomers ($x \rightarrow +\infty$), define the dimer states to be 8 kcal/mol more stable than monomer states, and place a modest 2 kcal/mol free energy barrier for association between the two states. Arbitrarily, we select $x = 20$Å as the dimer state for the spheres, and $x = 1.5$Å as the dimer state for the planes. The resulting reaction coordinates can be modeled as:

$$\text{Planes}: \quad \mu_{P,0} = \left(\frac{1}{x}\right)^6 - 8.51\mathrm{e}^{-(x-1.5)^2} + 2.02\mathrm{e}^{-\left(\frac{x-4}{2}\right)^2} \tag{4.2}$$

$$\text{Spheres}: \quad \mu_{P,0} = \left(\frac{15}{x}\right)^6 - 8.21\mathrm{e}^{-\left(\frac{x-20}{10}\right)^2} + 2.12\mathrm{e}^{-\left(\frac{x-40}{10}\right)^2} \tag{4.3}$$

which employ Gaussians for the energy minimum at the dimer state and for the energy maximum at the transition state and an inverse sixth power function for repulsion

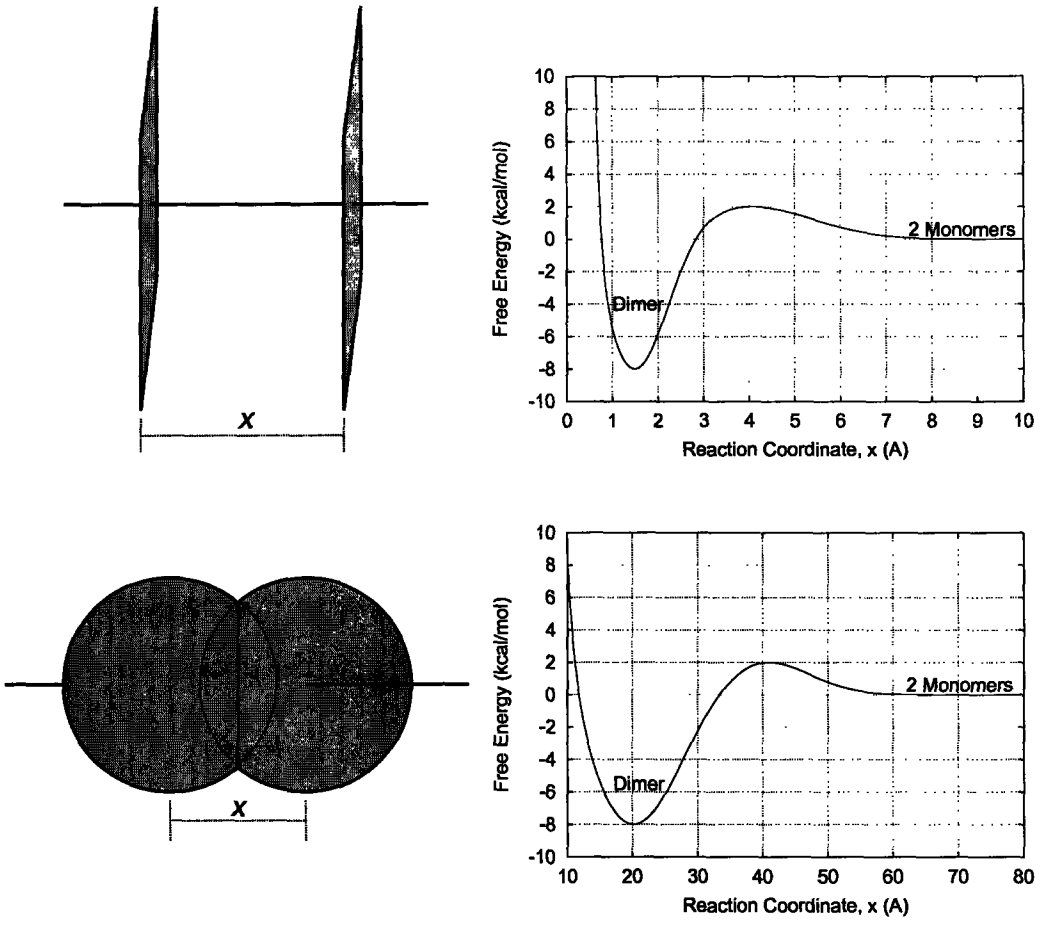

Figure 4-3: The definition of each reaction coordinate and the free energy diagrams (equations 4.2 and 4.3) are shown for the two model protein systems used in this work. For the spheres, the association/dissociation reaction coordinate, x, is the distance between the sphere centers. For the planes, it is the shortest distance between the planes (which are always parallel). x is zero when the two proteins overlap each other completely.

at distances closer than the dimer state. The reaction coordinate and free energy diagram for each model are shown in Figure 4-3.

4.1.2 Calculating the Effect of an Additive

We compute the free energy along the reaction coordinate in the presence of a solution additive by combining the free energy in the absence of additive (equations 4.2 and

4.3) with an appropriate transfer free energy:

$$\mu_P = \mu_{P,0} + \Delta\mu_P^{tr} \tag{4.4}$$

where $\mu_{P,0}$ is the free energy of a given protein state in the absence of additive, μ_P is the free energy of the same state when the additive is present, and $\Delta\mu_P^{tr}$ is the transfer free energy. The transfer free energy can be computed via the following equation:

$$\Delta\mu_P^{tr} = \int_0^{m_X} \left(\frac{\partial\mu_P}{\partial m_X}\right)_{T,P,m_P} dm_X \tag{4.5}$$

where m_X is the molality of the additive and m_P is the molality of the protein. The integrand above can be split into a derivative involving properties of only the additive and water, and one that describes binding:

$$\Delta\mu_P^{tr} = -\int_0^{m_X} \left(\frac{\partial\mu_X}{\partial m_X}\right)_{T,P,m_P} \Gamma_{XP}\, dm_X \tag{4.6}$$

where Γ_{XP}, the preferential binding coefficient of additive to the protein in water, is defined as:

$$\Gamma_{XP} \equiv \left(\frac{\partial m_X}{\partial m_P}\right)_{T,P,\mu_X} \tag{4.7}$$

Following our earlier work in all-atom molecular dynamics simulations of preferential binding [10], we introduce the relation:

$$\Gamma_{XP} = c_X \int (g_X - g_W)\, dV \tag{4.8}$$

where c_X is the bulk concentration of additive, g_X and g_W are the additive and water radial distribution functions with respect to the protein, respectively, and the integral is over the solvent volume around the protein. Note that g_X and g_W typically differ only in the first two solvation shells. For the association/dissociation reactions being modeled here, the protein state is a pair of protein molecules, and dV is a function of x.

The above relation can be substituted into equation 4.6 to yield:

$$\Delta\mu_P^{tr} = -\int_0^{m_X} c_X \left(\frac{\partial \mu_X}{\partial m_X}\right)_{T,P,m_P} \left(\int (g_X - g_W)\, dV\right) dm_X \tag{4.9}$$

We now invoke three assumptions that allow significant simplification of the above equation. These are:

1. The additive free energy (μ_X) is thermodynamically ideal, or $(\partial \mu_X/\partial m_X)_{T,P,m_P} \approx RT/m_X$, where R is the gas constant. Since the ternary system in question here is dominated by the water and additive, this assumption effectively means that water-additive interactions are ideal. For many solution additives of interest, such as NaCl, glycerol, sucrose, and urea, the experimental activities of water-additive mixtures are within 10% of ideality at molalities up to 1 mol/kg [74, 63]. For other real systems where experimental thermodynamic activity data on the binary system of water and additive are available, these can be used to more accurately evaluate the partial derivative in question. The assumption used here does not limit the complexity of the additive-protein and water-protein interactions, which ultimately will lead to changes in the transfer free energy via the Γ_{XP} term.

2. The concentrations of protein and additive are sufficiently low such that molar and molal concentration are equivalent ($c_X \approx m_X$).

3. The radial distribution functions of water and the additive are independent of the concentration of additive. This is expected to be true at low additive concentrations ($m_X < 1$ m) for additives that interact weakly with the protein [25, 35, 68].

Applying these approximations to equation 4.9 yields:

$$\Delta\mu_P^{tr} = -RTc_X \int (g_X - g_W)\, dV \tag{4.10}$$

To enable computation of the transfer free energies, we now require a model for

the radial distribution functions of the additive and water around the proteins. In prior studies of all-atom molecular dynamics of proteins in mixed solvents [10], we noted that the radial distribution function of water as a function of distance from the protein, $g_W(r)$, did not vary whether the protein was RNase A or RNase T1, two proteins with significantly different amino acid sequences. Thus, in this work, we use the water radial distribution function found previously, and assume it to be independent of the protein model and reaction coordinates employed here.

In the case of the additive, we wish to introduce a simple, physically-based model for the additive-protein radial distribution function, g_X. To do this, we relate g_X to the potential of mean force between the additive and protein, $\langle U_{XP} \rangle$:

$$g_X = \mathrm{e}^{-<U_{XP}>/k_bT} \tag{4.11}$$

We then choose the form of the potential of mean force as a standard intermolecular potential function. To select a suitable function, we fit the parameters of standard, physically-based intermolecular potentials, such as Lennard-Jones, Kihara, and exponential-6 (Exp-6) to the radial distribution functions of water, urea, and glycerol obtained from all-atom molecular dynamics simulations [10]. In each case, the intermolecular potential parameters were fit by nonlinear minimization (Marquardt method) of the squared residuals while constraining (Lagrange method) the radial distribution to give the same preferential binding coefficient (Γ_{XP}, via equation 4.8) as the actual additive radial distribution function. Preferential binding coefficients were preserved in the fitting procedure because of their tight relationship (via equation 4.6) to the transfer free energy, the property we ultimately wish to model.

After obtaining the best fits with each potential function, it was observed that the Lennard-Jones potential did not adequately fit the data, and the Kihara potential did not fit the data well at physically meaningful parameter values. Therefore, the three-parameter Exp-6 potential was selected as a model of the additive-protein potential

Molecule	ϵ (kcal/mol)	r_m (Å)	γ
Water	0.662	0.925	3.65
Urea	0.917	1.59	3.17
Glycerol	0.497	2.11	4.25

Table 4.1: Exponential-6 potential parameters for averaged interaction energies of water, urea, and glycerol with RNase A and T1. The parameters were obtained by constrained fitting to radial distribution functions obtained from all-atom molecular dynamics data.

of mean force:

$$\langle U_{XP} \rangle = \frac{\epsilon}{1 - 6/\gamma} \left(\frac{6}{\gamma} \exp(\gamma(1 - r/r_m)) - \left(\frac{r_m}{r} \right)^6 \right) \tag{4.12}$$

In the above equation, r is the distance between the solvent molecule and protein, and r_m, ϵ, and γ are the Exp-6 parameters, described below. Results of the fitting process are shown in Figure 4-4 and Table 4.1. Note that the first peak in the radial distribution functions occurs at a value of r smaller than what might be expected. This is because in our case, r is defined as the distance from the center of mass of the solvent or additive to the van der Vaals shell of the protein, not to the nucleus of the atoms at the protein surface. This also leads to a value of r_m that may be smaller than expected.

It is undoubtedly possible to obtain a tighter three-parameter fit to the radial distribution functions shown in Figure 4-4 by using a broader set of basis functions to model $\langle U_{XP} \rangle$. However, we wished to constrain ourselves to standard potential functions whose parameters had some physical meaning. Fits are also shown for water because water-protein radial distribution function data was available; however, the full radial distribution function for water was used for all of the calculations in this work.

The Exp-6 potential combines an exponential repulsive term with an inverse sixth-power attractive term and has a single minimum at $U(r = r_m) = -\epsilon$. The last Exp-6 parameter, γ, is related to the breadth of the minimum near $r = r_m$, and reflects the rigidity and shape of the additive.

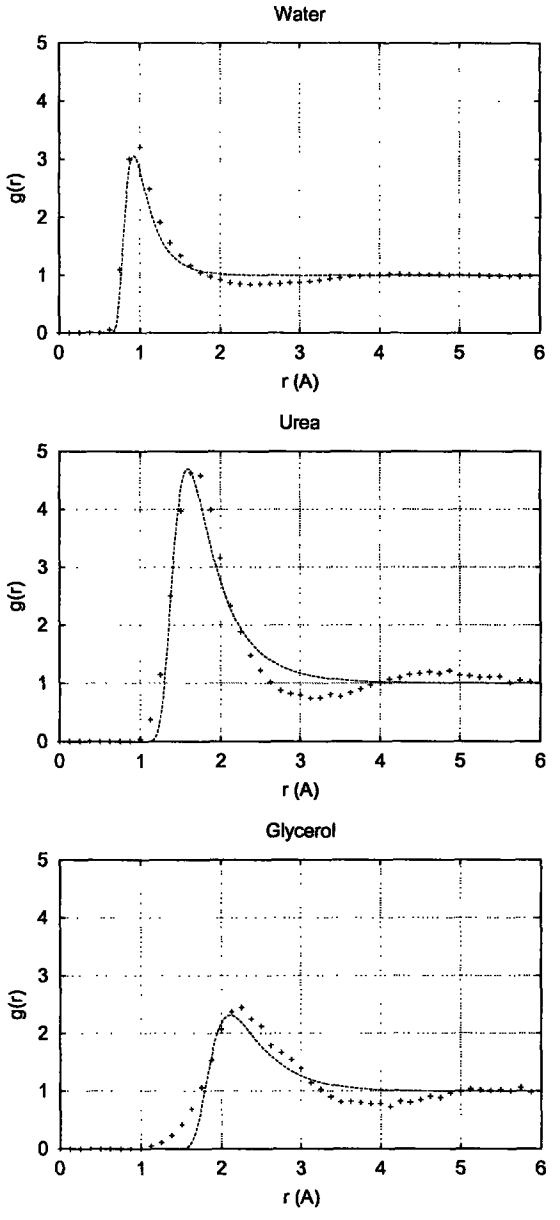

Figure 4-4: Fits of the protein-water and protein-additive radial distribution functions from molecular dynamics simulations for various additives with the protein RNase T1 using the exponential-6 intermolecular potential. Note that $r = 0$ is at the surface of the protein. The observed data are shown as crosses, the fits as lines. The corresponding parameters are shown in Table 4.1.

In extending the Exp-6 representation to neutral crowders and other additives for which no radial distribution functions are available, r_m is used as a measure of additive size; γ is held constant at 3.7, the mean of the observed values for water, glycerol, and urea; and ϵ is set to yield a desired preferential binding coefficient between the additive and dissociated protein state. For a neutral crowder, Γ_{XP} is set to 0 at $x \rightarrow \infty$ (the dissociated state) by the constraint that such an additive should not affect the free energy of isolated protein molecules.

