Understanding and Designing Protein Beta-Sheets

Thesis by

Arthur George Street

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2000

(submitted September 16, 1999)

Acknowledgements

I am immensely grateful to my advisor, Steve Mayo, for providing a stimulating and enjoyable laboratory in which to work, for his guidance and vision, and for being so willing to take on a physicist with no prior experience in a molecular biology "wet lab."

In the lab, I have particularly benefited, both scientifically and personally, from collaboration with Ben Gordon (and also Leah and son Akiva, although his input was largely constrained to pulling objects off tables and throwing M&M's over his shoulder). Sandy Malakauskas also requires special mention, as I am indebted to her for teaching me almost everything I know about molecular biology, and for never following up on our mock mutant blackmail threats. But I have learnt from and enjoyed the company of everyone in the lab: Bassil Dahiyat, Chantal Morgan, Scott Ross, Monica Breckow, Fred Lee, Dirk Bökenkamp, Barry Olafson, Marie Ary, Alyce Su, Cathy Sarisky, Andrei Marinescu, Dan Bolon, Pavel Strop, Shannon Marshall, Julie Archer, Niles Pierce, Deepshikha Datta, Julia Shifman, Chris Voigt, Hez McMurray, Possu Huang, John Love, Rhonda Digiusto and Cynthia Carlson.

My first years at Caltech were particularly trying, during the time I was still searching for a research direction, and being simultaneously pounded by heavy courseload requirements. Throughout this time, invaluable to me was the dynamic group we assembled to ease the homework burden: particularly Teviet Creighton, Jim Mason, David Vernooy, Alexa Harter and John Cortese.

There are many other friends I will be sorry to leave behind too. But here I will mention only a couple. My time in L.A. would have been unimaginably more barren without the flatmates I was so fortuitous to find.

Mike Levene, brimming with entrepreneurial ideas and who could always be found restless at 3 a.m. downstairs, ready to assail me with his latest idea. Christina Hood, fellow antipodean and physicist, Los Feliz hipster and invaluable companion on our journeys to comprehend Los Angeles and America. And Aaron Batista, Erik Winfree and Alli Magidsohn, whose conversations have been engrossing and insightful. Also, thanks to my Australian friends who made it over here to visit!

Most importantly I would like to thank my family, Margery, Ross and brother Gavin, for being patient despite the hardships of living so far apart for so long, and for taking every opportunity to visit. (And frequently whisking me off to some exotic location, the Grand Canyon, Costa Rica, British Columbia...) I look forward to being back together with you.

Looking back on my time here, it has certainly been an interesting five years. Los Angeles has seemed at times an intensely atomizing experience, largely fueled by its celebrated single-occupant-vehicle-strip-mall-parking-lot culture. Yet somewhere in the sprawling infinity of L.A., you can always find something to fascinate you. I am glad to have adopted some part of the creative energy of this city and to be able to call it, in some admittedly (thankfully?) small sense, home.

Abstract

Our goal is a quantitative algorithm for protein design which is not limited to particular protein folds. In this endeavor there have been previous successes designing protein cores, where van der Waals packing, and the tendency of hydrophobic amino acids to avoid contact with solvent, are the dominant forces. On the surfaces of proteins, efforts at α-helix surface design have also been successful, where hydrogen bonding and α-helix propensities are additionally important. However, there are no algorithmically designed stable β-sheet surfaces.

One of the energy terms expected to be important for β-sheet surface design is β-sheet propensity. No concise theory explaining the amino acids' differing β-sheet propensities has previously been developed. In this thesis, I examine the underlying physical-chemical basis for β-sheet propensities, and show that they are caused primarily by van der Waals interactions between the side chains and the local backbone.

I then consider an additional energy term, a penalty for the exposure of hydrophobic surface area. This is not a thermodynamic term, but rather one that can be justified through "negative design," in which alternative badly folded ground state structures are disfavored. I show experimentally that this term improves the algorithm's predictive ability, and determine its strength in the context of our previously published energy expression. In order to do this, I developed a two body approximation for buried and exposed surface area calculation which very closely reproduces the true surface areas.

