The Dissertation Committee for Anthony Manson Certifies that this is the approved version of the following dissertation:

The Impact Of Protein Fluctuations On Molecular Recognition

Committee:

Dr. Vince Hilser PhD., Supervisor

Dr. Wlodek Bujalowski PhD., Co-Supervisor

Dr. Werner Braun PhD.

Dr. B. Montgomery Pettitt PhD.

Dr. Mary Moslen PhD.

Dean, Graduate School

The Impact Of Protein Fluctuations On Molecular Recognition

by

Anthony Manson, BSME

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas Medical Branch

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas Medical Branch

December, 2008

Dedication

This work is dedicated to my daughter Sarah.

Acknowledgements

I would like to acknowledge my dissertation committee and the members of the Hilser lab for their inspiration and encouragement during the course of this work.

The Impact Of Protein Fluctuations On Molecular Recognition

Publication No._____

Anthony Manson, PhD

The University of Texas Medical Branch, 2008

Supervisor: Dr. Vincent Hilser

The effect of protein fluctuations on molecular recognition is poorly understood. Prediction of useful properties such as binding affinity using rigid structures has produced sporadic success. Although attempts have been made to model the effect of conformational fluctuations, capturing the impact of backbone relaxation has remained particularly elusive. In order to investigate these effects, a series of surface exposed Ala/Gly mutants were designed in the flexible RT loop of the C-terminal SH3 domain of SEM5. One set of mutations was designed to perturb the ensemble of accessible conformations in the unbound ensemble while leaving the interaction surface with the ligand unchanged. The other set was designed to perturb both the interaction surface as well as the ensembles of bound and free conformations. The effects of these mutations were investigated by generating random conformations of the RT loop and performing principal component analysis to organize the randomly generated conformational states into a coherent landscape. To predict the effect of these mutations, we developed a statistical mechanical technique using a simplified energy function that only applied the

effects of excluded volume and implicit solvation. This energy function was utilized to weight an ensemble of conformational states from which aggregate thermodynamic properties could be derived. The computed effects of the mutations on the binding affinity agreed with experimentally determined values (R= 0.97) from isothermal titration calorimetry. The results indicate that the bound state of SEM5 SH3 domain contains a considerable repertoire of conformational variants of the high-resolution structure and that the determinants of binding cannot be elucidated from the static structure of the bound complex.

Table of Contents

List of Tables

List of Figures

THE IMPACT OF PROTEIN FLUCTUATIONS ON MOLECULAR RECOGNITION

Introduction

STATEMENT OF PROBLEM

Proteins are dynamic molecules that often undergo conformational changes while performing their specific functions, such as an enzyme reaction or ligand binding. The dynamic properties intrinsic to a protein structure may provide information on the location and the energetics of the conformational change process, and are thus the focus of many biophysical studies.

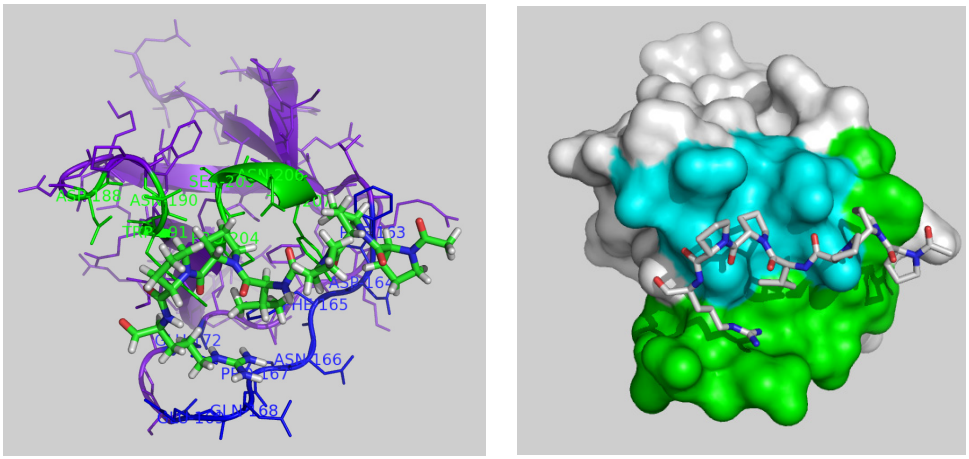


FIGURE 1: SUMMARY OF PROBLEM SYSTEM: Left: Composite view of SH3 domain with ligand. Right: Surface view of C-SH3 modular binding domain. Cyan region is rigid and hydrophobic. Green region is a mix of hydrophobic and hydrophilic residues and exhibits a high degree of conformational heterogeneity.

The effect of protein fluctuations on molecular recognition is poorly understood. Although attempts have been made to model the effect of fluctuations, capturing the impact of backbone relaxation has remained particularly elusive.

The objective of this project was to develop a model that would link the microscopic aspects of an all-atom molecular model to the bulk thermodynamic measurables and use this model to quantitatively predict measured trends in binding affinity.

BACKGROUND

SH3 Modular Binding Domain

The SH3 domain is probably the most widespread protein recognition module in the proteome and more than 1500 different SH3 domains (Mayer 2001) can be identified by search algorithms in protein databases FIGURE 1. It is found in proteins that have been implicated in signal transduction, cytoskeleton organization and membrane trafficking. All SH3 domains share a highly conserved fold that can be represented as a sandwich formed by two three-stranded β-sheets (Weng, Thomas et al. 1994; Weng, Rickles et al. 1995; Tsai, Levitt et al. 1999). One side of the sandwich is hydrophobic and constitutes the ligand binding surface.

Functional Contexts

The Sh3 modular binding domain occurs frequently in roles related to signal transduction where the SH3 domain acts as an adaptor. An **adaptor protein** is a protein which is accessory to main proteins in a signal transduction pathway FIGURE 2. These proteins tend to lack any intrinsic enzymatic activity themselves but instead mediate specific protein-protein interactions that drive the formation of protein complexes.

Examples of adaptor proteins include MyD88, Grb2 and Shc. Adaptor proteins usually contain several domains within their structure (e.g., Src homology 2 (SH2) and SH3 domains) which allow specific interactions with several other specific proteins. SH2 domains recognize specific amino acid sequences within proteins containing phosphotyrosine residues and SH3 domains recognize proline-rich sequences within specific peptide sequence contexts of proteins.

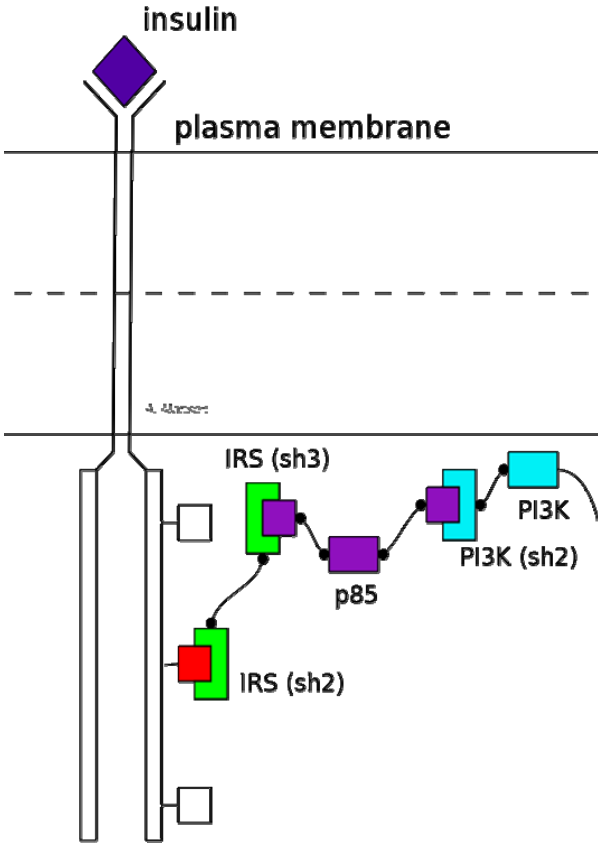


FIGURE 2: FUNCTIONAL CONTEXT OF THE SH3 DOMAIN: adapter protein within signaling cascade. The SH3 domain transduces a primary signal (insulin) and allows amplification of a secondary signal I3P.

There are many other types of interaction domains found within adaptor and other signaling proteins which allow a rich diversity of specific and coordinated protein-protein interactions to occur within the cell during signal transduction.

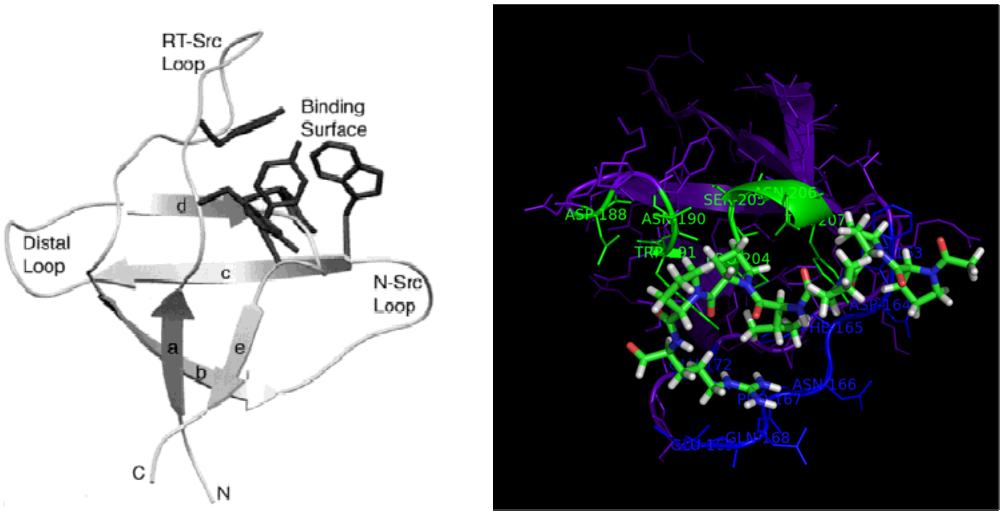
Proline Recognition Domains

Proline-rich sequences are widely distributed in distinct proteomes from prokaryotes to eukaryotes (Yu, Chen et al. 1994; Li 2005). For example, *Drosophila* is estimated to harbor 579 proline-rich regions, making them the most abundant sequence pattern in its proteome (Yu, Chen et al. 1994). Together with their binding proteins, proline-rich sequences play an indispensable role in mediating a multitude of protein–protein interactions that are essential for a host of cellular processes (Nguyen, Turck et al. 1998). Why are proline-rich motifs favoured in a cell? The answer to this intriguing question appears to lie within proline itself. First, for a peptide sequence to function in a binary interaction, it has to be exposed to the solvent and be accessible to the binding partner. Of the 20 naturally occurring amino acids, proline may be best suited for such a role. It is a well known breaker of regular secondary structures such as α-helices and β-sheets that are essential for protein folding and topology. Consequently, proline-containing sequences are often found on the surface of a protein, as opposed to being buried within the core (Li 2005). Secondly, the closure of the side chain of proline in a five-member ring restricts one of its dihedral angles, φ, at approx. -60°. This severely restrains the types of conformation that proline and proline-rich sequences can adopt. The most common structure formed by two or more proline residues in a row is PPII (polyproline type II), a left-handed helix with three residues per turn. This structure is more relaxed than an ideal α-helix that has a pitch of 3.6 residues. Thirdly, the PPII conformation can arise automatically from a stretch of proline residues of sufficient

length (Ferreon and Hilser 2003). It is believed that restraints in the side chain of proline and, in some cases, the pre-formed structure would significantly reduce the entropic cost associated with binding of a proline-rich sequence. Fourthly, since both the side chains and the backbone carbonyls in a PPII structure are projected outwards from the axis of the helix, they are poised to interact with another molecule. Moreover, the lack of an amide proton in proline to participate in intramolecular hydrogen bonding frees its carbonyl group for intermolecular interactions (Garrett and Grisham 1999). Finally, the PPII structure is stable and resistant, to a large extent, to amino acid substitutions. Therefore various combinations of non-proline and proline residues can be incorporated into a peptide sequence without compromising the integrity of the PPII structural frame. This unique property of the PPII helix might have played an important part in the evolution of modular domains that bind to proline-rich sequences.

Structural Overview

Comprised of approximately 60 residues, the SH3 domain fold is composed of five **β**-strands arranged into two sheets packed at right angles FIGURE 3. The first sheet is formed by **β**-strands *a*, *e*, and the first half of *b*, while the second is formed by **β**-strands *c*, *d*, and the second half of *b*. A kink in **β**-strand *b* allows it to participate in both sheets. **β**-Strands *a* and *b* are separated by the long RT-Src loop, which possesses an irregular antiparallel structure. The shorter N-Src and Distal loops are found between **β**-strands *b* and *c*, and *c* and *d*, respectively. The four residues N-terminal to **β**-strand *b* form a type II **β**-turn, and the three residues separating **β**-strands *d* and *e* are generally found in a 3_{10}-helical conformation. Peptide binding by the SH3 domain is mediated by a surface region rich in aromatic residues FIGURE 1, and by various polar residues located in the RT-Src and N-Src loops.



(Mayer 2001)

FIGURE 3: CARTOON DIAGRAM OF THE SH3 DOMAIN from the Fyn tyrosine kinase (1shf). Left: The β-strands are labeled a–e, and the loops are designated. Aromatic side chains involved in peptide binding are shown. Right: In the presence of the ligand.

Residue Conservation

The SH3 domain provides an excellent system for the examination of sequence and structural conservation. Hundreds of very diverse SH3 domain sequences are available for analysis (Mayer 2001), providing a broad sampling of sequences that are consistent with the SH3 domain fold. In addition, 44 structures of 19 different SH3 domains are available in the structural database.

From the analysis of side-chain burial described above, 15 positions were identified as playing a significant role in **target peptide binding**. Among these positions, 51(P) and 36(W) are the two most conserved positions in the SH3 domain alignment and five more, 8(Y), 10(Y), 35(G), 53(N), and 54(Y), possess positional entropy values below 7.5. At some positions, the increase in residue burial in the

presence of bound ligand varies widely from structure to structure. However, there is considerably less variability in the increase in burial seen at the seven most conserved positions involved in ligand interaction. Thus, the most conserved positions that contact ligand are seen to be consistently important in ligand binding in all the solved structures. On the other hand, positions that contact ligand that are not highly conserved in the sequence alignment [13(R), 14(E), 15(D), 16(E), 33(D), 34(D), and 49(L)] also vary a great deal more in their degree of burial upon ligand binding in different structures. The sequence and structural variation at these positions suggest that they may play a greater role in defining target specificity, as their importance in the binding reaction is dependent on the combination of SH3 domain and target peptide under examination.

The **hydrophobic core** of a protein is comprised of a group of hydrophobic side chains that are predominantly inaccessible to solvent due to their close packing in the interior of the protein FIGURE 4. Because the hydrophobic core is critical for protein stability [for review, see Dill, 1990], hydrophobic core positions in an alignment are generally highly conserved and are mostly occupied by hydrophobic residues (Bashford, Chothia et al. 1987). In the SH3 domain alignment, there are nine positions [4(V), 10(Y), 18(L), 20(F), 26(I), 28(V), 37(W), 50(F) and 55(V)] that are on average 0.85% buried and have an average hydrophobicity of 0.1. Analysis of the SH3 domain structures indicates that all of these positions except position 10(Y) are indeed hydrophobic core residues. Position 10(Y) is not considered to be a hydrophobic core residue because it contacts ligand and is not packed closely with most of the other core residues. Although they do not have high average hydrophobicities, the 6(A) and 39(G) positions are designated as hydrophobic core positions. They are 100% buried in all SH3 domain structures, they are almost never occupied by polar residues, and their side chains make extensive contacts with other hydrophobic core side chains. Although the 41(L) position

is highly buried and hydrophobic in a number of SH3 domains, it cannot be considered as part of the conserved hydrophobic core of the domain because it is one of the least conserved positions in our alignment and is often occupied by polar residues. Furthermore, a variety of amino acid substitutions at this position in the Fyn SH3 domain was found to have only small effects on stability and function (Maxwell and Davidson 1998).

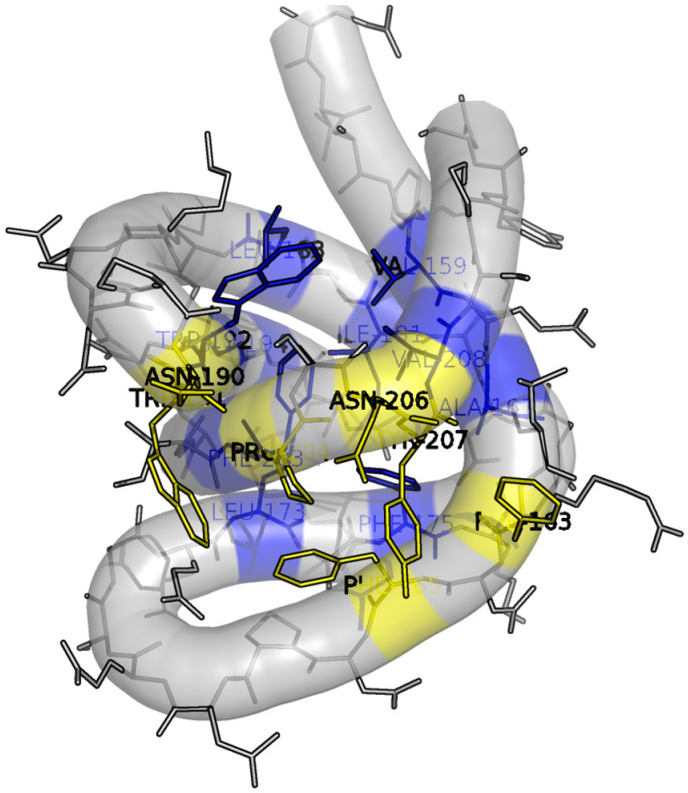
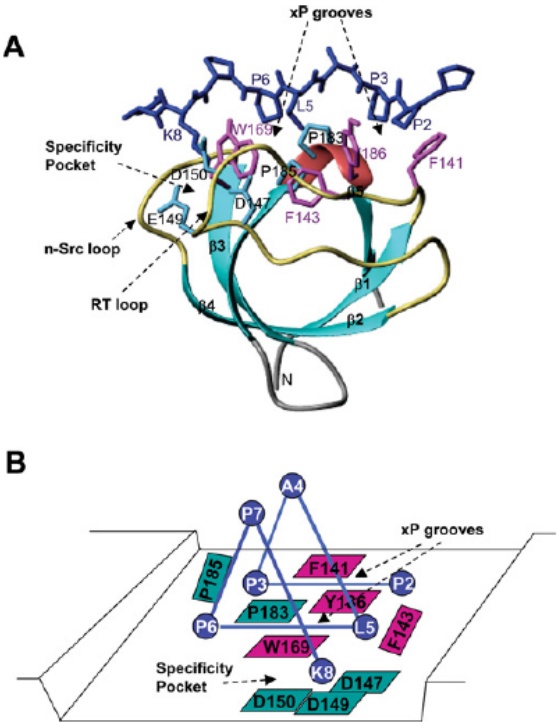


FIGURE 4: CONSERVED RESIDUES OF C-SH3: Conserved positions relevant to stability (blue) and relevant to binding the putative ligand (yellow). The binding residues are primarily hydrophobic **(F163, F165, W191, Y207)**

Binding Mechanism and Interface Structure

The discovery that most SH3 ligands are rich in prolines and the analysis of several structures of SH3 domains complexed with their peptide ligands (Mayer and Gupta 1998; Mayer 2001; Li 2005) led to the formulation of a general SH3-peptide binding model (Nguyen, Turck et al. 1998; Cesareni, Panni et al. 2002). SH3 ligands contain two XP dipeptides, separated by a scaffolding residue (often a proline). The two XP moieties in the core (XP-x-XP) motif occupy two hydrophobic pockets formed by residues that are conserved in most SH3 domains. The third binding pocket is lined by negative residues and can host a positively charged side chain flanking the core motif. SH3 ligands bind to their receptors in a left-handed polyproline type II (PPII) helical conformation in either of two opposite orientations depending on the position of a positive residue in the peptide sequence FIGURE 5. Peptides that bind in a type I orientation conform to the consensus RxLPP#P (where # is normally a hydrophobic residue), while peptides that are characterized by Px#PxR (type II) bind in the opposite orientation. The SH3 domain of the protein kinase Abl binds to ligands that have a tyrosine (or a large hydrophobic residue) in place of the positively charged side chain at position P3 of class I peptides (for residue nomenclature see Fig. 1). This model has served as a framework in the interpretation of SH3 binding experiments and in the identification of SH3 peptide targets on newly discovered proteins.

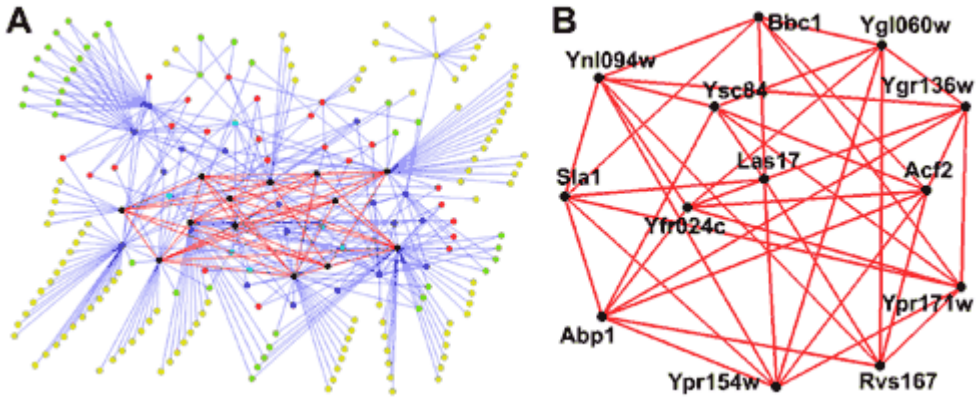


(Lim, Richards et al. 1994)

FIGURE 5: STRUCTURE OF THE CRK SH3-N DOMAIN IN COMPLEX WITH A HIGH-AFFINITY PEPTIDE FROM C3G [62] (A) The structure shown is based on PDB accession code 1CKB. The SH3 domain is depicted in ribbons with secondary-structural elements shown in different colors and labeled. The bound peptide, PPPALPPKKR, is shown in blue with side chains. Interface residues on the SH3 domain are shown in pink, for aromatic residues, and in light blue, for non-aromatic residues. The locations of the xP grooves and specificity pocket on the SH3 domain are identified by broken arrows. (B) A schematic representation of the same structure to highlight the characteristics of the ligand-binding surface on the SH3 domain such as the enrichment of aromatic residues. The same coloring scheme is used in (A) and (B) for purposes of comparison.

Metabolic and Regulatory Pathways

Protein Interaction Networks



(Li 2005)

FIGURE 6: YEAST SH3 DOMAIN PROTEIN-PROTEIN INTERACTION NETWORK predicted using phage display selected peptides. 394 interactions and 206 proteins are shown; a network with each gene name labeled is included in the supplementary material (7). The proteins are colored according to their k-core value (6-core = black, 5-core = cyan, 4-core = blue, 3-core = red, 2- core = green, 1-core = yellow), identifying subsets of interconnected proteins in which each protein has at least k interactions. By definition, lower core numbers encompass all higher core numbers (e.g. a 4-core includes all the nodes in the 4-core, 5-core and 6-core). The interactions of the 6-core subgraph are highlighted in red. (B) The 6-core subgraph derived from the phage display protein-protein interaction network, expanded to allow identification of individual proteins.

RELEVANCE

Drug and Gene Therapy

Gene Therapy is a field of medicine in which genes are introduced into the body to cure diseases. It involves: 1) Detection of a gene, 2) Determination of its role, 3) Its isolation and cloning, 4) Properly introducing the gene. There are two types; namely, germline gene therapy [done in germ cells] and somatic gene therapy [done in somatic cells].

Adapter proteins such as SH3 are good candidates for gene therapy due to their well understood role in signaling cascades. Subtle changes to its binding affinity can manifest significant changes in cellular behavior. Select mutants can be introduced that would have a stabilizing effect on a cellular system.

Understanding the detailed physical processes during binding can aid in designing drugs in the treatment of signaling diseases such as cancer. Designing effective inhibitors requires sufficient dynamic information about the binding process to estimate the thermodynamic impact of a drug.

MOLECULAR MODELING

Justification for All Atom Models

In the scientific method, an **experiment** (Latin: *ex- periri*, "of (or *from*) trying") is a set of observations performed in the context of solving a particular problem or question, to retain or falsify a hypothesis or research concerning phenomena. The experiment is a cornerstone in the empirical approach to acquiring deeper knowledge about the physical world.

The nature of many processes such as protein aggregation (limited structural order, insolubility in water, and involvement of cell membrane) renders its experimental study extremely difficult. Traditionally, one must know the positions and momenta of all particles within a system at a short time resolution to gain the insights needed to support prediction. This requirement has generally proved difficult and due to the uncertainty principle may well prove impossible in a wide number of cases. Thus, one needs to bridge this gap using logical [theoretical] methods. Such methods provide a rational framework in which empirical observations can be interpreted. Here the credibility of such an approach will depend on the accuracy of the rules and facts used to construct this logical framework.

Computation has the unbeatable edge in that it can describe protein dynamics completely: the precise position of each atom at any instant in time for a single protein molecule can be followed, along with the corresponding energies, provided that at least one high-resolution structure is known as a starting point. Although conformational substates (located in energy wells) and their rates of interconversion can be detected experimentally (as described earlier), an atomic-resolution structural description of the 'climb from one valley to another' (the transition pathway) is out of experimental reach, owing to the extremely low probability and short lifetime of the high-energy conformers. Computational methods would be able to overcome these limitations if a perfect description of the protein–solvent system could be provided by the force field (that is, parameter sets describing the potential energy of all atoms). Impressive progress has been made in the development of these force fields since their original conception, and they are used in molecular-dynamics simulations.

Unfortunately, protein dynamics on the microsecond-to-millisecond timescale is currently out of reach for conventional molecular-dynamics simulations. To overcome

this restriction, a large variety of approaches that simplify force fields have been developed, including normal mode analysis (Karplus and Kushick 1981; Ma and Karplus 1997), gaussian network models (Haliloglu, Bahar et al. 1997), FIRST (floppy inclusion and rigid substructure topography) (Jacobs, Rader et al. 2001), FRODA (framework rigidity optimized dynamic algorithm) (Wells, Menor et al. 2005) and Gō models (Scheraga, Khalili et al. 2007). Alternatively, the dynamic process is accelerated by external force to access this timescale (used in methods such as targeted, steered and accelerated molecular-dynamics simulations (Hamelberg, Mongan et al. 2004)), or prior knowledge about features of the reaction coordinate (umbrella sampling algorithms to construct a potential of mean force (Roux 1995)) or the transition end points (transition-path sampling (Dellago and Bolhuis 2007)) is necessary.

Knowledge of thousands of high-resolution protein structures, together with the growing accessibility of various computational methods, has resulted in a large body of computational studies of protein dynamics. Given the power of computation, on the one hand, and the stringent prerequisite for accurate energetic descriptions of the system (small energy differences must be calculated relative to the absolute sum of all energetic terms of the system), on the other hand, experimental validation is necessary. Ideally, this should be an iterative process, with experimental testing of computational predictions and extensions of current computational methodology. This process is particularly important for tier-0 motions, because extensive approximations are required to gain access to this timescale computationally.

Modeling Frameworks

Molecular modeling frameworks attempt to provide a robust description of the mechanical effects and influences on a system. They can be devised to provide a logical

basis that can be strongly correlated on a short time scale resulting in a deterministic description of a process [molecular dynamics]. Similarly, frameworks based on enumeration [Monte Carlo] can be used to provide a more computationally tractable characterization of a phenomenon.

Molecular Dynamics [Deterministic, Relaxation Based]

Molecular dynamics (**MD**) is a form of computer simulation wherein atoms and molecules are allowed to interact for a period of time under known laws of physics, giving a view of the motion of the atoms. Because molecular systems generally consist of a vast number of particles, it is impossible to find the properties of such complex systems analytically [in closed form]; MD simulation circumvents this problem by using numerical methods. It represents an interface between laboratory experiments and theory, and can be understood as a "virtual experiment". MD probes the relationship between molecular structure, movement and function. Molecular dynamics is a multidisciplinary method. Its laws and theories stem from mathematics, physics, and chemistry, and it employs algorithms from computer science and information theory. It was originally conceived within theoretical physics in the late 1950's, but is applied today mostly in materials science and biomolecules.

Molecular dynamics is a specialized discipline of molecular modeling and computer simulation based on statistical mechanics; the main justification of the MD method is that statistical ensemble averages are equal to time averages of the system, known as the ergodic hypothesis. MD has also been termed "statistical mechanics by numbers" and "Laplace's vision of Newtonian mechanics" of predicting the future by animating nature's forces and allowing insight into molecular motion on an atomic scale. However, long MD simulations are mathematically ill-conditioned, generating

cumulative errors in numerical integration that can be minimized with proper selection of algorithms and parameters, but not eliminated entirely. Furthermore, current potential energy functions are, in many cases, not sufficiently accurate to reproduce the dynamics of molecular systems, so the much more demanding Ab-Initio Molecular Dynamics method must be used. Nevertheless, molecular dynamics techniques allow detailed time and space resolution into representative behavior in phase space.

Statistical Thermodynamic [Enumeration Based]

Enumeration based methods are commonly called monte carlo methods in the literature. A **Monte Carlo method** is a computational algorithm that relies on repeated random sampling to compute its results. Monte Carlo methods are often used when simulating physical and mathematical systems. Because of their reliance on repeated computation and random or pseudo-random numbers, Monte Carlo methods are most suited to calculation by a computer. Monte Carlo methods tend to be used when it is infeasible or impossible to compute an exact result with a deterministic algorithm.

The **fundamental assumption of the Monte Carlo method** is that the molecular movement and collision phases can be decoupled over time periods that are smaller than the mean collision time.

Topology

The insight motivating topology is that some geometric problems depend not on the exact shape of the objects involved, but rather on the way they are put together. For example, the square and the circle have many properties in common: they are both one dimensional objects (from a topological point of view) and both separate the plane into two parts, the part inside and the part outside.

Molecular systems are typically characterized using molecular graphs. These graphs associate nodes with atoms and bonds with edges. The properties of atoms such as atom radii, charge and bonds such as bond length, angle and dihedral angles are attributed to each topological entity.

Geometry

For a given topology, the geometric attributes provide the information needed to provide a specific and unique embedding of a structure in 3d space.

Protein Data Bank Files

The Protein Data Bank format provides the de-facto standard for description of molecular structures. It directly specifies a 3d embedding of the atoms of a given structure. Topological connectivity must be inferred by an application processing the data.

Structural Data Model

A UML (Unified Modeling Language) structural data model was designed FIGURE 7 to facilitate the enumerative [monte carlo type] framework needed to generate the statistical mechanical ensembles used to calculate binding affinity.

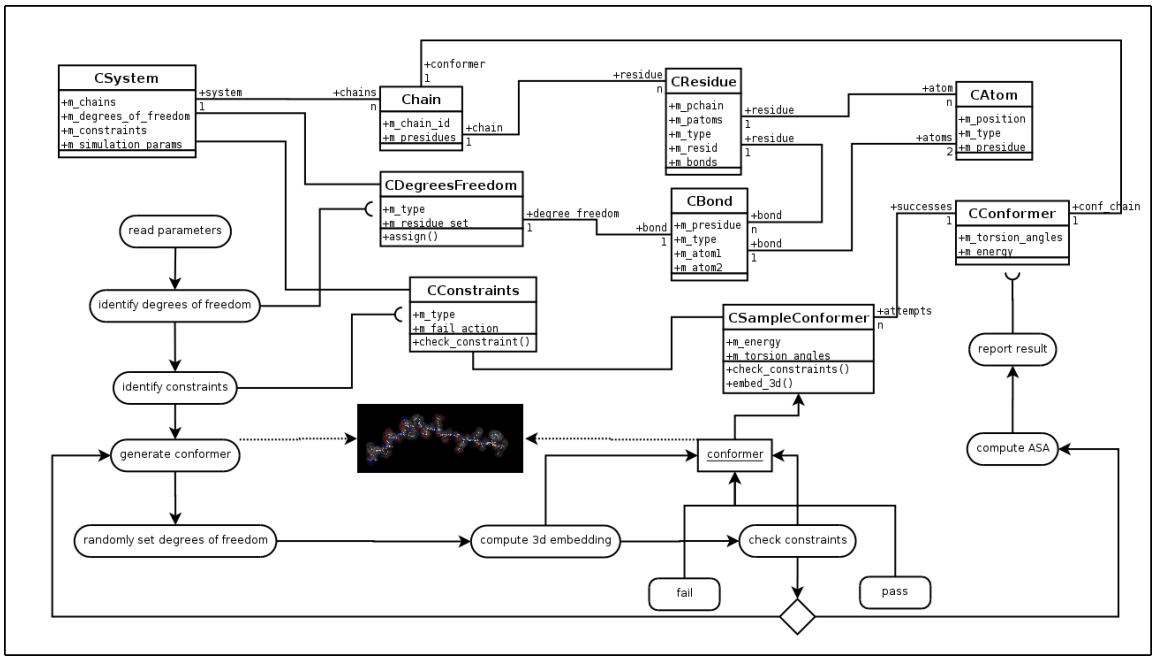


FIGURE 7: UML STRUCTURAL DATA MODEL: Object classes needed to model multi-chain proteins are given [CAtom, CResidue, CBond, Chain]. Each of these object classes directly correspond and can be derived from records within the PDB file. Object classes needed to define the simulation are [CDegreesFreedom, CConstraints, CSystem]. Object classes needed to store results are [CSampleConformer, CConformer].