4.1.3 Relation to Virial Coefficients

Given the approximations used in this work, the transfer free energy can be related to the additive-protein and water-protein second virial coefficients. This follows directly from the McMillan-Mayer formula for the second virial coefficient [58]:

$$B_{iP} = -\frac{1}{2}\int (\mathrm{e}^{-<U_{iP}>/kT} - 1)\, dV \tag{4.13}$$

where B_{iP} is the second virial coefficient and $\langle U_{iP} \rangle$ is the potential of mean force between a solvent component i (additive or water) and the protein. In terms of the radial distribution functions, this is:

$$B_{iP} = -\frac{1}{2}\int (g_i - 1)\, dV \tag{4.14}$$

The preceding equation can then be substituted into the integrand of equation 4.10 for the water and additive to yield:

$$\Delta\mu_P^{tr} = -RTc_X \int ((g_X - 1) - (g_W - 1))\, dV \tag{4.15}$$

$$= 2RTc_X(B_{XP} - B_{WP}) \tag{4.16}$$

Thus, the additive-protein and water-protein second virial coefficients are related to the radial distribution functions (equation 4.14) and to the transfer free energy (equation 4.16).

4.2 Results

4.2.1 The Gap Effect Can Contribute Significantly to Association and Dissociation Rate Constants

The transfer free energies for placing the model proteins into 1M solutions of neutral crowders were computed over the entire association/dissociation reaction coordinates (via equations 4.10, 4.11, and 4.12) and are presented in Figure 4-5. r_m was chosen to have values of 2, 4, 6, or 8 Å in order to give a range of crowder sizes. In each case, ϵ was set according to the neutrality condition ($\Gamma_{XP} = 0$ for the dissociated state, $x \rightarrow \infty$).

It is readily apparent that at constant Γ_{XP}, the gap effect on the transfer free energy increases proportionately with increasing additive size, r_m. For the additive sizes illustrated, the effect on transfer free energy ranges from 0 to almost 6 kcal/mol. At the same additive size and change in surface area of reaction, the planes exhibit about double the gap effect of the spheres. This is because the lack of curvature of the planes necessitates that their gap effect is concentrated over a narrower region of the reaction coordinate.

These transfer free energy effects can be superimposed onto a free energy diagram by simple addition (equation 4.4). The final free energy diagrams are shown in Figure 4-6. In the case of the spherical model, the transition state in the original free energy surface ($\mu_{P,0}(x)$) is near the maxima in the transfer free energies, so the transfer free energy effects make significant changes to the activation free energy of the association and dissociation reactions in an additive solution at all values of r_m. For the planar model, the location of the maximum in the transfer free energy depends on r_m. Consequently, the transfer free energy maximum for the planar model does not always build on the existing free energy barrier in $\mu_{P,0}(x)$. In fact at higher r_m, the transition state for association and dissociation results completely from the gap effect. The original transition state is "drowned out."

Using equation 4.1 to estimate the resulting changes in reaction rate, the relative

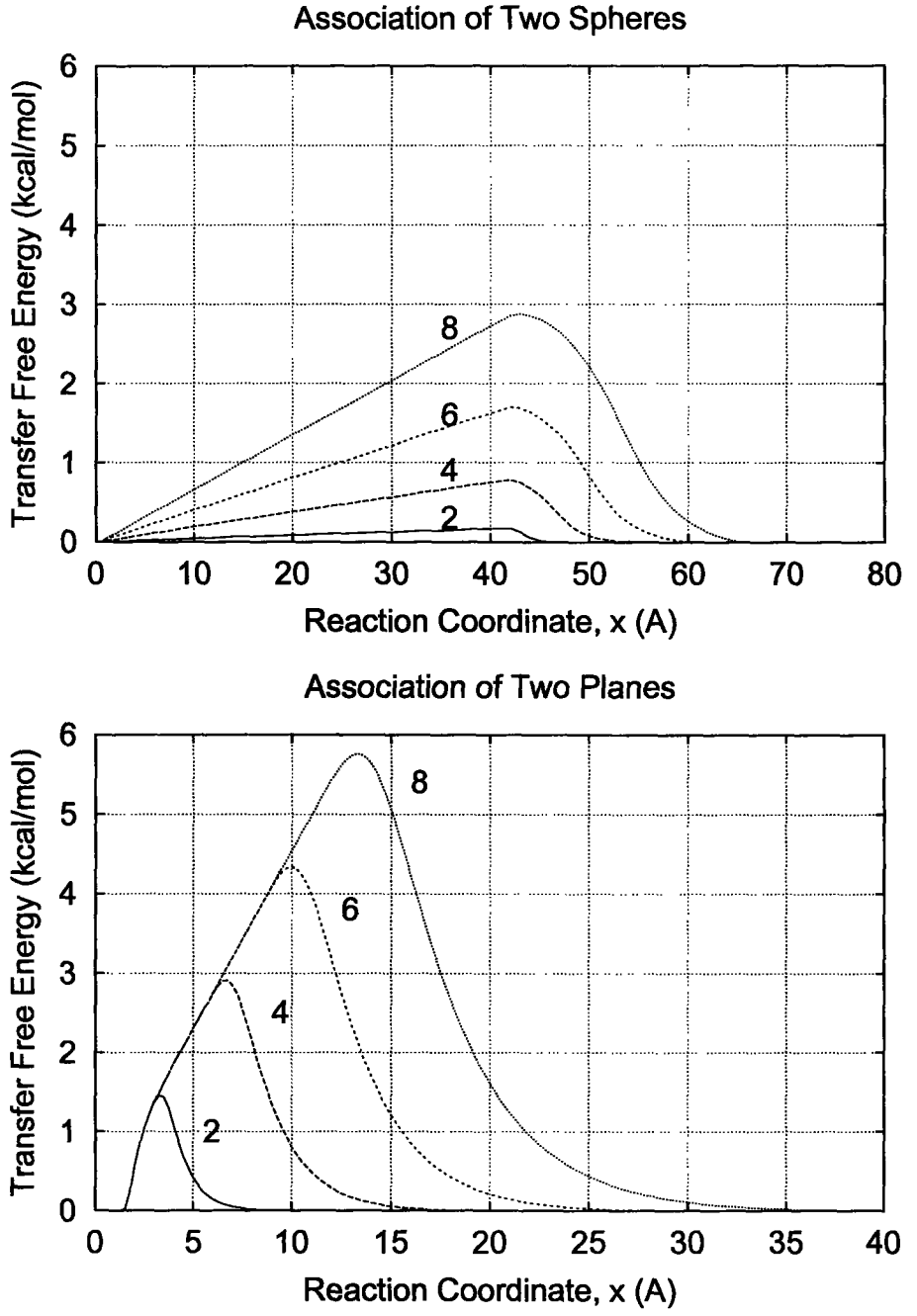

Figure 4-5: Transfer free energies for pairs of protein molecules transferred into 1M additive solution as a function of position along the association reaction coordinate, x, are shown. The sizes of the additives are varied while keeping the second virial coefficients constant. The curves are labeled with the additive sizes (r_m in Å) to which they correspond. The left-hand figure is for the association of two spherical proteins 20Å in radius, and the right-hand figure is for the association of two pseudo-infinite planar proteins with an area of 400πÅ^2 on each face.

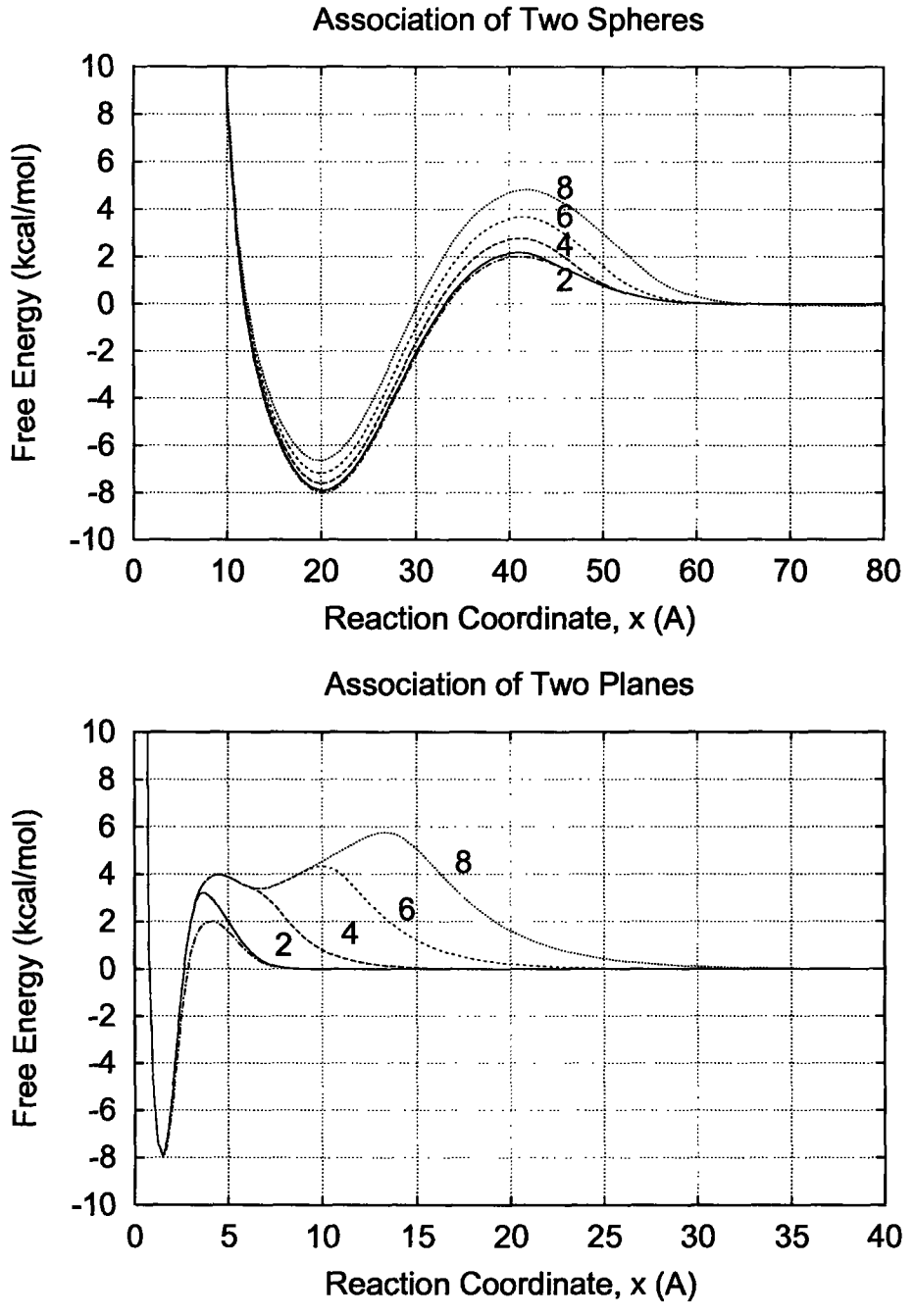

Figure 4-6: The protein free energies along the reaction coordinate for association/dissociation in the presence of neutral crowders at 1M concentration are shown. This combines $\mu_{P,0}$ (equations 4.2 and 4.3) with the transfer free energies shown in Figure 4-5. The curves are labeled with the additive sizes (r_m in Å) to which they correspond.

reaction rates (k/k_0) can be determined as a function of additive size (r_m) at constant Γ_{XP}, as shown in Figures 4-7 and 4-8. Increasing Γ_{XP} can also be seen to decelerate association and accelerate dissociation. This is a well-known result consistent with the fact that denaturants are used to slow protein association reactions. The magnitude of this change depends on whether the free surface area of the transition state is more similar to that of the monomer or to that of the dimer. If the transition state is similar to the dimer, as in the case of the planar proteins, there is a strong effect on k_a and almost no effect on k_d. For the spherical protein geometry, the transition state is closer to the monomer, and the effect of Γ_{XP} is larger on k_d than on k_a.

We also see that increasing additive size at constant Γ_{XP} decreases both the association and dissociation rates, consistent with the gap effect hypothesis. About a two order-of-magnitude drop in the association rate constant can be seen over the range of additive sizes shown (2 - 8Å). In the case of two associating planes, this effect does not appear at moderate additive sizes ($2.5 < r_m < 5.5$Å) because, although the maximum in transfer free energy keeps increasing, it moves away from the original transition state on the reaction coordinate.

4.2.2 Designing Additives for the Control of Aggregation

It may be possible to exploit the gap effect in designing solvent additives for the prevention of protein aggregation. Prevalent additives that work via a pure free surface effect, such as guanidinium and urea, have apparent radii (r_m) of about 2-3 Å. These have the disadvantage, however, that they can also enhance the unfolding or partial unfolding of proteins because of their positive preferential binding coefficients. The results of the preceding section suggest that if the size of these additives could be increased to about 8Å while maintaining their preferential binding coefficient with isolated protein molecules, the gap effect can potentially contribute another one to two order-of-magnitude depression in the association rate.