Finally, I develop a general method for calculation of the optimal energy expression for protein design, from theoretical lattice model studies, and apply it to real proteins. In particular the method is applicable to β-sheet

surfaces. The β-sheet surfaces of two real proteins are thus redesigned and made experimentally. The culmination is a protein of greater stability than the naturally occurring protein. This is the first time greater stability has been achieved solely through mutations to the β-sheet surface, and marks a major step towards an ability to completely design *de novo* arbitrary proteins of arbitrary size.

Successful protein design will lead to many practical applications, from new catalysts for industrial processes, to improved stability for existing medicines, to completely novel enzymes.

Table of Contents

Chapter 1.

Introduction

"I am, as I said, inspired by the biological phenomena in which chemical forces are used in repetitious fashion to produce all kinds of weird effects, one of which is the author."

Richard P. Feynman, "There's plenty of room at the bottom," 1959.

Introduction

In 1959 Richard Feynman outlined his hopes that one day physicists would be able to create materials by manipulating the locations of individual atoms, vastly superseding current chemical synthesis techniques. As a first step, he suggested two competitions. One was to take the information on the page of a book and shrink it twenty-five thousand times, but so that it could still be read. The second was to build a motor only half a millimetre across.

As Feynman was aware, biological systems encode information far more compactly than the winner of the first competition would. After all, practically every nucleus of every cell in our bodies contains all the information necessary (i.e., about three billion base pairs of DNA) for us to grow up from a zygote, stored in chromosomes only a few millionths of a metre across. For example, this thesis, written in the language of DNA, would be 10^{17} times smaller (and easily readable by current sequencing techniques) – a sizable improvement over Feynman's initial challenge!

The second challenge, it turns out, was also long ago taken up by nature. Every cell of our body also contains a huge number of

electrochemically-driven motors, which undergo a three-stroke cycle to generate a molecule (adenosine triphosphate, or ATP) which can be utilized by many other processes in the cell for their energy needs. These motors are part of a protein known as ATP synthase, and are a tiny ten billionths of a metre across – about fifty thousand times smaller than stipulated by the competition. Many other protein motors have also been discovered, including bacterial flagellar motors, which propel bacteria; kinesin, responsible for neuronal transport; and myosin, responsible for muscle contraction, and thus for our ability to move at all.

Nature having satisfactorily solved the two challenges, how might we achieve Feynman's grand vision of designing materials by manipulating the positions of individual atoms? He envisioned a purely "physics" approach to the problem, but given nature's proficiency at the job, it makes sense to consider instead a "biophysics" approach.

In this thesis I describe advances I have made in the field of computational protein design. Proteins are complex three-dimensional molecules which perform most of the tasks necessary for life. They can be viewed as tiny machines, only a few billionths of a metre across, whose abilities include everything from the synthesis of ATP described above, to metabolism, to oxygen transport in the blood, to regulating salt concentrations in our cells, to duplicating DNA. Despite proteins' wide-ranging capabilities, an individual protein can actually be uniquely specified by a one-dimensional sequence in DNA (i.e., a gene). This sequence is translated into a one-dimensional sequence of amino acids which, with no further external help, self-assembles into a functional complex three-dimensional protein. This correspondence between one-dimensional sequence and three-dimensional shape means something very complex arises

from something very simple, enabling a description of our entire bodies to be encoded in a few cubic micrometres, and thus making evolution through natural selection possible. Quite apart from the intellectual satisfaction of understanding such a critical element of life, a knowledge of the rules which govern protein self-assembly would enable us to make molecules with shapes of our own design, thereby achieving the spirit of Feynman's vision. An ability to design proteins thus opens up a whole realm of nanotechnology whose possibilities are almost unfathomable.

From these heady heights, let us examine the details.

What is a Protein?

Proteins are linear polymers of amino acids. Because it is a linear polymer, a protein (or polypeptide) consists of a sequence of amino acid side chains branching off an unbranched backbone. The twenty naturally occurring amino acids are shown in Figure 1. The sequence of amino acids is called the primary structure of a protein.

In the cell, particular amino acid sequences are specified by genes, in deoxyribonucleic acid (DNA). The DNA is first transcribed into ribonucleic acid (RNA), which is then translated into protein. Three neighboring RNA bases (of which there are four, adenine, cytosine, uracil and guanine, denoted A, C, U and G respectively) are read at a time, and interpreted as a particular amino acid to be appended to the protein, according to the "genetic code." In some cases a protein is then modified to better carry out its function; hemoglobin, for example, contains the prosthetic group heme. Additionally, functional proteins may result from assemblies of more than one polypeptide chain.