Conformational Space

Degrees of Freedom

In mechanics, **degrees of freedom** (DOF) are the set of independent displacements that completely specify the displaced or deformed position of the body or system. This is a fundamental concept relating to systems of moving bodies in mechanical engineering, robotics, structural engineering, etc.

A particle that moves in three dimensional space has three translational displacement components as DOFs, while a rigid body would have at most six DOFs, which include three rotations. Translation is the ability to move without rotating, while rotation is angular motion about some axis.

Phase Space

In a phase space (Hill 1960), every degree of freedom or parameter of the system is represented as an axis of a multidimensional space (Cullen 1972). For every possible state of the system, or allowed combination of values of the system's parameters, a point is plotted in the multidimensional space. Often this succession of plotted points is analogous to the system's state evolving over time. In the end, the phase diagram represents all that the system can be, and its shape can easily elucidate qualities of the system that might not be obvious otherwise. A phase space may contain very many dimensions. For instance, a gas containing many molecules may require a separate dimension for each particle's x, y and z positions and velocities as well as any number of other properties.

In classical mechanics the phase space co-ordinates are the generalized coordinates $\boldsymbol{q_i}$ and their conjugate generalized momenta $\boldsymbol{p_i}$. The motion of an ensemble of systems in this space is studied by classical statistical mechanics. The local density of points in such systems obeys Liouville's Theorem (Hill 1960), and so can be taken as constant. Within the context of a model system in classical mechanics, the phase space coordinates of the system at any given time are composed of all of the system's dynamical variables. Because of this, it is possible to calculate the state of the system at any given time in the future or the past, through integration of Hamilton's or Lagrange's

equations of motion (Hill 1960). Furthermore, because each point in phase space lies on exactly one phase trajectory, no two phase trajectories can intersect.

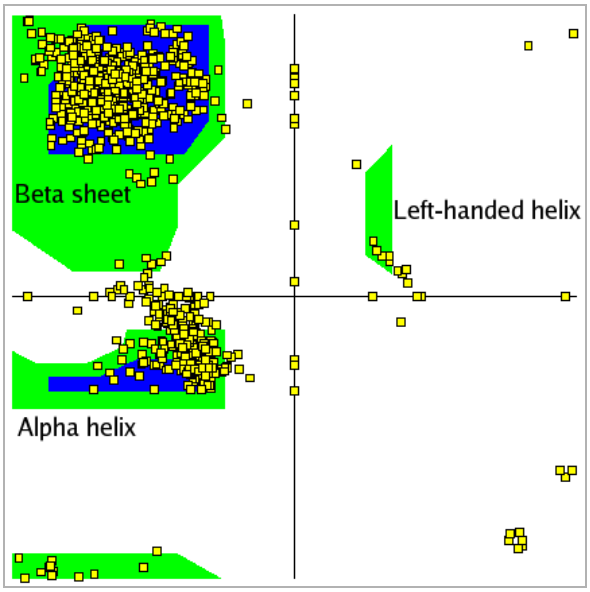
Residue Level [Ramachandran Plot]

A **Ramachandran plot** (Ramachandran, Ramakrishnan et al. 1963; Ramachandran and Sasisekharan 1968) (also known as a **Ramachandran map** or a **Ramachandran diagram**), developed by Gopalasamudram Narayana Ramachandran, is a way to visualize dihedral angles φ against ψ of amino acid residues in protein structure. It shows the possible conformations of φ and ψ angles for a polypeptide that are energetically favorable.

Mathematically, the Ramachandran plot is the visualization of a function FIGURE 8. The domain of this function is the torus. Hence, the conventional Ramachandran plot is a projection of the torus on the plane, resulting in a distorted view and the presence of discontinuities.

One would expect that larger side chains would result in more restrictions and consequently a smaller allowable region in the Ramachandran plot. In practice this does not appear to be the case; only the methylene group at the β position has an influence (Ramachandran, Ramakrishnan et al. 1963). Glycine has a hydrogen atom, with a smaller van der Waals radius, instead of a methyl group at the ***β*** position. Hence it is least restricted and this is apparent in the Ramachandran plot for Glycine for which the allowable area is considerably larger.

In contrast, the Ramachandran plot for proline shows only a very limited number of possible combinations of ψ and φ.



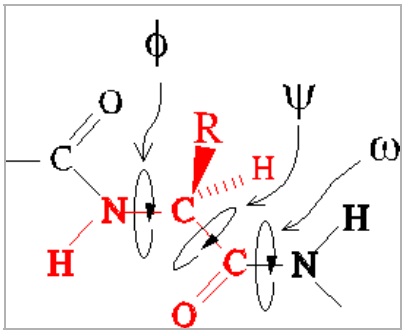


FIGURE 8 : A RAMACHANDRAN PLOT GENERATED FROM THE PROTEIN PCNA (a human DNA clamp protein that is composed of both beta sheets and alpha helices) (PDB ID 1AXC). Left: Points that lie on the axes indicate N- and C-terminal residues for each subunit. The green regions show possible angle formations that include Glycine, while the blue areas are for formations that don't include Glycine. Right: ***φ, ψ*** angles within the context of a left handed amino acid.

Effect of Ala/Gly Mutation on Conformational Heterogeneity

Chain Level [Multidimensional Allowed Space]

The **Ramachandran** map was conceived as a theoretical means of predicting the allowed conformational space of a single amino acid in a peptide by means of a hard sphere model that allowed for the steric coupling effects of both ***φ*** and ***ψ*** angles (Richards 1977). This work showed that protein conformations are substantially restricted, due to steric hindrances from what one might expect without considering the

coupling of φ and ψ angles. Conformations of experimental structures can be plotted into this φ - ψ space. If this plot is constructed from a database of protein structures that are well resolved, the 2D plot discriminates φ - ψ space into "allowed" and "disallowed" regions by outlining the most populated regions FIGURE 9. The experimentally observed conformations from well resolved structures basically correspond to those regions of φ - ψ space initially predicted by **Ramachandran**. New structures can be analyzed for the fraction of residues within allowable regions. This type of analysis is implemented in commonly used validation tools for protein structure, such as PROCHECK (Laskowski, Macarthur et al. 1993). In this way, the φ - ψ plot has proven itself as an unequalled tool in understanding the conformational space available for proteins and in the refinement and analysis of newly determined protein structures.

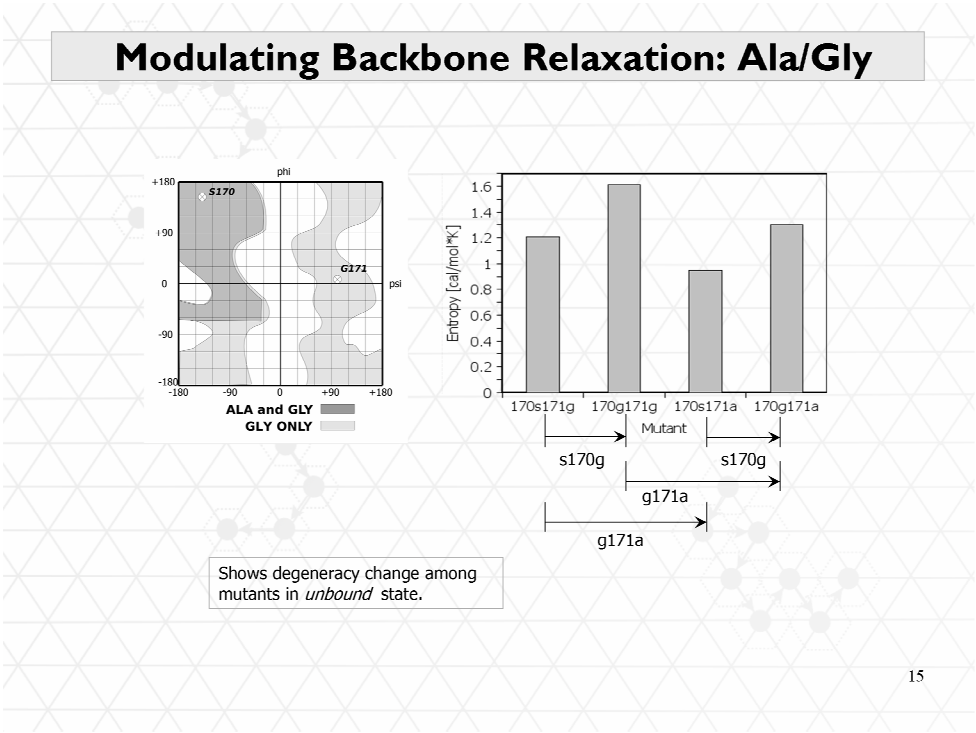


FIGURE 9: CONFORMATIONAL HETEROGENEITY MODULATED THROUGH ALA/GLY MUTATION Left: φ, ψ plot of Ala and Gly residues. Right: Conformational heterogeneity of unbound ensembles of Ala/Gly mutant cycles.

It is also possible to validate protein structures by means of longer fragment lengths. Protein substructures or building blocks have been used for modeling earlier by (Unger and Moult 1996) and (Jones and Thirup 1986). Most recently, (Micheletti, Seno et al. 1998) have also demonstrated that the conformational spaces for peptides are restricted, and almost any known protein structure can be reconstructed within 1 Å rms deviation by using a representative set of polypeptide units of four to seven residues in length with between 28 and 2,500 representative conformations, respectively (Micheletti, Seno et al. 1998), suggesting that it is possible to define an allowable space for polypeptides longer than three residues. The conformation of a dipeptide fragment, that is, two complete residues in length (with attached C- and N-terminal peptide bonds), can be described by four torsion angles (two pairs of φ - ψ values) around two central C_α atoms. We refer to polypeptide units of a given length by the number of φ,ψ pairs: $(\varphi,\psi)_1$, which is equivalent to a **Ramachandran** map, $(\varphi,\psi)_2$, $(\varphi,\psi)_3$, and so forth. Unfortunately, the 4D space of the $(\Phi, \Psi)_2$ unit (and the subsequent higher dimensional spaces of longer units) cannot be readily visualized in two or even three dimensions. However, the multidimensional scaling (MDS) method often allows one to reduce the number of dimensions and view the conformational space in a reduced (e.g., three) dimensional representation. A family of statistical methods exists that can be used for dimensional reduction, of which we have used classical MDS in interpreting the conformational space of each polypeptide length. The technique of mapping by means of dimensional reduction has been applied successfully in nucleic acid conformational space

as well as protein fold space (Banavar, Maritan et al. 2002). We have implemented MDS for extending conformational space analysis to peptide fragments of longer length beyond $(\varphi,\psi)_1$, the conventional φ,ψ map FIGURE 10.

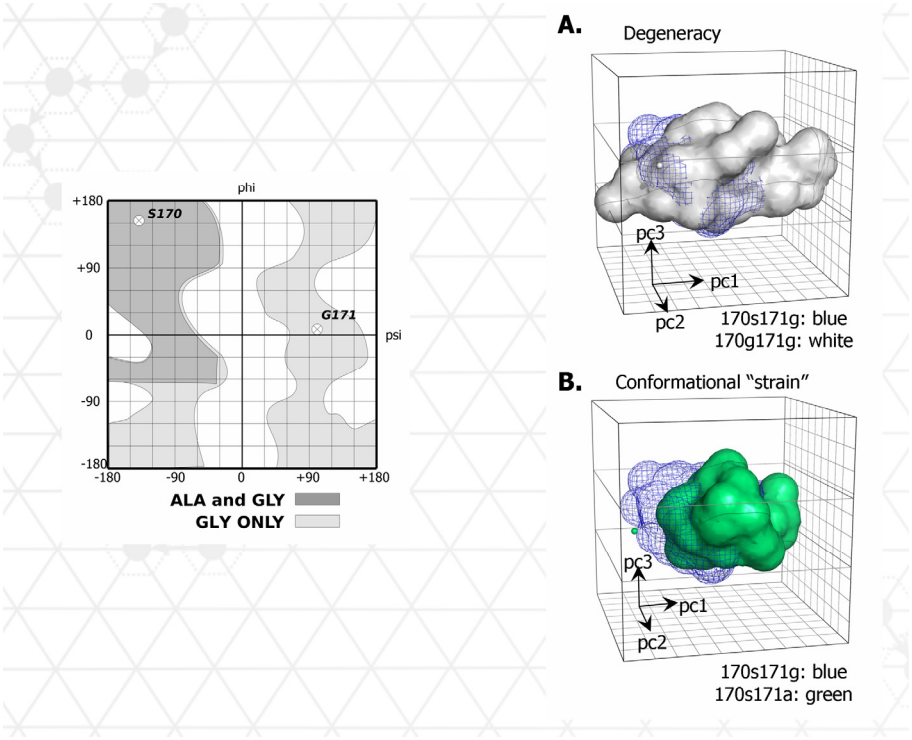


FIGURE 10: MULTIDIMENSIONAL PLOT OF ALLOWED SPACE FOR SYSTEM OF 10 RESIDUES:, A: Shows effect of Ala/Gly mutation that enlarges the allowed space. B: Shows Gly/Ala mutation that constrains (i.e. contracts) the allowed space. Residue level φ, ψ plot provides basis for the multi-residue perturbations.

THERMODYNAMIC PARAMETERS

State Function

In thermodynamics, a **state function**, *state quantity*, or a *function of state*, is a property of a system that depends only on the current state of the system, not on the way in which the system got to that state. A state function describes the equilibrium state of a system. For example, internal energy, enthalpy and entropy are *state quantities* because they describe quantitatively an equilibrium state of thermodynamic systems. At the same time, mechanical work and heat are process quantities because they describe quantitatively the *transition* between equilibrium states of thermodynamic systems.

Thermodynamic Potential

A **thermodynamic potential** is a scalar potential function used to represent the thermodynamic state of a system. One main thermodynamic potential which has a physical interpretation is the internal energy, U. It is the energy of configuration of a given system of conservative forces (that is why it is a potential) and only has meaning with respect to a defined set of references. Expressions for all other thermodynamic energy potentials are derivable via *Legendre* transforms from an expression for U. Thermodynamic potential functions include:

- Internal Energy
- Helmholtz Free Energy
- Enthalpy
- Gibbs Free Energy

Gibbs Free Energy

The free energy at constant pressure is defined by:

$$G = H - TS; \qquad (1)$$

It represents the amount of thermodynamic energy in a system that can be converted into work at a constant temperature and pressure.

Enthalpy

The **enthalpy** or **heat content** (denoted as H, h, or rarely as χ) is a quotient or description of thermodynamic potential of a system, which can be used to calculate the "useful" work obtainable from a closed thermodynamic system under constant pressure and entropy. It is defined by:

$$H = U + pV \qquad (2)$$

where U is the internal energy. The **internal energy** of a thermodynamic system, denoted by U, or sometimes E, is the total of the kinetic energy due to the motion of molecules (translational, rotational, vibrational) and the potential energy associated with the vibrational and electric energy of atoms within molecules or crystals. It includes the energy in all the chemical bonds.

Entropy

Entropy at constant pressure is defined by:

$$S = -\left(\frac{\partial G}{\partial T}\right)_p; \qquad (3)$$

the entropy is defined as (proportional to) the logarithm of the number of microscopic configurations (see text below on statistical mechanics) that result in the observed macroscopic description of the thermodynamic system. It

corresponds to the number of ways (i.e. degeneracy) that a system can be expressed at a given energy level. It is commonly expressed as:

$$S = -k_B \ln(\Omega); \qquad (4)$$

where Ω is known as the degeneracy level.

Heat Capacity

Heat capacity at constant pressure is defined by:

$$C_p = T\left(\frac{\partial S}{\partial T}\right)_p; \qquad (5)$$

It is the change in degeneracy of the system at the temperature is increased (i.e. as the energy level of the system increases).

ENERGY FUNCTIONS

Free energy is the most important property to consider when determining the probability of states within a thermodynamic ensemble. The viability of techniques such as the monte carlo technique will be highly dependent on the accuracy of the energy function used.

Electrostatics and solvation energies are important for defining protein stability, structural specificity, and molecular recognition. In the context of many protein systems they are considered to be the most dominant components of the energetic balance.

The energy function that describes the predicted stability of a sequence threaded onto a structure is:

$$\Delta G = E_{forcefield} + \Delta G_{solvation} - G_{reference} \qquad (6)$$

where ΔG is the predicted stability, $E_{\text{forcefield}}$ is the molecular mechanics force-field energy (van der Waals, torsion, and Coulombic electrostatics; equivalent to enthalpy in constant temperature, pressure, and volume simulations), $\boldsymbol{\Delta G_{solvation}}$ is the solvation energy, which $G_{\text{reference}}$ is the reference (unfolded) state energy, which includes the enthalpy and conformational entropy of the unfolded state. The $E_{\text{forcefield}}$ term describes the interactions between protein atoms, and is parameterized with quantum calculations and experiments performed on small molecules in vacuo (Jorgensen et al. 1996). Given that proteins are macromolecules dissolved in water, the energy function must also take into account the energy required to solvate the molecule in water ($\Delta G_{\text{solvation}}$). The solvation energy has two primary components, one due to the hydrophobic effect and the other due to solvation of charged/polar groups.

Solvation

The transfer energies of compounds from vacuum to water or from a nonpolar solvent to water are described by an (Solvent Accessible Surface Area) SASA-dependent energy function (Chothia 1975; Eisenberg and McLachlan 1986; Sitkoff, Sharp et al. 1994; Street and Mayo 1998):

$$\boldsymbol{\Delta G_{SASA}} = \sum_i \boldsymbol{\lambda_i A_i} \qquad (7)$$

where ΔG_{SASA} is the SASA-dependent hydrophobic solvation energy, γ_i is the atomic solvation parameter for atom i, and A_i is the SASA of i.

The SASA for a given atom is dependent on other atoms in a molecule. For pair energy calculations, a complete molecule never exists. As first described by (Wodak and Janin 1980) and later elaborated by (Street and Mayo 1998), one can approximate the surface area by adding up the surface area buried between a rotamer and the backbone and individual rotamer pairs, using empirical scale factors to account for the

overcounting of areas buried by multiple atoms. We describe here an approximation that is additive, and is therefore faster than the pairwise methods, because the number of surface-area calculations scales linearly rather than quadratically with the number of atomic positions.

The SASA of a given rotamer in a complete molecule can be estimated by calculating the SASA of that rotamer in the context of a molecule in which the "missing" side-chain atoms are mimicked with enlarged backbone pseudoatoms. In this scheme, the SASA of a rotamer is calculated once during the rotamer–backbone calculation step, and ΔG_{SASA} is calculated and added to $\Delta G_{i\_internal}$.

Structural Parameterization of Free Energy

The binding affinity is defined by the free energy of binding, which, in turn, is determined by the enthalpy and entropy changes. Because the binding enthalpy is the term that predominantly reflects the strength of the interactions of the ligand with its target relative to those with the solvent, it is desirable to develop ways of predicting enthalpy changes from structural considerations. Three terms need to be considered: (1) the intrinsic enthalpy change that reflects the nature of the interactions between ligand, target, and solvent; (2) the enthalpy associated with any possible conformational change in the protein or ligand upon binding; and, (3) the enthalpy associated with protonation/ deprotonation events, if present. As in the case of protein stability, the intrinsic binding enthalpy scales with changes in solvent accessible surface areas.

It is now well established that the enthalpy change for protein denaturation scales in terms of changes in solvent accessible surface areas (ASA) for different atom types associated with the transition between native and denatured states. In the most detailed

parametric equation, the scaling coefficients for the changes in solvent accessibility are a function of the atomic packing density (δ) of the native structure.

$$\Delta H(T,\delta) = \sum_i a_i(T,\delta) \times \Delta ASA_i \qquad (8)$$

where ΔASA_i represents the changes in solvent accessible surface area for atoms of type i and $a_i(T,\delta)$ their corresponding scaling coefficients. The $a_i(T,\delta)$ coefficients are a function of temperature and the atomic packing density.

Fully accounting for the free energy at different temperatures will require a heat capacity and an estimation of the solvation entropy. Changes in conformational entropy due to backbone relaxation and rotameric contributions must also be added. Adding in these effects results in the phenomenological model:

$$\begin{aligned} \Delta C_p &= 0.45\Delta ASA_{ap} - 0.26\Delta ASA_{pol} + 0.43\Delta ASA_{OH}; \\ \Delta H(60^\circ C) &= 31.4\Delta ASA_{pol} - 8.44\Delta ASA_{ap} + \Delta H_{ion}; \end{aligned} \qquad (9)$$

$$\begin{aligned} \Delta S(T) &= \Delta S_{solv}(T) + \Delta S_{conf}(T) + \Delta S_{trans} + \Delta S_{ion}; \\ \Delta S_{solv}(T) &= 0.45\Delta ASA_{ap} \ln\left(\frac{T}{384.15}\right) - 0.26\Delta ASA_{pol} \ln\left(\frac{T}{335.15}\right); \end{aligned} \qquad (10)$$

$$\begin{aligned} \Delta S_{conf}(T) &= \Delta S_{bu\to ex} + \Delta S_{ex\to bu} + \Delta S_{bb}; \\ \Delta S_{bu\to ex} &= \sum_i \left(\frac{\Delta ASA_i}{ASA_i}\right) \Delta S_{bu\to ex,T}; \end{aligned} \qquad (11)$$

(Murphy and Freire 1992)

At room temperature, T=25C, one can easily see that apolar surface burial is highly favorable entropically and modestly unfavorable enthalpically. Polar surface exposure is favored enthalpically and slightly unfavored entropically.

Conspicuously absent from these equations are the details of the electrostatic contributions to the free energy and effects such as hydrogen bonding and quantum effects such as π bonding. Electrostatic effects, which are more diffuse and can be dependent on extended 3-d structure, cannot be described in terms of localized solvation surface area.

The hydrophobic effect can be rationalized by considering the packing of water molecules at polar and apolar surfaces (Bennaim and Marcus 1984; Silverstein, Haymet et al. 1998; Choudhury and Pettitt 2007). As shown below, the water molecule cannot directly hydrogen bond with a Lennard-Jones (i.e. apolar) surface FIGURE 11. This means that the hydrogens and the lone pairs cannot be oriented perpendicularly to the apolar surface. This induces a tetrahedral structure in the neighborhood of the surface that optimizes the hydrogen bonding structure (similar to ice) in the neighborhood of the surface. This structure typically has a lower density than bulk water and because the region is structured, the entropy near the surface will be greatly reduced. On the other hand, the polar surface can interact favorably with the protons and lone pairs at a greater number of orientations; hence the entropy will increase for this case.

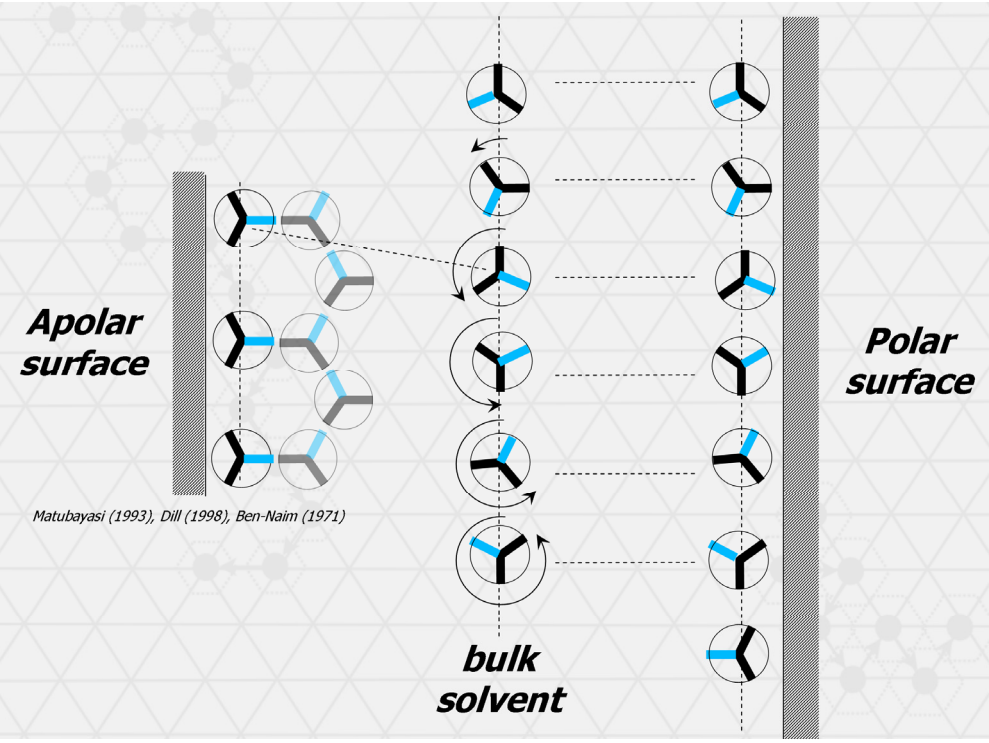


FIGURE 11: RATIONALE FOR SOLVENT ENTROPY VARIATION FOR APOLAR AND POLAR SURFACES: Left: shows limited solvent orientations (degrees of freedom) near an apolar surface. Middle: orientations in bulk solvent. Right: Possible orientations near a polar surface.

Dependence on Internal Degrees of Freedom

The variation of the solvation free energy described above with the internal molecular coordinates is not smooth. An example where two structurally similar conformers have a wide variation in free energy is shown in FIGURE 12.

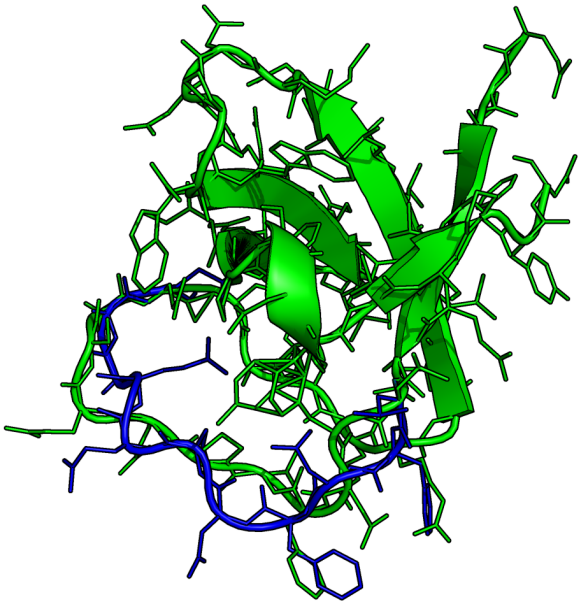


FIGURE 12: VARIATION OF FREE ENERGY WITH INTERNAL STRUCTURAL COORDINATES: Two conformations that are structurally similar differ in free energy by more that 5 kcal/mol. This example illustrates the modal [rugged] character of the thermodynamic landscape as a function of internal coordinates.

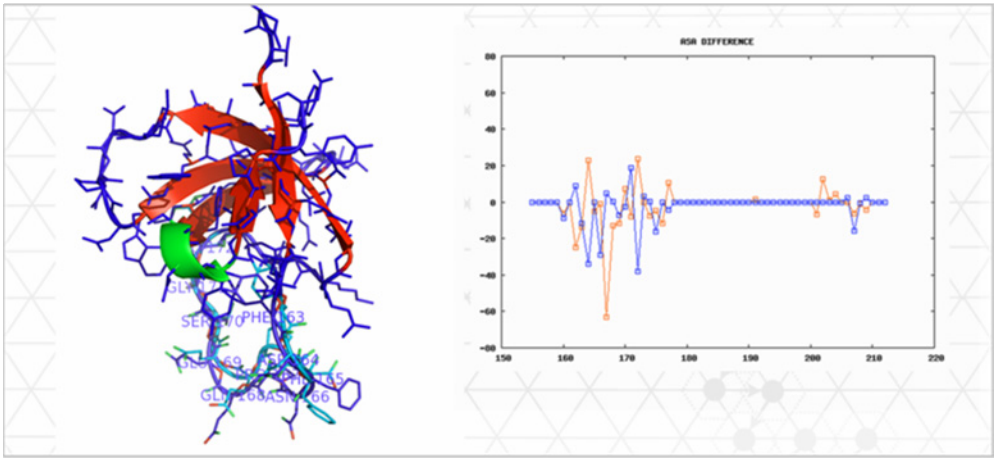


FIGURE 13 : POSITION DEPENDENT VARIATION OF FREE ENERGY WITH INTERNAL STRUCTURAL COORDINATES: Two conformations that are structurally similar (*0.5 rmsd*) differ in free energy by more that 6 kcal/mol. The residue level apolar (green) and polar (pink) differences in surface are for each position in flexible section of chain.

Electrostatic Interactions

Poisson Boltzmann Equation

The **Poisson-Boltzmann equation** is a differential equation that describes electrostatic interactions between molecules in ionic solutions. It is the mathematical base for the Gouy-Chapman Double layer (interfacial) theory; first proposed by Gouy in 1910 and complemented by Chapman in 1913. The equation is important in the fields of molecular dynamics and biophysics because it can be used in modeling implicit solvation, an approximation of the effects of solvent on the structures and interactions of proteins, DNA, RNA, and other molecules in solutions of different ionic strength. It is often difficult to solve the Poisson-Boltzmann equation for complex systems, but several computer programs have been created to solve it numerically.

The equation can be written as:

$$\nabla \cdot \varepsilon(X)\nabla\varphi(X) = -\frac{\rho(X)}{\varepsilon_0} - \frac{1}{\varepsilon_0}\sum_{i=1}^{N} q_i n_i^0 e^{-\frac{q_i\varphi(X)}{kT}} \qquad (12)$$

where $\varepsilon(X)$ represents the position-dependent dielectric, $\varphi(X)$ represents the electrostatic potential, $\rho(X)$ represents the charge density of the solute, represents the concentration of the ion i at a distance of infinity from the solute, z_i is the charge of the ion, q is the charge of a proton, k_B is the Boltzmann constant, T is the temperature, and is

a factor for the position-dependent accessibility of position *r* to the ions in solution. If the potential is not large, the equation can be linearized to be solved more efficiently.

Once the potential field for a system of particles has been calculated FIGURE 14, the free energy of the system can be computed from the following integral:

$$G[\phi] = \frac{1}{4\pi}\int\left[\rho_f\phi - \frac{\varepsilon}{2}(\nabla\phi) + \kappa^2(\cosh(\phi) - 1)\right]dx \qquad (13)$$

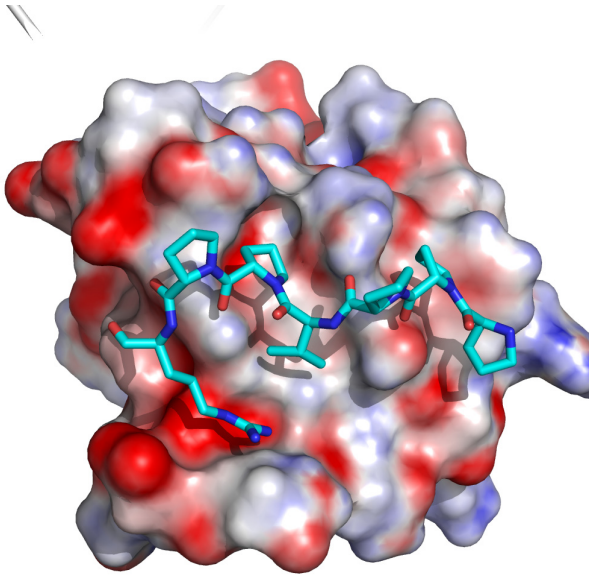


FIGURE 14 : ELECTROSTATIC POTENTIAL AT SOLVENT ACCESSIBLE SURFACE : Electrostatic potential distribution for C-SH3:SosY complex. Although much of the binding surface is neutral, there is an electronegative patch near position E172 where a salt bridge is believed to form with the R8 position of the ligand.

Force Fields

Hard Sphere Collision Model

The hard sphere model is an approximation based on the van der waals energy profile (Sowdhamini, Ramakrishnan et al. 1993). Here, the repulsive energy increases according to the twelfth power as two atoms are brought together, resulting in physically unrealistic energy levels below a given distance threshold. The primary effect of the hard sphere approximation is to highlight excluded volume effects.

Dominant Free Energetic Components

Thermodynamic Cycle of Folding

A **thermodynamic cycle** is a series of thermodynamic processes which returns a system to its initial state. Properties depend only on the thermodynamic state and thus do not change over a cycle. Variables such as heat and work are not zero over a cycle, but rather are process dependent. Thermodynamic cycles can be applied to processes such as protein folding:

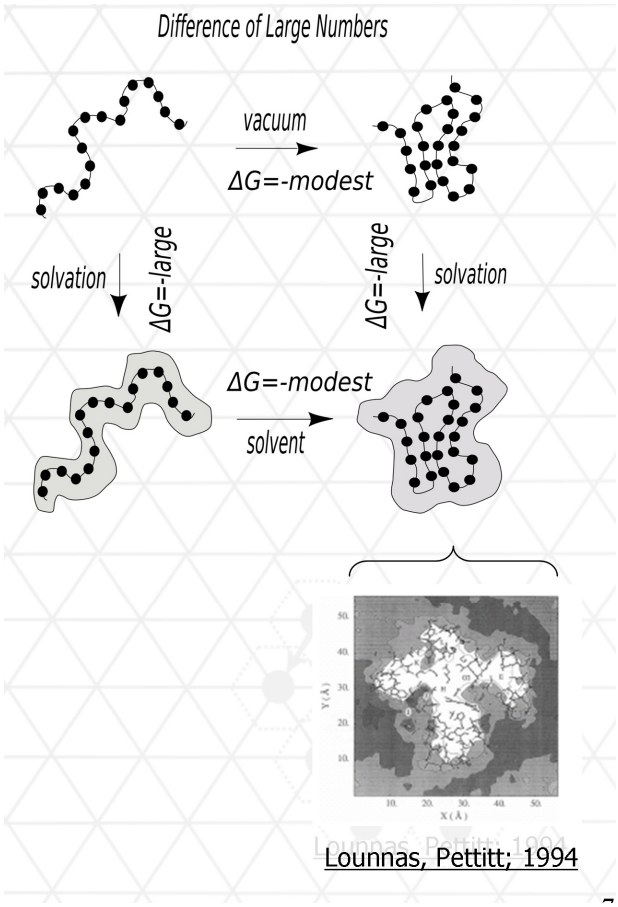


FIGURE 15: THERMODYNAMIC CYCLE OF PROTEIN FOLDING: Top transition shows protein folding in a vacuum. Bottom transition is the corresponding state change within a solvent environment. Solvation transitions indicate large change in free energy due to placement in solvent. Inset shows diffusion coefficient of solvent around solute myoglobin (Lounnas, Pettitt et al. 1994). This distribution suggests that the solvent is structured at the solute boundary.