As the size of an additive is increased, its preferential binding coefficient will tend to decrease as the third power of radius. This is because increasing additive size increases the excluded volume of additive and protein, which decreases the preferential

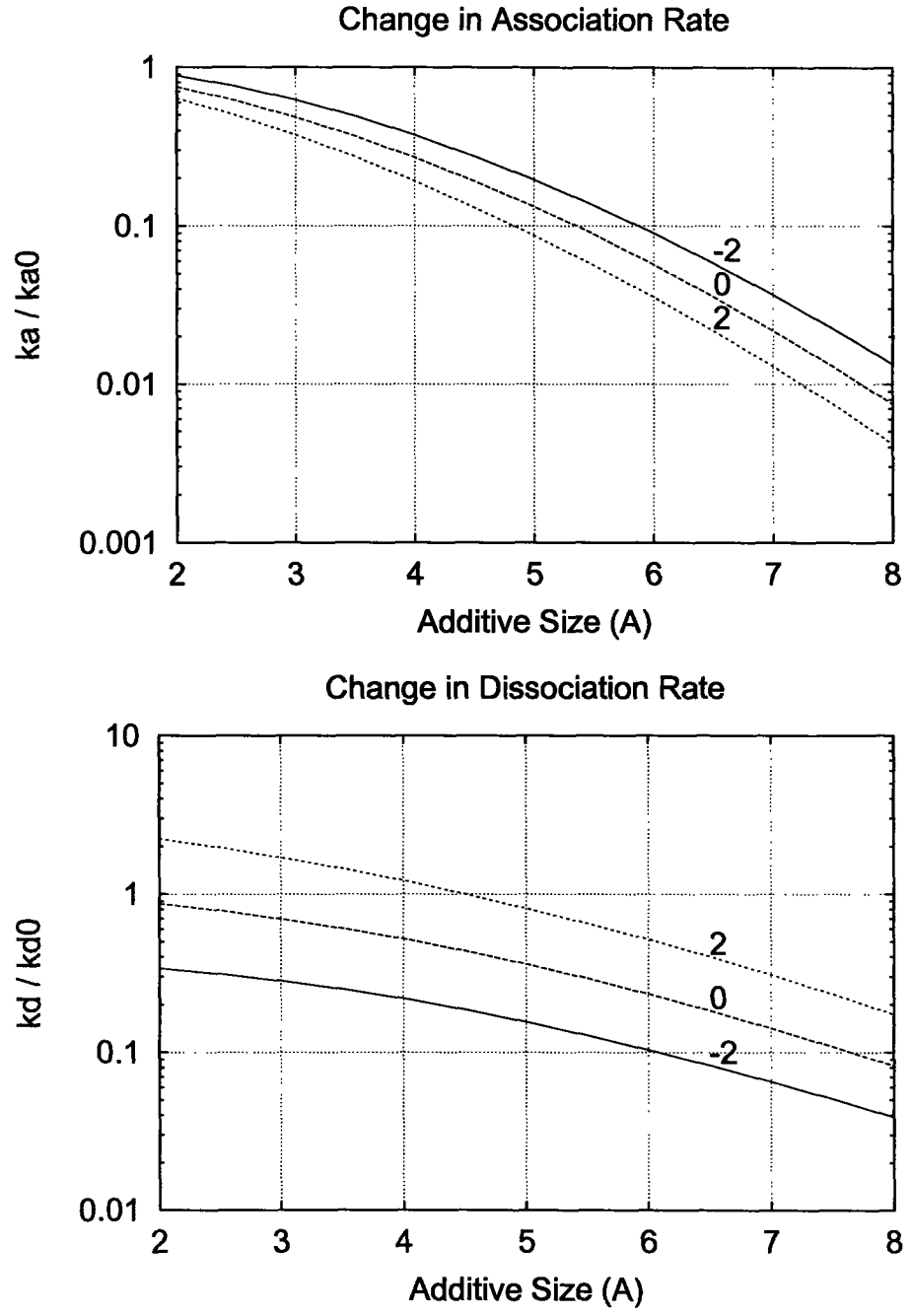

Figure 4-7: The change in association and dissociation rates for 20Å spherical proteins caused by a 1M additive is shown as a function of the additive size (x-axis) and additive-protein preferential binding coefficient, Γ_{XP} (labels).

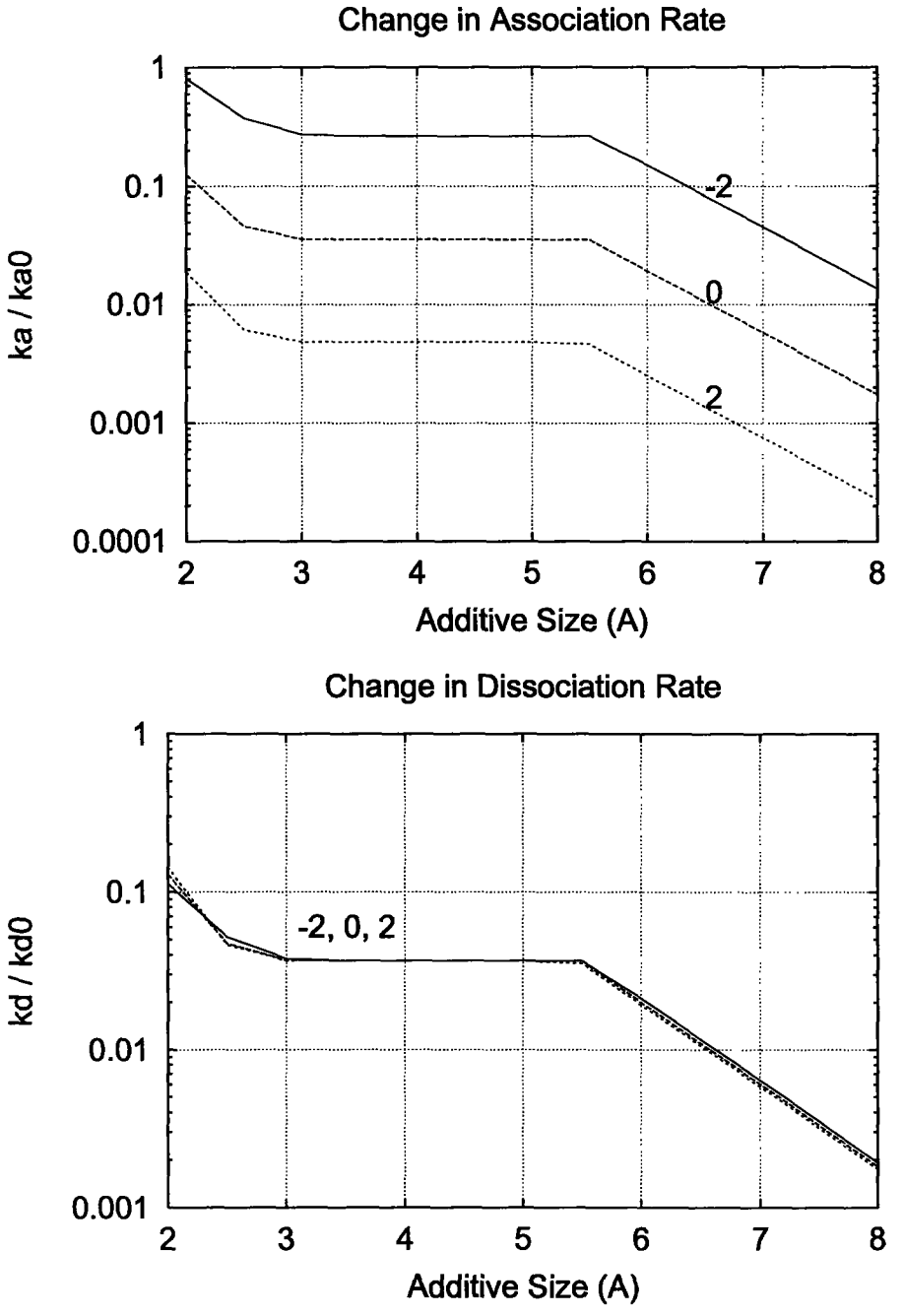

Figure 4-8: The change in association and dissociation rates for planar proteins (with $400\pi\text{Å}^2$ area on a face) caused by a 1M additive is shown as a function of the additive size (x-axis) and additive-protein preferential binding coefficient, Γ_{XP} (labels). For the dissociation rate, the Γ_{XP} dependence is negligible.

binding coefficient. In order to compensate for this excluded volume difference and return Γ_{XP} to its original value, an additional additive-protein attraction must be introduced into the molecule.

If it is not possible to increase the additive-protein attraction in some way as size is increased, the additive will have a large, negative preferential binding coefficient, and a gap effect will not appear. A gap effect arises for neutral crowders because there is a region of solvent that is inaccessible to the additive around encounter complexes but not around isolated protein molecules. In the case of an "excluded crowder" with a large, negative preferential binding coefficient, the volume of exclusion is actually larger in the dissociated state than in the encounter complex or associated state. Thus, in stark contrast to neutral crowders, excluded crowders like sugars, polyols, and large, hydrophilic polymers favor association [54, 49, 62].

4.3 Conclusions

In this work, a simple framework for modeling protein association and dissociation reactions in the presence of solution additives was developed and analyzed. Our model extends prior work in binding theory by considering various geometric models of the protein surface, the protein-protein association/dissociation transition states, and solvent radial distribution functions obtained from all-atom molecular dynamics simulations [10].

Our analysis of the model supports the hypothesis that a "gap effect," analogous to osmotic stress, will occur in association reactions when large solution additives with sufficient protein affinity are present. This gap effect affects the free energy of protein-protein encounter complexes, such as the association transition state, and has only a small effect on the end states. Thus, we have demonstrated how it is possible for an additive to exert a purely kinetic effect on protein association/dissociation. We call additives which have these properties "neutral crowders": they are neutral in that they do not significantly shift the free energy of isolated protein molecules, but they decrease the rate of protein association and dissociation by being excluded from

the inter-protein gap in protein-protein encounter complexes for steric reasons.

For an optimal effect, the maximum in the transfer free energy induced by the gap effect must be near the original association free energy barrier. When this is not the case, the gap effect will be strongest when the original barrier is small (less than 1-2 kcal/mol) or nonexistent, such as in diffusion-controlled reactions.

As the size of a neutral crowder is increased, the gap effect becomes proportionately larger, but maintaining neutrality is difficult as size increases. At a constant protein-additive interaction energy, increasing additive size would decrease the protein-additive preferential binding coefficient as the third power of additive size due to an excluded volume effect. Thus, to make a large neutral crowder, additive-protein interactions must become significantly more attractive as size is increased. If this cannot be achieved, the gap effect will diminish and ultimately disappear.

Today, the best known additives for suppressing aggregation are small denaturants such as urea and guanidinium chloride. Our gap effect model predicts that if a significantly larger additive, perhaps 4-5 times the size of these small additives, can be developed, and it were a "neutral crowder," it would depress association rates by a factor of 100-1000 times more than guanidinium or urea at the same molar concentration.

Chapter 5

Arginine is a Neutral Crowder

A prevalent anti-aggregation additive whose mechanism of function is unknown is the amino acid arginine. Arginine has very little effect on the folding equilibrium [6, 83, 82] yet it facilitates refolding of several proteins from the unfolded state [73, 8, 7, 70, 16]. While a mechanism which can explain fully how arginine functions has not been proposed [52], these results suggest that arginine selectively increases the barrier for protein-protein association while having little effect on protein folding.

These experimental observations are not consistent with any of the mechanisms reviewed earlier (section 2.4). They are, however, consistent with the proposed behavior of a neutral crowder (Figure 4-2).

If arginine is a neutral crowder, it should slow protein association reactions in general. There may be interactions between arginine and specific proteins which give rise to other kinetic effects, but on average, association rates should be lower in the presence of arginine than in its absence.

The purpose of the present investigation is to test the hypothesis that arginine is a neutral crowder via experiments designed to elucidate its effects on protein association reactions. Two types of protein association reactions have been selected for study: association of antibodies with their antigens (globular protein association) and association of folding intermediates (aggregation during refolding). By comparing the kinetic effects of arginine in these tests to other additives with known effects, we expect to be able to draw conclusions about arginine's performance in each of these cases.

5.1 Methodology

5.1.1 Proteins and Reagents

Human insulin (I8530), bovine carbonic anhydrase II (CA) (C2522), hen egg white lysozyme (L7651), and bovine serum albumin (B4287) were obtained from Sigma (St. Louis, MO). Monoclonal anti-insulin (10-I30 clone M322214) was obtained from Fitzgerald Industries (Concord, MA). Sheep myoglobin, monoclonal anti-myoglobin,

and consumable reagents for BIAcore experiments (NHS, EDC, ethanolamine, glycine, and HBS-EP buffer) were obtained from Biacore (Switzerland). Guanidinium chloride, arginine hydrochloride, sodium chloride, and lysine hydrochloride were attained from Sigma in the highest available grade.

Concentration of carbonic anhydrase in solution was determined by absorbance at 280nm using an extinction coefficient of 1.83 mL/mg/cm [65].

5.1.2 Globular Protein Association Kinetics

Protein association and dissociation rate constants, k_a and k_d, were measured for globular proteins via surface plasmon resonance on a BIAcore 3000 instrument. Each pair of proteins consisted of an antigen and a monoclonal antibody to that antigen. The antibody was immobilized on a BIAcore CM5 sensor chip via amine coupling. The amount of immobilized antibody was selected to give a detector response in the range of 50-100 RU when antigen was present. A reference surface was created by activating and deactivating the surface without coupling an antibody to it.

Different concentrations of antigen in the nanomolar range (1-1000 nM) were prepared by dilution and injected serially into the antibody-containing and reference flow cells. Lower concentrations of insulin (1-100nM) were used to ensure that multimerization of insulin did not affect the results [64]. For the myoglobin experiments, 5 μl of 50 mM glycine, pH 2.0, was used to regenerate the surface between injections of antigen. For the insulin experiments, the dissociation rate was sufficiently fast in buffer that a regeneration buffer was not required. Kinetic constants were extracted by simultaneous fitting of k_a and k_d to each set of sensorgrams using a 1:1 kinetic model in the BIAevaluation 3.0 software package.

5.1.3 Refolding of Carbonic Anhydrase

Refolding of carbonic anhydrase was accomplished by dilution from high concentrations of the denaturant guanidinium chloride (GuHCl) as done previously [21, 89]. High concentrations of carbonic anhydrase (>10 mg/ml) were denatured in

6M GuHCl and equilibrated overnight. Refolding was initiated by dilution to 0.5M GuHCl with 50 mM Tris-HCl buffer, pH 7.5. This final GuHCl concentration was selected because it yields a mixture of active, refolded protein and aggregates. The distribution of this mixture was analyzed via esterase activity, size exclusion HPLC, and dynamic light scattering as described below.

5.1.4 Carbonic Anhydrase Esterase Activity

Esterase activity of carbonic anhydrase was assessed using para-nitrophenylacetate (pNPA) as the substrate as described previously [65]. Briefly, 10 μl samples of carbonic anhydrase solution were added to 500 μl of Tris-HCl, pH 7.5 and 50 μl of 50 mM pNPA in acetonitrile. Kinetics of hydrolysis of pNPA was observed by the increase in absorbance at 400nm due to the appearance of the para-nitrophenolate ion (pNP^-). In all cases, the observed hydrolysis rate in absorbance units per second (AU/s) under these conditions was constant (pseudo-zero order). Hydrolysis rates were corrected for the hydrolysis of pNPA by the buffer for each type of buffer used. Hydrolysis rates were converted to concentration of active protein via a standard curve constructed from dilutions of known concentrations of native protein. The active protein concentration data was reproducible to within 5-8% in replicated experiments.

5.1.5 Size Exclusion HPLC

Size exclusion HPLC (SE-HPLC) experiments were performed on a Beckman System Gold HPLC instrument equipped with a Tosohaas G3000SWXL size exclusion column and a UV detector. 30 μl samples were introduced to the column by a constant flow of 1 ml/min mobile phase. Each sample ran for 15 minutes, with carbonic anhydrase eluting between 6 and 10 minutes, depending on its molecular weight and buffer. Protein was observed at the exit of the column via absorbance at 280nm. For samples that did not contain large submicron or micron-sized aggregates (which do not pass through the column), the total chromatogram areas at 280nm were consistent to within 2-3% during the entire refolding process, indicating that the extinction

coefficients of different sized aggregates did not vary significantly on a mass basis. A mixture of lysozyme, carbonic anhydrase, and bovine serum albumin (monomer and dimer) was used as a standard to calibrate molecular weight to retention time. Using this calibration curve and the breakthrough time of the column, the largest multimer that could pass through the column was a 15-mer. When significant mass was missing from a chromatogram, large multimers were quantitated by difference. The presence of large multimers was confirmed via turbidity or dynamic light scattering for each buffer. The instrument was cleaned with 30 μl injections of 4M GuHCl, a denaturing concentration found to dissociate and elute precipitates and large soluble carbonic anhydrase multimers.