As discussed above, proteins are perhaps best viewed as self-organizing molecular machines – the ultimate nanotechnology. They are the enzymes that make life possible, even responsible for the translation and transcription of DNA to form new proteins (although proteins are sometimes coupled with other molecules, including RNA). Proteins can also play structural roles, such as that of collagen, which maintains the cellular structure of connective tissue, and messenger roles, such as that of the polypeptide hormone insulin. Proteins are also ubiquitous – they make up 18% of the weight of a mammalian cell (water accounts for 70%).

Protein Folding

The incredible range of protein functions is possible because, depending on the sequence of amino acids which constitute it, a protein always folds into a particular "native" compact structure (under physiological conditions, and not considering exceptional cases such as prions). Only the sequence of amino acids comprising the protein is necessary to determine the ultimate structure of the protein – for example, dilute protein solutions may be heated until the protein unfolds (denatures), and cooled again, to form functional protein again. However, as discussed in more detail below, no theory can yet predict a protein's native structure from just its sequence, despite nature's ability to solve this problem in usually less than a second.

The ability of a protein to fold is demonstrated in Figure 2. The backbone has a degree of freedom at each single bond, at which rotation around the bond is possible. Such rotation is described by a dihedral angle. Each amino acid has three dihedral angles, not including those specific to the side chain, denoted ϕ, ψ and ω. The last of these, however, corresponds to rotation about the peptide bond (the bond formed by the polymerization

between two amino acids), which has a slight double bond character and is constrained to be flat, with $\omega \approx 180°$, as shown in the figure (although $\omega \approx 0°$ may also precede proline). Therefore, a folded protein backbone can be largely described by a sequence of (ϕ, ψ) pairs, with one pair for each amino acid (or "residue") in the protein.

Protein structures, while not as regular as the double helix of DNA, nevertheless show some regularity. The backbone often adopts conformations known as α-helices and β-sheets; these are known as elements of secondary structure (Figure 3). (Frequently, protein structures are depicted showing only the backbone, with stylized α-helices and β-sheets.) One feature which identifies units of secondary structure is their pattern of hydrogen bonding, as shown in the figure. They also have characteristic (ϕ, ψ) angles, as shown in Figure 4. Linking these elements are less well-defined turns, some of which are themselves common motifs.

The organization of secondary structure units in a protein is known as the protein's tertiary structure. The same tertiary structures can be seen even in proteins with markedly different sequences.

Protein Structure Prediction

No theory currently explains which physical and chemical forces are most important in the folding process, or is able to predict the structure that a particular amino acid sequence will adopt upon folding. One reason is that the process occurs in aqueous solution. Indeed, the desolvation of hydrophobic side chains into a hydrophobic core – often referred to as the "hydrophobic effect" – is characteristic of protein folding. Unfortunately, an accurate but computationally tractable way to model solvation has yet to

emerge – modeling every solvating water molecule is beyond current computing power.

Another reason is that protein stability is finely balanced between competing effects. A protein's stability is defined to be the change in free energy, ΔG, between its native (folded) and denatured (unfolded) states. Both enthalpy and entropy contribute to free energy. The change in entropy on folding is large and negative, as the polypeptide chain moves from a loosely restricted state to an essentially unique state. This opposing force almost exactly cancels the benefits gained from improved physical and chemical interactions in the native state. For example, at 25 °C and pH 2.5, the protein ribonuclease, which hydrolyzes RNA, has a change in enthalpy on folding, ΔH, of -238 kcal/mol, but an entropic contribution to free energy, $-T\Delta S$ (where T is the temperature), of 231 kcal/mol. The resulting protein stability is thus only -7 kcal/mol.

Protein Design

It makes sense to approach the issue of what forces are most important in protein folding from another direction. Instead of asking what structure a given amino acid sequence will adopt, we can ask what amino acid sequence will adopt a given protein structure. This approach has a number of advantages over the more direct approach. In particular, it is easier to experimentally test our method by actually constructing the predicted sequences and determining their structures. Another important advantage is provided by degeneracy – there are many sequences which will fold to a given structure, but only one structure per (foldable) sequence.