It is apparent that the free energy change due to solvation within this cycle is significant FIGURE 15. These large free energy changes suggest that solvation effects could be the dominant force in this and other processes.

STATISTICAL MECHANICS

Statistical mechanics is the application of probability theory, which includes mathematical tools for dealing with large populations, to the field of mechanics, which is concerned with the motion of particles or objects when subjected to a force. Statistical mechanics, sometimes called statistical physics, can be viewed as a subfield of physics and chemistry.

It provides a framework for relating the microscopic properties of individual atoms and molecules to the macroscopic or bulk properties of materials that can be observed in everyday life, therefore explaining thermodynamics as a natural result of statistics and mechanics (classical and quantum) at the microscopic level. In particular, it can be used to calculate the thermodynamic properties of bulk materials from the spectroscopic data of individual molecules.

This ability to make macroscopic predictions based on microscopic properties is the main asset of statistical mechanics over thermodynamics. Both theories are governed by the second law of thermodynamics through the medium of entropy. However, entropy in thermodynamics can only be known empirically, whereas in statistical mechanics, it is a function of the distribution of the system on its micro-states.

Maxwell-Boltzmann Distribution

The **Maxwell–Boltzmann distribution** is a probability distribution with applications in physics and chemistry FIGURE 16. The most common application is in the field of statistical mechanics. The temperature of any (massive) physical system is the result of the motions of the molecules and atoms which make up the system. These particles have a range of different velocities, and the velocity of any single particle constantly changes due to collisions with other particles. However, the fraction of a large

number of particles within a particular velocity range is nearly constant. The Maxwell distribution of velocities specifies this fraction, for any velocity range, as a function of the temperature of the system.

Ensembles

Partition Function

The classical partition function is the configuration integral over the degrees of freedom (positions and momenta) of the thermodynamic system with maxwell-boltzmann weighting of each microstate:

$$Q_{class} = \frac{1}{N!h^{3N}} \int_{-\infty}^{\infty}\int_{V} e^{(-\beta H(x_N, p_N))} d^{3N}x \, d^{3N}p; \qquad (14)$$

Where $H(x_N, p_N)$ is the Hamiltonian (energy functional) of the system at positions x_N and momenta p_N. The partition function represents the number of effective states within a thermodynamic system. The term effective applies because of the energetic weighting factor applied to each point (microstate) in the 6N dimensional phase space.

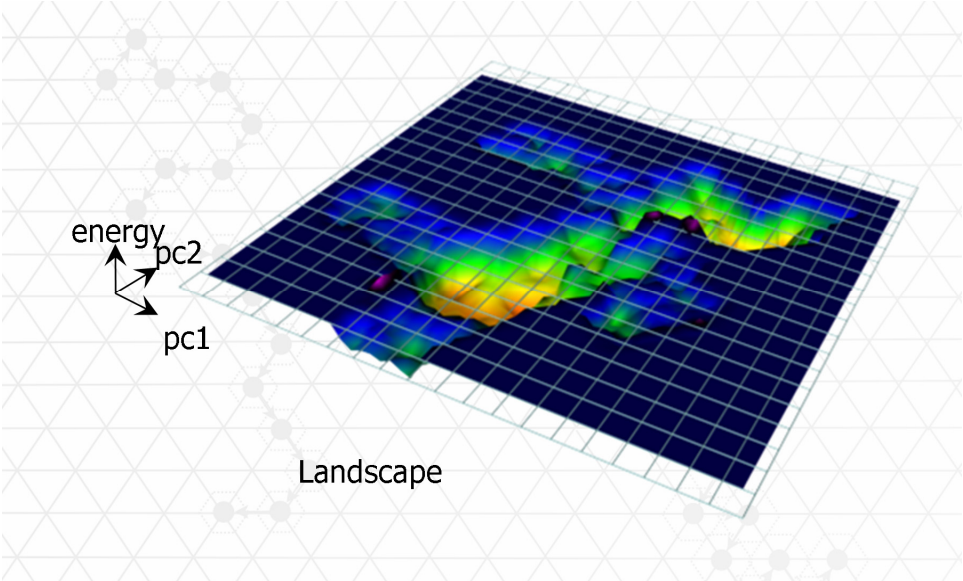


FIGURE 16: FREE ENERGY LANDSCAPE: Microscopic partition function plotted against two spatial degrees of freedom. Surface is color coded to show probability of thermal states within landscape. Overall partition function for this system is obtained by integrating over the microscopic landscape.

The canonical ensemble partition function of a system in contact with a thermal bath at temperature T is the normalization constant of the Boltzmann distribution function. The partition function integrated over the **energy levels** of the system is:

$$Z(T) = \int \Omega(E) \exp(-E/k_B T)\, dE, \qquad (15)$$

where $\Omega(E)$ is the density of states or degeneracy with energy E and k_B the Boltzmann constant.

In classical statistical mechanics, there is a close connection between the partition function and the configuration integral, which has played an important role in many applications (e.g., drug design).

The partition function of a system is related to the Helmholtz energy function through the formula

$$A = -k_B T \log Z.$$

This connection can be derived from the fact that $k_B \log \Omega(E)$ is the entropy of a system with total energy E. This is an extensive [additive] magnitude [an **extensive** property is one that cannot be defined for a specific spatial point, and whose value varies with the size of the system] in the sense that, for large systems (i.e. in the thermodynamic limit, when the number of particles $N \to \infty$ or the volume $V \to \infty$), it is proportional to N or V. In other words, if we assume N large, then

$$k_B \log \Omega(E) = N s(e), \qquad (16)$$

where $s(e)$ is the entropy per particle in the thermodynamic limit, which is a function of the energy per particle $e = E / N$. We can therefore write

$$Z(T) = N \int \exp\{N(s(e) - e/T)/k_B\}\, de. \qquad (17)$$

Since N is large, this integral can be performed through steepest descent, and we obtain

$$Z(T) = N \exp\{N(s(e_0) - e_0/k_B T)\}, \qquad (18)$$

where e_0 is the value that maximizes the argument in the exponential; in other words, the solution to

$$s'(e_0) = 1/T. \qquad (19)$$

This is the thermodynamic formula for the inverse temperature provided e_0 is the mean energy per particle of the system. On the other hand, the argument in the exponential is:

$$\frac{1}{k_B T}(TS(E_0) - E_0) = -\frac{A}{k_B T} \qquad (20)$$

the thermodynamic definition of the Helmholtz energy function. Thus, when N is large,

$$A = -k_B T \log Z(T). \qquad (21)$$

We have the aforementioned Helmholtz energy function,

$$A = -k_B T \log Z(T) \qquad (22)$$

we also have the internal energy, which is given by

$$U = k_B T^2 \left.\frac{\partial \log Z(T)}{\partial T}\right|_{N,V} \qquad (23)$$

and the pressure, which is given by

$$p = k_B T \left.\frac{\partial \log Z(T)}{\partial V}\right|_{N,T}. \qquad (24)$$

These equations provide the link between classical thermodynamics and statistical mechanics.

Microscopic (Mechanical) Properties and Collections

Partitions

A partition of a set X is a set of nonempty subsets of X such that every element x in X is in exactly one of these subsets.

Equivalently, a set P of nonempty sets is a partition of X if:

1. The union of the elements of P is equal to X. (We say the elements of P *cover* X.)

2. The intersection of any two elements of P is empty. (We say the elements of P are pairwise disjoint.)

The elements of P are sometimes called the **blocks** or **parts** of the partition.

The phase space of a statistical mechanical framework can be partitioned. A logical motivation for this is the definition of physical phases that can occur in a set of interacting particles .eg., the representation of the bound and unbound phases of a protein-ligand system. The border of these partitions is termed the encounter boundary (Camacho, Weng et al. 1999; Camacho and Vajda 2001).

Constraints

Constraint in statistical mechanics refers to the degree of statistical dependence between degrees of freedom.

In an equilibrium state there are no unbalanced potentials, or driving forces, within the system. A central aim in equilibrium thermodynamics is: given a system in a well-defined initial state, subject to accurately specified constraints, to calculate what the state of the system will be once it has reached equilibrium. An equilibrium state is obtained by seeking the extrema of a thermodynamic potential function, whose nature depends on the **constraints** imposed on the system. For example, a chemical reaction at constant temperature and pressure will reach equilibrium at a minimum of its components Gibbs free energy and a maximum of their entropy.

Excluded volume or self-avoidance stems from the physical requirement that no two particles of matter can occupy the same place at the same time. It is the constraint that must be adhered to when modeling on the microscopic and mesoscopic level. This constraint acts to limit states allowable to a system and defines the so-called allowable region within the phase space.

Two Component [Docking] Ensembles

In a system consisting of two interacting molecules, one will have to extend the landscape in order to accommodate the presence of the second molecule. The energy distribution for each of the conformations of the first molecule will be altered by the proximity of the second at all the allowable **inter**-molecular positions and orientations.

Conformational Coordinates

The N position vectors of the nuclei constitute a $3N$ dimensional linear space $\mathbf{R}^{3N}$: the *configuration space*. The Eckart conditions give an orthogonal direct sum decomposition of this space (Hill 1960).

$$\boldsymbol{R}^{3N} = \boldsymbol{R}_{ext} \oplus \boldsymbol{R}_{\text{int}} \qquad (25)$$

The elements of the $3N$-6 dimensional subspace $\mathbf{R}_{\text{int}}$ are referred to as *internal coordinates*, because they are invariant under overall translation and rotation of the molecule and, thus, depend only on the internal (vibrational) motions. The elements of the 6-dimensional subspace $\mathbf{R}_{\text{ext}}$ are referred to as *external coordinates*, because they are associated with the overall translation and rotation of the molecule.

Contact Topology

In mathematics, a **distance matrix** is a matrix (two-dimensional array) containing the distances, taken pairwise, of a set of points. It is therefore a symmetric $N \times N$ matrix containing non-negative real values as elements, given N points in Euclidean space. The number of pairs of points $\boldsymbol{N \times (N-1)/2}$ is the number of independent elements in the distance matrix.

Distance matrices are used to represent protein structures in a coordinate-independent manner, along with the pairwise distances between two sequences in

sequence space. Distance matrices are related to adjacency matrices, with the differences that (a) the latter only provides the information which vertices are connected but does not tell about *costs* or *distances* between the vertices and (b) an entry of an distance matrix is smaller if two elements are closer, while "close" (connected) vertices yield larger entries in an adjacency matrix. The adjacency matrix of a polymer defines its contact topology.

Interaction Order Parameter [Reaction Coordinate]

In chemistry, a **reaction coordinate** is an abstract one-dimensional coordinate which represents progress along a reaction (or process) pathway. It is usually a geometric parameter that changes during the conversion of one or more molecular entities.

These coordinates can sometimes represent a real coordinate system (such as bond length, bond angle...), although, for more complex reactions especially, this can be difficult (and non geometric parameters are used, e.g., bond order or *contact topology*).

Reaction coordinates are often plotted against free energy to demonstrate in some schematic form the potential energy profile (an intersection of a potential energy surface) associated to the reaction.

In the formalism of transition-state theory the reaction coordinate is that coordinate in set of curvilinear coordinates obtained from the conventional ones for the reactants which, for each reaction step, lead smoothly from the configuration of the reactants through that of the transition state to the configuration of the products. The reaction coordinate is typically chosen to follow the path along the gradient (path of shallowest ascent/deepest descent) of potential energy from reactants to products.

Potential of Mean Force

It is useful to know how the free energy changes as a function of reaction coordinates, such as the distance between two atoms or the torsion angle of a bond in a molecule. Here we introduce the concept of the potential of mean force (PMF). The PMF expresses the variation in free energy of a select set of the degrees of freedom while integrating out the remainder. When the system is in a solvent, the PMF incorporates solvent effects as well as the intrinsic interaction between the two particles. When the same two particles are brought together in the gas phase, the free energy would simply be the pair potential ***u(r)***, which has only a single minimum. But the PMF between two particles in liquid oscillates with maximum and minimum. For a given separation r between the two molecules, the PMF describes an average over all the conformations (positions and orientations) of the surrounding solvent molecules.

Various methods have been proposed for calculating potentials of mean force. The simplest representation of the PMF is to use the separation r between two particles as the reaction coordinate. The PMF is related to the radial distribution function using the following expression for the Helmholtz free energy

$$A(r) = -k_B T \ln(g(r)) + const.; \qquad (26)$$

The constant is chosen so that the most probable distribution corresponds to a free energy of zero. Unfortunately, the PMF may vary by several multiples of $k_B T$ over the relevant range of the distance r. The algorithmic relationship between the PMF and the radial distribution function means that a relatively small change in the free energy (i.e. a small multiple of $k_B T$) may correspond to $g(r)$ changing by an order of magnitude from its most likely value.

Chemical Potential

Consider a thermodynamic system containing n constituent species. Its total internal energy U is postulated to be a function of the entropy S, the volume V, and the number of particles of each species N_1,..., N_n:

$$U = U(S,V,N_1,..N_n) \qquad (27)$$

By referring to U as the *internal energy*, it is emphasized that the energy contributions resulting from the interactions between the system and external objects are excluded. For example, the gravitational potential energy of the system with the Earth are not included in U.

The chemical potential of the i-th species, μ_i is defined as the partial derivative:

$$\mu_i = \left(\frac{\partial U}{\partial N_i} \right)_{S,V,N_{j \neq i}} ; \qquad (28)$$

where the subscripts simply emphasize that the entropy, volume, and the other particle numbers are to be kept constant.

In **real** systems, it is usually difficult to hold the entropy fixed, since this involves good thermal insulation. It is therefore more convenient to define the Helmholtz free energy A, which is a function of the temperature T, volume, and particle numbers:

$$A = A(T,V,N_1,..N_n) \qquad (29)$$

In terms of the Helmholtz free energy, the chemical potential is:

$$\mu_i = \left(\frac{\partial A}{\partial N_i} \right)_{T,V,N_{j \neq i}} ; \qquad (30)$$

Laboratory experiments are often performed under conditions of constant temperature and pressure. Under these conditions, the chemical potential is the partial derivative of the Gibbs free energy with respect to number of particles:

$$\mu_i = \left(\frac{\partial G}{\partial N_i} \right)_{T,p,N_{j \neq i}} ; \qquad (31)$$

Grand Canonical Ensemble

In statistical mechanics, the **grand canonical ensemble** is a statistical ensemble (a large collection of identically prepared systems), where each system is in equilibrium with an external reservoir with respect to both particle and energy exchange. Therefore both the energy and the number of particles are allowed to fluctuate for each individual system in the ensemble. It is an extension of the canonical ensemble, where systems are only allowed to exchange energy (but not particles). And the chemical potential is introduced to control the fluctuation of the number of particles.

ENTROPY AND ORDER

Entropy, historically, has often been associated with the amount of order, disorder, and/or chaos in a thermodynamic system. The traditional definition of entropy is that it refers to changes in the status quo of the system and is a measure of "molecular disorder" and the amount of wasted energy in a dynamical energy transformation from one state or form to another. In this direction, a number of authors, in recent years, have derived exact entropy formulas to account for and measure disorder and order in atomic and molecular assemblies.

Relevance to Biology

Many processes require a process competent state before proceeding to a next phase. For instance, transcription factors must bind DNA before the RNA polymerase can produce mRNA. This assembly can occur in a multiplicity of ways before the process competent state is reached and the polymerase can commence its function. If the

process competent state is termed the ordered state while all others termed dis-ordered, then the ratio of the size of these sets will reflect the entropy change required to induce the action of the polymerase.

Entropy Cannot Be Measured Directly

Entropy cannot be measured directly; there are equations that relate it to other properties which can be measured, so it is possible to determine the entropy of the system indirectly. For instance, for a reversible system at constant temperature, $\Delta S = \Delta Q/T$. Measurement of the heat removed or added to the system, and dividing by the temperature of the system will yield the change in entropy. To determine an absolute number, we have to define some sort of standard system to which the current system of study is compared.

Microstates, Macrostates and Degeneracy

In statistical mechanics, a **microstate** describes a specific detailed microscopic configuration of a system, which the system visits in the course of its thermal fluctuations. For a system of N particles, a microstate is described by 3N positions and 3N momenta. Each microstate can therefore be described by a point in a 6N dimensional space.

In contrast, the **macrostate** of a system refers to its macroscopic properties such as its temperature and pressure. A macrostate is defined by a set microstates and its properties are computed by integrating over that set of microstates. In statistical mechanics, a macrostate is characterized by a probability distribution on a certain ensemble of microstates.

This distribution describes the probability of finding the system in a certain microstate as it is subject to thermal fluctuations.

In the case of large systems, even if those systems are theoretically able to fluctuate between very different microstates, observing such a fluctuation becomes less and less likely as the size of the system increases. This makes up for the thermodynamic limit. In this limit, the microstates visited by a system during its fluctuations all have the same bulk (or macroscopic) properties.

The definitions of this section link the thermodynamic properties of a system to its distribution on its ensemble (or set) of microstates. Note that all definitions and expressions of this section are valid even far away from thermodynamic equilibrium.

We will consider a system which is distributed on an ensemble of N microstates. p_i is the probability associated to the microstate i, and E_i is its energy FIGURE 17. Here microstates form a discrete set, which means we are working in quantum statistical mechanics, and E_i is an energy level of the system.

Internal energy: The internal energy is the mean of the system's energy:

$$\boldsymbol{U = \langle E \rangle = \sum_{i=1}^{n} p_i E_i;} \qquad \textbf{(32)}$$

This definition is the traduction of the first law of thermodynamics.

Entropy: The absolute entropy exclusively depends on the probabilities of the microstates. Its definition is the following:

$$\boldsymbol{S = -k_B \sum_i p_i \ln(p_i);} \qquad \textbf{(33)}$$

where k_B is Boltzmann's constant

Entropy evaluates according to the second law of thermodynamics. From the definition, it is clear that entropy is maximized as the probability of the microstates becomes more evenly distributed. The third law of thermodynamics is consistent with

this definition, since an absolute entropy of 0 means that the macrostate of the system reduces to a single microstate.

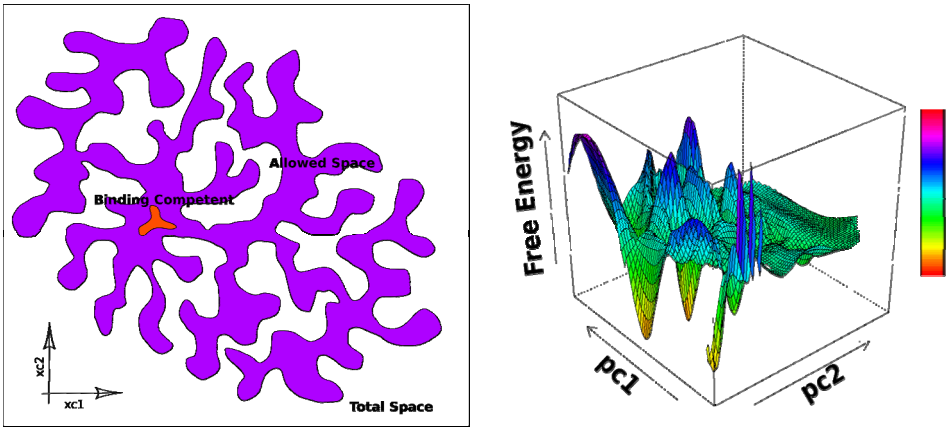


FIGURE 17: MICROSCOPIC DEGENERACY: Left: Landscape showing states (blue) at a given energy level E. The landscape is projected onto two reduced dimensional (principal) spatial coordinates. The complex boundary of the region at energy E is typical of polymeric systems. The ergodic hypothesis states that over a "long" enough time frame, each point within the degenerate region is equally probable. The small purple region is a configuration that is sufficiently ordered to be process competent. Right: Free energy landscape for local unfolding transitions for Sh3 modular binding domain.

Order Parameters

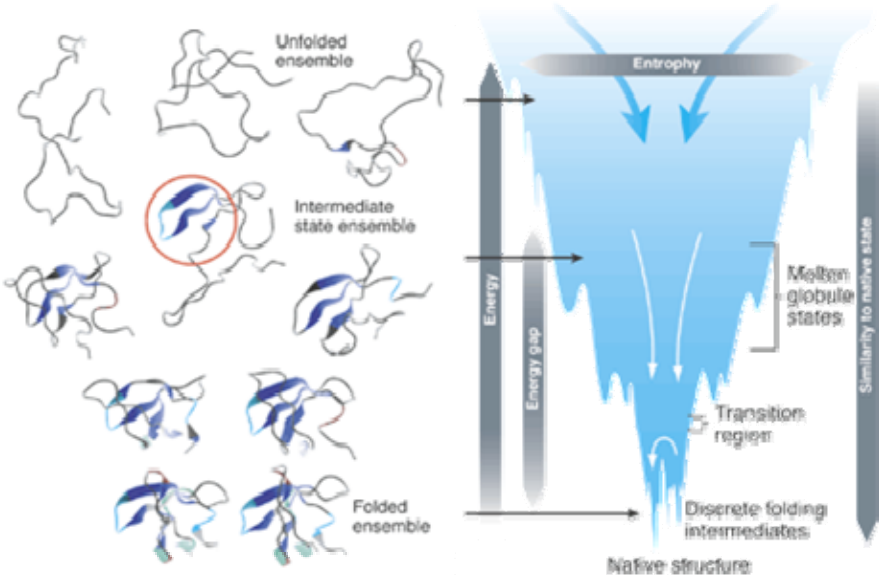
An order parameter measures some degree of order in a system; the values typically range from zero for total disorder to one for complete order. For example, an order parameter can indicate the degree of order in a liquid crystal or a folded protein. However, note that order parameters can also be defined for non-symmetry-breaking transitions such as reaction progress coordinates.

THERMODYNAMIC LANDSCAPES

The folding of proteins, the complex behavior of glasses, and the structure and dynamics of atomic and molecular clusters has long been studied in separate disciplines. In recent years, energy landscape theory has emerged as a unifying language for experimentalists and theorists to describe structure formation and dynamics in these complex systems.

The energy landscape describes how the energy of a system changes with geometry, as defined by the coordinates of atoms, molecules, or side chains. At a minimum, a small displacement in any direction increases the potential energy, just as in a basin surrounded by mountains; a step in any direction is uphill. Potential energy surfaces of complex systems usually have vast numbers of local minima; the lowest one--the deepest basin--is the global minimum FIGURE 18. Energy landscapes for different systems may differ widely. These differences are responsible for the fact that natural proteins and crystals can reliably locate one particular structure from many possible ones, whereas glasses fail to do so.

Landscape Organization and Terrain Features



(Brooks and Karplus 1989)

FIGURE 18: THE FREE ENERGY LANDSCAPE FOR PROTEIN FOLDING Folding occurs through the progressive organization of ensembles of structures [shown here for the src-SH3 domain (left)] on a funnel-shaped free energy landscape (right). Conformational entropy loss during folding is compensated by the free energy gained as more native interactions are formed. Kinetics is determined by the local roughness of the landscape, relative to thermal energy. Key interactions in early folding (dashed circle) coincide, for this protein, with experimentally determined regions.

Energy Landscape [Microscopic]

In physics, an **energy landscape** is a pair (X, f) consisting of a topological space X representing the physical states or parameters of a system together with a continuous

function f: $X \rightarrow \mathbf{R}^n$ representing the energies associated to these states or parameters such that the image of f represents a hypersurface in $\mathbf{R}^n$.

Landscape Coordinates

The coordinates of an energy landscape are generally the degrees of freedom of the system. However coordinate mappings can be used to reconfigure a space to another space that may make the landscape more transparent. One example is Cartesian coordinates for each particle in a molecule to the internal ***BAT*** [bond, angle torsion] coordinates. One could proceed to determine even better decompositions that may help reduce the effective dimensionality of the space. Such a technique [based on principle components] will be discussed later.

Probabilistic Roadmap [Landscape]

A landscape can be built and described using a so-called disconnectivity graph (Becker 1997; Becker and Karplus 1997; Becker 1998). One instance of this type of graph is the **Probabilistic Roadmap (PRM)** (Amato, Bayazit et al. 2000). It is a method used in motion planning algorithm in robotics, which solves the problem of determining a path between a starting configuration of the robot and a goal configuration while avoiding collisions.

The basic idea behind PRM is to take random samples from the configuration space of the robot, testing them for whether they are in the free space, and use a local planner to attempt to connect these configurations to other nearby configurations. The starting and goal configurations are added in, and a graph search algorithm is applied to the resulting graph to determine a path between the starting and goal configurations.

PRM is provably probabilistically complete, meaning that as the number of sampled points increases without bound, the probability that the algorithm won't find a path if one exists approaches zero (Amato and Song 2002).

Linear Regions of Landscape

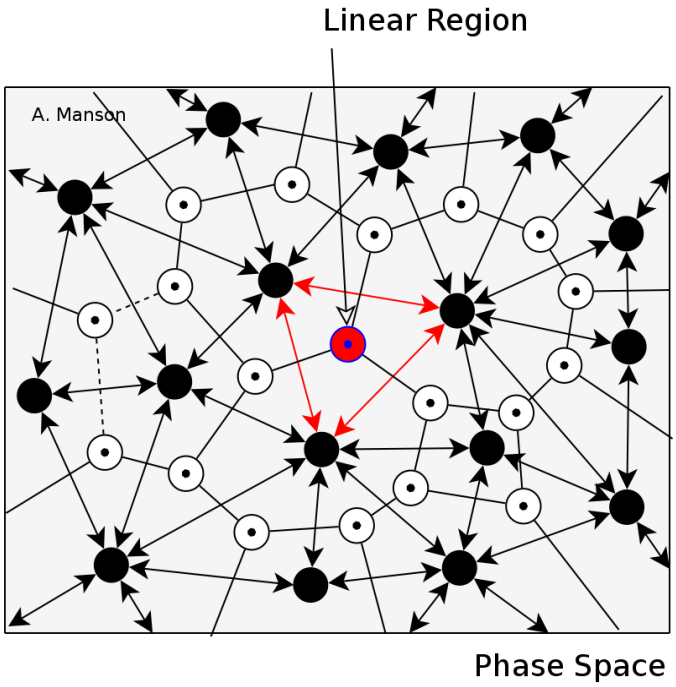
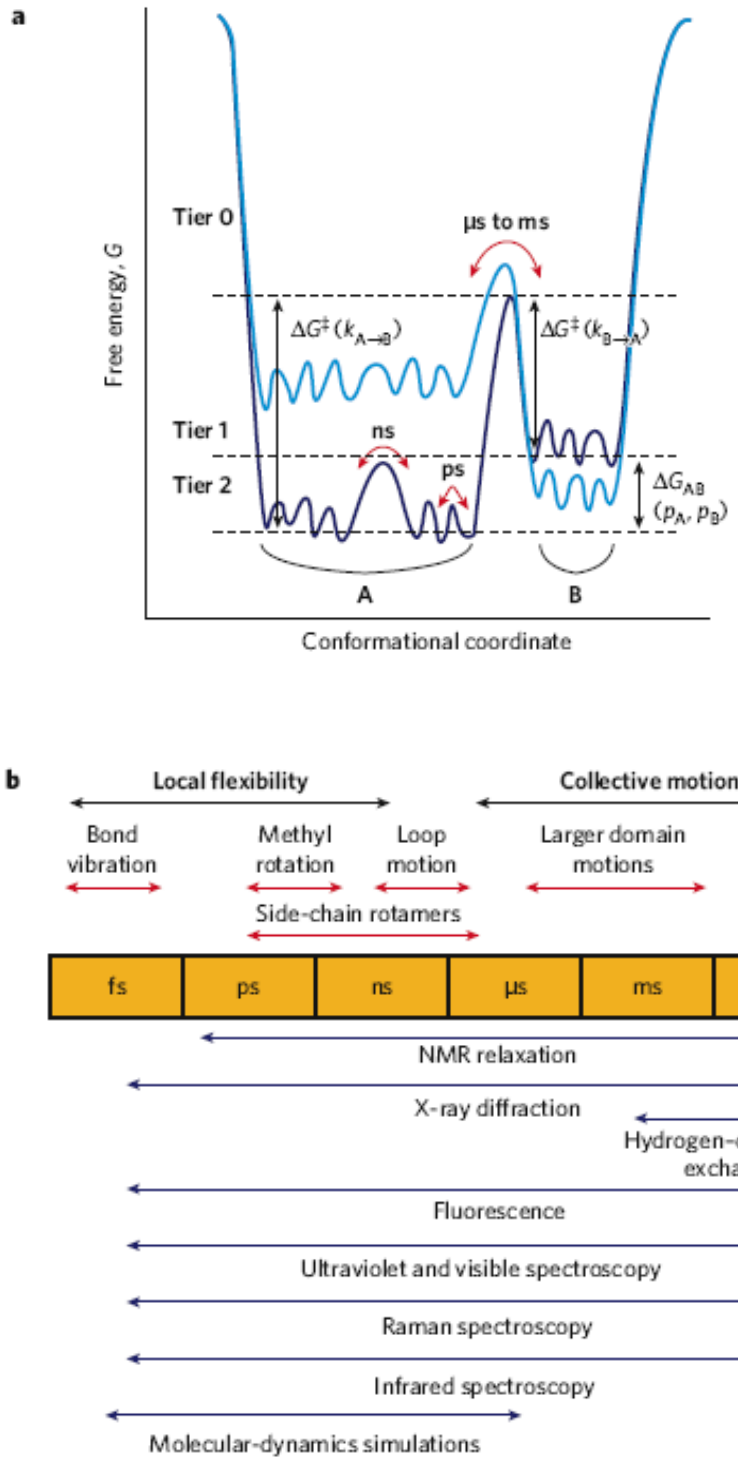


FIGURE 19: TOPOLOGICAL DEPICTION OF A GENERAL LANDSCAPE SHOWING A LINEAR SUB-REGION. The black nodes are maxima in the landscape and the nodes with a dot in their center correspond to minima. The intersection of the edges between pairs of maxima and pairs of minima are saddle points. The edges connecting adjacent pairs of maxima correspond to the activation barriers between the dual pair of minima. A quasi-linear region of the landscape is shown in the red hues. Here, the basin surrounding the red minimum can be described using linear techniques.

The pattern of hills and valleys in a potential energy landscape can be generally described by a graph. In cases where the local unfolding approximation is valid, the effective configuration space can be described by a sub-region of the overall space. In this region the conformations can be represented as linear perturbations from the minimum within the basin FIGURE 19. Random linear [additive] perturbations within this system can be analyzed using established statistical techniques.

PROTEIN DYNAMICS

Based on the observation of multiple energy barriers and non-exponential kinetics below a temperature of 230 K, an energy-landscape model was developed (Frauenfelder, Sligar et al. 1991). Frauenfelder and colleagues insightfully connected this energy-landscape concept to myoglobin function and characterized the features of the landscape FIGURE 20, including the heights of the barriers between energy wells and the existence of multiple conformational substates (Frauenfelder, Petsko et al. 1979). Subsequent studies on myoglobin led to the idea that substates are in thermal equilibrium and that both solvent (Brooks and Karplus 1989) and ligands influence the landscape. At the glass transition temperature (Frauenfelder, Sligar et al. 1991), an increase in anharmonic dynamics occurs in proteins, and this is interpreted as the protein no longer being trapped in a single energy well. This transition has recently been attributed to a solvent relaxation effect in the hydration shell of proteins (Fenimore, Frauenfelder et al. 2004). Since these early studies, many more details of protein energy landscapes have been characterized as a result of advances in experimental and computational techniques.



(Henzler-Wildman and Kern 2007)

FIGURE 20: THE ENERGY LANDSCAPE DEFINES THE AMPLITUDE AND TIMESCALE OF PROTEIN MOTIONS. a: One-dimensional cross-section through the high dimensional energy landscape of a protein showing the hierarchy of protein dynamics and the energy barriers. Each tier is classified following the description introduced by Frauenfelder and co-workers 93. A state is defined as a minimum in the energy surface, whereas a transition state is the maximum between the wells. The populations of the tier-0 states A and B (pA, pB) are defined as Boltzmann distributions based on their difference in free energy (ΔG_{AB}). The barrier between these states ($\Delta G^{\ddagger}$) determines the rate of interconversion (k). Lower tiers describe faster fluctuations between a large number of closely related substates within each tier-0 state. A change in the system will alter the energy landscape (from dark blue to light blue, or vice versa). For example, ligand binding, protein mutation and changes in external conditions shift the equilibrium between states. b: Timescale of dynamic processes in proteins and the experimental methods that can detect fluctuations on each timescale.