5.1.6 Dynamic Light Scattering

Dynamic light scattering (DLS) experiments were performed with a SpectraPhysics stabilite 2017 laser (514nm) at a measurement angle of 90°. Brookhaven BI-200SM light scattering software was used for data acquisition and analysis. The autocorrelation function was fit with an exponential fitting software program to extract the distribution of diffusion coefficients, and the Stokes-Einstein equation was used to convert the diffusion coefficients to a distribution of hydrodynamic diameters. Most samples were filtered with a 0.45 μm syringe filter to remove dust.

5.1.7 Molecular Simulation

Molecular dynamics simulation of RNase A (PDB code: 1fs3) solvated by water and arginine were performed according to a protocol developed previously for simulation of proteins in mixed solvents (chapter 3).

5.2 Effect on Globular Protein Association

Surface plasmon resonance experiments were conducted to measure the effect of added ArgHCl on the kinetics of globular protein association and dissociation versus an

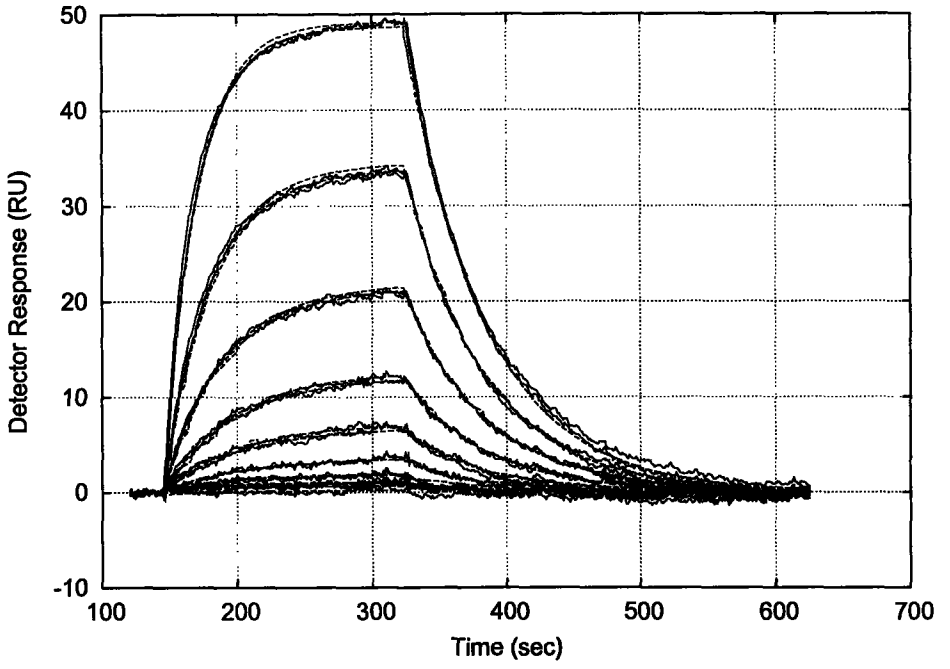

Figure 5-1: Biacore 3000 surface plasmon resonance data for insulin binding to immobilized anti-insulin. Raw binding data (solid curves) are shown with a three-parameter, least squares fit to all the data (dashed curves). The detector response is proportional to the mass of antigen bound to the antibody immobilized in the flow cell.

equimolar salt control (NaCl). Two model antibody-antigen protein pairs were tested as model systems: insulin and a monoclonal antibody to insulin, and myoglobin and a monoclonal antibody to myoglobin. A typical experimental data set for a binding interaction at one buffer condition is shown in Figure 5-1. The data set shown in the figure is a composition of 8 different concentration runs plus a replicate, for a total of 16 runs. At t=140 sec, the flow cell with immobilized anti-insulin was exposed to a constant concentration of insulin in the range of 2 to 188nM for 3 minutes. During this 3 minutes, the antibody and antigen were free to associate and dissociate. The net reaction is the binding of free antigen in solution, resulting in an increase in detector response proportional to the mass of antigen bound. At t=320 sec, the insulin concentration is returned to zero, and the bound antigen then dissociates from the surface. All 16 runs were simultaneously fit to a binding model by minimizing the squared residuals to yield the association and dissociation rate constants, k_a and k_d.

Antigen	Buffer additive	k_a ($M^{-1}s^{-1}$)	k_d (s^{-1})	$K_d(\mu M)$	k_a/k_{a0}	k_d/k_{d0}
Insulin	0.5M NaCl	4.4 x 10^4	1.4 x 10^{-2}	0.32		
Insulin	0.5M ArgHCl	1.2 x 10^4	2.2 x 10^{-2}	1.8	0.27	1.6
Insulin	0.5M GuHCl	4.0 x 10^4	9.4 x 10^{-2}	2.4	0.91	6.7
Myoglobin	0.5M NaCl	1.8 x 10^4	4.1 x 10^{-3}	0.23		
Myoglobin	0.5M ArgHCl	8.3 x 10^3	7.1 x 10^{-3}	0.86	0.46	1.7

Figure 5-2: Effect of arginine on association and dissociation rate constants for two model proteins, insulin and myoglobin, with monoclonal antibodies to each. The base buffer was BIAcore HBS-EP (10mM HEPES, 0.15M NaCl, 3mM EDTA, 0.005% polysorbate 20, pH 7.4) buffer. k_{a0} and k_{d0} are the association and dissociation rate constants in added 0.5M NaCl for each protein. $K_d \equiv k_d/k_a$. The estimated error in the absolute values of k_a and k_d is about 15%.

This process was repeated to yield association, dissociation, and equilibrium constant data for the model systems in various buffers as shown in Figure 5-2.

Relative to the 0.5M NaCl control, 0.5M GuHCl significantly increases the dissociation rate of insulin and anti-insulin and has an insignificant effect on the association rate. This effect of GuHCl on dissociation rate is consistent with its well-known behavior as a strong denaturant. Small denaturants such as guanidinium chloride and urea bind uniformly to protein surfaces and thermodynamically favor protein states which have the largest solvent-accessible area, such as denatured states (in folding equilibria) and dissociated states (in association equilibria). Since GuHCl does not significantly affect the rate of association of insulin and anti-insulin, it is likely that the association transition state does not have a significantly different solvent-accessible area than the dissociated state.

In stark contrast to GuHCl, ArgHCl has a much smaller effect on the dissociation rate, and induces a large decrease in the association rate. This difference in "kinetic signatures" suggests that ArgHCl and GuHCl act via distinct mechanisms. Considering only free surface binding effects like that of the denaturant GuHCl, it is not possible to account for both the change in association and dissociation rates seen in the ArgHCl case. The kinetic signature of ArgHCl is, however, consistent with that of a neutral crowder, a large additive that has little effect on the free energy of isolated protein molecules [11].

Using gap effect theory (chapter 4), it is possible to estimate the association rate

depression (k_a/k_{a0}) that can be caused by a model neutral crowder the size of arginine. Since the structure of the antibody-antigen complexes are not known, and we only require an approximate value for the association rate change, a simplified reaction coordinate model can be utilized. Approximating the antigens and the binding domain of each antibody as having a characteristic dimension of 20Å, the model reaction coordinates shown in Figure 4-3 can be utilized. Taking $r_m \approx 4$Å as the size of arginine, Figure 4-5 gives a range of 0.8 - 2.8 kcal/mol/M for the maximum increase in the free energy barrier to association. For 0.5M arginine solution, this is 0.4 - 1.4 kcal/mol, or a rate effect of $k_a/k_{a0} = e^{-\Delta\Delta\mu_{\ddagger}^{tr}/RT} = 0.51$ to 0.10, which covers the range of association rate effects observed for insulin and myoglobin with their monoclonal antibodies.

5.3 Effect on Refolding of Carbonic Anhydrase

Carbonic anhydrase was refolded in 0.5M GuHCl plus 0.5M of three different chloride salts to assess the effect of the cations on the refolding process. The cations tested were sodium (a control), arginine, and guanidine, all at a constant pH of 7.5. The yield of native protein was assessed via the recovery of carbonic anhydrase's esterase activity. The molecular weight distribution of aggregates formed during the refolding processes was assessed via size exclusion HPLC.

5.3.1 Yield of Native Protein

Esterase activity assays were performed as a function of initial unfolded protein concentration and buffer composition (Figure 5-3) to determine how each cation affected refolding yield. The yield of active protein as a function of buffer additive increases in the following order: NaCl << ArgHCl < GuHCl.

The relation between the refolding yield and initial protein concentration suggests a simple model for predicting yield. The data are consistent with an aggregation model in which there is a "decision point" species, X, which can either refold to the

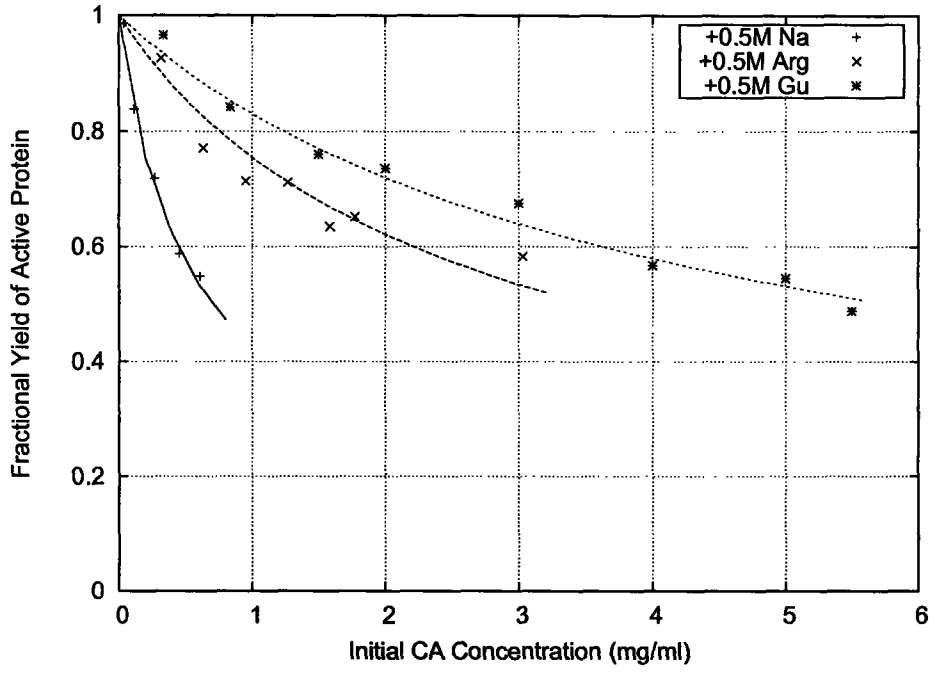

Figure 5-3: Effect of refolding buffer composition on carbonic anhydrase refolding yield. The points are experimental esterase activity data, and the lines are the best fit to a one-parameter, first versus second order kinetic model (equation 5.5).

native form or aggregate irreversibly and remain inactive:

$$\text{Active Monomer} \overset{k_1}{\leftarrow} X \overset{k_2}{\rightarrow} \text{Inactive Aggregate} \tag{5.1}$$

where k_1 is the rate constant for activation and k_2 is the rate constant for inactivation. In the simplest case, X is the unfolded or partially-unfolded monomer, which either folds to the native form or associates irreversibly. In that case, $[X]_0 = [U]_0$. More generally, if a large fraction of the monomer is in a state with mer number m before the critical inactivation decision, then $[X]_0 \approx [U]_0/m$. The yield of active protein is then:

$$\begin{aligned} \text{Yield} &= \frac{m}{[U]_0} \int_0^{[U]_0/m} \frac{r_1}{r_1 + r_2} d[X] \\ &= \frac{m}{[U]_0} \int_0^{[U]_0/m} \frac{d[X]}{1 + r_2/r_1} \end{aligned} \tag{5.2}$$

where $[U]_0$ is the initial concentration of the unfolded protein, r_1 is the rate of acti-

vation, r_2 is the rate of inactivation, and concentrations and rates are expressed on a mass basis. Since the activation process is likely to be either an intramolecular rearrangement (folding) or a dissociation event (from a multimeric form), it should exhibit first-order kinetics and follow a rate law $r_1 = k_1[X]$. If the primary inactivation process is association, it will exhibit higher-order kinetics, as $r_2 = k_2[X]^n$, where n is the order of the inactivation reaction. Substituting the assumed rate laws into the above equation produces:

$$\text{Yield} = \frac{m}{[U]_0} \int_0^{[U]_0/m} \frac{d[X]}{1 + k_2[X]^{n-1}/k_1} \tag{5.3}$$

Without knowledge of n, the solution to the above integral can only be expressed as an infinite series. If the inactivation process is simple association, n=2. If it is a combination of many parallel association reactions, or the crossing of a high free-energy barrier at a high mer number as in nucleation-dependent polymerization, n will be greater than 2. Such higher-order inactivation kinetics ($n = 3$) have been observed for refolding of other proteins such as lysozyme [40]. In an earlier study of carbonic anhydrase refolding via dynamic light scattering, Cleland and Wang [23] proposed a 2.6-power relationship between initial protein concentration and monomer depletion rate at short times (30-60 sec). Thus, we expect a value of n between 2 and 3 to be applicable in this case.

Model cases for $n = 2$ and $n = 3$ were fit to the data and revealed that $n = 2$ gave a much better fit for all three buffer conditions. The activity data with added 0.5M GuHCl and 0.5M ArgHCl are suggestive of slightly higher inactivation order than the added 0.5M NaCl case, but because of the uncertainty ($\pm 5\%$) in the esterase activity data, it is not possible to determine n to better than about ± 0.5 by direct fitting. For $n = 2$, the model reduces to:

$$\text{Yield} = \frac{mk_1}{k_2[U]_0} \ln\left(1 + \frac{k_2[U]_0}{mk_1}\right) \tag{5.4}$$

Since the constants k_1, k_2, and m appear only as a quotient, they can be condensed to a single "refolding selectivity parameter," $\alpha \equiv mk_1/k_2$, having units of concentration

Additive	α (mg/ml)	α/α_0
0.5M NaCl	0.28	1
0.5M ArgHCl	1.4	5.0
0.5M GuHCl	2.3	8.2

Figure 5-4: Refolding selectivity parameters (α) and parameters relative to 0.5M NaCl (α/α_0) are shown for refolding of carbonic anhydrase with three different buffer additives. The base buffer composition was 0.5M GuHCl.

to the $n-1$ power and resulting in a working equation:

$$\text{Yield} = \frac{\alpha}{[U]_0} \ln\left(1 + \frac{[U]_0}{\alpha}\right) \tag{5.5}$$

Each of the data sets in Figure 5-3 were fit to the above model equation, yielding the values of α shown in Figure 5-4. The functional forms of the model at these values of α are shown in Figure 5-3.