As discussed earlier, successful protein design will lead to practical applications, from new catalysts for industrial processes, to improved stability

for existing medicines, to completely *de novo* enzymes. It may also pave the way towards a general theory for designing self-organizing macro-molecules.

The Road Ahead

Computational protein design is taken up in more detail in Chapter 2, giving an overview of the forces thought to be important for protein design, and the main algorithms we use to search through the enormous number of sequences available. Important issues such as backbone flexibility and negative design are also discussed.

In Chapter 3, I take a step back from designing proteins and ask what causes some amino acids to occur more frequently than others in β-sheets. A concise physical theory behind this phenomenon had previously been lacking. I show that it is largely due to the van der Waals interactions between the side chain and the local backbone. This result is interesting because the non-local nature of β-sheets (i.e., β-strands from different regions of the sequence fold up adjacent to each other) had suggested that non-local effects might play a dominant role.

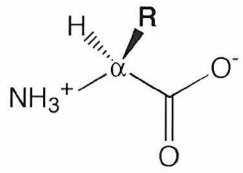
In Chapter 4, I embark on a study of the role of negative design in the design of β-sheet surfaces. Specifically, I examine whether penalizing exposure to solvent of hydrophobic surface area can improve β-sheet surface design efforts. Chapter 5 details my development of a new two body treatment of surface area determination, which was required for accurate calculation of the exposed hydrophobic surface area.

Finally, in Chapter 6, I develop a general theoretical approach for incorporating negative design into the design of real proteins (again, specifically their β-sheet surfaces), and apply it to the design of two real proteins. The culmination is a protein of greater stability than the naturally

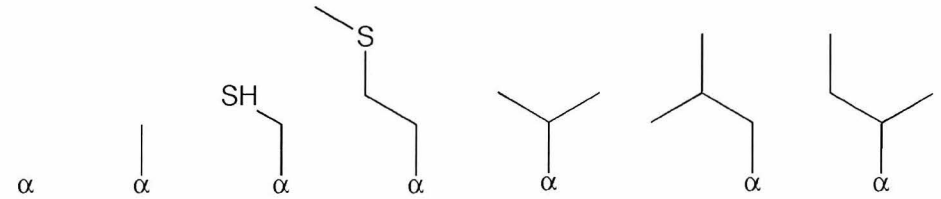
occurring protein. This is the first time greater stability has been achieved solely through mutations to the β-sheet surface, and marks a major step towards an ability to completely design *de novo* arbitrary proteins of arbitrary size.

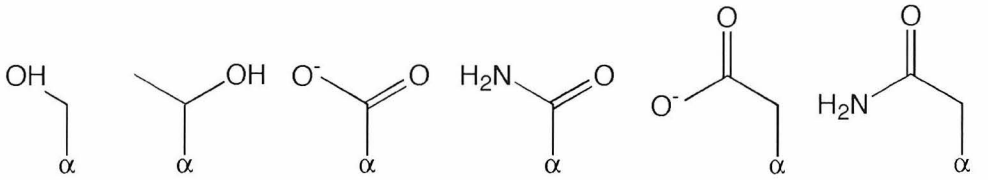
Figure 1-1. Amino acids. a) The naturally occurring biological amino acids are all α-amino acids, meaning they have one carbon (called the α-carbon) between the amino (NH_3^+) and acid (COO^-) termini. They differ from one another only in the side chain **R**. The α-carbon has four different substituents, so that sterically different molecules result from the two possible placements of the side chain and the hydrogen; all naturally occurring amino acids have the same left-handed placement shown (with the side chain coming out of the page and the hydrogen going into the page). b) The side chains of the 20 naturally occurring amino acids, and their three-letter abbreviations. The respective full names and one-letter codes are glycine G, alanine A, cysteine C, methionine M, valine V, leucine L, isoleucine I, serine S, threonine T, aspartic acid D, aspargine N, glutamic acid E, glutamine Q, lysine K, arginine R, histidine H, proline P, phenylalanine F, tyrosine Y and tryptophan W. Proline is unique in that it reconnects to the backbone nitrogen.