MOLECULAR RECOGNITION

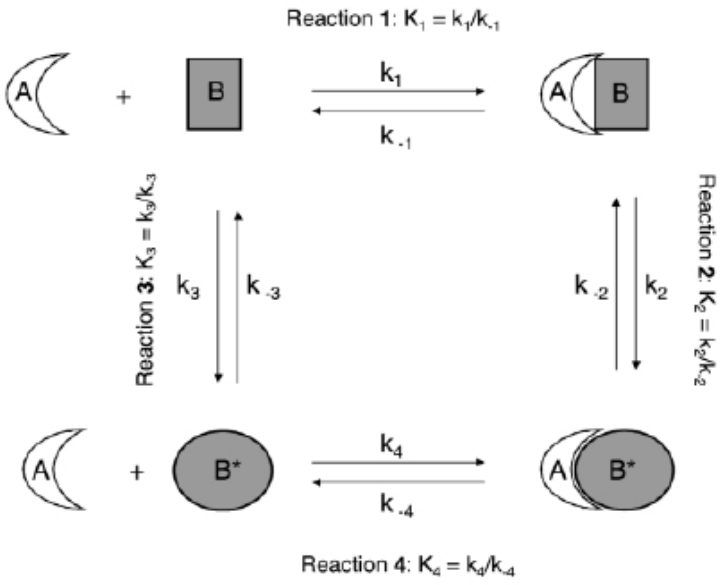
The term **molecular recognition** refers to the specific interaction between two or more molecules through noncovalent bonding such as including hydrogen bonding, metal coordination, hydrophobic forces, van der Waals forces, pi-pi interactions, and/or electrostatic effects. The host and guest involved in molecular recognition exhibit molecular complementarity.

Binding as a Unit Operation in Biological Processes

Molecular recognition plays an important role in biological systems and is observed in between receptor-ligand, antigen-antibody, DNA-protein, sugar-lectin, RNA-ribosome, etc. An important example of molecular recognition is the antibiotic vancomycin that selectively binds with the peptides with terminal D-alanyl-D-alanine in bacterial cells through five hydrogen bonds. The vancomycin is lethal to the bacteria since once it has bound to these particular peptides they are unable to be used to construct the bacteria's cell wall.

Binding Models

Virtually all biological phenomena depend in one way or another on specific molecular recognition. At the end of the 19th century, Emil Fischer coined his famous lock-and-key analogy to picture the specificity of enzyme reactions, which are a molecular premise of life (Garrett and Grisham 1999). The enzyme was considered to be a rigid template in which the substrate had to fit as a key into a lock. Over the years, however, it became apparent that a rigid fit between preformed molecular structures cannot explain all aspects of enzyme catalysis. It is in this context that, over 40 years ago, Daniel Koshland formulated the concept of the induced fit (Koshland 1994). To facilitate the enzymatic reaction in the absence of a precise fit, he postulated that "the substrate may cause an appreciable change in the three-dimensional relationship of the amino acids at the active site" (Koshland 1994). The idea of a precise fit was retained from the lock-and-key image, but it was stated explicitly that the fit "occurs only after the changes induced by the substrate itself" FIGURE 21.



(Bosshard 2001)

FIGURE 21: THERMODYNAMIC CYCLE FOR THE REACTION OF MOLECULES A AND B TO COMPLEX AB*, where B and B* are different conformational states of the same molecule. The induced-fit pathway follows reactions 1 and 2. The initial complex AB formed in reaction 1 is not stable because the conformation of B is not optimized. Induced fit reaction 2 brings B into the fitting conformation B*. The conformational selection pathway follows reactions 3 and 4. Reaction 3 describes the conformational equilibrium between the nonfitting conformation B and the fitting conformation B*. Reaction 4 is the binding of the fitting conformation B* to A. The induced-fit pathway is kinetically competent only if complex AB has appreciable stability so that the induced fit has a reasonable chance to occur. If this is not the case and a small amount of the fitting conformation B* is present in the absence of A, the conformational selection pathway dominates.

Other researchers, (Leder, Berger et al. 1995; Berger, Weber-Bornhauser et al. 1999; Tsai, Kumar et al. 1999) have pointed out that there is an alternative mechanism to induced fit. The essence of conformational selection, described by reactions 3 and 4 (figure above), is that the conformation change is not assumed to occur after initial binding. Folded proteins do not have a single unique structure but are better regarded as a large ensemble of similar structures having similar energy contents. These so-called conformers are in rapid fluctuation with each other (Onuchic, Nymeyer et al. 2000). If the energy landscape is smooth, the many conformers interchange rapidly. If it is rugged, the ensemble may include conformers that may be quite different and interchange more slowly. Thus selection between the structures B and B* is a grossly oversimplified view. In reality it is more like a selection among very many more-or-less fitting structures (Tsai, Ma et al. 1999). However, the end result of conformational selection is the same: those conformers that show the best fit bind best.

Affinity

In biochemistry, a **ligand** is a substance that is able to bind to and form a complex with a biomolecule to serve a biological purpose. In a narrower sense, it is a signal triggering molecule binding to a site on a target protein, by intermolecular forces such as ionic bonds, hydrogen bonds and Van der Waals forces. The docking (association) is usually reversible (dissociation). Actual irreversible covalent binding between a ligand and its target molecule is rare in biological systems. Ligand binding to receptors alters the chemical conformation, i.e. the three dimensional shape of the receptor protein. The conformational state of a receptor protein determines the functional state of a receptor. The tendency or strength of binding is called affinity.

Specificity

Specificity (Altman and Bland 1994) is the state of being specific rather than general. In a statistical context, it is the probability, in a binary test, of a true negative being correctly identified (Altman and Bland 1994). In biochemistry the binary test is binding affinity. The probability relates to the set of candidate binding partners. Simply stated, specificity reflects the potential of a molecule to interact with some partners while not interacting with others.

N K Graph Organization (Fitness Landscape)

Life has been described as order at the edge of chaos. Move into the chaotic regime and life would not survive. Move too far in the opposite direction, towards stability, and life would not evolve. Relevant to this viewpoint is the NK landscape. In NK landscapes, N is the number of members and K is the number of other members that each member interacts with. The maximum K is N-1. At K=N-1, all possible states are possible and you have maximum chaos. At K=0 there would be no interactions and the system would freeze, life would hardly evolve as there would be no co-evolution. Every member would only change based on random mutations. A fairly low K seems to produce the maximum level of fitness. At K=0, fitness is low. As K increases so does fitness but after a while it starts to decrease until the system becomes chaotic. In this context, binding specificity is directly related to the magnitude of the K parameter.

DOCKING

In the field of molecular modeling, **docking** is a method which predicts the preferred orientation and position of one molecule to a second when bound to each other to form a stable complex. Knowledge of the preferred orientation in turn may be used to

predict the strength of association or binding affinity between two molecules using for example scoring functions.

The associations between biologically relevant molecules such as proteins, nucleic acids, carbohydrates, and lipids play a central role in signal transduction. Furthermore, the relative orientation of the two interacting partners may affect the type of signal produced (e.g., agonism vs. antagonism). Therefore docking is useful for predicting both the strength and type of signal produced.

Docking is frequently used to predict the binding orientation of small molecule drug candidates to their protein targets in order to in turn predict the affinity and activity of the small molecule. Hence docking plays an important role in the rational design of drugs. Given the biological and pharmaceutical significance of molecular docking, considerable efforts have been directed towards improving the methods used to predict docking.

The focus of molecular docking is to computationally stimulate the molecular recognition process. The aim of molecular docking is to achieve an optimized conformation for both the protein and ligand and relative orientation between protein and ligand such that the free energy of the overall system is minimized.

Rigid Body Docking

Molecular docking can be thought of as a problem of *"lock-and-key"*, where one is interested in finding the correct relative orientation of the *"key"* which will open up the *"lock"* (where on the surface of the lock is the key hole, which direction to turn the key after it is inserted, etc.). Here, the protein can be thought of as the "lock" and the ligand can be thought of as a "key". Molecular docking may be defined as an optimization problem, which would describe the "best-fit" orientation of a ligand that binds to a

particular protein of interest FIGURE 22. However since both the ligand and the protein are flexible, a *"hand-in-glove"* analogy is more appropriate than *"lock-and-key"*. During the course of the process, the ligand and the protein adjust their conformation to achieve an overall "best-fit" and this kind of conformational adjustments resulting in the overall binding is referred to as **"induced-fit"**. In this thesis this concept of induced fit will be further refined using the framework of statistical ensembles.

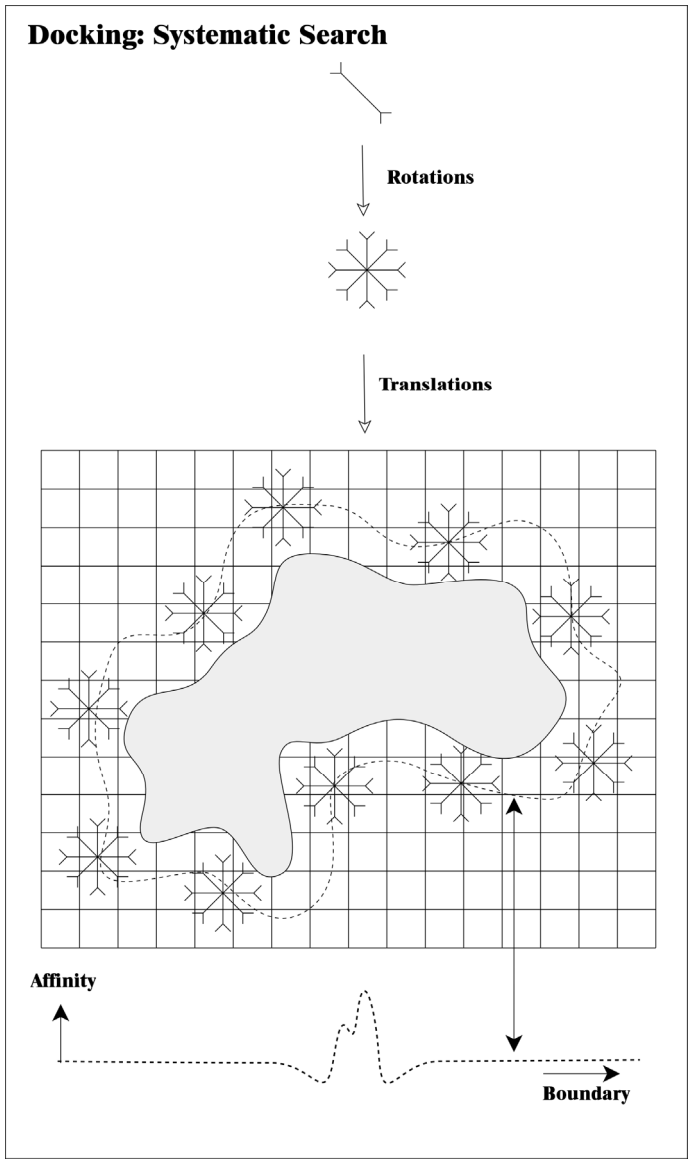


FIGURE 22: SCHEMATIC OF MOLECULAR DOCKING: A ligand is shown being docked to its putative receptor protein in a two dimensional rendering. For each position on the surface of the receptor, the ligand must sample all orientations. The free energy of the complex changes [inset below] as the ligand is moves over the receptor surface. Areas where the complex free energy is low (high affinity) show preferred orientations.

Positional and Orientational Degrees of Freedom

The position of a rigid body can be described by a combination of a translation and a rotation from a given reference position. For this purpose a reference frame is chosen that is rigidly connected to the body. This is typically referred to as a *"local"* reference frame (*L*). The position of its origin and the orientation of its axes with respect to a given *"global"* or *"world"* reference frame (*G*) represent the position of the body. The position of *G* not necessarily coincides with the initial position of *L*.

Thus, the position of a rigid body has two components: **linear** and **angular**, respectively. Each can be represented by a vector. The angular position is also called orientation. There are several methods to describe numerically the orientation of a rigid body (see orientation). In general, if the rigid body moves, both its linear and angular position varies with time. In the kinematic sense, these changes are referred to as translation and rotation, respectively.

Specificity of Orientation

This refers to an orientation of a ligand relative to a receptor such that the free energy change of interaction is clearly maximized. Due to the diverse shapes and energy distributions amongst receptors, high orientation specificity is common among interactions within biologic systems.

Scoring Functions

Shape Complementarity

Geometric matching / shape complementarity methods describe the protein and ligand as a set of features that make them dockable (Shoichet, Bodian et al. 1992). These features may include molecular surface / complementary surface descriptors. In this case, the receptor's molecular surface is described in terms of its solvent-accessible surface area and the ligands molecular surface is described in terms of its matching surface description. The complementarity between the two surfaces amounts to the shape matching description that may help finding the complementary pose of docking the target and the ligand molecules. Another approach is to describe the hydrophobic features of the protein using turns in the main-chain atoms. Yet another approach is to use a Fourier shape descriptor technique described in (Katchalskikatzir, Shariv et al. 1992). Whereas the shape complementarity based approaches are typically fast and robust, they cannot usually model the movements or dynamic changes in the ligand/ protein conformations accurately, although recent developments allow these methods to investigate ligand flexibility. Shape complementarity methods can quickly scan through several thousand ligands in a matter of seconds and actually figure out whether they can bind at the protein's active site, and are usually scalable to even protein-protein interactions. They are also much more amenable to pharmacophore based approaches, since they use geometric descriptions of the ligands to find optimal binding.

Residue Pair Potential

In protein structure prediction, a **statistical potential** (also **knowledge-based potential**, **empirical potential**, or **residue contact potential**) is an energy function

derived from an analysis of known structures in the Protein Data Bank. Typical measures could be phi / psi backbone torsion angles, binned by residue pairs or triplets, solvent accessibility, hydrogen bond characteristics, or empirical observations about the likelihood of native contacts between any two amino acid residues in the native state tertiary structure of a protein. Taking the last case as our example, in its simplest form, a statistical potential is formulated as an interaction matrix that assigns a weight or energy value to each possible contact pair of standard amino acids. The energy of a particular structural model is then the combined energy of all the residue-residue contacts (often defined as residues within 4Å) identified in the structure. The probabilities or weights are determined by statistical examination of native contacts present in a database of structures represented in the Protein Data Bank. According to the energy landscape view of protein folding, structures that closely resemble the native state will be distinguishably lower in free energy than those that are different from the native state.

Statistical potentials are used as energy functions in the assessment of an ensemble of structural models produced by homology modeling or protein threading - predictions for the tertiary structure assumed by a particular amino acid sequence made on the basis of comparisons to one or more homologous proteins whose structures have been experimentally determined. Many differently parameterized statistical potentials have been shown to successfully identify the native state structure from an ensemble of "decoy" or non-native structures (Park, Vendruscolo et al. 2000). In response to criticism that statistical potentials capture only the tendency of hydrophobic amino acids to pack closely in the hydrophobic core of a globular protein, refinements have included the creation of two interaction matrices parameterized separately for residues in the core and those on the solvent-accessible surface of the protein (Richards 1977). The primary alternative method for assessing ensembles of models and identifying the lowest-energy

structure represented relies on direct energy calculations, which are more computationally expensive than statistical potentials (Park, Vendruscolo et al. 2000) due to the necessity of calculating long-range electrostatic interactions.

Flexible Docking

Molecular docking algorithms suggest possible structures for molecular complexes. They are used to model biological function and to discover potential ligands. A present challenge for docking algorithms is the treatment of molecular flexibility. Molecular docking faces several methodological problems. These include predicting the relative binding affinities of different possible complexes, identifying binding sites on receptors, and allowing for molecular flexibility in the docking event.

BINDING AFFINITY

Components of Binding Free Energy

The binding process could be broken down into several components (Finkelstein and Janin 1989; Murphy, Xie et al. 1994) each incurring a thermodynamic cost FIGURE 23.

$$\Delta G_{binding} = \Delta E_{elec} + \Delta E_{mm} + \Delta G_{solv} - T\Delta S_{position} - T\Delta S_{orientation} - T\Delta S_{conf} \quad (34)$$

where ΔE_{elec} is the change in internal electrostatic energy; ΔE_{mm} is the change in internal molecular mechanical energy; ΔG_{solv} is the change in solvation free energy; $\Delta S_{position}$ is the change in positional entropy; $\Delta S_{orientation}$ is the change the entropy of orientation; ΔS_{conf} is the change in conformational entropy. Positional and rotation entropy components account for the cost of aligning the protein-ligand complex. These are weakly dependent on molecular weight (Janin and Chothia 1978). This weak dependence allows these contributions to be subtracted out with little error when

comparing the binding free energy from mutant to mutant. As a result the relative free energy changes are primarily due to the effects of conformational entropy and solvation free energy.

The loss of translational and rotational entropy is the major term unfavorable to association. This loss of entropy is proportional to the logarithm of the molecular weights and so does not vary much with their size: for the associations of two proteins, it is 23-30 kcal/mol (Janin and Chothia 1978) and for the binding of a small molecule by a protein 17-22 kcal/mol. The translational entropy loss of binding is:

$$\Delta S_{trans}^{bind} = -RT \ln\left[\alpha(RT)^{3/2} \frac{M_A M_B}{M_A + M_B}\right]; \qquad (35)$$

where M_A and M_B are the molecular weights of species A and B respectively.

The corresponding rotational entropy loss of binding is:

$$\Delta S_{rot}^{bind} = -RT \ln\left[\beta(RT)^{3/2} \frac{M'_A M'_B}{M'_{AB}}\right]; \qquad (36)$$

where M'_A and M'_B are the moments of inertia of species A and B respectively.

It has been shown (Janin and Chothia 1978) that hydrophobicity provides the major source of stabilization free energy in protein-protein complexes.

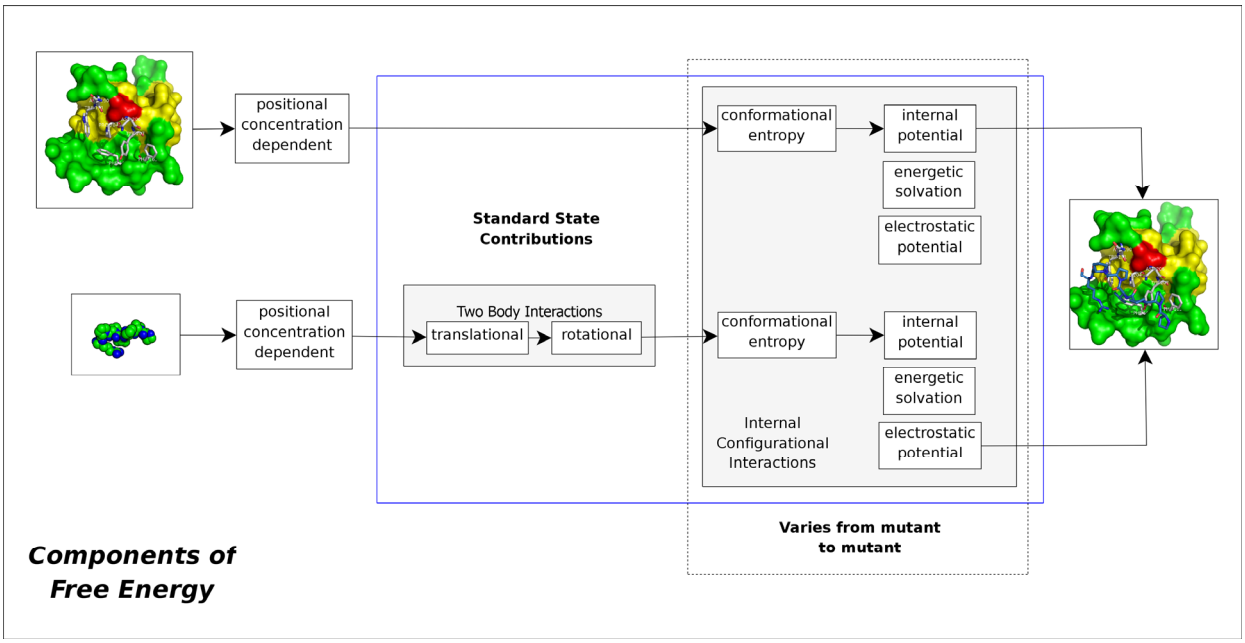
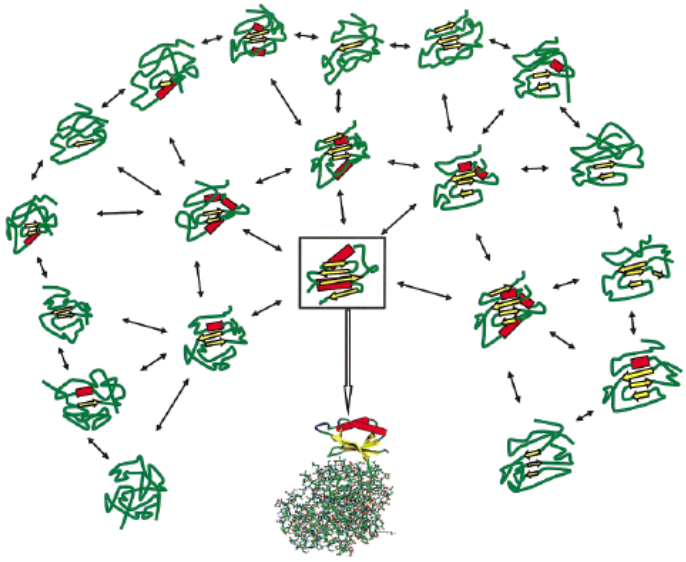


FIGURE 23: COMPONENTS OF BINDING FREE ENERGY: Overall binding free energy contributions are broken down into concentration dependent terms and those determined under the conditions of standard state [blue box]. The two body terms indicate the terms positional and orientation entropy changes. The shaded box on the right accounts for conformational heterogeneity of the receptor and the changes in internal and solvent free energy changes.

Ensemble Based Binding [Conformational Selection]

Traditionally, molecular disorder has been viewed as local or global instability. Molecules or regions displaying disorder have been considered inherently unstructured. The term has been routinely applied to cases for which no atomic coordinates can be derived from crystallized molecules. Yet, even when it appears that the molecules are disordered, prevailing conformations exist, with population times higher than those of all alternate conformations FIGURE 24. Disordered molecules are the outcome of rugged

energy landscapes away from the native state around the bottom of the funnel. Ruggedness has a biological function, creating a distribution of structured conformers that bind via conformational selection, driving association and multimolecular complex formation, whether chain linked in folding or unlinked in binding.



(Tsai, Ma et al. 2001)

FIGURE 24: POTENTIAL CONFORMERS OF THE PROREGION OF SUBTILISIN. Rather than there being a flexible, disordered conformation that becomes ordered upon binding, the native conformer is already present, albeit with a low population. This conformer is selected in the binding process, with the equilibrium shifting in its direction. The subtilisin is depicted on the right-hand side of the figure, with the native conformer of the proregion bound to it. Potential conformers are depicted in a half-moon form, with the native conformer boxed. Other conformers are also compact and structured, with different extents of secondary structures, and are in equilibrium with the native conformer. Subtilisin belongs to the first case type.

Binding Competence

Members of a ensemble of randomly generated conformers that have a low energy when docked to the ligand are termed **binding competent** FIGURE 25.

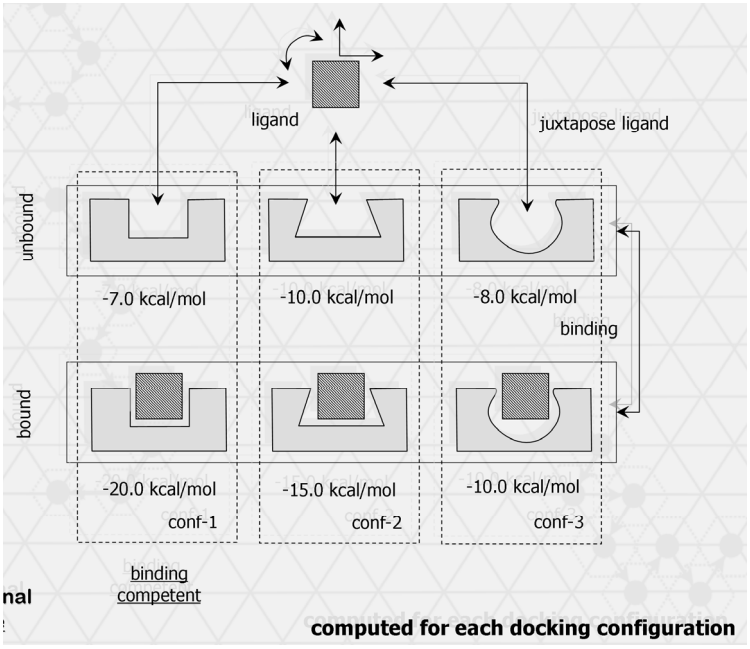


FIGURE 25: ENSEMBLE BASED BINDING: Conformational variation of ensemble showing free energy of unliganded and liganded species for a given mutant. Conformational variation results in changes in levels of complementarity and desolvation free energy. Three conformations in the unbound and bound microphases are shown. A free energy is assigned to each microstate. The free energy of the unbound conformations has been determined at high protein-ligand separations. The corresponding bound states have had the ligand docked to the protein at the optimal [crystal structure] orientation. The aggregate [ensemble weighted] change in free energy is the binding affinity. Conformations that experience large stabilization upon ligand binding are termed binding competent.

EXPERIMENTAL METHODS [RELATED TO BINDING]

Isothermal Titration Calorimetry

Isothermal titration calorimetry (ITC) is a biophysical technique used to determine the thermodynamic parameters of (biochemical) interactions. It is most often used to study the binding of small molecules (such as medicinal compounds) to larger macromolecules (proteins, DNA etc.). ITC is a quantitative technique that can directly measure the binding affinity (K_a), enthalpy changes (ΔH), and binding stoichiometry (n) of the interaction between two or more molecules in solution. From these initial measurements Gibbs energy changes (ΔG), and entropy changes (ΔS), can be determined using the relationship:

$$\Delta G = -RT\ln K = \Delta H - T\Delta S$$

(where R is the gas constant and T is the absolute temperature).

NMR HSQC

The ^{15}N HSQC experiment is probably the most frequently recorded experiment in protein NMR. The HSQC experiment can be performed either using the natural abundance of the ^{15}N isotope, or using isotopically labeled protein. The latter can be recorded on much lower concentrations of protein, but requires recombinant expression of the protein.

Each residue of the protein (except proline) has an amide proton attached to a nitrogen in the peptide bond. If the protein is folded, the peaks are usually well dispersed, and most of the individual peaks can be distinguished. Being a relatively cheap and quick experiment, the HSQC is useful to screen candidates for structure determination by NMR. The number of peaks in the spectrum should match the number

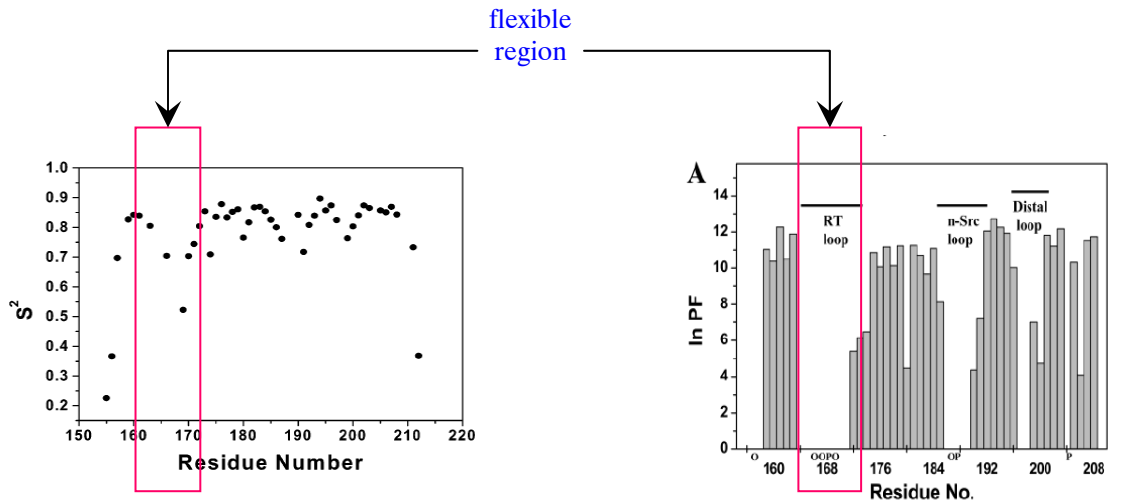
of residues in the protein (though sidechains with nitrogen-bound protons will add additional peaks). It will probably be difficult to solve the structure of the protein if this is not the case. The labour-intensive process of structure determination is usually not undertaken until a good HSQC is obtained.

It is not possible to assign the HSQC spectrum by itself; in other words to determine which peaks correspond to a particular residue in the protein. This process can be done in different ways. The assignment of the spectrum is usually the first step in a structure determination, and is essential for a meaningful interpretation of more advanced NMR experiments.

The HSQC experiment is also useful for detecting interactions with ligands, such as other proteins or drugs. By comparing the HSQC of the free protein with the one bound to the ligand, it is possible to find the changes in the chemical shifts of the peaks, which is most likely to occur in the binding interface.

NMR Order Parameter Analysis

Protein motions often play a critical role in enzyme catalysis and protein–ligand interactions. These atomic-level motions serve as the basis for many important biological processes, such as muscle contraction, cellular metabolism, antigen–antibody interactions, gene regulation and virus assembly. Obtaining detailed experimental descriptions of protein motions continues to be a challenging task due to the limitations and complexity of existing methods.



Ferreon and Hilser, Protein Sci. (2003) Ferreon et.al., Biochemistry (2003)

FIGURE 26: ORDER PARAMETER AND HYDROGEN EXCHANGE ANALYSIS OF C-SH3: The red box shows the candidate flexible region. This region spans 11 contiguous positions (162-172) in the solvent exposed section of the RT loop.

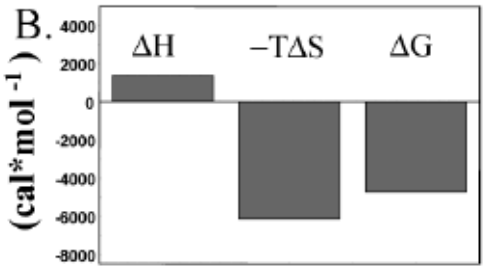
Fast protein dynamics.

Information about site-specific internal dynamics on the subnanosecond timescale derives primarily from model-free analyses. From such analyses, one obtains a set of parameters for each site (for example, the backbone amides) monitored, the most informative of which is the order parameter (S2). The analysis is termed model-free because the parameters are derived without the need to invoke a specific model for internal motion. The order parameter measures the magnitude of the angular fluctuation of a chemical bond vector (Yang and Kay 1996) such as the NH bond in a protein, and thus reflects the flexibility of the polypeptide chain at these sites. The magnitude of this parameter depends on interactions responsible for relaxing nuclear spins in proteins, including (i) magnetic dipole−dipole (ii) chemical shift anisotropy (CSA) and (iii)

electric quadrupole (from nuclei such as deuterium with spin quantum number I ≥1) interactions. To obtain quantitative information related to protein functions, such as the amplitude of structural variation or the configurational entropy of the polypeptide chain, it is necessary to determine the order parameter as accurately as possible FIGURE 26.

PREVIOUS STUDIES OF SEM5 C-SH3 BINDING AFFINITY

The free energy change upon binding was calculated with a simple model based upon changes in free energy as a function of ASA changes. The free energy change was calculated using the rigid protein and ligand of the crystal structure and simply taking the difference in which only C-SH3 was present and then when the Sos peptide was complexed as in the crystal structure (Lim, Richards et al. 1994). Calculated in this way, the enthalpy is slightly positive (unfavorable) and the entropy change due to solvation is markedly favorable. Since the overall free energy change is negative, this suggests that the burial of the hydrophobic residues in the binding cleft drives the reaction.



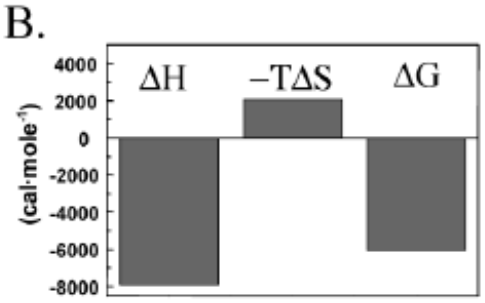
Ferreon et.al., Biochemistry (2003)

FIGURE 27: CALCULATED BINDING AFFINITY: Overall binding energetics calculated using surface area free energy function. The results are due solely to the change in the ASA of the SEM5:WT and Sos ligand as they are assembled at the orientation in the crystal structure complex.

The measured components of binding affinity are shown next. Here, the overall change in free energy is slightly more negative than calculated above, but the components of the free energy have exactly the opposite trends as shown above

Ferreon et.al., Biochemistry (2003)

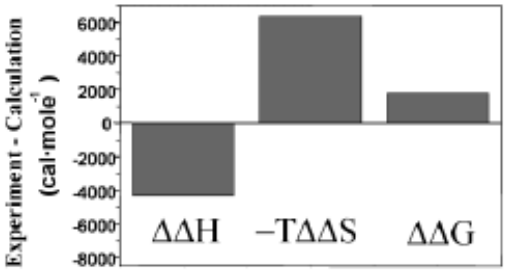
FIGURE 27. The enthalpy change upon binding is large and negative indicating that it is the driving force of binding and the entropy is negative in contrast to the simple calculation shown above.



Ferreon et.al., Biochemistry (2003)

FIGURE 28: MEASURED BINDING AFFINITY Overall thermodynamic parameters obtained from a fit of the data to calorimetric titration of the Sem-5 C-SH3 domain with the SosY peptide (Ac-VPPPVPPRRRY-NH2) in 20 mM Tris and 200 mM NaCl at pH 7.5 and 15 °C.

The difference in the calculated and the measured components of binding free energy change are shown below. This difference can be rationalized by considering many factors ranging from the validity of the energy function to the methodology used in the initial calculation. A principle goal of this work is to account for the individual components of free energy (see methods below) and explain the differences amongst these results.



Ferreon et.al., Biochemistry (2003)

FIGURE 29: DIFFERENCE BETWEEN EXPERIMENTAL AND COMPUTED: Overall binding energetics calculated using surface area free energy function. The results are due solely to the change in the ASA of the SEM5 and Sos ligand as they are assembled in the crystal structure complex.