The parameter α is a direct measure of the performance of a refolding additive. It is equal to the concentration of unfolded protein at which the refolding yield will be ln(2), or about 70%.

The relative refolding selectivity values (α/α_0) for ArgHCl and GuHCl are qualitatively consistent equilibrium shifts effects seen in globular protein association (section 5.2). This implies that formation of irreversible aggregates is at least partially equilibrium-controlled.

5.3.2 Multimer Distribution

Size exclusion HPLC experiments were performed to analyze the distribution of multimers formed during refolding. CA was refolded with three different additives, 0.5M NaCl, 0.5M GuHCl, and 0.5M ArgHCl, relative to a base refolding buffer of 0.5M GuHCl, as done above in the esterase activity assays. The 0.5M NaCl refolding experiment was performed at 4-fold lower concentration (5 μM) because visible aggregates were formed within seconds at concentrations comparable to the other two experiments (20 μM). Other than this protein concentration difference, these experiments allow direct comparison of how an additional 0.5M of the three different cations

(a) Additive: 0.5M NaCl, $[U]_0 = 5\ \mu M$

Time (min)	M	A_2	A_{3-5}	A_{6-15}	Large
2	56	0	0	0	44%
20	56	0	0	0	44%
38	58	0	0	0	42%

(b) Additive: 0.5M ArgHCl, $[U]_0 = 20\ \mu M$

Time (min)	M	A_2	A_{3-5}	A_{6-15}	Large
2	22	30	25	21	2%
20	54	7	14	26	-1%
38	62	4	11	24	-1%
1500	80	0	0	19	1%

(c) Additive: 0.5M GuHCl, $[U]_0 = 20\ \mu M$

Time (min)	M	A_2	A_{3-5}	A_{6-15}	Large
2	42	39	8	0	11%
20	82	3	6	0	9%
38	85	1	5	0	9%
1500	89	0	2	0	9%

Figure 5-5: HPLC analysis of multimers formed during refolding of carbonic anhydrase in different buffers, expressed as a percentage of the total carbonic anhydrase. The time reported is the time between injection onto the HPLC column and dilution of the denatured carbonic anhydrase into the refolding buffer. The base refolding buffer contained 0.5M GuHCl. M indicates monomer, and A_{i-j} indicates multimers of mer number i through j. The amount of "Large" multimers which do not pass through the column is inferred from the difference between the amount of protein injected onto the column and the total chromatogram area. The reproducibility of any peak area determination from experiment to experiment is $\pm 1\%$.

affect refolding.

After initiating refolding by diluting denatured CA with an appropriate buffer, three samples were run in sequence. A final sample was run after one day at each condition. The molecular weight distributions observed as a function of time are shown in Figure 5-5.

In all three cases, rapid association of the unfolded or partially unfolded protein occurs at a time scale shorter than that which can be probed via HPLC. At time zero, all of the protein is unfolded and monomeric, but by the time the first protein peaks elute from the HPLC (about 12 minutes after dilution and refolding commence), multimer assembly has occurred to a significant extent and the major reactions observed

are dissociation of small multimers. This is consistent with observations in previous studies [20].

In 0.5M NaCl, the refolded carbonic anhydrase is partitioned entirely between monomers and large aggregates, with no significant isomerization between these states observed on a time scale of minutes. With added 0.5M ArgHCl or GuHCl, the yield of monomeric protein at 1500 min (about 1 day) is significantly increased. Also, small multimeric intermediates are observed at short times (minutes to hours). These intermediates predominately dissociate into monomeric protein, consistent with previous studies [20].

In all three refolding buffers, significant amounts of large aggregates form which do not dissociate into monomeric protein within one day. With longer refolding times, the average aggregate molecular weight and hydrodynamic radii continue to increase (data not shown), with only very slow loss of the monomer. This implies that the native protein and large aggregate states are separated by a large free energy barrier.

5.3.3 Models of Additive Effects on Aggregation during Refolding

Changing the refolding buffer additive from 0.5M NaCl to 0.5M GuHCl or 0.5M ArgHCl had significant effects on the refolding yield and on the observed multimer distribution. In this section, we propose simple kinetic models that are consistent with the experimental data and the effects of the three solution additives.

0.5M NaCl

With added 0.5M NaCl as in the other two buffer conditions, the yield of active protein can be predicted from a model where a first order reaction is in competition with a reaction of order 2-2.5. Since no multimeric intermediates are observed on a time scale of minutes (HPLC), the simplest kinetic model consistent with this data

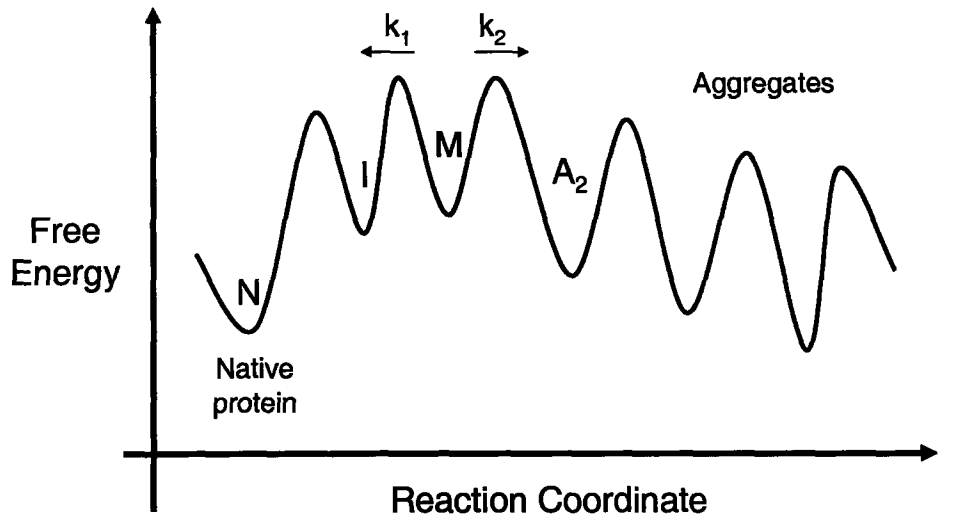

Figure 5-6: Proposed kinetic model for carbonic anhydrase refolding in 0.5M GuHCl + 0.5M NaCl. The unfolded protein rapidly collapses to the molten state (M) from which it can either refold (via k_1 to the intermediate state I and then N) or irreversibly aggregate (via k_2).

and previous studies [81, 80, 20] is the following:

$$U \overset{\text{fast}}{\rightarrow} M \overset{k_1}{\rightarrow} I \rightarrow N \tag{5.6}$$

$$M + M \overset{k_2}{\rightarrow} \text{aggregates} \tag{5.7}$$

where U is the unfolded protein; M is the association-competent, molten intermediate; I is an folding intermediate close to the native state; and N is the native, active protein. This kinetic model is shown in terms of a free energy-reaction coordinate diagram in Figure 5-6. This model will be considered the base case, and changing the additive from 0.5M NaCl to 0.5M GuHCl or 0.5M ArgHCl will be analyzed as perturbations to this base case in the following sections.

0.5M GuHCl

With 0.5M GuHCl added in the refolding buffer, the refolding yield is higher, and small aggregates which dissociate to monomer are visible on the time scale of minutes to an hour. Mechanistically, guanidinium chloride acts by binding to and reducing

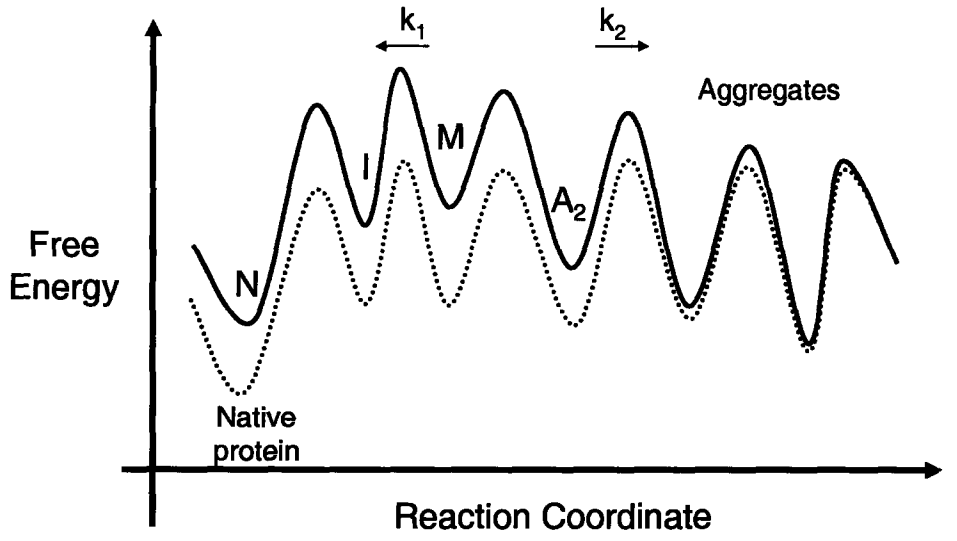

Figure 5-7: Proposed kinetic model for carbonic anhydrase refolding with 0.5M GuHCl added to the refolding buffer. The base case model for 0.5M NaCl added is shown as a solid line, and the new free energy landscape for 0.5M GuHCl added as a dotted line. GuHCl shifts the landscape toward the smaller mers by increasing the dissociation rates. The net effect is an increase in the yield of active protein.

the free energy of protein states in proportion to their solvent accessible area [5, 34]. The postulated effect of added GuHCl on the free energy-reaction coordinate diagram for carbonic anhydrase refolding is shown in Figure 5-7. Since there is a loss of solvent accessible area upon association, guanidinium chloride shifts the free energy landscape toward the dissociated species. Also, because a significant amount of dimer which ultimately refolds is observed at short times via HPLC, there must be a major barrier to further association after the dimerization reaction. The NaCl experiments do not reveal this barrier because no small aggregates are observed. These observations can be reconciled by assuming that guanidinium chloride decreases the free energy of the monomeric species relative to the dimer such that a new maximum in free energy for association occurs after the dimer species, as shown in Figure 5-7. After formation of a species larger than the dimer, association becomes more rapid and only a small amount of intermediate multimers are observed between the dimer and very large aggregates.

0.5M ArgHCl

Similar to added 0.5M GuHCl, the presence of an additional 0.5M ArgHCl increases the yield of active carbonic anhydrase relative to 0.5M NaCl, and small aggregates which dissociate to monomer are visible on the time scale of minutes to an hour.

The additives differ significantly in their performance at higher mer number, however. In added 0.5M NaCl or 0.5M GuHCl, only a small amount of mass between the dimer and very large multimers ($A_1 6+$) is observed. This indicates there is a "downhill polymerization" regime between these states where association is very rapid, the beginning of which is shown in Figures 5-6 and 5-7. In added 0.5M ArgHCl, this downhill polymerization is significantly attenuated, and the largest multimers are smaller than a 15-mer. This occurs despite the fact that the refolding yield is lower in 0.5M ArgHCl than in 0.5M GuHCl.

This difference indicate that, like in the globular protein association studies (section 5.2), the kinetic signatures of ArgHCl and GuHCl are quite different. In the downhill polymerization regime, association is rapid, and dissociation is negligible. Thus, additives which deter aggregation by increasing multimer dissociation rate will not significantly affect the downhill polymerization regime. Since guanidinium primarily acts by increasing k_d, guanidinium cannot attenuate this downhill polymerization, and mass which escapes from the free energy well at low mer number proceeds rapidly to very large mers. In the case of arginine, however, this phase of aggregation is significantly attenuated. This indicates that arginine slows association of the intermediate multimers. The kinetic signature of arginine in refolding thus also includes a significant k_a decrease, as observed in the globular protein association studies.

A postulated free energy-reaction coordinate diagram for refolding in the presence of 0.5M ArgHCl is shown in Figure 5-8.

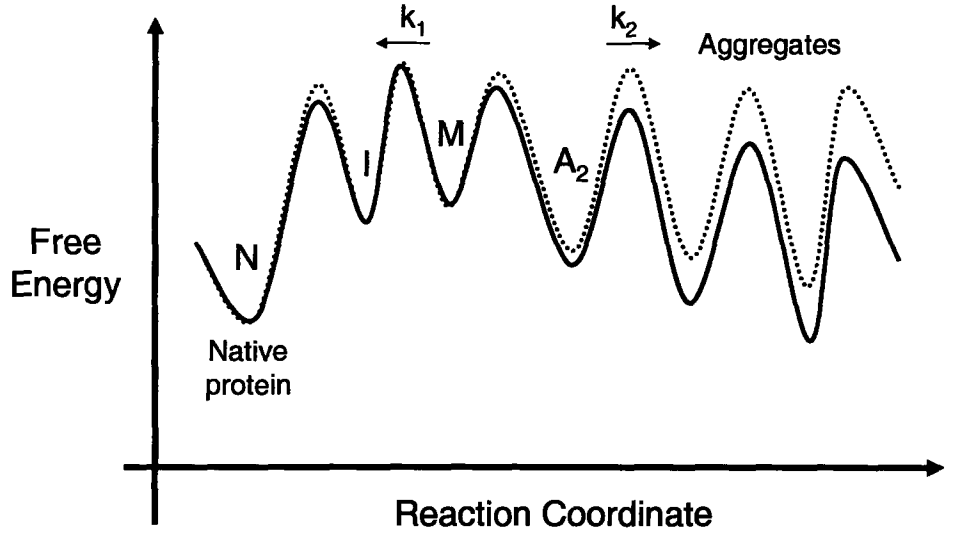

Figure 5-8: Proposed kinetic model for carbonic anhydrase refolding with 0.5M ArgHCl added to the refolding buffer. The base case model for 0.5M NaCl added is shown as a solid line, and the new free energy landscape for 0.5M ArgHCl added as a dotted line. ArgHCl shifts the landscape toward the smaller mers by decreasing association rates and slightly increasing dissociation rates. The net effect is smaller than that of GuHCl but still results in an increase in the yield of active protein.

5.3.4 Effect of Different Kinetic Signatures on Surrogate Assays

We have seen that arginine affects association reactions primarily by decreasing association rate, and guanidinium affects association reactions primarily by accelerating dissociation rate. This difference in kinetic signatures may have important consequences when using simple surrogate assays to detect protein aggregation in solution.

As seen in the differences in yield and aggregate molecular weight distribution between the refolding buffer additives ArgHCl and GuHCl (sections 5.3.1 and 5.3.2), a decrease in the average aggregate molecular weight may not be indicative of increased refolding yield. Simple aggregation assays such as turbidity and dynamic light scattering, which roughly measure the amount of large particles in solution, will also not correlate with yield when comparing additives that affect association with those that affect dissociation.

5.4 Molecular-level Interactions of Arginine with a Model Protein

Arginine is composed of an amino acid backbone, a guanidino group, and a trimethylene linker, chemical moieties which have very different affinities for proteins. The closest analog of the amino acid backbone is the amino acid glycine. Glycine stabilizes proteins against unfolding, increases the free energy of proteins in proportion to their solvent accessible area, and is preferentially-excluded from their vicinity. The guanidino group, a strong denaturant, has exactly the opposite behavior. Thus, we suspected that if the trimethylene linker is neutral, arginine will be preferentially-orientated at the protein interface, with its guanidino group tending to face the protein.

To test this hypothesis and probe the interactions of arginine with proteins in more detail, a detailed molecular simulation of the model protein RNase A (PDB code: 1fs3) in 1M arginine solution was conducted. RNase A was selected as a model protein because its small size (14kD) made the simulation more economical in terms of computational time.

5.4.1 Orientation of Arginine

The orientation of an arginine molecule relative to the protein was defined as illustrated in Figure 5-9. Two vectors drawn from the center of mass of an arginine molecule, one normal to the protein, the other through the arginine guanidino carbon, define an orientation angle θ. When $\theta < 90°$, the guanidino group is facing toward the protein. When $\theta > 90°$, the guanidino group is facing away from the protein. Because the side chain of arginine remains fairly extended in solution, the alpha carbon is on the opposite side of the molecule. Thus, $\theta < 90°$ means that the alpha carbon and amino acid backbone are facing bulk solution, and $\theta > 90°$ means that they face the protein.

As observed in molecular dynamics simulations of additive effects on protein free

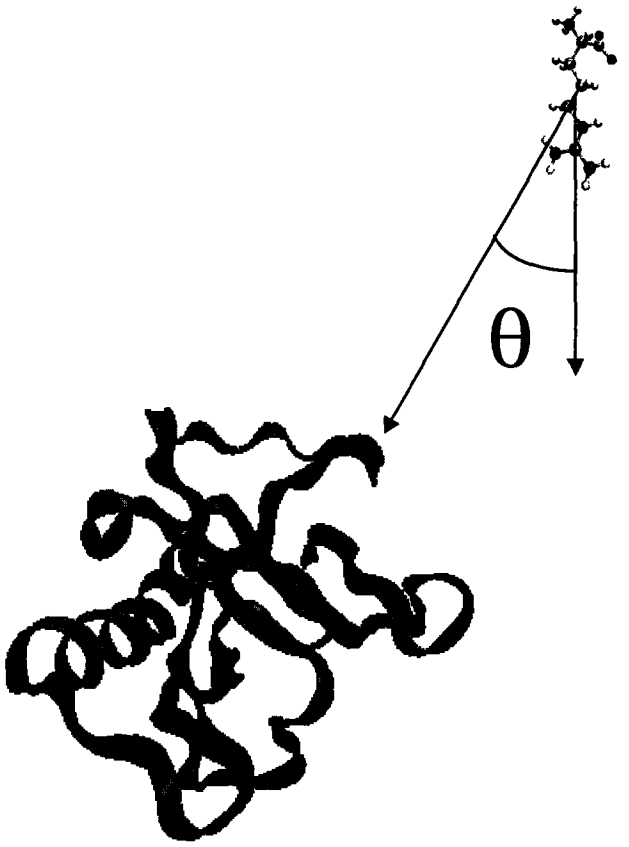

Figure 5-9: The definition of the orientation angle of arginine (θ) relative to a protein is shown. The vertex of the angle θ is at the center of mass of the arginine molecule. One vector is normal to the protein's van der Waals surface. The other goes through the zeta (guanidino) carbon.

energy (chapter 3), it is expected that arginine molecules that are a sufficient distance from the protein will not show any orientation preference relative to the protein. It is important to classify arginine molecules as either "local" or "bulk," and then analyze the preferential orientation of only the local arginine molecules in order to maximize the statistical significance of this calculation.

To determine the minimum distance between an arginine molecule and protein at which the orientation of an arginine is not influenced by the protein, the ensemble average of arginine orientation as a function of distance from the protein was computed. Statistical differences observed during the simulation can be expressed succinctly in terms of an orientation free energy which captures the free energy difference between the state with the guanidino carbon pointing away from the protein ($\theta = 90 - 180°$) and facing the protein ($\theta = 0 - 90°$). The ratio of the populations of each state can be used to define an orientation free energy:

$$\Delta G_\theta = -RT \ln \left(\frac{\int_0^{90} f_\theta \, d\theta}{\int_{90}^{180} f_\theta \, d\theta} \right) \tag{5.8}$$

where ΔG_θ is the orientation free energy, and f_θ is the normalized orientation probability density. The orientation free energy (ΔG_θ) for an arginine as a function of distance from the protein is shown in Figure 5-10. At short distances ($r < 2$Å), arginine has significant preferences to either orient its guanidino group toward the protein or away from the protein. We take these arginine molecules as the "first solvation shell" and assume they are the only ones significantly influenced by the protein. Further from the protein ($r > 2$Å), there is no significant statistical preference for either orientation. These arginine molecules are classified as "bulk." The fact that there is no orientation preference in the bulk solution is a control on the simulation method, particularly the possibility of edge effects.

The distribution of orientation angles observed during the molecular dynamics simulation for the "first shell" and "bulk" arginine molecules are shown together with the theoretical random probability distribution in Figure 5-11. The probability densities (f_θ) shown are normalized so that their integrals over the range of orientations

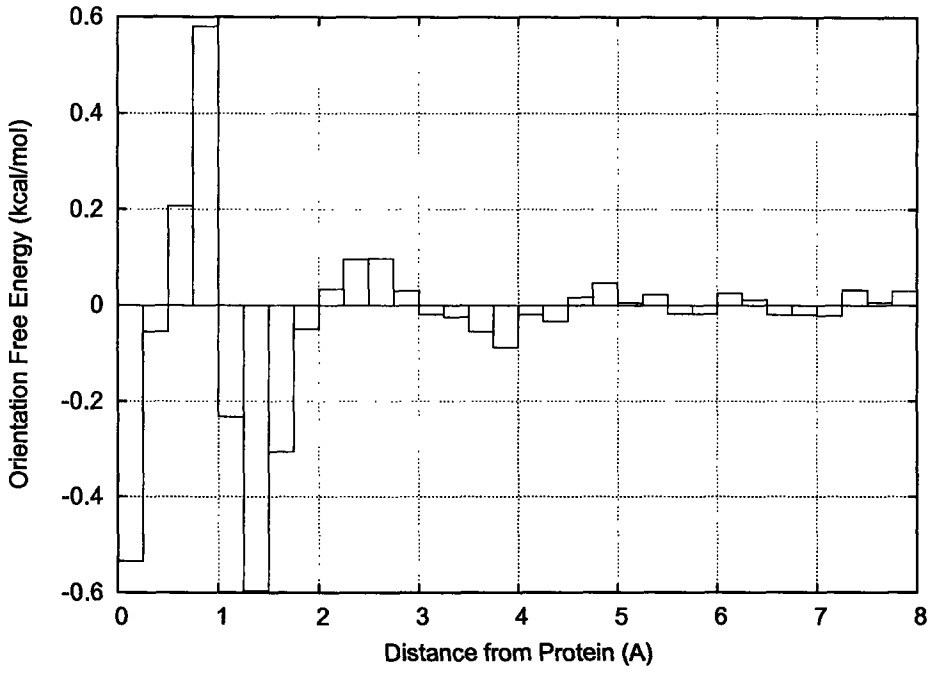

Figure 5-10: The orientation free energy of arginine (ΔG_θ) as a function of distance from the protein (r) is shown. The orientation free energy is the free energy of flipping an arginine molecule from a state where its guanidino group faces away from the protein to a state where its guanidino group faces the protein.

is equal to 1.

The random probability density can be derived by considering the probability density of a randomly oriented vector relative to any fixed vector. For convenience, we can select the x-axis as the fixed vector, and uniformly sample all orientations of a vector whose tail is at the origin relative to this fixed vector. The (non-normalized) probability density is then:

$$f_\theta = \frac{d}{d\theta}\left(\int_0^\theta \int_0^{2\pi} \sin(\theta) d\phi d\theta\right) \tag{5.9}$$

where ϕ is the angle in the y-z plane, and θ is the angle relative to the x-axis. The result is simply $f_\theta = \sin(\theta)$, which with an appropriate normalization factor for the unit of degrees becomes $f_\theta = \pi/360 \sin(\theta)$.

The arginine molecules in the bulk solution do not show any significant deviations from this random distribution, as expected. In contrast, arginine molecules in the

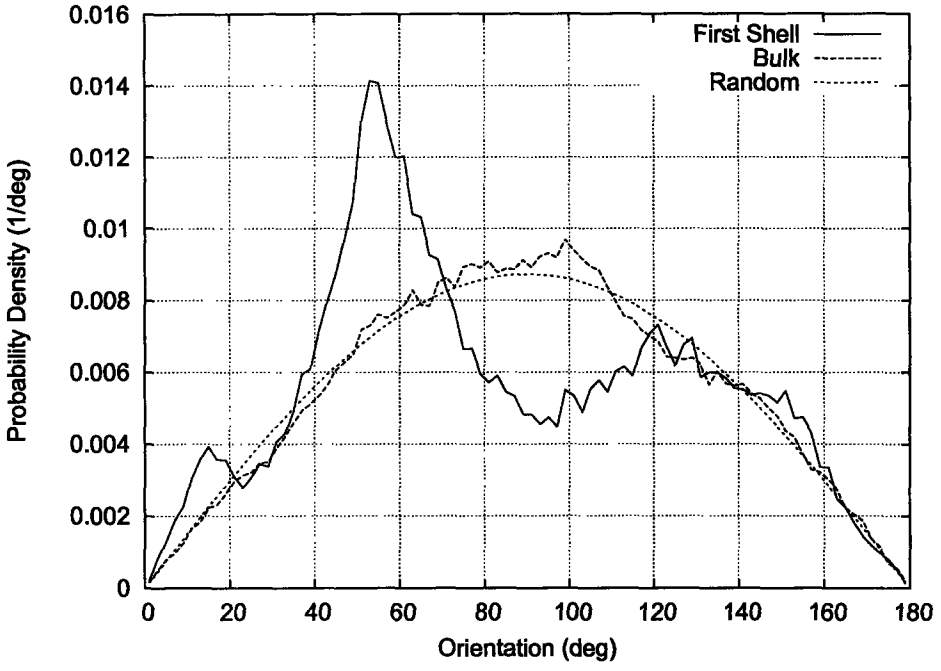

Figure 5-11: The probability density of arginine orientation (θ) relative to RNase A in solution is shown. Arginine molecules are divided into two classes, first shell and bulk, depending on whether any of their atoms lie within 2Å of the protein's van der Waals surface. The random probability density is shown for comparison. The deviations from the random distribution in the first shell imply that arginine is preferentially-oriented at the protein's surface.

first solvation shell, have a slight preference for $\theta \approx 40 - 70°$, and slightly avoid the orientations $\theta \approx 70 - 120°$.

Using the probability density in Figure 5-11, the orientation free energy of an arginine molecule in the first shell is -0.19 ± 0.04 kcal/mol. This means that there is a slight preference for arginine to be oriented so its guanidino group faces the protein, as hypothesized. The error (± 0.04 kcal/mol) is an estimate of the statistical error introduced by using a finite length trajectory to perform this calculation. This estimate was obtained by dividing the trajectory into four pieces of equal length, computing ΔG_θ in each piece, and taking the standard deviation of these.

5.5 Conclusions

Arginine slows protein association in two types of model systems when used as a solution additive. Added 0.5M ArgHCl slowed association in two globular protein systems, insulin and myoglobin with their corresponding monoclonal antibodies, by a factor of 2.2 to 3.7 versus a 0.5M NaCl control. In contrast, denaturants like guanidinium affect association reactions primarily by accelerating the dissociation rate.

0.5M ArgHCl also promoted the refolding of carbonic anhydrase by a factor of 5 versus 0.5M NaCl. When both arginine and guanidine are used as refolding buffer additives, an additional 0.5M GuHCl is superior to 0.5M ArgHCl, promoting refolding by a factor of 1.8. Despite this difference, the formation of large submicron aggregates is significantly suppressed by added 0.5M ArgHCl relative to added 0.5M GuHCl. Thus, arginine attenuates aggregation of large mers more effectively than guanidinium, even though small aggregates form more rapidly in arginine.

The observation that ArgHCl attenuates large aggregate formation better than GuHCl is consistent with arginine acting primarily by slowing protein-protein association, and GuHCl acting primarily by accelerating multimer dissociation. GuHCl increases refolding yield more than ArgHCl because the magnitude of its effect on the association equilibrium constant is larger than the ArgHCl effect.

In molecular simulations of a model protein (RNase A) in arginine solution, arginine was observed to interact most favorably with the protein through its guanidino group. When in the first solvation shell around a protein, arginine preferentially orients itself with its guanidino group tending to face the protein. The free energy of this configuration is 0.19 kcal/mol lower than an orientation in which the guanidino group faces away from the protein.

Taken together, these results suggest that the mechanism by which arginine prevents aggregation is that a loose, preferentially-oriented shell of arginine molecules surround each protein and slow protein-protein association via a gap effect. The order-of-magnitude of arginine's effects on globular protein association and on protein aggregation during refolding are consistent with that predicted by gap effect theory for a neutral crowder of 4Å radius.

Chapter 6

Conclusions

6.1 Calculation of Thermodynamic Properties of Proteins in Mixed Solvents

A quantitative method for calculating preferential binding coefficients without adjustable parameters has been developed and validated. The preferential binding coefficient is linked directly to the fundamental thermodynamic property of interest for stabilization studies, the transfer free energy. Our calculation method is not a derivative of thermodynamic integration or thermodynamic perturbation methods and requires only a single trajectory to compute the transfer free energy of a protein into a weak-binding cosolvent system. Our results match experimental data well for glycerol and urea solutions, covering a range of positive and negative binding behavior.

This represents, to our knowledge, the first time such quantities have been calculated without adjustable parameters for proteins in mixed solvents. In addition to allowing rapid, direct computation of thermodynamic properties, the detail in the simulation employed here provides insights into the molecular-level origins of the observed free energy effects.