Previous work on the C-Terminal SH3 domain includes calorimetric and NMR studies. The calorimetric studies establish the basis for the aggregate observed binding affinities and the corresponding changes in thermodynamic properties. The NMR studies attempt to experimentally quantify the conformational entropy of the flexible parts of C-SH3 FIGURE 26.

Surface Mutants Generated

In order to investigate the effects of changes in backbone rigidity on binding affinity, a four element Ala/Gly mutant cycle was generated. The mutations were made at two surface exposed positions that were known not to affect the topology of the bound complex. These surface exposed ala/gly mutants selectively modulated the conformational heterogeneity of the peptide backbone.

Justification for Approach

One of the objectives for these mutations was to induce subtle changes to the binding affinity of the system. Since the SH3 domain mediates the generation of a secondary signal by co-localizing other factors such as GNEF, excessive changes to its binding affinity may cause the system swing to the opposite extreme of its operational envelope.

Measurements Made

Calorimetric [ITC] binding assays were made to the four mutants FIGURE 30. A rank order of binding affinity was observed.

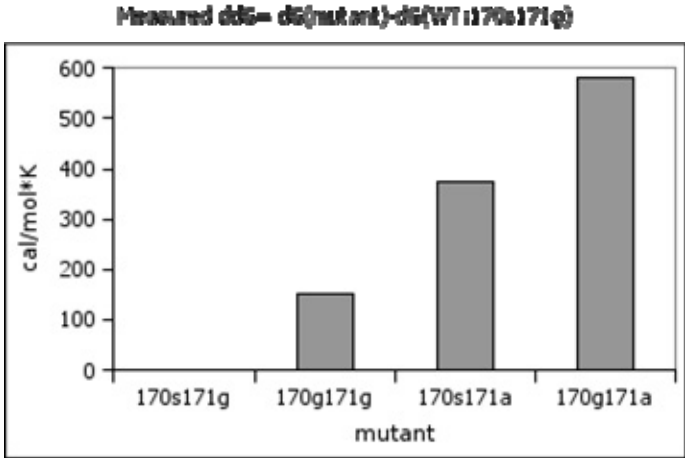


FIGURE 30: CALORIMETRIC BINDING TRENDS: Calorimetric binding assays with the putative SosY ligand were made with the four surface ala/gly mutants [170S171G, 170G171G, 170S171A, 170G171A] .

Hypothesis

PROTEIN DYNAMICS

The effect of protein fluctuations on molecular recognition is poorly understood. Prediction of useful properties such as binding affinity using rigid structures has produced sporadic success. Although attempts have been made to model the effect of fluctuations, capturing the impact of backbone relaxation has remained particularly elusive.

ENSEMBLE ORGANIZATION

The thermodynamic characteristics (binding affinity etc.) of the model system can be modeled and explained in terms of statistical mechanical ensembles. In order to investigate these effects, a series of surface exposed ala/gly mutants were designed in the flexible RT loop of the C-terminal SH3 domain of SEM5. One of the mutations was designed to perturb the ensemble of accessible conformations in the unbound ensemble and leaving the interaction surface with the ligand unchanged, while the other was designed to perturb both the interaction surface as will as the ensembles of bound and free conformations. The effects of these mutations were investigated by generating random conformations of the loop and performing principle component analysis was used to organize the randomly generated conformational states into a coherent landscape. To predict the effect of these mutations, we have developed a monte-carlo technique using a simplified energy function that only applied the effects of excluded volume and implicit solvation. This energy function was utilized to weight an ensemble of conformational states from which aggregate thermodynamic properties could be derived. The computed effects of the mutations on the binding affinity agreed with experimentally determined

values (R= 0.95) from isothermal titration calorimetry. The results indicate that the bound state of SEM5 SH3 domain contains a considerable repertoire of conformational variants of the high-resolution structure and that the determinants of binding cannot be elucidated from the static structure of the bound complex.

Specific Aims

SPECIFIC AIM ONE: FRAMEWORK TO STUDY ENTROPY

In order to estimate the entropy of a molecular system, one must first develop a framework sufficient to model and enumerate the microscopic states of the system. Since entropy can only be calculated (Hill 1960), a suitable **bottom-up** approach must be implemented. Such an approach must capture the molecular mechanical details necessary to calculate the ensemble based quantities needed to derive the bulk thermodynamic properties.

As discussed above, two essential approaches can be taken: 1) modeling of states with dynamics (molecular dynamics), 2) enumeration of states using Monte Carlo techniques.

SPECIFIC AIM TWO: LINK FRAMEWORK TO BULK THERMODYNAMICS THROUGH ENERGY FUNCTION

In order to link the mechanical states to thermodynamic states, an energy function is required. Such a function must be capable of linking mechanical/geometric/structural quantities to equilibrium thermodynamic properties. It is desirable that such a function be computationally tractable in order to handle the large number of states that are needed to reconstitute an equilibrium thermodynamic landscape.

SPECIFIC AIM THREE: APPLY TO SH3 MODULAR BINDING DOMAIN

The framework outlined in specific aims 1 and 2 will be applied to the SH3 recognition domain. The principle objective here is to model the binding process and attempt to rationalize select experimental binding affinity observations.

Methodology

STATISTICAL THERMODYNAMIC BASIS FOR BINDING AFFINITY

In order to calculate the binding affinity, one must perform the calculation at a suitable reference state. This reference state, once defined, allows the comparison of the affinities of diverse physico-chemical reactions.

Standard State

In chemistry, the **standard state** of a material is its state at 1 bar (100 kilopascals exactly). This pressure was changed from 1 atm (101.325 kilopascals) by IUPAC in 1990. The standard state of a material can be defined at any given temperature, most commonly 25 degrees Celsius. When the standard state is referred to in a chemical reaction, it also includes the condition that the concentrations of all solutions are at unity (or another designated quantity) for whatever measure of concentration is specified. For molarity that would be 1 $mol \cdot dm^{-3}$ and for molality 1 $mol \cdot kg^{-1}$. If mole fraction is used, the pure liquid or solid is the standard state (x=1). In biochemistry the pH is set to 7 (neutral physiologic conditions).

EQUILIBRIUM BINDING AFFINITY

The system was modeled using an equilibrium ensemble of states. This assumption was justified by the low molecular weight of the system components and the temperature

[298K] at which the experiments and simulations were undertaken. As a result, it was expected that the ergodic hypothesis would hold and that the conformational states would repeat over a "small" time scale.

Components of Free Energy

The binding process was broken down into several components (Finkelstein and Janin 1989; Murphy, Xie et al. 1994) each incurring a thermodynamic cost.

$$\Delta G_{binding} = \Delta E_{elec} + \Delta E_{mm} + \Delta G_{solv} - T\Delta S_{position} - T\Delta S_{orientation} - T\Delta S_{conf} \qquad (37)$$

where ΔE_{elec} is the change in internal electrostatic energy; ΔE_{mm} is the change in internal molecular mechanical energy; ΔG_{solv} is the change in solvation free energy; $\Delta S_{position}$ is the change in positional entropy; $\Delta S_{orientation}$ is the change the entropy of orientation; ΔS_{conf} is the change in conformational entropy. Positional and rotation entropy components account for the cost of aligning the protein-ligand complex. These are weakly dependent on molecular weight (Janin and Chothia 1978). This weak dependence allows these contributions to be subtracted out with little error when comparing the binding free energy from mutant to mutant. As a result the relative free energy changes are primarily due to the effects of conformational entropy and solvation free energy. If we also assume that the molecular mechanical energy is same for the ensemble of conformers and the electrostatic energy is implicitly contained in the structural parameterization of free energy, we find that difference in free energy of binding for each mutant i relative to the wild-type is:

$$\Delta\Delta G^{i,wt}_{binding} = \Delta\Delta H_{solv} - T\Delta\Delta S_{solv} - T\Delta\Delta S_{conf}; \qquad (38)$$

Dependence on Internal Degrees of Freedom

The central modeling strategy required enumeration of a representative set of conformational states, each having a free energy and statistical weight. The internal energy was calculated using a hard-sphere [excluded volume] approximation. States with steric overlap are assumed to have unrealistically high energy levels and are rejected. The electrostatic and solvation effects were modeled using a structural parameterization of free energy (Murphy and Freire 1992; Freire, Haynie et al. 1993; Freire and Xie 1994; Lee, Xie et al. 1994; Hilser, Gomez et al. 1996) which is based upon the solvent accessible surface area, with the accessible surface area being calculated using the method of Richards (Richards 1977).

Degrees of Freedom

The flexible region consists of 10-12 contiguous residue positions within the RT loop. Each position has 3+n [where n depends on the sidechain] torsional degrees of freedom. Combinations of torsion angles [known to produce self avoiding sidechains] are assigned from a rotamer table. These internal Bond Angle Torsion coordinates can be embedded in 3d space resulting in a corresponding Cartesian description. This description can be reduced by considering only the 3d positions of select atoms in the chain.

Constraint Satisfaction

The degrees of freedom within the model system are not independent. They are correlated by the physical constraints that apply to the system. Common constraints are bond length, bond angle and non-collision on the atomic level. Additional constraints such as positioning the ends of the model chain with respect to its protein context must also be enforced.

CONFORMATIONAL STATE GENERATION

It is impossible to exhaustively enumerate the conformational states of a system having many degrees of freedom using a nested traversal [grid based] technique (Lange, Lakomek et al. 2008). Instead, an approach that randomly samples this multi dimensional conformational space must be used. As these conformational states are generated, they are tested for self-avoidance and other constraints. The set of conformational states can then be ordered to form a map of the allowed region. The boundary of the allowed region will effectively allow the energetic character of the ensemble to emerge.

Mini Protein Modeler

Ensembles were created (see FIGURE 36 below) using a program called MPMOD [Mini-Protein MODeler]. It generates a set of random conformations that satisfy a range of constraints. Generating a self avoiding protein fragment requires two steps. First, a self avoiding backbone is generated using random phi-psi torsion angles. Second, randomly chosen sidechain rotamers are added such that self avoidance is maintained (FIGURE 36). For cases where a fragment must be inserted into the context of a protein, loop closure is

performed (FIGURE 36) ensuring that the N and C termini are positioned and oriented properly.

X-PLOR

In addition to MPMOD, other tools such as X-PLOR (Brunger, Adams et al. 1998) and Gromacs (van Buuren) were employed for conformer generation. MPMOD was used to generate the ensembles reported in this study.

Choosing the Flexible Region to Simulate

NMR order parameter measurements can indicate regions of conformational heterogeneity (on the ns/ps timescale) in the presence and absence of ligand (Yang and Kay 1996; Ferreon, Volk et al. 2003), and can be used to select and confirm an optimal region to simulate. Here positions with order parameter ($\mathbf{S^2 < 0.8}$) were considered to be flexible. Hydrogen exchange measurements further confirm the relative stability of a candidate region; with protection factors [PF] such that ($\boldsymbol{ln(PF)} < 1$) corresponding to surface exposed [exchange competent] regions. The NMR data suggest that positions 163-172 are heterogeneous. This suggests that a good chain to simulate would include this region and the positions on its sequence boundary (162, 173).

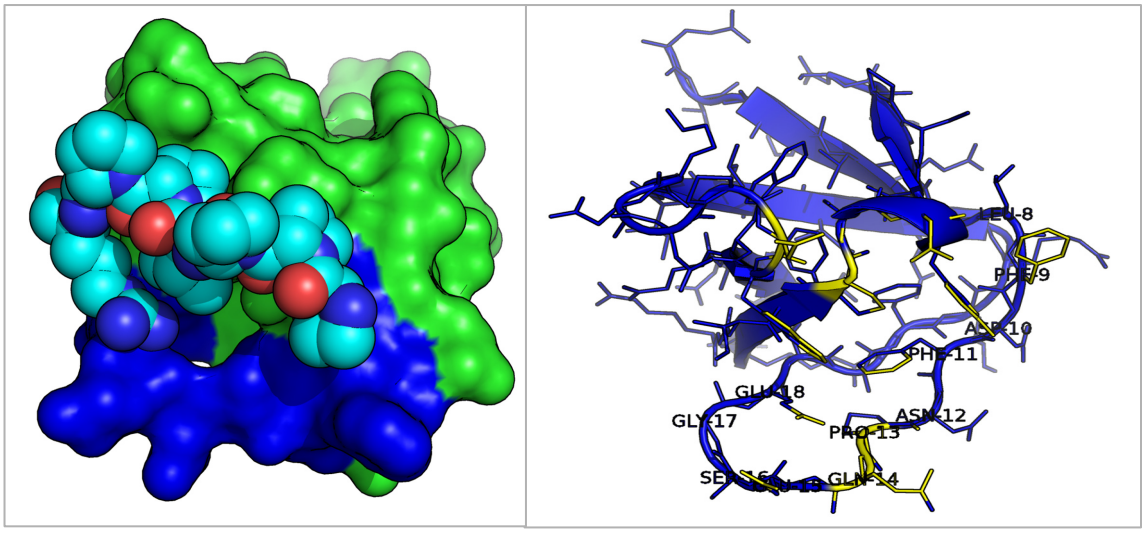


FIGURE 31: FLEXIBLE REGION: Space filling (left) and cartoon (right) views of flexible region of C-SH3. Left: The green region is the predominantly rigid region and the blue region is the flexible region of C-SH3. Right: The flexible positions are identified.

CHARACTERIZING THE ENSEMBLES

Aggregate Properties

The standard entropy change upon binding is the temperature derivative of the standard free energy of binding at constant pressure:

$$\Delta S^0 = -\left.\frac{\partial \Delta G^0}{\partial T}\right|_p ; \qquad (39)$$

The entropy change can be partitioned without approximation into a configurational part, ΔS^0_{config} which is associated with only the degrees of freedom of the protein and the

ligand, and the solvent part, ΔS_{solv}, which is the ensemble averaged change in the mean solvation entropy upon binding:

$$\Delta S^0 = \Delta S^0_{config} + \Delta S_{solv}; \qquad (40)$$

The conformational entropy can be partitioned into a conformational part which reflects the number of occupied energy wells(Karplus, Ichiye et al. 1987), and a vibrational part which reflects the average width of the occupied wells (Chang, Chen et al. 2007).

$$\Delta S^0_{config} = \Delta S_{conf} + \Delta S^0_{vib}; \qquad (41)$$

The vibrational part can also be thought of as the micro-degeneracy in the neighborhood of an energy minimum. In the case of M equally probable wells, the conformational entropy becomes $-R\ln(M)$; and when the states are not equally probable one must apply the general formula (Pettitt and Karplus 1988):

$$S_{conf} = -R\sum_{j=1}^{M} p_j \ln(p_j);$$
$$where: p_j = \frac{z_j}{Z_P}; \qquad (42)$$

is the probability of occupying energy state j and R is the gas constant. Z_p is the total partition function and each z_i represents a sub-partition function in the neighborhood of an energetic minimum.

The entropy for each ensemble [bound, unbound] is calculated and the aggregate difference is taken to get the change upon binding.

Characterizing the Thermodynamic Landscape

Each mutant was analyzed according to an orderly progression of energetic and interaction effects. First, the self-avoiding conformational states [un-weighted by solvation energy] were analyzed using statistical techniques. Second, solvation and other energetic effects were incorporated; effectively applying free energy based statistical weights to the ensemble members. Third, the interaction with the ligand was added in order to predict the binding affinity which could then be correlated to the corresponding experimental measurements.

The Unweighted Conformational States (Principle Component Analysis)

Principle component analysis (Fukunaga and Keinosuke 1990) takes a multidimensional data set and reduces it to a set of characteristic dimensions reflecting the intrinsic variation within the data set (Cullen 1972). In a system having many degrees of freedom, this technique effectively reduces and correlates these degrees of freedom to a new, more descriptive, basis FIGURE 32. In favorable instances, it allows the majority of the variation to be reduced to a manageable number of components and acts to intrinsically order a random set of conformational states such that they may be organized into a novel landscape. This ordering, which is based on the Euclidean distance between each conformational state, can provide a much greater level of transparency.

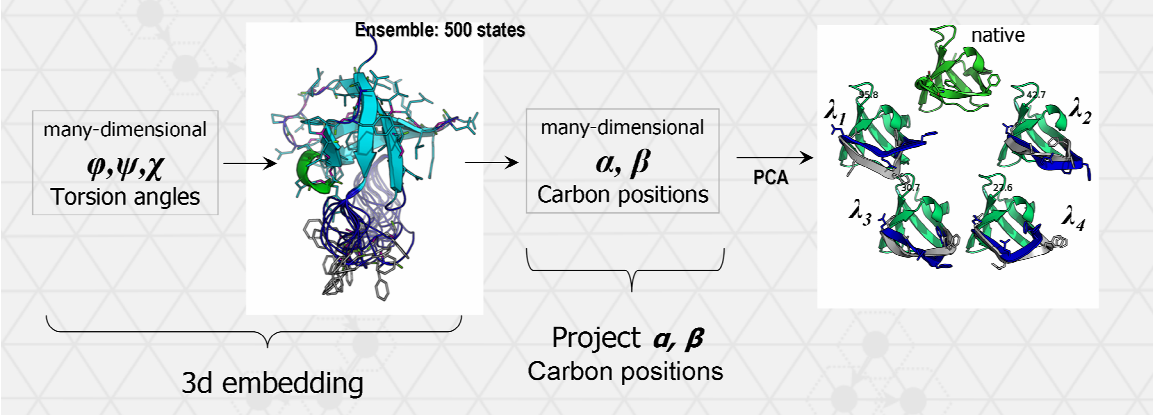


FIGURE 32: PRINCIPAL COMPONENT PIPELINE: 3D embedding through ensemble generation followed by projection of α and β carbon positions which are feature vector to PCA.

Conditions of Applicability

Feature Vector

If the distribution of microstates along each principle axis is Gaussian, then the [linear] technique of principle components is meaningful. If the eigenvalues, sorted in descending order, decrease rapidly in magnitude, one can characterize the system by using the first few principle components (FIGURE 32). The technique is summarized by the following set of equations:

$$\boldsymbol{c_i} = [\boldsymbol{r}_1, \ldots, \boldsymbol{r_n}] \{\boldsymbol{i} \in 1, \boldsymbol{n}\}; \qquad (43)$$

Where $\boldsymbol{c_i}$ is the feature vector of the flexible chain; each $\boldsymbol{r_j}$ is a component of the 3d vectors describing the cartesian positions of alpha and beta carbons and the cardinal

index ***j:1,m*** spans the contiguous residues of the flexible subchain and ***i*** indexes the conformers within the ensemble.

This equation shows a feature vector suitable for describing a relevant geometric aspect of the system. In this study, the cartesian positions of alpha and beta carbons were used. This metric captures the position of the residues and their orientation along the backbone. Each microstate in the ensemble was characterized by a corresponding feature vector. A similar approach using backbone dihedral angles was employed in (Sims, Choi et al. 2005). Direct use of the dihedral angles, however, becomes non-linear when the number of residues exceeds 4-5 (Sims, Choi et al. 2005). Use of cartesian positions maintains linearity for a larger number of residues provided that the topology of the segment context is maintained.

Metric Distance Function

For principle component analysis to work, the "distance" between data points must satisfy the criteria of a metric distance function. A metric distance function on a *space* ***M*** measures the distance ***d(x,y)*** between different feature vectors and is defined by the conditions:

For all ***x, y*** in ***M*****:** ***d*(*x*, *y*) >= 0**, with equality if and only if ***x* = *y***.

For all ***x, y*** in ***M*****:** ***d*(*x*, *y*) = *d*(*y*, *x*)**.

For all ***x, y, z*** in ***M*****:** ***d*(*x*, *z*) <= *d*(*x*, *y*) + *d*(*y*, *z*);** this is the <u>*triangle inequality*</u>.

$$\boldsymbol{d}_{i,j} = \left(\sum_{k} (\boldsymbol{r}_k^i - \boldsymbol{r}_k^j)^2 \right)^{1/2}; \qquad \textbf{(44)}$$

where $d_{i,j}$ is the Euclidean distance between conformers i and j; $i,j=1,n$; $k=1,m$ is the cardinal index of the components of the feature vector c_i defined above.

Distance Matrix

The distance matrix captures the distances between all the feature vectors of the ensemble. It epitomizes the intrinsic relationships amongst the members of the ensemble up to their overall chirality (Crippen and Havel 1988).

$$\boldsymbol{D}_{n,n} = \begin{bmatrix} d_{1,1} & \cdots & d_{1,n} \\ \vdots & \ddots & \vdots \\ d_{n,1} & \cdots & d_{n,n} \end{bmatrix}; \qquad (45)$$

where each $d_{i,j}$ is the distance from conformer i to conformer j for n conformers in the ensemble.

This distance matrix can be diagonalized using the eigenvalue decomposition method from linear algebra (Cullen 1972). The transformed vector space applies an affine rotation such that the new orthogonal basis vectors [eigenvectors] maximize the variances of the original data set (Cullen 1972).

$$\boldsymbol{D}_{n,n} = \boldsymbol{PAP}^{\boldsymbol{T}};$$

$$\boldsymbol{A} = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}; \qquad (46)$$

where $\boldsymbol{A}$ is the diagonal matrix of eigenvalues $\boldsymbol{\lambda_i}$ $\boldsymbol{[i=1,n]}$ and $\boldsymbol{P}$ is the matrix of eigenvectors and $\boldsymbol{P^T}$ is its transpose.

The strength of the principle component technique in this context is that it captures the coordinated variation of complex structural transformations. Small variations within the principle space will correspond to non-trivial coordinated displacements of the internal molecular coordinates [torsion angles].

The principle coordinates can be used to map the un-weighted conformational states into a thermodynamic landscape. These states can be arranged in a space of reduced dimension if the largest components capture most of the variation.

THE WEIGHTED CONFORMATIONAL STATES

The un-weighted states could be resolved using principle components due to their predominantly linear distribution. However, the corresponding Boltzmann weighted states cannot be similarly treated owing to a clustered [multi-modal] expression. Here, the distribution of the fractional occupancy of states within the allowed region [when plotted within the principle space] will tend to be characterized by tightly localized clusters (FIGURE 43). These clusters correspond to regions where the solvation free energy is minimized as a function of the principle coordinates. Therefore, these energetic effects will have to be portrayed using plots of the fractional occupancy versus the dominant principle coordinates (FIGURE 43).

Boltzmann weighting is accomplished by using the standard expression:

$$e^{-(E(r))/RT};$$

Where E(r) is the free energy which is the sum of all its energetic components and $\boldsymbol{r}$ is a vector describing the conformation of the solute.

$$E(r) = U(r) + W(r);$$

Here U(r) is the internal mechanical energy and W(r) is the free energy due to solvation. Combining these expressions gives:

$$e^{-(U(r)+W(r))/RT};$$

This expression has the characteristic of taking on an appreciable value when **both** U(r) and W(r) have "low" values. In this sense, this expression functions as a type of NAND gate commonly seen in Boolean logic circuits. In our formulation, U(r) is taken to be zero for a self-avoiding peptide and infinity when a steric collision occurs. W(r), on the other hand, has a low value when the solute has a favorable surface area profile. For a state r to have a high weight, **both** energy values must be low. The subspace where U(r) is zero is simply the allowed space. This allowed space, when combined (NANDED) with the solvation favorable space, yields the physically probable landscape.

DESCRIBING THE IMPACT OF THE LIGAND

In order to determine the thermodynamic impact of the ligand, an ensemble based binding process was applied to each mutant (FIGURE 25). Briefly, an ensemble of unbound conformational states was first generated; followed by docking of the ligand to each of these states and calculating the resulting free energy of the protein-ligand complex. This defined a two partition landscape (Camacho, Weng et al. 1999) in which

one partition contains the unbound conformational states, and the other, disjoint partition, contains the bound [complexed] states. Calculation of the binding affinity is performed by computing the aggregate free energy difference of these two partitions.

FORMULATION OF BINDING FREE ENERGY CHANGE

Chemical Potential of Bound and Unbound Partitions

The equilibrium binding process can be modeled by using the chemical potentials of two partitions corresponding to the [bound, unbound] phases (Gilson, Given et al. 1997; Gilson and Zhou 2007). The chemical potential μ_X for a given partition [phase] X is given by the standard formula:

$$\mu_X = -RT\ln(Z_x); \qquad (47)$$

$$Z_x \equiv \int J(r_{\text{int}}) e^{-E_x(r_{\text{int}})/RT} dr_{\text{int}}; \qquad (48)$$

$$E_x(r_{\text{int}}) = U_x(r_{\text{int}}) + W_x(r_{\text{int}}); \qquad (49)$$

Here μ_X is the chemical potential for species X where $X = PL, P \text{ or } L$; [P= protein, L=Ligand] R is the gas constant and T is absolute temperature. Z_X is the partition function for species X .ie. the configuration integral over the internal coordinates of solute species X. $J(r_{\text{int}})$ is the Jacobian determinant [for mapping internal to external coordinates]; and $E(r_{\text{int}})$ is the sum of the internal *(U)* and solvation free energy *(W)* of the molecule or its complex as a function of its conformation.

ENSEMBLE BASED BINDING FREE ENERGY CHANGE

General Formulation

The binding free energies can be modeled in terms of changes in chemical potentials. These chemical potentials correspond to the bound [protein-ligand] phase and the free protein and free ligand phases.

$$\Delta G_{bind} = -RT\ln\left(K_{bind}\right); \qquad (50)$$

$$\Delta G_{bind} = \mu_{PL} - \mu_P - \mu_L; \qquad (51)$$

These chemical potentials can be expressed in terms of the partition functions for each of the separate phases. Within a baseline of N_s solvent molecules and N_P solute molecules , the free energy change due to addition of a single solute molecule is given by (Gilson, Given et al. 1997):

$$\begin{aligned} \mu_P &= -RT\ln\left(\frac{Q_{N_P+1,N_s}}{Q_{N_P,N_s}}\right); \\ &= -RT\ln\left(\frac{8\pi^2}{\sigma_P C_0} Z_P\right); \end{aligned} \qquad (52)$$

where $Q_{Np,Ns}$ is the partition function of Np solute molecules in a solution of Ns solvent molecules; σ_P is the symmetry number for species P ; C_0 is the standard concentration; Z_P is the partition function for species P which can be expressed as:

$$Z_P = \int e^{-(U(r_P)+W(r_P))/RT} dr_P; \qquad (53)$$

Here, the multidimensional vector r_P spans the conformational space .ie. each r_p corresponds to a distinct conformer for species P.

Specialization

If one uses an implicit solvation approximation for $W(r_p)$, the above set of equations simplify to:

$$\Delta G_{bind} = -RT\ln\left(K_{bind}\right); \qquad \Delta G_{bind} = \mu_{PL} - \mu_P - \mu_L; \qquad (54)$$

$$\mu_P = -RT\ln\left(\frac{8\pi^2}{\sigma_P C_0} Z_P\right); \qquad (55)$$

The calculation of the partition function, however, will be done using a discrete approximation using the set of randomly generated states:

$$Z_P \approx \sum_{j=1}^{M} z_j; \qquad z_j \equiv \Delta vol_{gen} * e^{-(U(r_{gen})+W(r_{gen}))/RT}; \qquad (56)$$

The vector r_{gen} spans the conformational space .ie., each r_{gen} corresponds to a distinct conformer. The total partition function Z_P is approximated by a set of disjoint sub-partitions z_i. Each z_i corresponds to the integral of the sub-partition neighboring a randomly generated conformation and distributes the Boltzmann weight throughout the surrounding hyper-volume Δvol_{gen}.

The calculation of the free energy of binding is summarized in the following figure:

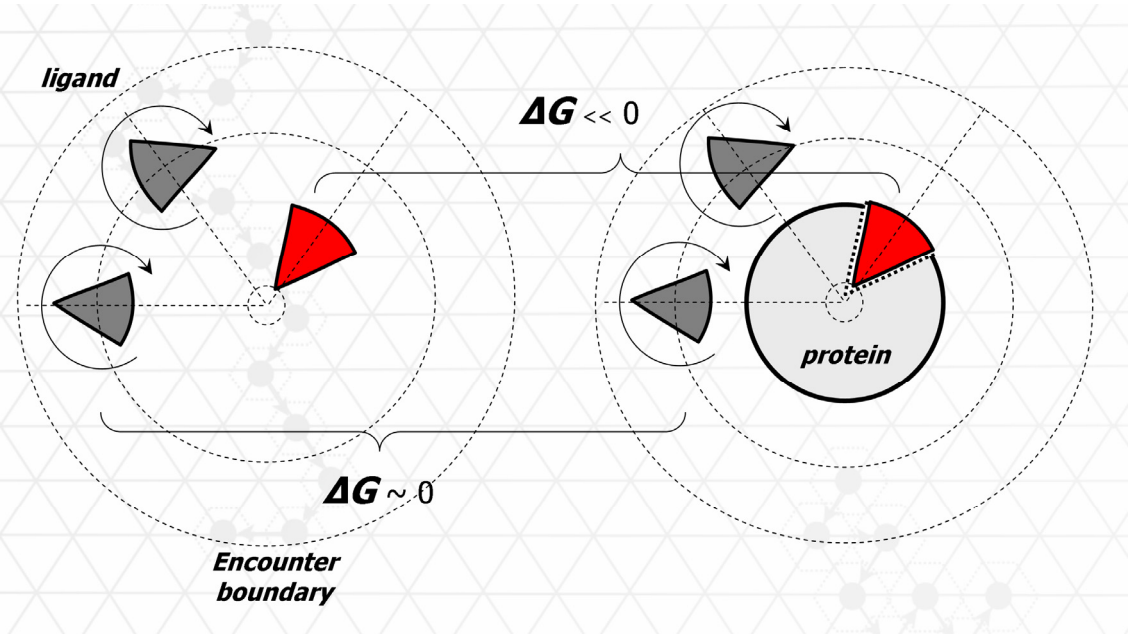


FIGURE 33: CALCULATION OF FREE ENERGY CHANGE UPON BINDING: Left: shows the unbound ensemble for the ligand. The free energy is calculated within a spherical coordinate frame where the free energy is summed for all radii, and roll pitch and yaw of the ligand relative to the origin. The calculation is repeated with the receptor protein present. Each summation is the partition function of the respective macrostates (unbound, bound). The red orientation shows the optimal position of the ligand in the binding pocket. Due to the large buried surface area, this orientation will make a dominant contribution to the bound partition.

OPTIMAL BINDING ORIENTATION

Scoring Complex Formation Using FTDOCK

Docking simulations were performed using the system FTDOCK. A total of 10000 orientations were generated and the position resulting in the highest surface complementarity was found for each. The resulting complexes were then rescored using the residue pair potential matrix. The docking orientation of the crystal structure was found to have the highest score by a factor of 100 FIGURE 34.

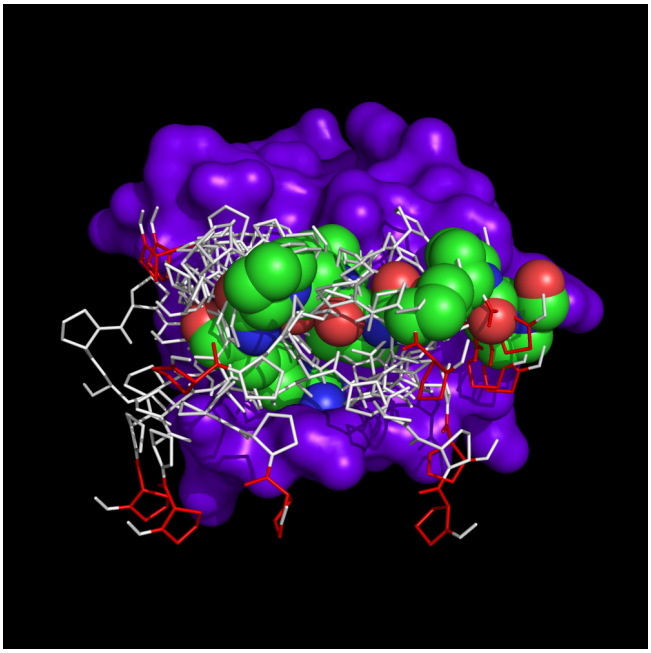


FIGURE 34: MOST FAVORABLE DOCKING ORIENTATION: Space filling ligand shows orientation with highest residue pair potential score.

CORRELATION TO EXPERIMENT

ITC relative binding affinities

Isothermal Titration Calorimetry measurements were used to establish the relative binding free energies for a series of surface Ala/Gly mutants (Ferreon, Hamburger et al. 2004).

Results for the SEM5 C-SH3 System

Conformational changes have been observed to contribute prominently to the energetics of binding between the SH3 domains and their putative binding partners (Ferreon and Hilser 2003; Ferreon, Volk et al. 2003; Bauer and Sticht 2007). In order to quantitatively determine the impact of this conformational heterogeneity on binding affinity, a model system was devised with the goal of simulating and ultimately predicting the effect of these fluctuations on binding. The model system investigated in this work was the SEM5 C-Terminal SH3 domain (C-SH3) which is shown in (FIGURE 35) (Lim, Richards et al. 1994; Ferreon, Volk et al. 2003). Common among the SH3 domains, the binding site in SEM5 C-SH3 is situated between the core of the β sandwich structure and the highly flexible RT loop (named for the arginine and threonine residues occurring within the loop). SEM5 C-SH3 recognizes a stretch of prolines in the son of sevenless (Sos) protein that adopt the polyproline II (P_{II}) conformation (Yu, Chen et al. 1994). Previous NMR studies in our laboratory have revealed that only a subset of the positions in SEM5 that

are located in the SH3 binding site are affected conformationally upon interacting with the Sos ligand (Ferreon, Hamburger et al. 2004) (FIGURE 35).