This work also led to the theoretical formulation of a simple integral equation theory that relates the preferential binding coefficient to the additive-protein and solvent-protein radial distribution functions.

6.2 The Gap Effect

Through the integral equation theory for solvent effects and the molecular-level insights gained in the simulation studies above, a simple framework for modeling protein association and dissociation reactions in the presence of solution additives was developed and analyzed. This model extends prior work in binding theory by considering various geometric models of the protein surface, the protein-protein association/dissociation transition states, and solvent radial distribution functions obtained from all-atom molecular dynamics simulations.

The model supports the hypothesis that a "gap effect," analogous to osmotic stress, will occur in association reactions when large solution additives with sufficient protein affinity are present. This gap effect may significantly perturb the free energy of protein-protein encounter complexes, such as the association transition state, and have only a small effect on the end states. Thus, we have demonstrated how it is possible for an additive to exert a purely kinetic effect on protein association/dissociation.

6.3 Neutral Crowders: A Class of Additives that Deter Aggregation

We call additives which slow association reactions via the gap effect without affecting the free energy of isolated protein molecules, "neutral crowders." Neutral crowders represent a potentially important class of anti-aggregation additive because they can affect protein-protein association reactions with only a small effect on protein folding. Thus they can significantly improve the selectivity toward proper folding versus aggregation in a refolding process. They can also deter aggregation from the native state when such aggregation exhibits second or higher-order kinetics. These molecules may be useful in other applications that involve association. Such applications are discussed briefly in Section 7.2 (Future Work).

6.4 Arginine is Preferentially-oriented

Arginine is a combination of two chemical moieties, the amino acid backbone and guanidinium, which have very different affinities for proteins. It was hypothesized that when combined into a single molecule, these differing affinities would give rise to a molecule that is preferentially-oriented when solvating proteins.

This assertion was validated by detailed molecular simulation. When in the first solvation shell around a protein, arginine preferentially orients itself with its guanidino group tending to face the protein. The free energy of this configuration is 0.19 kcal/mol lower than an orientation in which the guanidino group faces away

from the protein.

6.5 Arginine is a Neutral Crowder

Based on the theory of neutral crowders and available experimental data, it was postulated that arginine's propensity to prevent aggregation when used as a solution additive (0.2 - 0.5M) is due to the fact that it is a neutral crowder. Three experimental tests of this hypothesis designed to probe different aspects of neutral crowder behavior were performed.

Arginine was shown to slow protein association in two types of model systems when used as a solution additive. 0.5M arginine slowed association in two globular protein systems, insulin and myoglobin with their corresponding monoclonal antibodies, by a factor of 2.2 to 3.7. 0.5M arginine also promoted the refolding of carbonic anhydrase by a factor of 5.

When both arginine and guanidine are used as refolding buffer additives in refolding of carbonic anhydrase, 0.5M guanidine is superior to 0.5M arginine, promoting refolding by a factor of 8.3. Despite this difference, the average aggregate molecular weight is decreased in 0.5M arginine relative to 0.5M guanidinium. Because arginine affects association reactions primarily by decreasing the association rate, it can attenuate a downhill polymerization, such as that which is responsible for large aggregate formation during refolding. Guanidinium chloride, which acts primarily by increasing the dissociation rate, does not affect such processes.

Chapter 7

Future Work

The work of this thesis suggests several new areas for future research. These are summarized briefly in the following sections.

7.1 Design of Large Neutral Crowders

Our gap effect model predicts that if a neutral crowder significantly larger than arginine, perhaps 8Å in radius or larger, can be developed, it would depress association rates by orders of magnitude more than arginine. Such an additive could revolutionize protein stabilization. Because the mechanism of its action is general and its potency could be so large, a large neutral crowder could be used broadly.

As the size of a neutral crowder is increased, the gap effect becomes proportionately larger, but maintaining neutrality is difficult as size increases. At a constant protein-additive interaction energy, increasing additive size would decrease the protein-additive preferential binding coefficient as the third power of additive size due to an excluded volume effect. Thus, to make a large neutral crowder, additive-protein interactions must become significantly more attractive as size is increased. If this cannot be achieved, the gap effect will diminish and ultimately disappear.

One strategy to make additive-protein interactions more favorable is to put multiple protein binding groups such as guanidinium on the additive's surface. If these binding groups orient themselves such that they can simultaneously bind to the same protein molecule, an avidity effect leading to improved binding affinity will be created. This avidity strategy is used in nature to significantly increase the affinity of otherwise low affinity interactions [57].

7.2 Control of Other Association Processes, such as Crystallization and Adsorption

There may be other applications involving protein association or assembly where molecules akin to neutral crowders (or molecules with opposite properties, such that they accelerate association) can be further studied and utilized. Two such applications

are protein crystallization and protein adsorption.

Protein crystallization, much like stabilization of proteins against aggregation, involves protein association and solution additives, the selection of which is an empirical art. The gap effect hypothesis and the assertion that arginine is a neutral crowder suggest that additives such as arginine may be useful additives. Because of their ability to slow protein association reactions, they may allow only the most stable crystal form to grow, or affect the selectivity between nucleation and growth. Because of the central role of crystallization in protein structure studies (via x-ray crystallography) and separations (via selective crystallization), improvements in current heuristic approaches via increased mechanistic understanding are likely to be significant.

Similarly, protein adsorption on surfaces is an association reaction-driven process. Adsorption can be favorable, in the case of a separation such as chromatography, or unfavorable, as in the case of handling low concentrations of protein where losses to the surface are not desirable. A greater understanding of how solvent conditions affect these processes will be beneficial. One possible outcome might be the use of an additive such as a large neutral crowder when using proteins in microfluidic devices to minimize losses to the device's internal surfaces.

In either case, a more fundamental understanding of how solution conditions affect these processes will be beneficial. Further, they will provide experimental tests of the gap effect hypothesis in new areas.

7.3 Molecular Mechanism of the Hofmeister series

The Hofmeiseter lyotropic series is one of the earliest empirical observations of how ions affect protein stability, dating back to the late 1800s [41]. Despite continued study over the past 120 years, no one has been able to explain the Hofmeister series on the basis of fundamental, physical-chemical properties [17].

With the free energy simulation technique developed in this work, it should be possible to reproduce the thermodynamic effects of Hofmeister series ions on a model protein. Then, the molecular-level detail afforded by this all-atom, statistical-mechanical

model can be probed further to ascertain the molecular-level mechanism behind the thermodynamic effects seen.

7.4 Testing Gap Effect Theory with Model Compound Studies

We have introduced a new theory about how additives affect protein association reactions, called "gap effect" theory (chapter 4), that relates an additive's size and protein affinity to its effects on association and dissociation rates. This theory was tested experimentally via studies on guanidine and arginine (chapter 5) and computationally using a simple statistical-mechanical model (chapter 4).

Further experimental validation of gap effect theory could be performed by analyzing the association and dissociation rate effects of a homologous series of additives, using Biacore (as in chapter 5) or another suitable kinetic test. The additives selected must have only weak, nonspecific interactions with proteins. Also, some of the additives should be neutral or nearly-neutral crowders, so that a gap effect will be observable. Lastly, it would be ideal if only the additive's size (r_m, see section 4.1.2) or its interaction energy with proteins (ϵ or Γ_{XP}) varied within the series.

One homologous series which meets these requirements is the series of arginine analogs with different length methylene units between the amino acid alpha carbon and the guanidino group. Compounds are available with at least 1, 2, 3 (arginine), and 4 methylene groups.

Another possible strategy is to perform more detailed tests of gap effect theory using molecular dynamics simulation. This would augment the mean-field statistical-mechanical models used in chapter 4. Molecular dynamics simulations of proteins along an association reaction coordinate in different additives could be performed to compute the transfer free energy as done in chapter 3. The additives tested in this fashion could be those that are tested experimentally (as above), or model additives designed to probe specific aspects of gap effect theory in more detail, such as size

(r_m) and affinity (Γ_{XP}) effects.

7.5 Osmotic Stress Effects of Large Additives on Protein Folding Equilibria

Large additives increase the free energy of protein species that contain narrow gaps into which the large additive cannot fit but water can. Such free energy effects of pure steric origin, called "osmotic stress," have been shown to be important in conformational transitions of proteins [67, 24].

When an additive induces other free energy effects on an equilibrium, such as by binding to the states in equilibrium to different extents in regions other than the narrow gaps, osmotic stress effects can be difficult to decouple from these other effects. One ubiquitous protein equilibrium where both types of effects are likely to be present is the folding equilibrium between native state, unfolded state, and other intermediately folded states such as the molten globule state. Typical additives used in osmotic stress studies are large sugars and polyols such as sorbitol and sucrose. These classes of additives are also typically preferentially-excluded from the vicinity of proteins for steric reasons, and have negative preferential binding coefficients with proteins. Because the solvent accessible area of the species on the folding pathway are often quite disparate, the magnitude of this effect is not the same at all points along the pathway. Because a negative preferential binding is unfavorable, such additives shift the folding equilibrium toward the most compact (native) state.

Partially-unfolded states, such as the molten state, are likely to have a large number of small intramolecular gaps into which water can easily fit but a large additive such as sucrose will not. This perspective suggests that these additives may also shift the folding equilibrium toward the native state via an osmotic stress effect in addition to the "free surface effect" described above.

This hypothesis can be explored via molecular simulation as done in this work. It is likely that a useful conclusion can be reached with a simplified model, such as

a worm-like chain model, for the protein, and a simple solvent-protein interaction potential such as the Lennard-Jones or Exponential-6 potential. To separate the free surface and osmotic stress effects, ideal neutral crowder additives of varying sizes can be used. This should allow the unfavorable free energy of osmotic stress to be determined as a function of additive size and protein radius of gyration, or other appropriate order parameter for the folding-unfolding reaction coordinate.

7.6 Rational Additive Selection

The ability to predict thermodynamic and kinetic properties of proteins in mixed solvents without adjustable parameters opens the door to truly rational additive selection. A simulation protocol similar to the one developed in this work can be used to pre-screen additive candidates for a desired thermodynamic or kinetic effect, thus diminishing the number of combinations that need to be tested experimentally.

Further reduction of the computational complexity of the free energy calculation would be of significant benefit in this effort. In particular, a method that can achieve more rapid sampling of additive position space may significantly increase the size of a protein for which this approach is computationally practical.

Bibliography

[1] Darwin O. V. Alonso and Valerie Daggett. Molecular dynamics simluations of protein unfolding and limited refolding: Characterization of partially unfolded states of ubiquitin in 60methanol and in water. *J. Mol. Biol.*, 247:501–520, 1995.

[2] Charles F. Anderson and M. Thomas Record Jr. Salt dependence of oligoion-polyion binding: A thermodynamic description based on preferential interaction coefficients. *J. Phys. Chem.*, 97:7116–7126, 1993.

[3] T. Arakawa and S. N. Timasheff. The stabilization of proteins by osmolytes. *Biophys. J.*, 47:411–414, 1985.

[4] Tsutomu Arakawa and Serge N. Timasheff. Stabilization of protein structure by sugars. *Biochemistry*, 21:6536–6544, 1982.

[5] Tsutomu Arakawa and Serge N. Timashefff. Protein stabilization and destabilization by guanidinium salts. *Biochemistry*, 23:5924–5929, 1984.

[6] Tsutomu Arakawa and Kouhei Tsumoto. The effects of arginine on refolding of aggregated proteins: Not facilitate refolding, but suppress aggregation. *Biochem. Biophys. Res. Comm.*, 304:148–152, 2003.

[7] Neali Armstrong, Alexandre de Lencastre, and Eric Gouaux. A new protein folding screen: Application to the ligand binding domains of a glutamate and kainate receptor and to lysozyme and carbonic anhydrase. *Protein Sci.*, 8:1475–1483, 1999.

[8] Dipti Arora and Navin Khanna. Method for increasing the yield of properly folded recombinant human gamma interferon from inclusion bodies. *J. Biotechnol.*, 52:127–133, 1996.

[9] P. A. Bash, U. C. Singh, R. Langridge, and P. A. Kollman. Free energy calculations by computer simluation. *Science*, 236:564–569, 1987.

[10] Brian M. Baynes and Bernhardt L. Trout. Proteins in mixed solvents: A molecular-level perspective. *J. Phys. Chem. B*, 107:14058–14067, 2003.

[11] Brian M. Baynes and Bernhardt L. Trout. Rational design of solution additives for the preventing of protein aggregation. *Biophys. J.*, 87:In Press, 2004.

[12] Brian J. Bennion and Valerie Daggett. The molecular basis for the chemical denaturation of proteins by urea. *PNAS*, 100:5142–5147, 2003.

[13] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliand, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Res.*, 28:235–242, 2000.

[14] D. Wayne Bolen. *Protein Stabilizaiton by Naturally Occurring Osmolytes.* Protein Structure, Stability, and Folding, 2001.

[15] Bernard R. Brooks, Robert E. Bruccoleri, Barry D. Olafson, David J. States, W. Swaminathan, and Martin Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.*, 4:187–217, 1983.

[16] Johannes Buchner and Rainer Rudolph. Renaturation, purification and characterization of recombinant fab-fragments produced in escherichia coli. *Biotechnology*, 9:157–162, 1991.

[17] M. G. Cacace, E. M. Landau, and J. J. Ramsden. The hofmeister series: Salt and solvent effects on interfacial phenomena. *Q. Rev. Biophys.*, 30:241–277, 1997.

[18] Amedeo Caflisch and Martin Karplus. Structural details of urea binding to barnase: A molecular dynamics analysis. *Struct. Fold. Des.*, 7:477–488, 1999.

[19] Edward F. Casassa and Henryk Eisenberg. Thermodynamic analysis of multicomponent solutions. *Adv. Protein Chem.*, 19:287–395, 1964.

[20] Jeffrey L. Cleland. *Mechanisms of Protein Aggregation and Refolding.* PhD thesis, MIT, 1991.

[21] Jeffrey L. Cleland, Chester Hedgepeth, and Daniel I. C. Wang. Polyethylene glycol enhanced refolding of bovine carbonic anhydrase b. *J. Biol. Chem.*, 267:13327–13334, 1992.

[22] Jeffrey L. Cleland, Michael F. Powell, and Steven J. Shire. The development of stable protein formulations: A close look at protein aggregation, deamidation, and oxidation. *Crit. Rev. Ther. Drug Carrier Systems*, 10:307–377, 1993.

[23] Jeffrey L. Cleland and Daniel I. C. Wang. Refolding and aggregation of bovine carbonic anhydrase b: Quasi-elastic light scattering analysis. *Biochemistry*, 29:11072–11078, 1990.

[24] Marcio F. Colombo, Donald C. Rau, and Adrian Parsegian. Protein solvation in allosteric regulation: A water effect on hemoglobin. *Science*, 256:655–659, 1992.

[25] E. S. Courtenay, M. W. Capp, C. F. Anderson, and M. T. Record Jr. Vapor pressure osmometry studies of osmolyte-protein interactions: Implications for the action of osmoprotectants in vivo and for the interpretation of osmotic stress experiments in vitro. *Biochemistry*, 39:4455–4471, 2000.

[26] Ken A. Dill. Dominant forces in protein folding. *Biochemistry*, 29:7133–7155, 1990.

[27] Yong Duan and Peter A. Kollman. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 282:740–744, 1998.

[28] Erin M. Duffy, Daniel L. Severance, and William L. Jorgensen. Urea: Potential functions, log p, and free energy of hydration. *Israel J. Chem.*, 33:323–330, 1993.

[29] Mirelle Dumoulin, Alexander M. Last, Allne Desmyer, Klaas Decanniere, Denis Canet, Goran Larsson, Andrew Spencer, David B. Archer, Jurgen Sasse, Serge Muyldermans, Lode Wyns, Christina Redfield, Andre Matagne, Carol V. Robinson, and Chrisopher M. Dobson. A camelid antibody fragment inhibits the formation of amyloid fibrils by human lysozyme. *Nature*, 424:783–788, 2003.

[30] Stephen P. Edgecomb and Kenneth P. Murphy. Variability in the pka of histidine side-chains correlates with burial within proteins. *Proteins*, 49:1–6, 2002.

[31] William R. Forsyth, Jan M. Antosiewicz, and Andrew D. Robertson. Empirical relationships between protein structure and carboxyl pka values in proteins. *Proteins*, 48:388–403, 2002.

[32] Kunihiko Gekko and Serge N. Timasheff. Mechanism of protein stabilization by glycerol: Preferential hydration in glycerol-water mixtures. *Biochemistry*, 20:4667–4676, 1981.

[33] Kunihiko Gekko and Serge N. Timasheff. Thermodynamic and kinetic examination of protein stabilization by glycerol. *Biochemistry*, 20:4677–4686, 1981.

[34] Julius A. Gordon. Denaturation of globular proteins. interaction of guanidinium salts with three proteins. *Biochemistry*, 11:1862–1870, 1972.

[35] Raymond F. Greene Jr. and C. Nick Pace. Urea and guanidine hydrochloride denaturation of ribonuclease, lysozyme, alpha-chymotrypsin, and beta-lactoglobulin. *J. Biol. Chem.*, 249:5388–5393, 1974.

[36] Sookhee N. Ha, Ann Giammona, Martin Field, and John W. Brady. A revised potential-energy surface for molecular mechanics studies of carbohydrates. *Carbohydrate Res.*, 180:207–221, 1988.

[37] Tanford Charles Hade, E. P. Kirby. Isopiestic compositions as a measure of preferential interactions of macromolecules in two-component solvents. application to proteins in concentrated aqueous cesium chloride and guanidine hydrochloride. *J. Am. Chem. Soc.*, 89:5034–5040, 1967.

[38] F. Ulrich Hartl and Manajit Hayer-Hartl. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science*, 295:1852–1858, 2002.

[39] John E. Hearst. Determination of the domainant factors which influence the net hydration of native sodium deoxyribonuclease. *Biopolymers*, 3:57–68, 1965.

[40] Diane L. Hevehan and Eliana De Bernardez Clark. Oxidative renaturation of lysozyme at high concentrations. *Biotechnol. Bioeng.*, 54:221–230, 1997.

[41] F. Hofmeister. Zur lehre von der wirkung der salze. ii. *Arch. Exp. Pathol. Pharmakol.*, 24:247–260, 1888.

[42] W. Humphrey, A. Dalke, and K. Schulten. Vmd - visual molecular dynamics. *J. Molec. Graphics*, 14:33–38, 1996.

[43] Amy R. Hurshman, Joleen T. White, Evan T. Powers, and Jeffrey W. Kelly. Transthyretin aggregation under partially denaturing conditions is a downhill polymerization. *Biochemistry*, 43:7365–7381, 2004.

[44] William L. Jorgensen, Jayaraman Chandrasekhar, Jeffrey D. Madura, Roger W. Impey, and Michael L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926–935, 1983.

[45] Martin Karplus and J. Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nature. Struct. Biol.*, 9:646–652, 2002.

[46] Brent S. Kendrick, John F. Carpenter, Jeffrey L. Cleland, and Theodore W. Randolph. A transient expansion of the native state precedes aggregation of recombinant human interferon-gamma. *PNAS USA*, 95:14142–14146, 1998.

[47] John G. Kirkwood and Richard J. Goldberg. Light scattering arising from composition fluctuations in multi-component systems. *J. Chem. Phys.*, 18:54–57, 1950.

[48] Peter Kollman. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.*, 93:2395–2417, 1993.

[49] Danuta Kosk-Kosicka, Maria M. Lopez, Ioulia Fomitcheva, and Virgilio L. Lew. Self-association of plasma membrane ca2+-atpase by volume exclusion. *FEBS Letters*, 371:57–60, 1995.

[50] James C. Lee and Serge N. Timasheff. Partial specific volumes and interactions with solvent components of proteins in guanidine hydrochloride. *Biochemistry*, 13:257–265, 1974.

[51] James C. Lee and Serge N. Timasheff. The stabilization of proteins by sucrose. *J. Biol. Chem.*, 256:7193–7201, 1981.

[52] Hauke Lilie, Elisabeth Schwarz, and Rainer Rudolph. Advances in refolding of proteins produced in e. coli. *Curr. Opin. Biotech.*, 9:497–501, 1998.

[53] T. Y. Lin and S. N. Timasheff. Why do some organisms use a urea-methylamine mixture as osmolyte- thermodynamic compensation of urea and trimethylamine n-oxide interactions with protein. *Biochemistry*, 33:12695–12701, 1994.

[54] Robyn Linder and Gregory Ralston. Effects of dextran on the self-association of human spectrin. *Biophys. Chem.*, 57:15–25, 1995.

[55] Y. F. Liu and D. W. Bolen. The peptide backbone plays a dominant role in protein stabilization by naturally-occurring osmolytes. *Biochemistry*, 34:12884–12891, 1995.

[56] R. Lumry and H. Eyring. Conformation changes of proteins. *J. Phys. Chem.*, 58:110–120, 1954.

[57] Mathai Mammen, Seok-Ki Choi, and George M. Whitesides. Polyvalent interactions in biological systems: Implications for design and use of multivalent ligands and inhibitors. *Angew. Chem. Int. Ed.*, 37:2754–2794, 1998.

[58] William G. McMillan and Joseph E. Mayer. The statistical thermodynamics of multicomponent systems. *J. Chem. Phys.*, 13:276–305, 1945.

[59] Donald A. McQuarrie. *Statistical Mechanics.* HarperCollins Publishers, 1973.

[60] Pamela Mills, Charles F. Anderson, and M. Thomas Record Jr. Grand canonical monte carlo calculations of thermodynamic coefficients for a primitive model of dna-salt solutions. *J. Phys. Chem.*, 90:6541–6548, 1986.

[61] Hans Neurath, Jesse P. Greenstein, Frank W. Putnam, and John A. Erikson. The chemistry of protein denaturation. *Chem. Rev.*, 34:157–265, 1944.

[62] Lawrence W. Nichol, Alexander G. Ogston, and Peter R. Willis. Effect of inert polymers on protein self-association. *FEBS Letters*, 126:18–20, 1981.

[63] L. Ninni, M. S. Camargo, and Antonio J. A. Meirelles. Water activity in polyol systems. *J. Chem. Eng. Data*, 45:654–660, 2000.

[64] Y. Pocker and B. Biswas, Subhasis. Self-association of insulin and the role of hydrophobic bonding: A thermodynamic model of insulin dimerization. *Biochemistry*, 20:4354–4361, 1981.

[65] Y. Pocker and J. T. Stone. The catalytic versatility of erythrocyte carbonic anhydrase. iii. kinetic studies of the enzyme-catalyzed hydrolysis of p-nitrophenyl acetate. *Biochemistry*, 6:668–678, 1967.

[66] Natasa Poklar, Nina Petrovcic, Miha Oblak, and Gorazd Vesnaver. Thermodynamic stability of ribonuclease a in alkylurea solutions and preferential solvation changes accompanying its thermal denaturation: A calorimetric and spectroscopic study. *Protein Sci.*, 8:832–840, 1999.

[67] R. P. Rand. Raising water to new heights. *Science*, 256:618, 1992.

[68] M. Thomas Record Jr., Wentao Zhang, and Charles F. Anderson. Analysis of effects of salts and uncharged solutes on protein and nucleic acid equilibria and processes: A practical guide to recognizing and interpreting polyelectrolyte effects, hofmeister effects, and osmotic effects of salts. *Adv. Protein Chem.*, 51:281–353, 1998.

[69] T. J. Richmond. Solvent accessible surface-area and excluded volume in proteins-analytical equations for overlapping spheres and implications for the hydrophobic effect. *J. Mol. Biol.*, 178:63–89, 1984.

[70] Ursula Rinas, Bernhard Risse, Rainer Jaenicke, Karl-Josef Abel, and Gerd Zettlmeissl. Denaturation-renaturation of the fibrin-stabilizing factor xiii a-chain isolated from human placenta. *Biol. Chem. Hoppe-Seyler*, 371:49–56, 1990.

[71] Christopher J. Roberts. Kinetics of irreversible protein aggregation: Analysis of extended lumry eyring models and implications for predicting protein shelf life. *J. Phys. Chem. B*, 107:1194–1207, 2003.

[72] Christopher J. Roberts, Richard T. Darrington, and Maureen B. Whitley. Irreversible aggregation of recombinant bovine granulocyte-colony stimulating factor (bg-csf) and implications for predicting protein shelf life. *J. Pharm. Sci.*, 92:1095–1111, 2003.

[73] R. Rudolph, S. Fischer, and R. Mattes. *Patent DE3537708, Process for the Activation of tPA after Expression in Prokaryotic Cells.* 1985.

[74] George Scatchard, W. J. Hamer, and S. E. Wood. Isotonic solutions. i. the chemical potential of water in aqueous solutions of sodium chloride, potassium chloride, sulfuric acid, sucrose, urea and glycerol at 25 deg. *J. Am. Chem. Soc.*, 60:3061–3070, 1938.

[75] John A. Schellman. Solvent denaturation. *Biopolymers*, 17:1305–1322, 1978.

[76] John A. Schellman. Fifty years of solvent denaturation. *Biophys. Chem.*, 96:91–101, 2002.

[77] John A. Schellman. Protein stability in mixed solvents: A balance of contact interaction and excluded volume. *Biophys. J.*, 85:108–125, 2003.

[78] Gideon Schreiber. Kinetic studies of protein-protein interactions. *Curr. Opin. Struct. Biol.*, 12:41–47, 2002.

[79] Tzvia Selzer and Gideon Schreiber. New insights into the mechanism of protein-protein association. *Proteins*, 45:190–198, 2001.

[80] G. V. Semisotnov, V. N. Uversky, I. V. Sokolovsky, A. M. Gutin, O. I. Razgulyaev, and N. A. Rodionova. 2 slow stages in refolding of bovine carbonic anhydrase-b are due to proline isomerization. *J. Mol. Biol.*, 213:561–568, 1990.

[81] Gennady Semisotnov, Natalya A. Rodionova, Victor P. Kutyshenko, Bernd Ebert, Jurgen Blanck, and Oleg B. Ptitsyn. Sequential mechanism of refolding of carbonic anhydrase b. *FEBS Letters*, 224:13–Sep, 1987.

[82] Kentaro Shiraki, Monotori Kudou, Shinsuke Fujiwara, Tadayuki Imanaka, and Masahiro Takagi. Biophysical effect of amino acids on the prevention of protein aggregation. *J. Biochem.*, 132:591–595, 2002.

[83] Superna Taneja and Faizan Ahmad. Increased thermal stability of proteins in the presence of amino acids. *Biochem. J.*, 303:147–153, 1994.

[84] Charles Tanford. Isothermal unfolding of globular proteins in aqueous urea solutions. *J. Am. Chem. Soc.*, 86:2050–2059, 1964.

[85] Karen E. S. Tang and Victor A. Bloomfield. Assessing accumulated solvent near a macromolecular solute by preferential interaction coefficients. *Biophys. J.*, 82:2876–2991, 2002.

[86] Julian Tirado-Rives, Modesto Orozco, and William L. Jorgensen. Molecular dynamics simulations of the unfolding of barnase in water and 8 m aqueous urea. *Biochemistry*, 36:7313–7329, 1997.

[87] Wei Wang. Instability, stabilization, and formulation of liquid protein pharmaceuticals. *Int. J. Pharm.*, 185:129–188, 1999.

[88] Jonathan N. Webb, Serena D. Webb, Jeffrey L. Cleland, John F. Carpenter, and Theodore W. Randolph. Partial molar volume, surface area, and hydration changes for equilibrium unfolding and formation of aggregate transition state: High-pressure and cosolute studies on recombinant human ifn-gamma. *PNAS*, 98:7259–7264, 2001.

[89] D. B. Wetlaufer and Y. Xie. Control of aggregation in protein refolding: A variety of surfactants promote renaturation of carbonic anhydrase ii. *Protein Sci.*, 4:1535–1543, 1995.

[90] Jeffries Wyman and S. J. Gill. *Binding and Linkage: Functional Chemistry of Biological Macromolecules.* University Science Books, 1990.

[91] Xie and Timasheff. The thermodynamic mechanism of protein stabilization by trehalose. *Biophys. Chem.*, 64:25–43, 1997.

[92] Guifu Xie and Serge N. Timasheff. Mechanism of the stabilization of ribonuclease a by sorbitol: Preferential hydration is greater for the denatured chain than for the native protein. *Protein Sci.*, 6:211–221, 1997.

[93] Guifu Xie and Serge N. Timasheff. Temperature dependence of the preferential interactions of ribonuclease a in aqueous co-solvent systems: Thermodynamic analysis. *Protein Sci.*, 6:222–232, 1997.

[94] Gerd Zettlmeissl, Rainer Rudolph, and Rainer Jaenicke. Reconstitution of lactic dehydrogenase. noncovalent aggregation vs. reactivation. 1. physical properties and kinetics of aggregation. *Biochemistry*, 18:5567–5571, 1979.

[95] Qin Zou, Brian J. Bennion, Valerie Daggett, and Kenneth P. Murphy. The molecular mechanism of stabilization of proteins by tmao and its ability to counteract the effecfts of urea. *J. Am. Chem. Soc.*, 124:1192–1202, 2002.