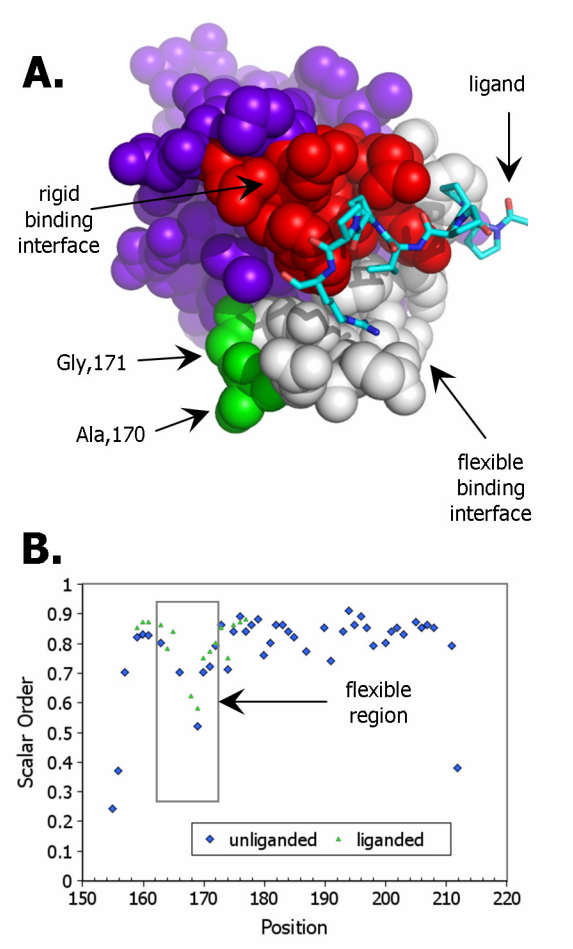


FIGURE 35: IDENTIFICATION OF DYNAMIC REGION OF BINDING POCKET: A. View of binding cleft of C-Terminal SH3 domain with rigid [red] and flexible [white] regions. Mutated positions shown at tip of RT loop. B. NMR order parameters versus position number. Positions 162-172 show low order parameters indicating conformational heterogeneity.

Positions in this subset are generally located in the RT loop near the cleft of the binding site. Because the NMR-derived order parameters near the binding interface change upon binding, it follows that the conformational manifold of the RT loop is modulated by the binding process. These changes in degeneracy of the RT loop can affect the entropic contributions to binding affinity (Ferreon, Hamburger et al. 2004). Here we use a combined experimental and computational strategy that changes the conformational manifold of the RT loop upon binding. Our goal is to introduce a well-understood set of perturbations (Matthews, Nicholson et al. 1987; Caves, Evanseck et al. 1998), and to use the effects on the binding affinity to elucidate the impact of the mutations on the conformational ensemble.

Our approach is to target for mutation two surface-exposed residues (i.e. the side chains of each residue project into solution), where the sidechain makes no contact with ligand in the bound state. The resulting mutations therefore should have little effect on the structure of the protein in its bound conformation (although, see below). In fact, because the strategy involves Ala to Gly mutations at each site, the primary effect should be in modulating the conformational ensemble. Specifically, our approach allows us to directly evaluate the extent to which fluctuations resemble order/disorder transitions (Tsai, Ma et al. 2001), wherein significant backbone relaxation processes are observed.

Our experimental approach was twofold. First we designed double Ala and Gly mutants at positions 170 and 171, and measured the binding affinity using isothermal titration calorimetry. Second the effects of the mutations were modeled using a computer algorithm specifically designed to sample large-scale backbone relaxation processes

(FIGURE 36). Positions which showed low-order parameters (< 0.8) in our previously published relaxation studies and were presumed to be conformationally heterogeneous, (Positions 162-172: See FIGURE 35B) were allowed to randomly sample alternative backbone phi, psi and sidechain torsion angles. The resultant conformational ensembles for the various single and double Gly and Ala variants (i.e. G171A170, G171G170, A171A170, A171G170) were generated. The result for each mutant was a conformational ensemble, wherein alternative conformations of the flexible part of the RT loop were generated. By evaluating the computed ability of each state to bind ligand, it should be possible to evaluate the physical basis for the differences in binding affinity for each mutant as well as the role of conformational heterogeneity in the binding process.

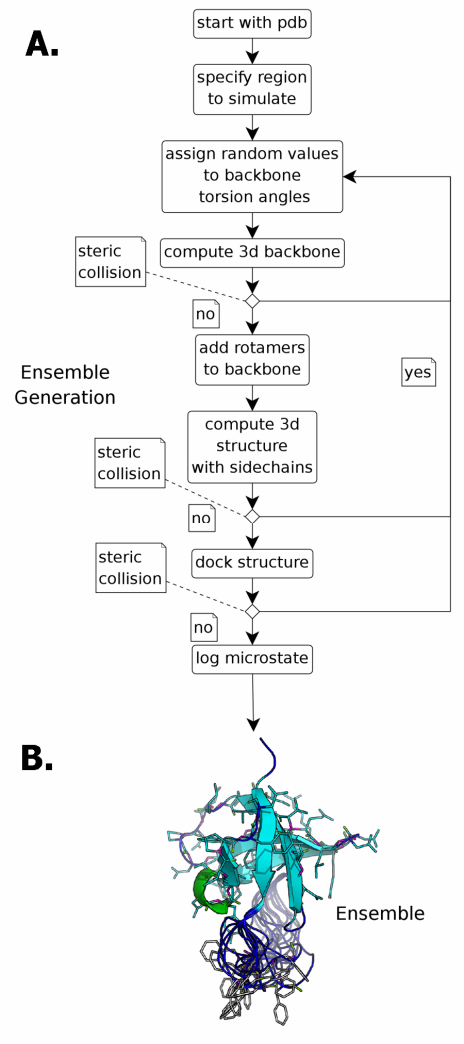


FIGURE 36: CONFORMATIONAL ENSEMBLE GENERATION STRATEGY: A. Flowchart showing strategy for [Hard Sphere Collision] HSC ensemble generation through random assignment of backbone torsion angles and sidechain rotamers followed by calculation of 3d embedding and tests for self avoidance. B. Overlay of 20 conformations randomly selected from among 500 successful conformations. The conformational diversity of the RT loop is evident.

THE RT LOOP CONFORMATIONAL ENSEMBLE

Conformational ensembles of the RT loop were generated using the hard sphere collision (HSC) model (Ramachandran, Ramakrishnan et al. 1963; Richards 1977; Lee, Xie et al. 1994; Daquino, Gomez et al. 1996) (FIGURE 36). Several features were apparent upon direct inspection of each ensemble. First, the RT loop ensemble exhibited a rich (i.e. spatially diverse) conformational manifold (FIGURE 36).

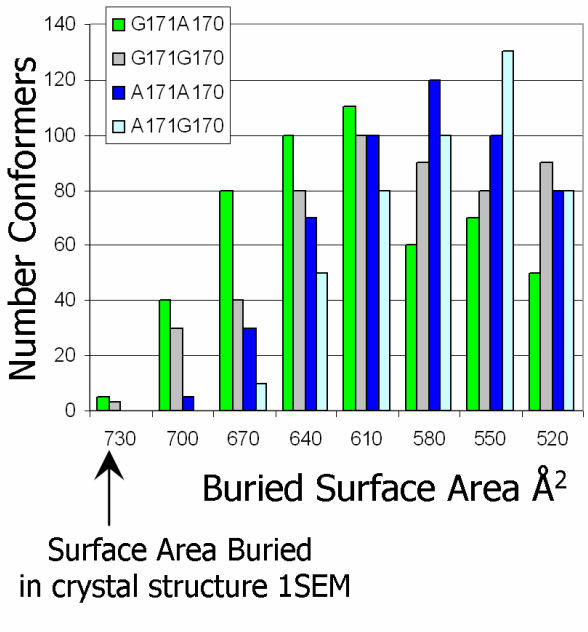


FIGURE 37: RECEPTOR:LIGAND SURFACE COMPLEMENTARITY: Histogram showing surface area burial distribution for four mutants. Maximum burial is found to occur for bound crystal structure. Buried surface area decreases for all mutants indicating sub-optimal surface burial for majority of each ensemble.

Second, the ensemble showed a remarkable absence of conformations in which the RT loop overlapped with the volume that is occupied by the ligand in the bound complex.

This latter result suggests that most conformations tend to enlarge the binding pocket, producing a greater number of states with sub-optimal receptor-ligand surface complementarity (insert). Further conclusions required the use of statistical analyses.

To classify and quantify the conformational ensembles of each mutant, we applied principal component analysis (PCA) to the cartesian positions of alpha and beta carbons, a procedure similar to that previously employed in the analysis of dihedral angles (Sims, Choi et al. 2005). Although direct use of the dihedral angles becomes non-linear when the number of amino acids exceeds 4 or 5 (Sims, Choi et al. 2005), we show here that use of cartesian positions maintains linearity for a larger number of amino acids provided that the topology (contact profile) of the segment context is maintained (Plotkin, Wang et al. 1997; Shea, Onuchic et al. 2002). The suitability of PCA for the analysis of conformations in the RT loop was determined by two criteria. First, the histogram of the microstates along each principal axis was found to have a single mode Gaussian appearance (see supplemental material). In other words, the distribution of data points within the principal space itself took on a predominantly ellipsoidal shape, expressing only one maximum along each principal axis. Second, the first three principal components expressed >90% of the total variation. This condition allowed for an acute reduction in the dimensionality of the phase space. To illustrate this, the envelope of the allowed region has been plotted against the first three principal components (FIGURE 40A).

Inspection of the principal component space of the first three eigenvectors revealed differences in the envelope of the allowed region between the ensembles for the mutants

with highest and lowest binding affinity (i.e. 170A171G and 170G171A, respectively) (FIGURE 40A).

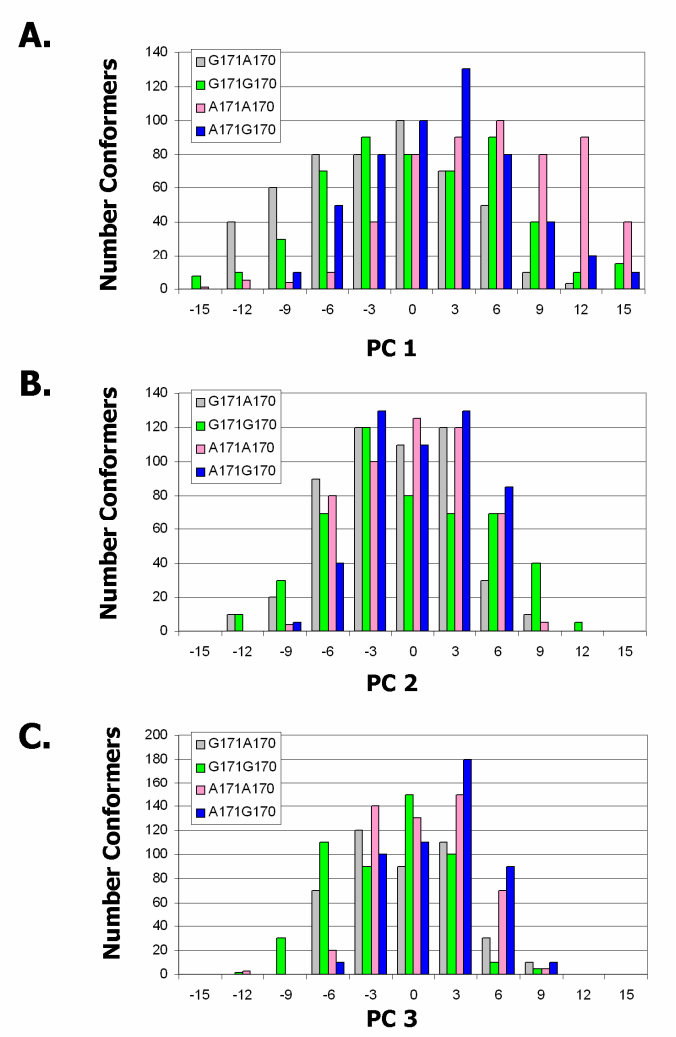


FIGURE 38: GAUSSIAN DISTRIBUTION OF CONFORMERS: Histograms showing distribution of conformers along each principal axis. A: distribution along most dominant principal axis, B: second most dominant axis, C: third most dominant axis.

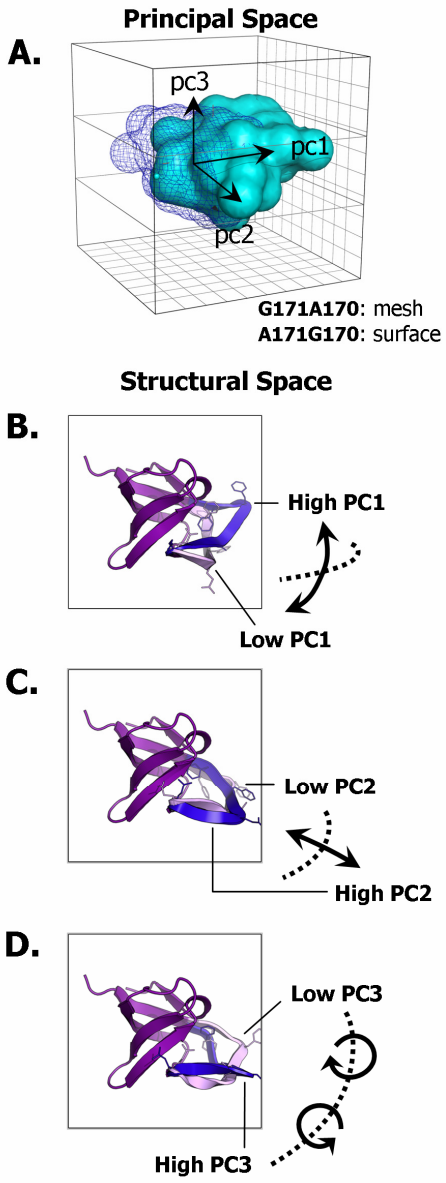


FIGURE 39: MODES OF VARIATION OF FIRST THREE PRINCIPAL COMPONENTS: A: Plot of mutants having highest and lowest binding affinity in principal space. B: Mode of variation of PC 1, C: Mode of PC2, D: Mode of PC3 in structural (3d space).

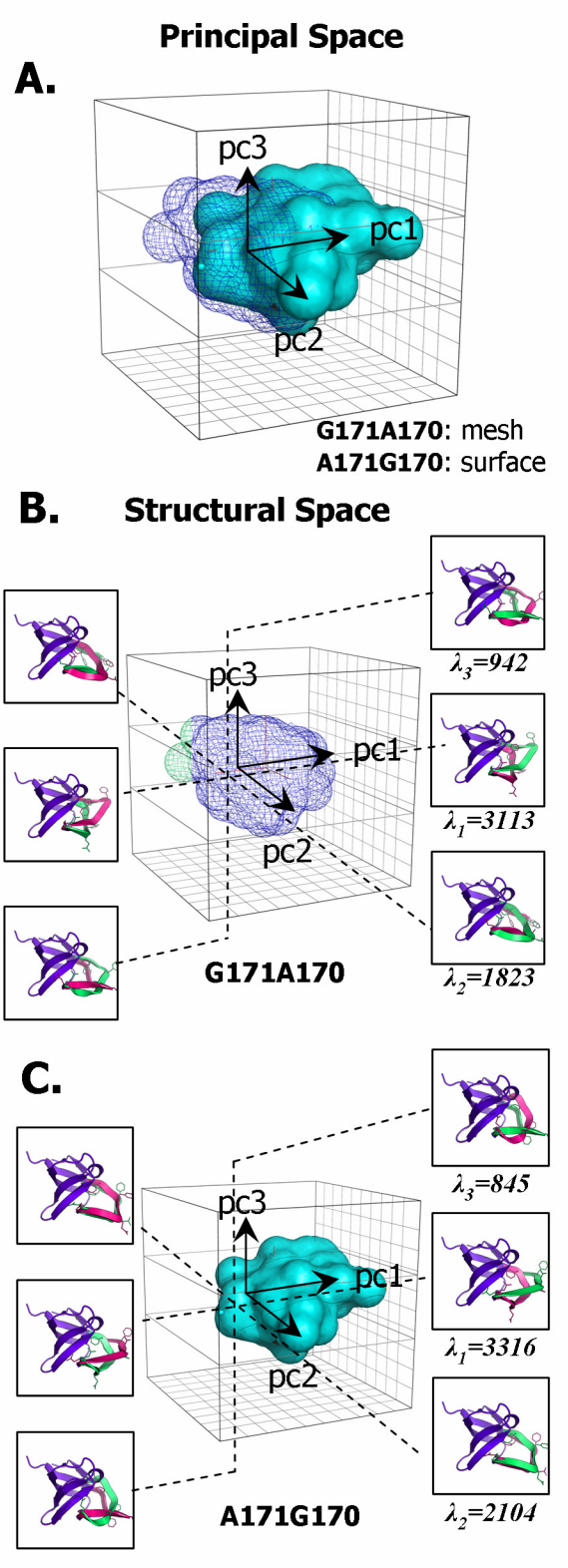

Principal Space
A.
pc3
pc1
pc2
G171A170: mesh
A171G170: surface
B.
Structural Space
pc3
pc1
pc2
G171A170
λ3=942
λ1=3113
λ2=1823
C.
pc3
pc1
pc2
A171G170
λ3=845
λ1=3316
λ2=2104

FIGURE 40: VOLUMETRIC PLOT OF ENSEMBLES IN PRINCIPLE CONFORMATIONAL SUB-SPACE for the RT loop of SEM5: A. Comparison of conformational envelopes in principle space [showing first three principle coordinates] of mutants with the highest [blue mesh] and lowest [cyan surface] binding affinity. Variances and corresponding structural extrema (see text) along principle axes of the first three principle components [$\lambda_{1\text{-}3}$] for the; B) highest affinity [.ie.wt:G171A170] and; C) lowest affinity [A171G170] mutant. The structures with the gray flexible region correspond to the highest extreme along the given eigenvector and conversely for the structures with the blue flexible region.

Each eigenvector described a concerted conformational variation of the RT loop. In order to illustrate the physical significance of each eigenvector, the conformers at both extremes along each eigenvector were determined (FIGURE 40B-C). These extreme structures reflect the breadth of the conformational envelope along the given eigenvector and can be interpreted as one mode of concerted variation. Importantly, the three eigenvectors were found to vary only subtly between the different mutants. For instance, the structures of the first eigenvector for each mutant exhibited a backbone displacement orthogonal to the path of the native chain (FIGURE 40B-C). The second eigenvector exhibited an "in and out" scissor type variation along the chain, toward and away from the centroid of the protein, and the third eigenvector showed a twisting variation along the chain. These similarities resulted from the relatively subtle backbone conformational perturbations introduced through the surface Ala/Gly mutations. Despite these similarities, however, there were (as expected from the fact that G171 is in a disallowed

region of φ, ψ space) perturbations to the binding surface that affected changes in binding affinity.

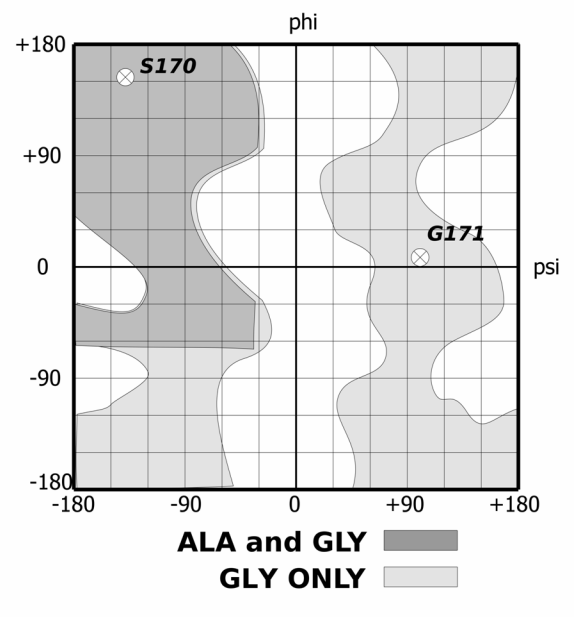
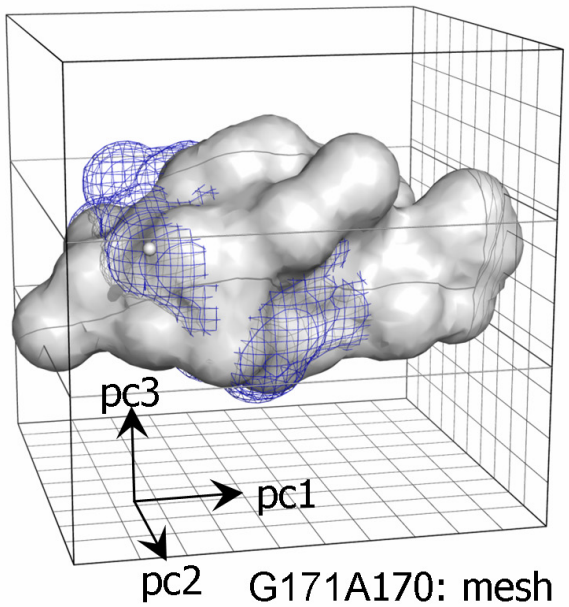


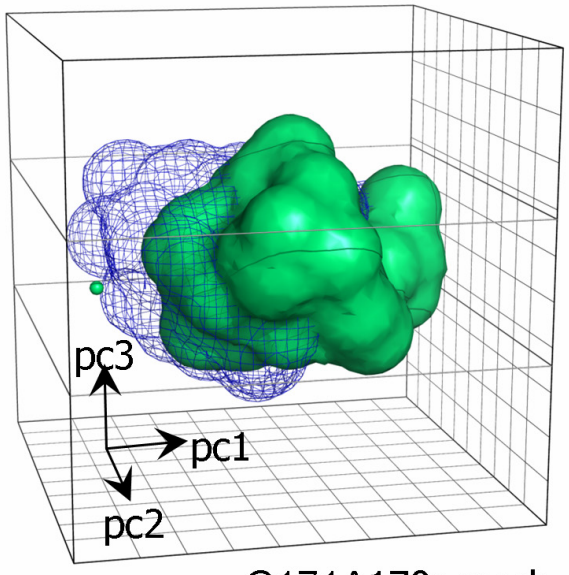
FIGURE 41: RAMACHANDRAN PLOT DEPICTING ALLOWED SPACE OF ALA AND GLY: Comparative plots of the allowed regions for Ala and Gly illustrating the difference in allowed space for a single amino-acid position. The phi/psi coordinates of positions 170 and 171 found within the crystal structure for the wild-type are labeled. Note that position 171 falls within the positive phi region for the complexed crystal structure.

A. Degeneracy



G171A170: mesh
G171G170: surface

B. Conformational "strain"



G171A170: mesh
A171A170: surface

FIGURE 42: COMPARATIVE VOLUMETRIC PLOTS OF ALLOWED SPACE OF FLEXIBLE REGION Within Principle Component Space: The allowed regions for select pairs of mutations. These plots show the cooperative allowed space for a set of contiguous flexible positions. A. Ala to Gly at position 170. Glycine residue in position 170 expands the allowable region. The A170 mutant is shown in blue mesh and G170 mutant is shown using white surface. B. Gly to ala mutation in residue 171. Ala residue in position 171 shifts allowable region in principle space, reflecting mechanical bias.

Conformational Ensembles in Principle Component Space

As described above, conformational heterogeneity can be modulated through Ala/Gly mutations. To understand the effects of such mutations, we note that the absence of a β-carbon in glycine dramatically increases the accessible φ and ψ space for a random peptide (Ramachandran and Sasisekharan 1968) (FIGURE 41). As such, an unfolded peptide or protein with a Gly at a particular position will be expected to have more accessible unfolded conformations (~3.4) (D'aquino, Hilser; 1996), relative to that same protein or peptide with an Ala at that position. For the two positions selected for mutation in this study, two different scenarios are expected. First, according to the high-resolution X-ray and NMR structures of the apo and holo protein (Lim, Richards et al. 1994; Weng, Rickles et al. 1995; Ferreon, Volk et al. 2003), the φ and ψ angles for A170 are in a region that is accessible to both Ala and Gly (i.e. φ = -140, ψ = 150) (FIGURE 41). In the second case, the φ and ψ angles for G171 are in a region that is only accessible to Gly. Consequently mutation at position 171 to a residue type other than Gly cannot accommodate the high-resolution structure of the WT protein without introducing

a steric overlap. Such a mutation would induce conformational strain on the ensemble and would shift the allowable space. Learning how the system responds to this mutation is one of the goals of the current study. In summary, the double Ala to Gly mutation strategy allowed us to simultaneously evaluate the effects of expanding and shifting the conformational ensemble with select positions having distinct effects on the distribution of states (FIGURE 42B).

BOLTZMANN WEIGHTING OF THE ENSEMBLE

The principal component analysis of the randomly generated states described above suggests that our analysis has adequately covered the allowable conformational space. In order to evaluate the impact on the ensemble probability, each state of the enumerated ensemble for each mutant was weighted by its free energy. To energetically weight each state, an empirical model of free energy was used that relates enthalpy, entropy and heat capacity changes to changes in solvent accessible surface area (ASA) (Murphy and Freire 1992; Freire, Haynie et al. 1993; Freire and Xie 1994; Murphy 1994; Xie and Freire 1994; Xie and Freire 1994). Although conceptually simple, this surface area based energy function has been employed successfully for the analysis of numerous phenomena and in spite of this simplicity, has proven remarkably robust.

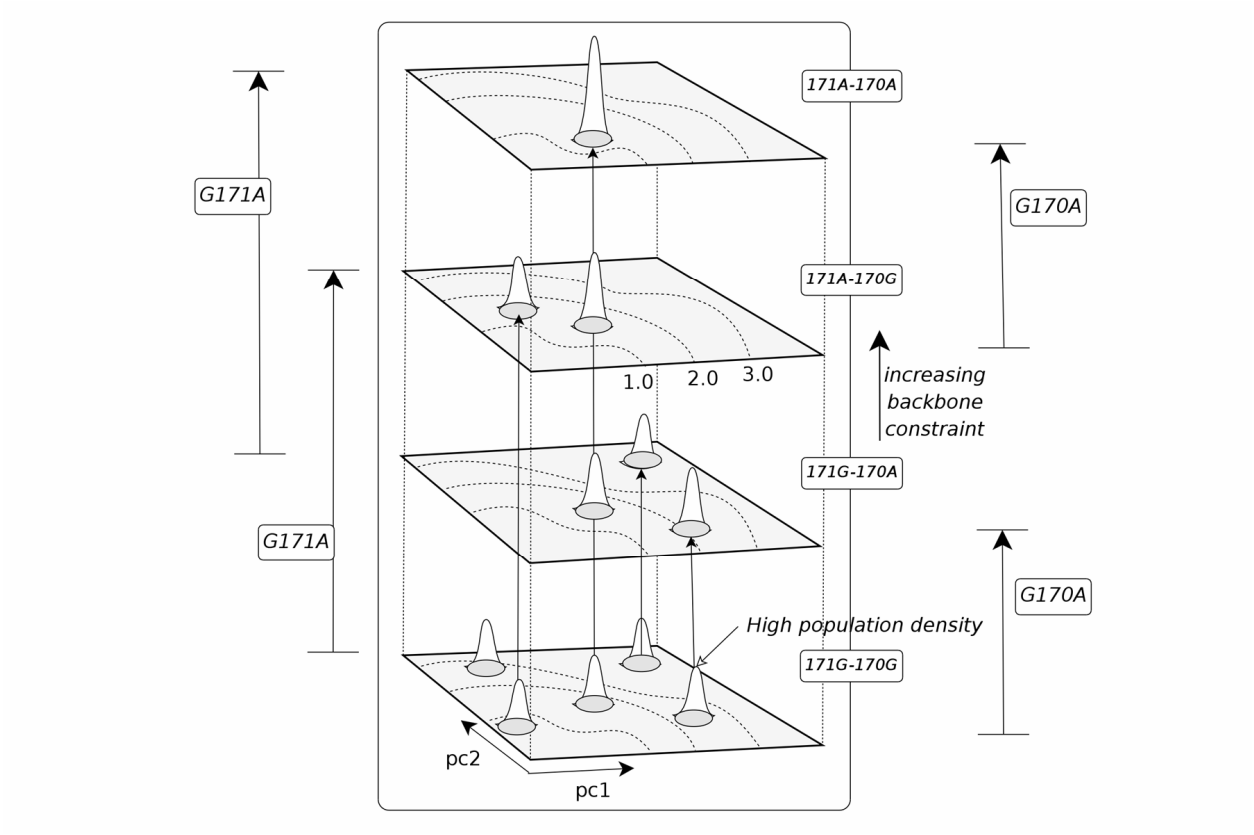


FIGURE 43: COMPARISON OF LANDSCAPES OF UNBOUND ENSEMBLES: The fractional occupancy [Boltzmann probability] plots against the first two principle components are shown one of top of the other. The plots are ordered from the most constrained mutant G171A170 [top] to the least constrained G171G170 [bottom]. Energetically favorable clusters are seen to appear and disappear as the allowable region is changed through mutation. Contours in each two dimensional plane represent increasing structural distance from crystal structure complex located at origin.

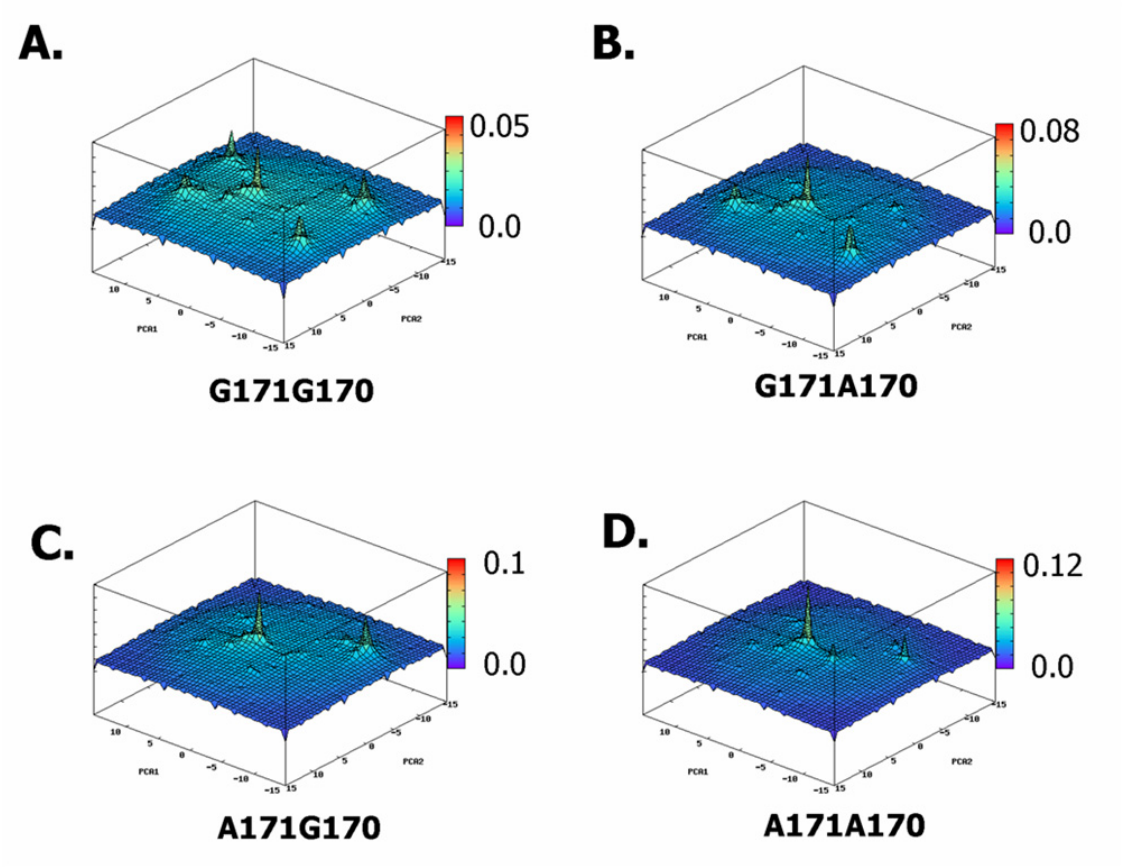


FIGURE 44: RAW FRACTIONAL OCCUPANCY PLOTS FOR MUTANT CYCLE: The fractional occupancy [Boltzmann probability] plots against the first two principle components are shown one of top of the other. The plots are ordered from the most constrained mutant G171A170 [**A**] to the least constrained G171G170 [**D**]. Energetically favorable clusters are seen to appear and disappear as the allowable region is changed through mutation. Contours in each two dimensional plane represent increasing structural distance from crystal structure complex located at origin.

The weighted conformers obtained by applying the energy function formed low energy (i.e. high probability) clusters within the principal space (FIGURE 43). Conformational

clusters having a statistical weight above 10% of the maximum observed weight were selected and plotted against the first two principal components. As is evident, the distribution of high occupancy clusters varied from mutant to mutant. The least constrained mutant G171G170 exhibited the most clusters. This was expected because the lower backbone constraint allowed the chain to sample a greater portion of conformational space, resulting in an enlargement of the allowed region. Accordingly, the G171G170 mutant can be considered as having the set of "basis" clusters for the mutant cycle. The G171A170 mutant showed fewer clusters reflecting a partial contraction of the allowed region of G171G170 with a concomitant removal of some clusters. The A171G170 mutant, which is similar in overall degeneracy to the wild-type (i.e. one Ala and one Gly), provided access to one more cluster than the wild-type and did not have access to two others present in the wild-type ensemble, reflecting a shift of the allowed space in response to the transposition in the location of the Gly residue (FIGURE 41A, FIGURE 43). The most constrained mutant, A171A170, showed exclusion of all but one cluster and exhibited the smallest allowed space (FIGURE 43). In short, the effect of the energy weighting was to differentially induce basins in the free energy landscapes of each mutant.

THE EFFECT OF THE LIGAND ON THE ENSEMBLE

As expected, addition of the Sos peptide perturbs the energetic landscape of C-SH3 due to the mutual exclusion (burial) of solvent accessible surface on the protein and the ligand (Ferreon and Hilser 2003).

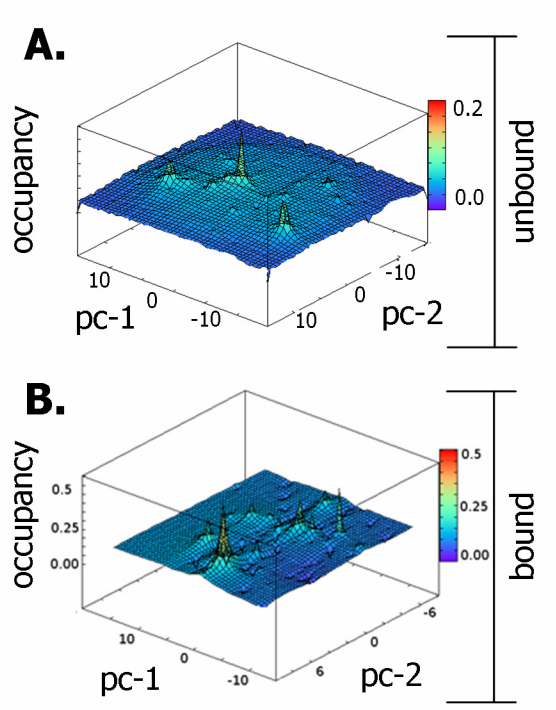


FIGURE 45: FRACTIONAL OCCUPANCY OF UNLIGANDED AND LIGANDED ENSEMBLES for wild-type G171A170: Fractional occupancy of unbound and bound partition as a function of the first two principle components [pc-1, pc-2]. Each partition was plotted such that corresponding points within each partition correspond to the same protein conformation. Lighter [reddish] hues give highest occupancy and darker [blue] hues give lowest populations. A. Fractional occupancy plot of unbound protein ensemble, showing two prominent peaks. B. Fractional occupancy plot of bound ensemble at optimal position and orientation showing three prominent peaks. The increase in the number of peaks reflects an increase in conformational entropy, which favors higher binding affinity.

We employed a protein-protein (intermolecular) free energy landscape to quantify this effect (Camacho, Weng et al. 1999; Nymeyer, Socci et al. 2000; Camacho and Vajda 2001; Cho, Levy et al. 2006; Ruvinsky and Vakser 2008).

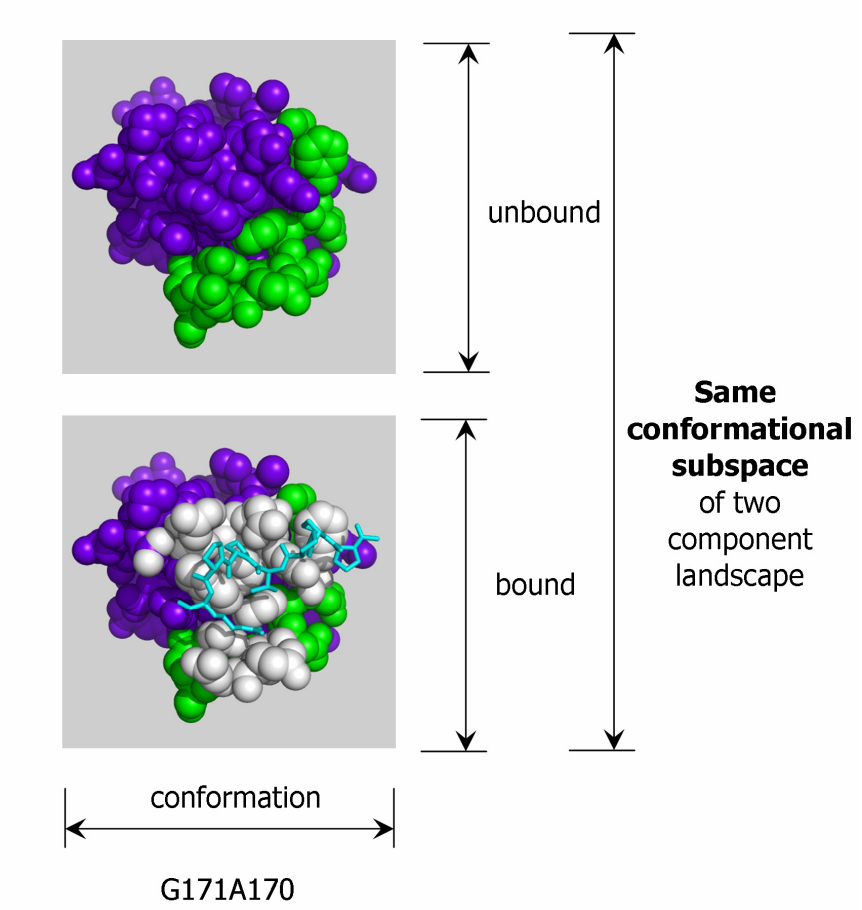


FIGURE 46: SURFACE AFFECTED BY LIGAND BINDING: for wild-type G171A170: Upper picture shows unbound SH3. Lower picture shows area affected (white atoms) by ligand binding.

To illustrate the effect of ligand on the landscape, the unliganded and the liganded ensembles were identified and their fractional occupancy plotted (FIGURE 45A,B). Each

ensemble was plotted against the first two principal components such that corresponding points within each ensemble corresponded to the same protein conformation.

The average free energy level of the liganded ensemble was found to be significantly lower than its corresponding unliganded ensemble, favoring the complexed state (see Table 1), a result that is not unexpected given the known burial of apolar surface involved in the SEM5:Sos interaction (Lim, Richards et al. 1994; Ferreon, Volk et al. 2003). Interestingly, the number and position of the high occupancy clusters of the bound and unbound ensembles were different (FIGURE 45), suggestive of a change in conformational degeneracy (i.e. entropy) upon binding. Indeed, partition of the energetics into their enthalpic and entropic components reveals that the conformational entropy is predicted to be slightly positive for the mutant most closely resembling the WT (i.e. G171A170) as well as the transposed mutant (A171G170), slightly negative for the double Gly, and significantly positive for the double Ala (FIGURE 47A). The importance of this observation becomes apparent when considering our previous results investigating the effect of the P_{II} conformational bias in the Sos peptide ligand. Specifically, the binding thermodynamics for C-SH3 domains have long been enigmatic, with the predominantly hydrophobic binding site producing binding energetic parameters that are uncharacteristic of classic hydrophobic binding (i.e. positive favorable entropy). Our previous results showed that some, but not all of the discrepancy could be attributed to unfavorable entropy of P_{II} formation in the peptide ligand upon binding. The remaining difference was presumed to arise from conformational heterogeneity in the SEM5 ensemble. The slightly positive entropy for binding for the G171A170 and A171G170, when combined with the P_{II} results, allow us to reconcile (at least qualitatively) the

experimentally observed entropy of binding with the values calculated from the structural studies.

Mutant	Calculated Energies (cal/mol)				$\Delta\Delta G_{bind}$ (cal/mol)	
	ΔH	$-T\Delta S_{solv}$	$-T\Delta S_{conf}$	ΔG_{bind}	*Calculated*	*Measured (ITC)*
G171A170	-390	-11888	-127	-12405	0.0	0.0±25
G171G170	-3449	-9043	192	-12300	104	107±25
A171A170	2763	-14218	-533	-11987	417	385±25
A171G170	-1310	-10481	-30	-11822	583	575±25

Table 1: Measured and Predicted Binding Free Energy Changes

Shown in Table 1 are the experimental and predicted binding affinities for the four Ala/Gly double mutants along with the solvation and conformational entropy components. The correlation of the experimental ΔΔG of binding (Ferreon, Hamburger et al. 2004) with the calculated changes is striking, producing a slope of 1.05, and a correlation coefficient of R=0.97 (FIGURE 48). Detailed comparison between the calculated and experimental energetics reveal additional trends. For instance, the impact of A170G is greater in the context of A171, meaning that the net effect of A171 is to amplify the Ala to Gly mutation at position 170. Conversely, the impact of the G171A is greater in the context of G170. The fact that the calculated values capture the detailed trends in the experimental data suggests that the PCA can be used as a tool to characterize the structural and thermodynamic basis for the mutational effects.

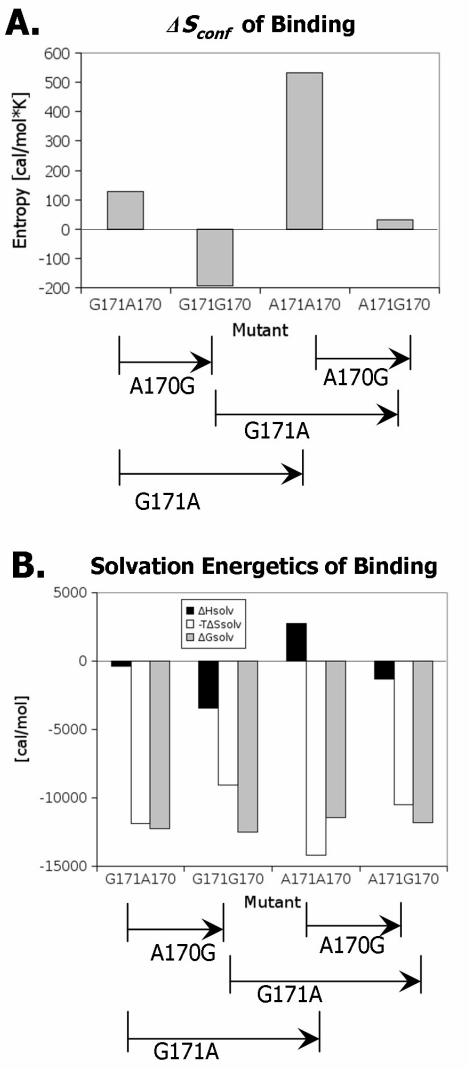


FIGURE 47: SOLVATION ENERGY AND CONFORMATIONAL ENTROPY CHANGES OF BINDING: A. Predicted conformational entropy change of binding. B. Corresponding changes in ensemble weighted solvation energy components (***ΔH, -TΔS, ΔG***). for each mutant.

STRUCTURAL AND THERMODYNAMIC CHARACTER OF MUTATION EFFECTS

The ensemble weighted thermodynamic properties relate the structural details of the conformational states to the experimental observables. Each mutant expressed a different ensemble weighted surface which resulted in a unique profile of solvation energy components (FIGURE 47B).

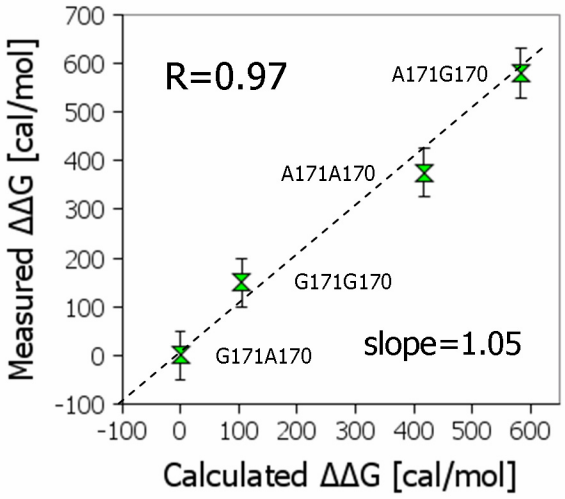


FIGURE 48: PREDICTED FREE ENERGY CHANGES UPON BINDING: Correlation of predicted free energy values [abscissa] to measured [ordinate]. Calculated points are shown as diamonds; reference points [perfect correlation] are shown as green triangles. Error bars shown for measured binding affinities. The correlation of predicted to measured binding affinities is [0.97].

For all mutants, the solvation entropy change for the binding process was large and positive (typical of hydrophobic burial). Here, the A170 mutants demonstrated the

largest increase in solvation entropy in contrast to the double Gly, which had the lowest. Interestingly, the solvation enthalpy of the double Ala mutant increased, whereas for both G171 mutants and the transposed mutant A171G170, the enthalpy decreased Despite these differences, the net decrease in solvation free energy for all mutants was similar in magnitude (~11 kcal), with a slightly greater decrease for the two G170 mutants (FIGURE 47B). These baseline changes were further refined by the conformational entropy changes (discussed above) which when applied, reproduced the rank order observed by the ITC experiments (FIGURE 48A).

Mutation affects each ensemble by shifting the boundaries of the allowed space (FIGURE 40, FIGURE 42, FIGURE 43); an effect termed strain re-distribution. Shifting the allowed space permits expression of different regions of the solvation energy landscape. To illustrate this, consider the landscapes of G171A170 and A171A170, where the allowed space of G171A170 is a super set of A171A170 (FIGURE 49). The regions common to both these mutants have similar terrain features, which suggest that the differences in the thermodynamic properties are caused by the non-overlapping regions of G171A170 (FIGURE 49).

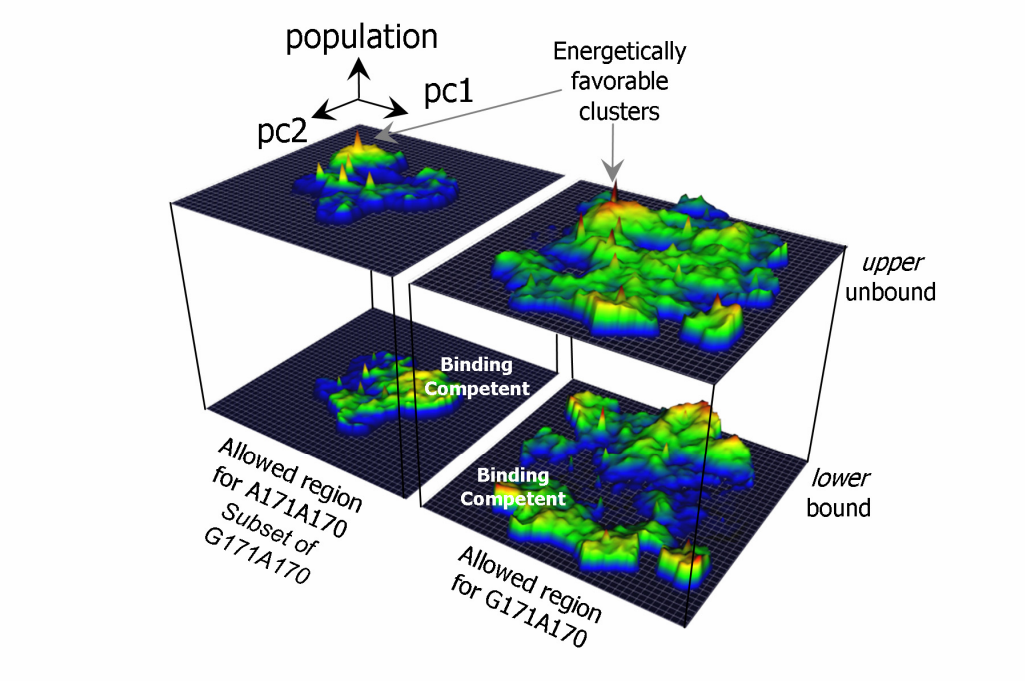


FIGURE 49: EFFECT OF STRAIN REDISTRIBUTION ON THE FREE ENERGY LANDSCAPE. The fractional occupancy of the unbound (upper) and bound (lower) ensembles of the two mutants A171A170 (left), G171A170 (right) are mapped onto the first two principal coordinates. The region delineating the allowed space (for each mutant) has been raised for clarity. The regions common to both these mutants have similar (but not identical) terrain features (within the corresponding bound and unbound ensembles), which suggest that the predominant differences in the thermodynamic properties are caused by the non-overlapping region of G171A170. The expression of the solvation landscape is determined by the allowed space which varies with mutation. Note the expression of an energetically favorable cluster (termed binding competent) in the bound ensemble of the G171A170 mutant. This cluster in the bound ensemble will tend to induce a higher binding affinity (see text for details).

This non-overlapping region manifests a prominent cluster in the bound ensemble that increases the magnitude of its partition function. This then decreases the free energy of its bound ensemble which favors binding.

The highest and lowest affinity mutants exhibited significant differences due to strain redistribution (FIGURE 40A). Comparison of both mutants showed a concomitant shift in the median position and size of the allowed space along each principal axis (FIGURE 40A). This resulted in a perturbation of the extreme principal structures (FIGURE 40B-C), which induced changes in the ensemble weighted surface expression changing the binding free energy.

A STRUCTURAL INTERPRETATION OF THE ALA AND GLY MUTATIONAL POSITIONS

Binding affinity trends were position dependent. In order to understand the effect of position 171 on binding affinity, the dominant conformational states, having (probability > 20%) were selected from within the unbound partition of mutants G171A170 and A171A170 (FIGURE 50A-B). Within these states, the presence of the methyl group of the Ala in position 171 induced a significant concerted displacement in the backbone throughout the flexible region (FIGURE 50A). Instead of allowing the chain to extend outward from the center of the protein as in the pseudo wild type, G171A170, the chain was directed in a tangential fashion away from the binding site (FIGURE 50A), an effect also seen in the principal plots as a shift of the allowed space (FIGURE 42B).

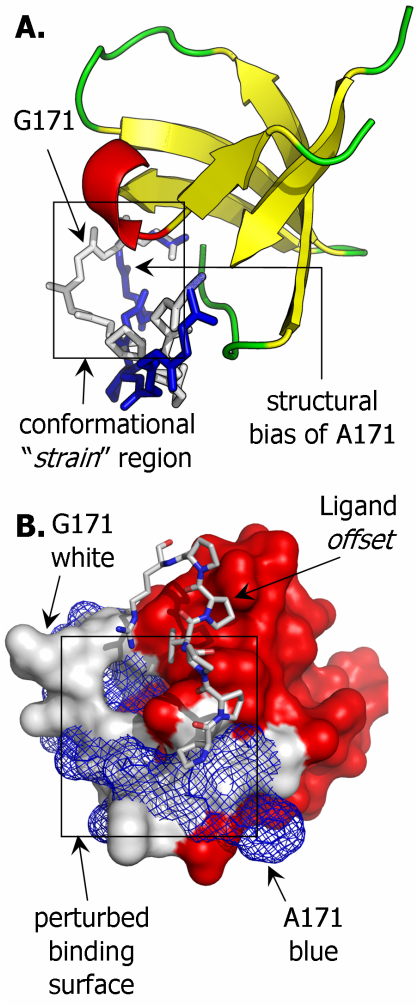


FIGURE 50: EFFECT OF RESIDUE 171: A. Cartoon rendering of the two most probable conformers [within unbound partition] for mutants G171A170 and A171A170 showing the effect of the mutation on the flexible chain segment. The A171 mutant shows a pronounced conformational [strained] bias away from the backbone structure of the wild type. B. The corresponding molecular surface diagram shows significant perturbation of the binding surface. The surface of the A171 mutation [blue mesh] shows much lower surface complementarity to the ligand than the WT [white surface].

As such, this conformation opened the binding cleft (FIGURE 50B, FIGURE 51) preventing optimal surface and residue-pair complementation to occur, which lowered the binding affinity for the A171 mutants. Thus in structural terms, the mutation from Gly to Ala at position 171 could be interpreted as "strain transfer" (Fox strain transfer ref). We note however we note that the determinants for the energy differences between the Gly and Ala variants at 171 cannot be reconciled solely in terms of any structural differences that may be observed. As detailed in the current analysis, the energetic impact of these mutations requires consideration of the effect of the mutations on the breadth as well as energies of the entire conformational manifold.

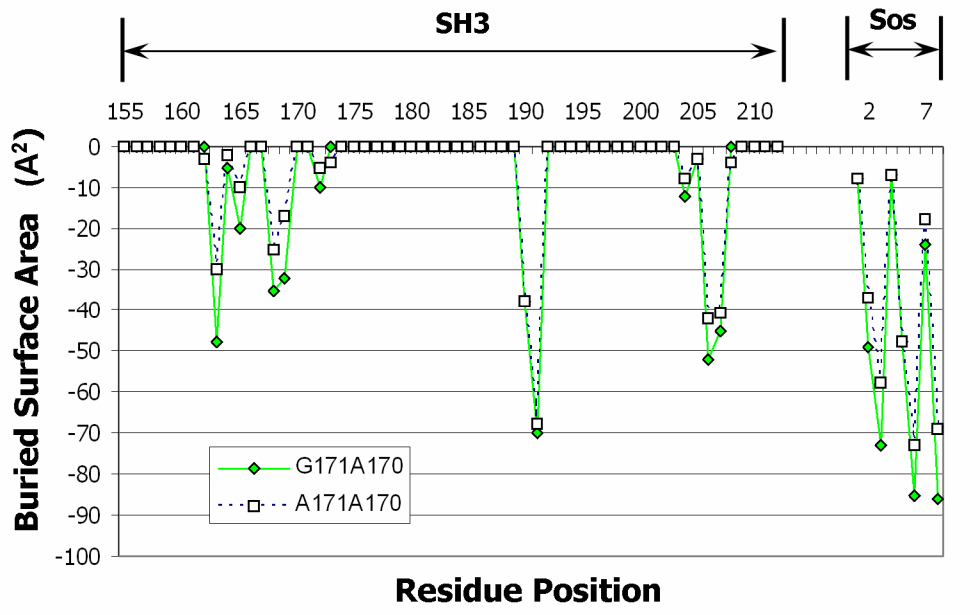


FIGURE 51: COMPARISON OF BOUND SURFACE BURIAL FOR MOST PROBABLE CONFORMERS: Buried surface area as a function of residue position for the most probable conformers in the G171A170 and A171A170 ensembles.

EFFECT OF ELECTROSTATICS

Binding affinity measurements taken with ITC at lower pH (~5.0) show a decrease in binding affinity of (~1 kcal/mol) (Ferreon 2002). This suggests that the ionization of surface groups has an impact on the thermodynamic landscape. Indeed, a salt bridge is formed within the crystal structure between receptor and ligand that is expected to increase the binding affinity (Lim, Richards et al. 1994). An important question is what effect do modulations in the electrostatic surface profile have an effect on binding affinity within the mutant cycle.

It is possible that the charged group on E172 could be one of the primary determinants of the variations in binding affinity of the mutant cycle studied. If the carboxyl group forms a salt bridge with the basic groups on the ARG-8 of the SOSY ligand, then changing its ensemble averaged position in the unbound state will probably impact the binding affinity. The g171a mutation will tend to strain the RT loop such that the binding cleft is made more wide and shallow. This mechanical bias could also affect the relative stability of a salt bridge. MPMOD simulations on the flexible region 162-173 can provide insight as to the effect of the mutations on the electrostatic profile. These simulations show that position 173 undergoes a slight shift in orientation but that the orientation itself is conformationally homogeneous.

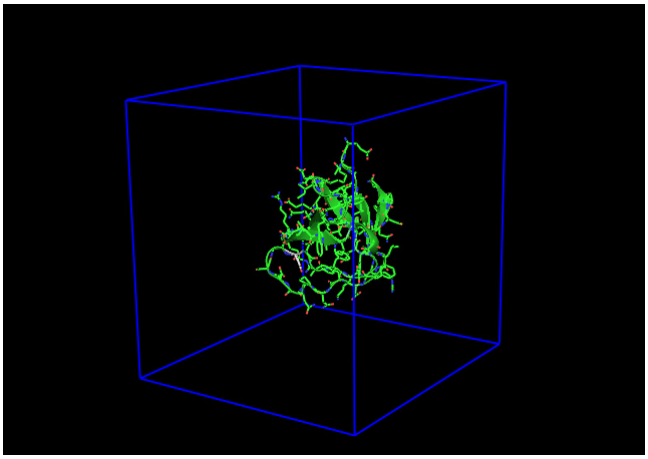


FIGURE 52: UNLIGANDED C-TERMINAL SH3 DOMAIN: Cartoon diagram of SH3 domain with GLU172 shown in white.

The following figures show a large electronegative region at the boundary of the E172 residue in the ensemble of the unliganded protein. This region will tend to repulse the flexible and negatively charged E172, but conversely could raise the pKa of the carboxyl group if that group could be conformationally stabilized near this boundary region.

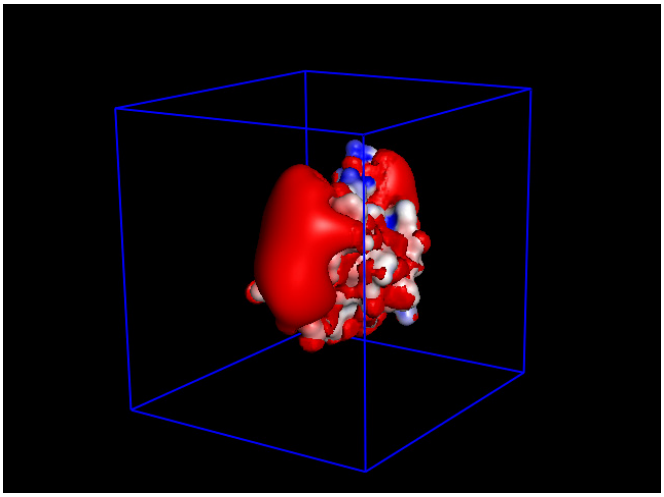


FIGURE 53: ELECTROSTATIC POTENTIAL OF UNLIGANDED C-TERMINAL SH3 DOMAIN: Surface diagram of SH3 domain [at same aspect angle as previous figure] showing electronegative surface at 3kT cutoff in red. Structure shows a large electronegative region near to the distal loop region juxtaposing the GLU172 position.

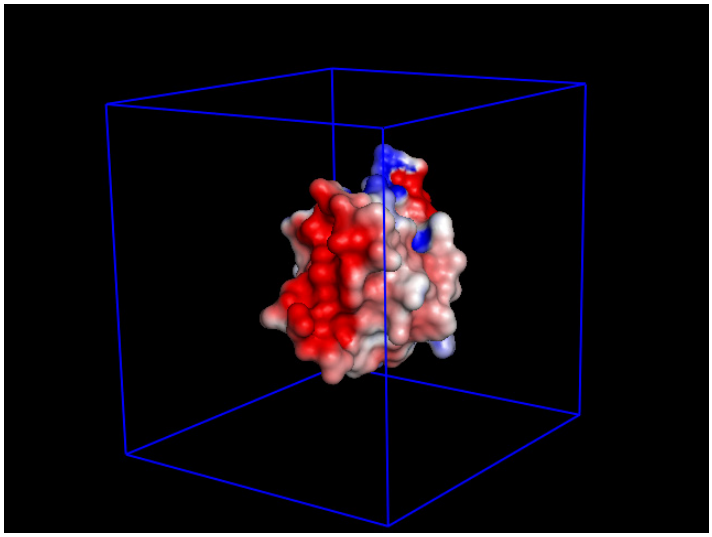


FIGURE 54: ELECTROSTATIC POTENTIAL OF UNLIGANDED C-TERMINAL SH3 DOMAIN AT SOLVENT ACCESSIBLE SURFACE: Surface diagram of SH3 domain [at same aspect angle as previous figure] showing electronegative surface at 3kT cutoff in red. Structure shows a large electronegative region juxtaposing the GLU172 position.

The following figures show 1) cartoon rendering of the protein with the ligand and 2-3) the corresponding electrostatic profiles of the bound complex. A neutral region results when the salt bridge is formed between the mobile E172 and the R-8 of the Sos peptide.

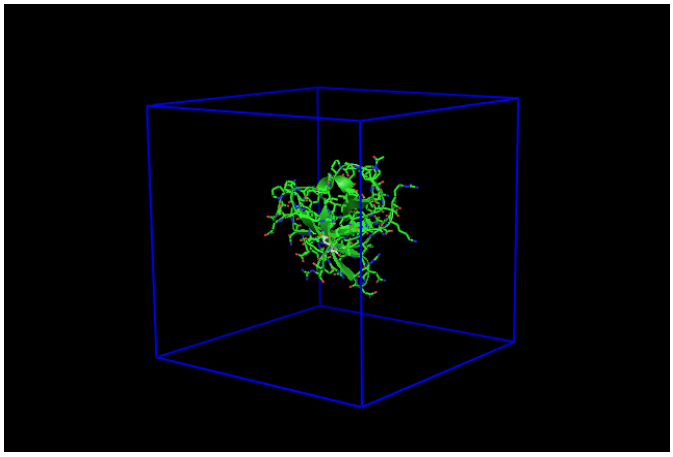


FIGURE 55: LIGANDED C-TERMINAL SH3 DOMAIN: Cartoon diagram of SH3 domain with GLU172 shown in white.

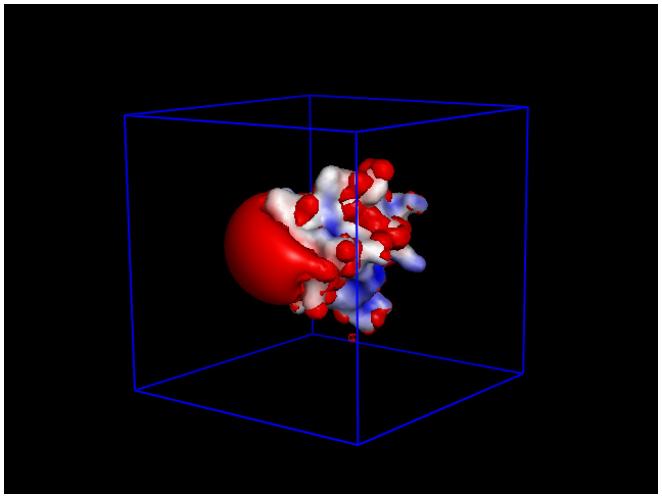


FIGURE 56: ELECTROSTATIC POTENTIAL OF LIGANDED C-TERMINAL SH3 DOMAIN: Surface diagram of SH3 domain [at same aspect angle as previous figure] showing electronegative surface at 3kT cutoff in red. Structure shows a large electronegative region near distal loop region juxtaposing the GLU172 position. Compared to the Unliganded electrostatic profile, the electronegative region appears to be diminished somewhat due to the formation of a salt bridge between E172 and R8 of the Sos peptide. Note also the introduction of a basic region where the ligand has been introduced.

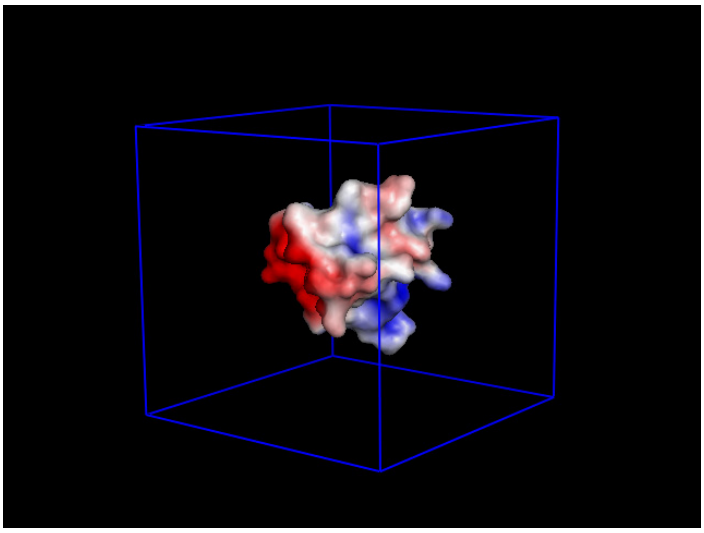


FIGURE 57: ELECTROSTATIC POTENTIAL OF LIGANDED C-TERMINAL SH3 DOMAIN AT SOLVENT ACCESSIBLE SURFACE: Surface diagram of SH3 domain [at same aspect angle as previous figure] showing electronegative surface at 3kT cutoff in red. Structure shows a large electronegative region juxtaposing the GLU172 position.

Electrostatic Effects on Unbound Ensemble

The pKa of the carboxyl group at position 172 in the crystal structure is estimated to be 4.5. At pH=7.0 this group will be fully ionized, so there will only be a 1:1000 chance of this group being neutral. As a result, biasing of this group into the protein [away from the position in the crystal structure] will be repulsed by the large electronegative region. The potential energy in the neighborhood of this carboxyl will be topologically similar to what would exist with steric repulsion only, meaning that the HSC approximation to the landscape in this region will probably be reasonable. If the pH were much closer to 4.5, the effect of electrostatics on the differential binding affinity of the mutant cycle would be more pronounced.

To quantify the impact of ionization on the charge distribution of the unbound ensemble, the following energetic formulation is given:

$$\begin{aligned} \Delta G_{base} &= \Delta G_{solv} + \Delta G_{elec} + \Delta G_{mm}; \\ \Delta G_{tot} &= \Delta G_{base} + \Delta G_{ion}; \\ \Delta G_{ion} &= \Delta H_{ion} - T\Delta S_{ion}; \end{aligned} \qquad (57)$$

Note the entropic cost of protonation in the last equation. This is the penalty of ordering the E172 such that the corresponding estimated enthalpy of ionization (below) is valid. Accounting for this effect cannot be done using a single rigid structure. Instead one must model a **region** of phase space.

The enthalpy change due to protonation is given by:

$$\Delta H_{ion} = 2.3RT\left(pH - pK_a\right) \qquad (58)$$

Note that as the pKa value exceeds the system pH, that protonation becomes favorable.

The partition function must consider the added states due to protonation:

$$\begin{aligned} Q_{tot} &= Q_{unprotonated} + Q_{protonated}; \\ Q_{tot} &= \sum_{i=1}^{nconf} e^{-(\Delta G_{base,i}/RT)}\left(1 + e^{-(\Delta G_{ion,i}/RT)}\right); \end{aligned} \qquad (59)$$

The fraction of E172 neutral states is given by:

$$neutral = \frac{Q_{unprotonated}}{Q_{unprotonated} + Q_{protonated}}; \qquad (60)$$

$$= \frac{\sum_{i=1}^{nconf} e^{-(\Delta G_{base,i}/RT)} e^{-(2.3*(pH - pK_{a,i}) - \Delta S_{ion}/R)}}{\sum_{i=1}^{nconf} e^{-(\Delta G_{base,i}/RT)} \left(1 + e^{-(2.3*(pH - pK_{a,i}) - \Delta S_{ion}/R)}\right)};$$

The web server H++ (Gordon, Myers et al. 2005) was used to estimate the pKa values of select conformers in the unbound state.

Flexible Segment 163-173			
G171A170		A171G170	
Weight	pKa	Weight	pKa
0.15	4.7	0.12	4.5
0.05	4.8	0.06	5.1
0.03	8.2	0.02	6.4
0.03	10.5	0.02	7.0
0.01	11.2	0.01	10.2
0.01	11.8	0.01	10.7

Table 2: pKa as a Function of Mutation

Table 2 shows that the pseudo WT mutant presents a more neutral E172 on the surface due to the net increase in the weighted pKa of the E172 position. This is in opposition to both the trends measured in the experiment and computed using the SASA free energy function. Also the effect of the entropy reduction associated with the increased pKa will significantly oppose the protonation of the carboxyl group on E172. These results suggest that the variation in binding affinity due to mutation is not due to changes in the pKa of the E172 group. To better understand the overall perspective on

binding affinity, one must consider the influence of the salt bridge found in the bound complex.

Bound Ensemble

The free energy change of binding is determined by the difference in the free energies of the bound and unbound ensembles. To gain insight to the electrostatic effects on binding affinity one needs to formulate the partition function for the bound state.

The partition function of the bound ensemble can be expressed as:

$$Q_{bound} = \frac{1}{N!h^{3N}} \int_{-\infty}^{\infty} \int_{V} e^{(-\beta H_{bound}(x_N, p_N))} d^{3N}x\, d^{3N}p; \qquad (61)$$

Where $H_{bound}(x_N, p_N)$ is the Hamiltonian (energy functional) of the system at positions x_N and momenta p_N. The Hamiltonian for this system can be expressed as:

$$H_{bound}(x_N, p_N) = V_{elec}(x_N) + E_{mm}(x_N, p_N) + W_{solv}(x_N, p_N); \qquad (62)$$

which allows the partition function to be decomposed as:

$$\begin{aligned} Q_{bound} &= Q_{elec} * Q_{mm,solv}; \\ Q_{mm,solv} &= \frac{1}{N!h^{3N}} \int_{-\infty}^{\infty} \int_{V} e^{(-\beta(E_{mm}(x_N, p_N) + E_{solv}(x_N, p_N))} d^{3N}x\, d^{3N}p; \\ Q_{elec} &= \frac{1}{N!h^{3N}} \int_{-\infty}^{\infty} \int_{V} e^{(-\beta(V_{elec}(x_N))} d^{3N}x\, d^{3N}p; \end{aligned} \qquad (63)$$

The free energy of the bound partition can then be expressed as:

$$\begin{aligned} G_{bound} &= -kT\ln(Q_{bound}); \\ &= -kT\ln(Q_{elec} * Q_{mm,solv}); \\ &= -kT\ln(Q_{elec}) - kT\ln(Q_{mm,solv}); \end{aligned} \qquad (64)$$

allowing the partition function to be factored into two terms. If the term Q_{elec} is similar for each mutant, then it will induce a similar shift in the free energy of binding for each mutant, meaning that the perturbations in binding free energy will be determined primarily by the solvation effects.

Due to the diffuse nature of the electrostatic effects, and the similarity of the conformational distributions of each mutant, the electrostatic energy profile will be similar for all four mutants. Subsequently, the contribution to the free energy of the bound state, which is about 1kcal/mol (Ferreon 2002), will be uniform for all mutants, and will be the same order of magnitude as the solvation mediated perturbations to binding affinity induced by backbone relaxation.

Salt Bridge Formation (MD Simulation)

A robust way to assay the energetics of salt bridge formation is to simulate using a molecular dynamics force field such as that used in AMBER9. These force fields have been extensively tested and calibrated and can be used to dissect the dominant forces during select phases of complex formation.

The formation of the Sos:R8 Sh3:E172 salt bridge was simulated using the AMBER-9 molecular dynamics system. A series of 25 starting orientations/positions in which the ligand was beyond the cutoff distance of 12 A were simulated for each of the four mutants. In each instance, the salt bridge was found to be fully formed within 30 ps at 25C and pH 7. These simulations show that at pH 7 the salt bridge formation is a constant energetic bias for all mutants.

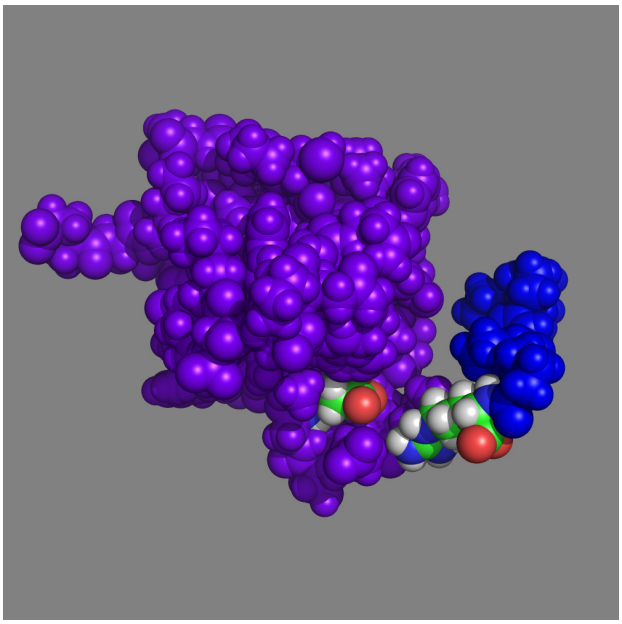


FIGURE 58: SALT BRIDGE FORMATION LIGAND:R8 PROTEIN:E172: Space filling diagram showing protein (purple) and ligand (blue) with salt bridge forming early in simulation.

Discussion

JUSTIFICATION OF ASSUMPTIONS

Optimal Orientation

The C-SH3 binding cleft is concave in shape (see left panel of figure below). This type of surface geometry has an intrinsically high orientational specificity i.e. has a well defined binding funnel. The extension of the potential range of the "steric" interactions results in low-resolution molecular shapes (the molecular shape is determined by the repulsion part of the steric interactions). The energy landscape is determined by the shape of the interacting proteins. Thus, the funnel usually corresponds

to the most prominent shape feature (e.g., a deep active site of an enzyme, or a flat multi-subunit interface, shown below).

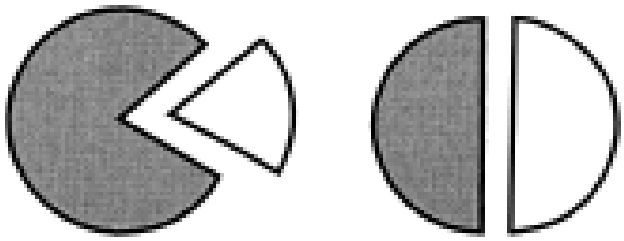


FIGURE 59: PROMINENT SHAPE BASED LANDSCAPE FEATURES: A geometric illustration of the concept of low-resolution protein recognition. The correct position of the ligand within the complex (global minimum of energy) corresponds to a prominently shaped feature. This geometric recognition factor (often the largest one) is still preserved after most of the smaller features are deleted by decreasing the resolution of the molecular image.

By lowering the molecular image resolution, we can reach the point at which only the largest shape characteristic is preserved. This would eliminate all minima but the funnel.

Conclusions for SEM5 C-SH3 Flexible Binding

Conformational dynamics of the surface exposed RT loop of the SEM5-CSH3 domain was simulated using a conformational sampling method based on hard sphere collision and analyzed using principal component analysis. Our results support several interesting conclusions. First, the effect of Ala to Gly mutation can be quantitatively reconciled with

experimental data, capturing the amplification of backbone relaxation imparted by the Gly residue. Second, the general effect of the mutations is to expand and shift the conformational manifold of accessible space. Third, that this shift in accessible space is responsible for exposing energetically favorable regions in the binding landscape. Fourth, that PCA applied to the α and β carbons can be used to facilitate a massive reduction in the dimensionality of the problem, thus providing a vehicle for characterizing the energy landscape of the protein and the effects of the ligand on the landscape. Lastly, our result demonstrates that the average solution structure is insufficient to account for the effects of the mutations and that an understanding of the effect of the mutations on the entire conformational manifold is a prerequisite to an understanding of molecular recognition.

The statistical thermodynamic model of binding affinity (Gilson, Given et al. 1997) proved a valuable framework in the formulation and prediction of binding affinity for this instance. An empirical continuum solvation model (Murphy and Freire 1992; Freire, Haynie et al. 1993; Freire and Xie 1994; Murphy 1994; Xie and Freire 1994; Xie and Freire 1994) provided an energy function that when applied to each conformational ensemble, correlated with the measured binding affinities. The relationship between time averaged conformational dynamics and binding affinity was determined within the context of a local unfolding model (Miller and Dill 1995; Bahar, Wallqvist et al. 1998). The implicit solvation based energy function induced basins in the free energy landscape.

However, within the context of equilibrium ensembles, dynamic fluctuations are averaged out because there is sufficient time to achieve the Boltzmann distribution.

Two theories of molecular recognition have been proposed to explain the impact of protein dynamics on binding affinity. The induced fit model claims that conformational motion occurs only after the initial binding interaction. Conformational selection (Bosshard 2001), on the other hand, says binding is like an ensemble based version of the lock and key paradigm. A statistical thermodynamic version of conformational selection, in which each complex was energetically weighted, provided a high correlation to the experimental data.

Protein dynamics has a rational and non-trivial effect on molecular recognition. Through its impact on binding affinity, it will, by definition, have an effect on the overall specificity of a protein. It is important to dissect the relationship between affinity and specificity. When proteins themselves are flexible, part of the affinity is used to adjust the conformations to best fit the binding partners. Therefore, flexibility through conformational change usually gives a good opportunity for realizing the specificity for molecular recognition, but often with the price of sacrificing certain amounts of affinity to adjust the conformations. The resulting lower affinity can give molecules the ability to both bind specifically and unbind easily, which is essential for cell signaling relay and gene regulation. Further studies could concentrate on understanding the overall impact of these effects on protein interaction networks as a whole.

Redistribution of the allowed space through surface ala/gly mutation explained the variations in binding affinity of the C-Terminal SH3 domain with its putative ligand, with select residue positions having distinct effects on the distribution of the conformational ensembles. Modulation of binding affinity, achieved in this study, was able to produce a subtle [10%] variation in binding affinity relative to the wild type (Ferreon, Hamburger et al. 2004). Shifting the boundaries of the allowed space changed both the size and shape of select regions of the landscape and modulated the expression or suppression of energetic minima. Enlarging the allowed region increased the conformational degeneracy which was an entropic manifestation. Expression of energetically favorable regions through boundary shifting was mediated by concerted changes in solvent based enthalpy and entropy as self avoiding conformations emerged from the phase space.

Model of Molecular Recognition

In this study, we first constructed a thermodynamic energy landscape for molecular recognition and address the roles of flexibility in determining the binding affinity and functional specificity. Affinity and specificity are the two key factors in molecular recognition. Affinity measures the stability resulting from the association of two molecules; specificity is the ability of one molecule to bind with another molecule while discriminating against others. For rigid binding, affinity, and specificity are often correlated. Yet, in flexible binding, flexibility can enable molecules to adjust their conformations to reach the best fit (e.g., high specificity). Quantifying the specificity as well as affinity in flexible binding is crucial in uncovering the mechanism of flexible binding. Flexible binding involves both binding and conformational degrees of freedom. Thus we need at least three reaction coordinates to describe it: Q_b, fraction of native

spatial contacts for interface binding; Q_{f1} and Q_{f2}, fraction of native spatial contacts for flexible conformational change. Based on this, we can construct an energy function and derive a free energy landscape. From the thermodynamic analysis, we expect that the requirement of stable binding against trapping would lead to a funneled binding landscape to guarantee both affinity and specificity. Only binding with landscape funneled against traps can survive natural evolution, be relatively stable, and perform specific biological functions. With this approach, the role of the interplay between binding and flexibility can be uncovered. Biomolecules need some affinity to be stable but they also need flexibility to adjust to achieve optimal fit and perform specific biological functions. The reality is a balance between the two. We will find an optimal criterion of binding specificity. It can be used for guiding further atomic detailed studies and flexible drug design.

FUTURE DIRECTIONS

The analysis of the effects of ionization is cursory. It is based on a monte-carlo type analysis which is weighted predominantly by the empirical solvation energy function. Missing from this energy function are explicit electrostatics and quantum mechanical effects. These energetic contributions would have to be modeled in cases where chemical reactions occur that are commonly mediated by charged regions subject to large electrostatic forces. Due to the high level of dielectric shielding afforded by the solvent, the more long range electrostatic contributions will be similar among the mutants studied in this work.

Methodological Extensions

There are some methodological generalizations that could be made although they were not needed in this analysis. First, one could generalize the ability to handle

conformations with arbitrary contact topology; i.e. expanding the scope of conformational variation. Second, the analysis could be extended to routinely enumerate all possible orientations in the calculation of the bound ensemble. Third, the energy function could be extended to handle electrostatic effects and other molecular mechanical components. Finally, quantum mechanical effects could be incorporated to account for discrete chemical changes such as protonation. All these additions, however, will come at a higher developmental and computational price.

References

Altman, D. G. and J. M. Bland (1994). "Statistics notes - diagnostic-tests-1 - sensitivity and specificity .3." British Medical Journal **308**(6943): 1552-1552.

Amato, N. M., O. B. Bayazit, L. K. Dale, C. Jones and D. Vallejo (2000). "Choosing good distance metrics and local planners for probabilistic roadmap methods." Ieee Transactions on Robotics and Automation **16**(4): 442-447.

Amato, N. M. and G. Song (2002). "Using motion planning to study protein folding pathways." Journal of Computational Biology **9**(2): 149-168.

Bahar, I., A. Wallqvist, D. G. Covell and R. L. Jernigan (1998). "Correlation between native-state hydrogen exchange and cooperative residue fluctuations from a simple model." Biochemistry **37**(4): 1067-1075.

Banavar, J. R., A. Maritan, C. Micheletti and A. Trovato (2002). "Geometry and physics of proteins." Proteins-Structure Function and Genetics **47**(3): 315-322.

Bashford, D., C. Chothia and A. M. Lesk (1987). "Determinants Of A Protein Fold - Unique Features Of The Globin Amino-Acid-Sequences." Journal of Molecular Biology **196**(1): 199-216.

Bauer, F. and H. Sticht (2007). "A proline to glycine mutation in the Lck SH3-domain affects conformational sampling and increases ligand binding affinity." Febs Letters **581**(8): 1555-1560.

Becker, O. M. (1997). "Geometric versus topological clustering: An insight into conformation mapping." Proteins-Structure Function and Genetics **27**(2): 213-226.

Becker, O. M. (1998). "Principal coordinate maps of molecular potential energy surfaces." Journal of Computational Chemistry **19**(11): 1255-1267.

Becker, O. M. and M. Karplus (1997). "The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics." Journal of Chemical Physics **106**(4): 1495-1517.

Bennaim, A. and Y. Marcus (1984). "Solvation Thermodynamics Of Nonionic Solutes." Journal of Chemical Physics **81**(4): 2016-2027.

Berger, C., S. Weber-Bornhauser, J. Eggenberger, J. Hanes, A. Pluckthun and H. R. Bosshard (1999). "Antigen recognition by conformational selection." Febs Letters **450**(1-2): 149-153.

Bosshard, H. R. (2001). "Molecular recognition by induced fit: How fit is the concept?" News in Physiological Sciences **16**: 171-173.

Brooks, C. L. and M. Karplus (1989). "Solvent Effects On Protein Motion And Protein Effects On Solvent Motion - Dynamics Of The Active-Site Region Of Lysozyme." Journal of Molecular Biology **208**(1): 159-181.

Brunger, A. T., P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J. S. Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, et al. (1998). "Crystallography & NMR system: A new software suite for macromolecular structure determination." Acta Crystallographica Section D-Biological Crystallography **54**: 905-921.

Camacho, C. J. and S. Vajda (2001). "Protein docking along smooth association pathways." Proceedings of the National Academy of Sciences of the United States of America **98**(19): 10636-10641.

Camacho, C. J., Z. P. Weng, S. Vajda and C. DeLisi (1999). "Free energy landscapes of encounter complexes in protein-protein association." Biophysical Journal **76**(3): 1166-1178.

Caves, L. S. D., J. D. Evanseck and M. Karplus (1998). "Locally accessible conformations of proteins: Multiple molecular dynamics simulations of crambin." Protein Science **7**(3): 649-666.

Cesareni, G., S. Panni, G. Nardelli and L. Castagnoli (2002). "Can we infer peptide recognition specificity mediated by SH3 domains?" Febs Letters **513**(1): 38-44.

Chang, C. E. A., W. Chen and M. K. Gilson (2007). "Ligand configurational entropy and protein binding." Proceedings of the National Academy of Sciences of the United States of America **104**(5): 1534-1539.

Cho, S. S., Y. Levy and P. G. Wolynes (2006). "P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes." Proceedings of the National Academy of Sciences of the United States of America **103**(3): 586-591.

Chothia, C. (1975). "Structural Invariants In Protein Folding." Nature **254**(5498): 304-308.

Choudhury, N. and B. M. Pettitt (2007). "The dewetting transition and the hydrophobic effect." Journal of the American Chemical Society **129**(15): 4847-4852.

Crippen, G. and T. Havel (1988). Distance Geometry and Molecular Conformation. New York, NY, Wiley.

Cullen, C. G. (1972). Matrices and Linear Transformations. Reading, MA, Addison-Wesley.

Daquino, J. A., J. Gomez, V. J. Hilser, K. H. Lee, L. M. Amzel and E. Freire (1996). "The magnitude of the backbone conformational entropy change in protein folding." Proteins-Structure Function and Genetics **25**(2): 143-156.

Dellago, C. and P. G. Bolhuis (2007). Transition path sampling simulations of biological systems. Atomistic Approaches in Modern Biology: from Quantum Chemistry to Molecular Simulations. **268:** 291-317.

Eisenberg, D. and A. D. McLachlan (1986). "Solvation Energy In Protein Folding And Binding." Nature **319**(6050): 199-203.

Fenimore, P. W., H. Frauenfelder, B. H. McMahon and R. D. Young (2004). "Bulk-solvent and hydration-shell fluctuations, similar to alpha- and beta-fluctuations in glasses, control protein motions and functions." Proceedings of the National Academy of Sciences of the United States of America **101**(40): 14408-14413.

Ferreon, J. C. (2002). Biophysical Characterization of the SEM-5 C-SH3 Domain and Investigation of Ala/Gly Mutants: Role of Conformational Fluctuations on Binding. Unpublished dissertation. Galveston, Texas, University of Texas Medical Branch. **PhD**.

Ferreon, J. C., J. B. Hamburger and V. J. Hilser (2004). "An experimental strategy to evaluate the thermodynamic stability of highly dynamic binding sites in proteins using hydrogen exchange." Journal of the American Chemical Society **126**(40): 12774-12775.

Ferreon, J. C. and V. J. Hilser (2003). "The effect of the polyproline II (PPII) conformation on the denatured state entropy." Protein Science **12**(3): 447-457.

Ferreon, J. C., D. E. Volk, B. A. Luxon, D. G. Gorenstein and V. J. Hilser (2003). "Solution structure, dynamics, and thermodynamics of the native state ensemble of the Sem-5 C-terminal SH3 domain." Biochemistry **42**(19): 5582-5591.

Finkelstein, A. V. and J. Janin (1989). "The Price Of Lost Freedom - Entropy Of Bimolecular Complex-Formation." Protein Engineering **3**(1): 1-3.

Frauenfelder, H., G. A. Petsko and D. Tsernoglou (1979). "Temperature-Dependent X-Ray-Diffraction As A Probe Of Protein Structural Dynamics." Nature **280**(5723): 558-563.

Frauenfelder, H., S. G. Sligar and P. G. Wolynes (1991). "The Energy Landscapes And Motions Of Proteins." Science **254**(5038): 1598-1603.

Freire, E., D. T. Haynie and D. Xie (1993). "Molecular-Basis Of Cooperativity In Protein-Folding .4. Core - A General Cooperative Folding Model." Proteins-Structure Function and Genetics **17**(2): 111-123.

Freire, E. and D. Xie (1994). "Thermodynamic Prediction Of Structural Determinants Of The Molten Globule State Of Barnase." Biophysical Chemistry **51**(2-3): 243-251.

Fukunaga and Keinosuke (1990). Introduction to Statistical Pattern Recognition. Amsterdam, Elsevier.

Garrett, R. H. and C. M. Grisham (1999). Biochemistry, Saunders College Publishing.

Gilson, M. K., J. A. Given, B. L. Bush and J. A. McCammon (1997). "The statistical-thermodynamic basis for computation of binding affinities: A critical review." Biophysical Journal **72**(3): 1047-1069.

Gilson, M. K. and H. X. Zhou (2007). "Calculation of protein-ligand binding affinities." Annual Review of Biophysics and Biomolecular Structure **36**: 21-42.

Gordon, J., J. Myers, T. Folta, V. Shoja, L. Heath and A. Onufriev (2005). "H++: A Server for Estimating pKas and Adding Missing Hydrogens to Molecules." Nucleic Acids Research **33**: 369-371.

Haliloglu, T., I. Bahar and B. Erman (1997). "Gaussian dynamics of folded proteins." Physical Review Letters **79**(16): 3090-3093.

Hamelberg, D., J. Mongan and J. A. McCammon (2004). "Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules." Journal of Chemical Physics **120**(24): 11919-11929.

Henzler-Wildman, K. and D. Kern (2007). "Dynamic personalities of proteins." Nature **450**(7171): 964-972.

Hill, T. (1960). An Introduction to Statistical Thermodynamics. New York, NY, Addison-Wesley.

Hilser, V. J., J. Gomez and E. Freire (1996). "Enthalpy change in protein folding and binding: Refinement of parameters for structure-based calculations." Proteins-Structure Function and Genetics **26**(2): 123-133.

Jacobs, D. J., A. J. Rader, L. A. Kuhn and M. F. Thorpe (2001). "Protein flexibility predictions using graph theory." Proteins-Structure Function and Genetics **44**(2): 150-165.

Janin, J. and C. Chothia (1978). "Role Of Hydrophobicity In Binding Of Coenzymes." Biochemistry **17**(15): 2943-2948.

Jones, T. A. and S. Thirup (1986). "Using Known Substructures In Protein Model-Building And Crystallography." Embo Journal **5**(4): 819-822.

Karplus, M., T. Ichiye and B. M. Pettitt (1987). "Configurational Entropy Of Native Proteins." Biophysical Journal **52**(6): 1083-1085.

Karplus, M. and J. N. Kushick (1981). "Method For Estimating The Configurational Entropy Of Macromolecules." Macromolecules **14**(2): 325-332.

Katchalskikatzir, E., I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo and I. A. Vakser (1992). "Molecular-Surface Recognition - Determination Of Geometric Fit Between Proteins And Their Ligands By Correlation Techniques." Proceedings of the National Academy of Sciences of the United States of America **89**(6): 2195-2199.

Koshland, D. E. (1994). "The Key-Lock Theory And The Induced Fit Theory." Angewandte Chemie-International Edition **33**(23-24): 2375-2378.

Lange, O. F., N. A. Lakomek, C. Fares, G. F. Schroder, K. F. A. Walter, S. Becker, J. Meiler, H. Grubmuller, C. Griesinger and B. L. de Groot (2008). "Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution." Science **320**(5882): 1471-1475.

Laskowski, R. A., M. W. Macarthur, D. S. Moss and J. M. Thornton (1993). "Procheck - A Program To Check The Stereochemical Quality Of Protein Structures." Journal of Applied Crystallography **26**: 283-291.

Leder, L., C. Berger, S. Bornhauser, H. Wendt, F. Ackermann, I. Jelesarov and H. R. Bosshard (1995). "Spectroscopic, calorimetric, and kinetic demonstration of conformational adaptation in peptide antibody recognition." Biochemistry **34**(50): 16509-16518.

Lee, K. H., D. Xie, E. Freire and L. M. Amzel (1994). "Estimation Of Changes In Side-Chain Configurational Entropy In Binding And Folding - General-Methods And Application To Helix Formation." Proteins-Structure Function and Genetics **20**(1): 68-84.

Li, S. S. C. (2005). "Specificity and versatility of SH3 and other proline-recognition domains: structural basis and implications for cellular signal transduction." Biochemical Journal **390**: 641-653.

Lim, W. A., F. M. Richards and R. O. Fox (1994). "Structural Determinants Of Peptide-Binding Orientation And Of Sequence Specificity In Sh3 Domains." Nature **372**(6504): 375-379.

Lounnas, V., B. M. Pettitt and G. N. Phillips (1994). "A Global-Model Of The Protein-Solvent Interface." Biophysical Journal **66**(3): 601-614.

Ma, J. P. and M. Karplus (1997). "Ligand-induced conformational changes in ras p21: A normal mode and energy minimization analysis." Journal of Molecular Biology **274**(1): 114-131.

Matthews, B. W., H. Nicholson and W. J. Becktel (1987). "Enhanced Protein Thermostability From Site-Directed Mutations That Decrease The Entropy Of Unfolding." Proceedings of the National Academy of Sciences of the United States of America **84**(19): 6663-6667.

Maxwell, K. L. and A. R. Davidson (1998). "Mutagenesis of a buried polar interaction in an SH3 domain: Sequence conservation provides the best prediction of stability effects." Biochemistry **37**(46): 16172-16182.

Mayer, B. J. (2001). "SH3 domains: complexity in moderation." Journal of Cell Science **114**(7): 1253-1263.

Mayer, B. J. and R. Gupta (1998). Functions of SH2 and SH3 domains. Protein Modules in Signal Transduction. **228:** 1-22.

Micheletti, C., F. Seno, A. Maritan and J. R. Banavar (1998). "Design of proteins with hydrophobic and polar amino acids." Proteins-Structure Function and Genetics **32**(1): 80-87.

Miller, D. W. and K. A. Dill (1995). "A Statistical-Mechanical Model For Hydrogen-Exchange In Globular-Proteins." Protein Science **4**(9): 1860-1873.

Murphy, K. P. (1994). "Hydration And Convergence Temperatures - On The Use And Interpretation Of Correlation Plots." Biophysical Chemistry **51**(2-3): 311-326.

Murphy, K. P. and E. Freire (1992). "Thermodynamics Of Structural Stability And Cooperative Folding Behavior In Proteins." Advances in Protein Chemistry **43**: 313-361.

Murphy, K. P., D. Xie, K. S. Thompson, L. M. Amzel and E. Freire (1994). "Entropy In Biological Binding Processes - Estimation Of Translational Entropy Loss." Proteins-Structure Function and Genetics **18**(1): 63-67.

Nguyen, J. T., C. W. Turck, F. E. Cohen, R. N. Zuckermann and W. A. Lim (1998). "Exploiting the basis of proline recognition by SH3 and WW domains: Design of n-substituted inhibitors." Science **282**(5396): 2088-2092.

Nymeyer, H., N. D. Socci and J. N. Onuchic (2000). "Landscape approaches for determining the ensemble of folding transition states: Success and failure hinge on the degree of frustration." Proceedings of the National Academy of Sciences of the United States of America **97**(2): 634-639.

Onuchic, J. N., H. Nymeyer, A. E. Garcia, J. Chahine and N. D. Socci (2000). "The energy landscape theory of protein folding: Insights into folding mechanisms and scenarios." Advances in Protein Chemistry **53**: 87-152.

Park, K., M. Vendruscolo and E. Domany (2000). "Toward an energy function for the contact map representation of proteins." Proteins-Structure Function and Genetics **40**(2): 237-248.

Pettitt, B. M. and M. Karplus (1988). "Conformational Free-Energy Of Hydration For The Alanine Dipeptide - Thermodynamic Analysis." Journal of Physical Chemistry **92**(13): 3994-3997.

Plotkin, S. S., J. Wang and P. G. Wolynes (1997). "Statistical mechanics of a correlated energy landscape model for protein folding funnels." Journal of Chemical Physics **106**(7): 2932-2948.

Ramachandran, G. N., C. Ramakrishnan and V. Sasisekharan (1963). "Stereochemistry Of Polypeptide Chain Configurations." Journal of Molecular Biology **7**(1): 95-105.

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

Richards, F. M. (1977). "Areas, Volumes, Packing, And Protein-Structure." Annual Review of Biophysics and Bioengineering **6**: 151-176.

Roux, B. (1995). "The Calculation Of The Potential Of Mean Force Using Computer-Simulations." Computer Physics Communications **91**(1-3): 275-282.

Ruvinsky, A. A. and I. A. Vakser (2008). "Interaction cutoff effect on ruggedness of protein-protein energy landscape." Proteins-Structure Function and Bioinformatics **70**(4): 1498-1505.

Scheraga, H. A., M. Khalili and A. Liwo (2007). "Protein-folding dynamics: Overview of molecular simulation techniques." Annual Review of Physical Chemistry **58**: 57-83.

Shea, J. E., J. N. Onuchic and C. L. Brooks (2002). "Probing the folding free energy landscape of the src-SH3 protein domain." Proceedings of the National Academy of Sciences of the United States of America **99**(25): 16064-16068.

Shoichet, B. K., D. L. Bodian and I. D. Kuntz (1992). "Molecular Docking Using Shape Descriptors." Journal of Computational Chemistry **13**(3): 380-397.

Silverstein, K. A. T., A. D. J. Haymet and K. A. Dill (1998). "A simple model of water and the hydrophobic effect." Journal of the American Chemical Society **120**(13): 3166-3175.

Sims, G. E., I. G. Choi and S. H. Kim (2005). "Protein conformational space in higher order phi-psi maps." Proceedings of the National Academy of Sciences of the United States of America **102**(3): 618-621.

Sitkoff, D., K. A. Sharp and B. Honig (1994). "Accurate Calculation Of Hydration Free-Energies Using Macroscopic Solvent Models." Journal of Physical Chemistry **98**(7): 1978-1988.

Sowdhamini, R., C. Ramakrishnan and P. Balaram (1993). "Modeling Multiple Disulfide Loop Containing Polypeptides By Random Conformation Generation - The Test Cases Of Alpha-Conotoxin Gi And Endothelin-I." Protein Engineering **6**(8): 873-882.

Street, A. G. and S. L. Mayo (1998). "Pairwise calculation of protein solvent-accessible surface areas." Folding & Design **3**(4): 253-258.

Tsai, C. J., S. Kumar, B. Y. Ma and R. Nussinov (1999). "Folding funnels, binding funnels, and protein function." Protein Science **8**(6): 1181-1190.

Tsai, C. J., B. Y. Ma and R. Nussinov (1999). "Folding and binding cascades: Shifts in energy landscapes." Proceedings of the National Academy of Sciences of the United States of America **96**(18): 9970-9972.

Tsai, C. J., B. Y. Ma, Y. Y. Sham, S. Kumar and R. Nussinov (2001). "Structured disorder and conformational selection." Proteins-Structure Function and Genetics **44**(4): 418-427.

Tsai, J., M. Levitt and D. Baker (1999). "Hierarchy of structure loss in MD simulations of src SH3 domain unfolding." Journal of Molecular Biology **291**(1): 215-225.

Unger, R. and J. Moult (1996). "Local interactions dominate folding in a simple protein model." Journal of Molecular Biology **259**(5): 988-994.

van Buuren, A. "Gromacs Documentation." from http://www.gromacs.org/.

Wells, S., S. Menor, B. Hespenheide and M. F. Thorpe (2005). "Constrained geometric simulation of diffusive motion in proteins." Physical Biology **2**(4): S127-S136.

Weng, Z. G., R. J. Rickles, S. B. Feng, S. Richard, A. S. Shaw, S. L. Schreiber and J. S. Brugge (1995). "Structure-Function Analysis Of Sh3 Domains - Sh3 Binding-Specificity Altered By Single Amino-Acid Substitutions." Molecular and Cellular Biology **15**(10): 5627-5634.

Weng, Z. G., S. M. Thomas, R. J. Rickles, J. A. Taylor, A. W. Brauer, C. Seideldugan, W. M. Michael, G. Dreyfuss and J. S. Brugge (1994). "Identification Of Src, Fyn, And Lyn-Sh3 Binding-Proteins - Implications For A Function Of Sh3 Domains." Molecular and Cellular Biology **14**(7): 4509-4521.

Wodak, S. J. and J. Janin (1980). "Analytical Approximation To The Accessible Surface-Area Of Proteins." Proceedings of the National Academy of Sciences of the United States of America-Physical Sciences **77**(4): 1736-1740.

Xie, D. and E. Freire (1994). "Structure-Based Prediction Of Protein-Folding Intermediates." Biophysical Journal **66**(2): A180-A180.

Xie, D. and E. Freire (1994). "Thermodynamic And Structural Conditions For The Stabilization Of Compact Denatured States." Biophysical Journal **66**(2): A178-A178.

Yang, D. W. and L. E. Kay (1996). "Contributions to conformational entropy arising from bond vector fluctuations measured from NMR-derived order parameters: Application to protein folding." Journal of Molecular Biology **263**(2): 369-382.

Yu, H. T., J. K. Chen, S. B. Feng, D. C. Dalgarno, A. W. Brauer and S. L. Schreiber (1994). "Structural Basis For The Binding Of Proline-Rich Peptides To Sh3 Domains." Cell **76**(5): 933-945.