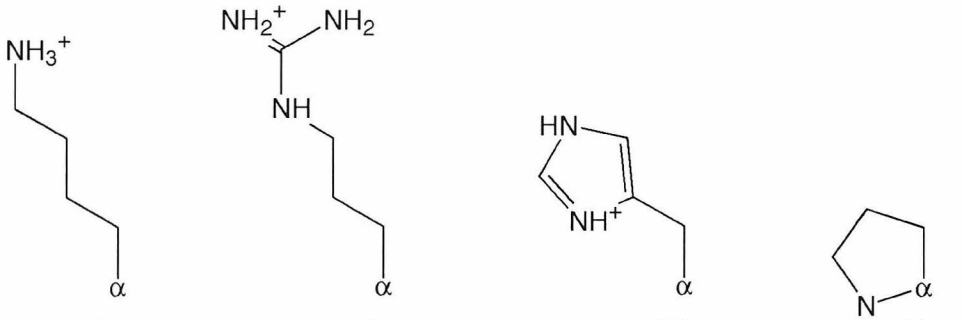
A


H
R
α
O⁻
NH3+
O

B


S
SH
α
Gly
Ala
Cys
Met
Val
Leu
Ile


OH
OH
O⁻
O
H2N
O
O⁻
O
H2N
O
α
Ser
Thr
Asp
Asn
Glu
Gln


NH3+
NH2+
NH2
NH
HN
NH+
N
α
Lys
Arg
His
Pro

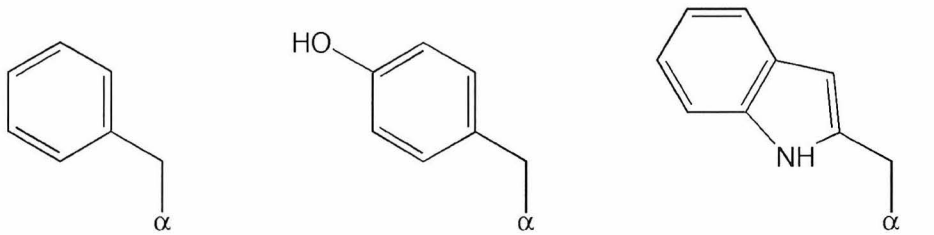

HO
NH
α
Phe
Tyr
Trp

Figure 1-2. A section from a sequence of polymerized amino acids (peptide) showing the backbone dihedral angles ϕ and ψ, whose rotation lead to the phenomenon of protein "folding." The direction defined as positive rotation is shown. The extended conformation of the chain is shown, when both dihedral angles are defined to be 180°. Figure copyright Irving Geis.

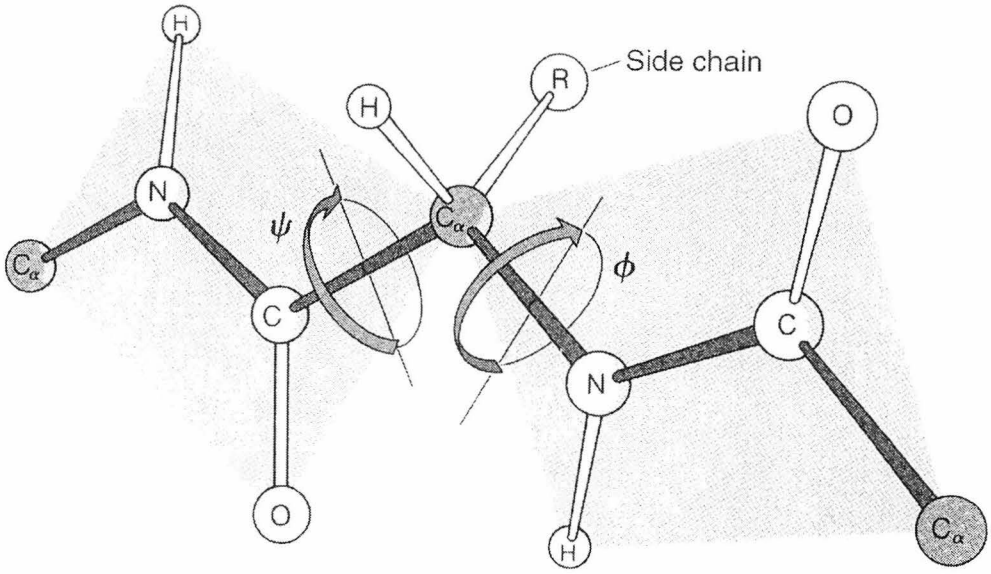
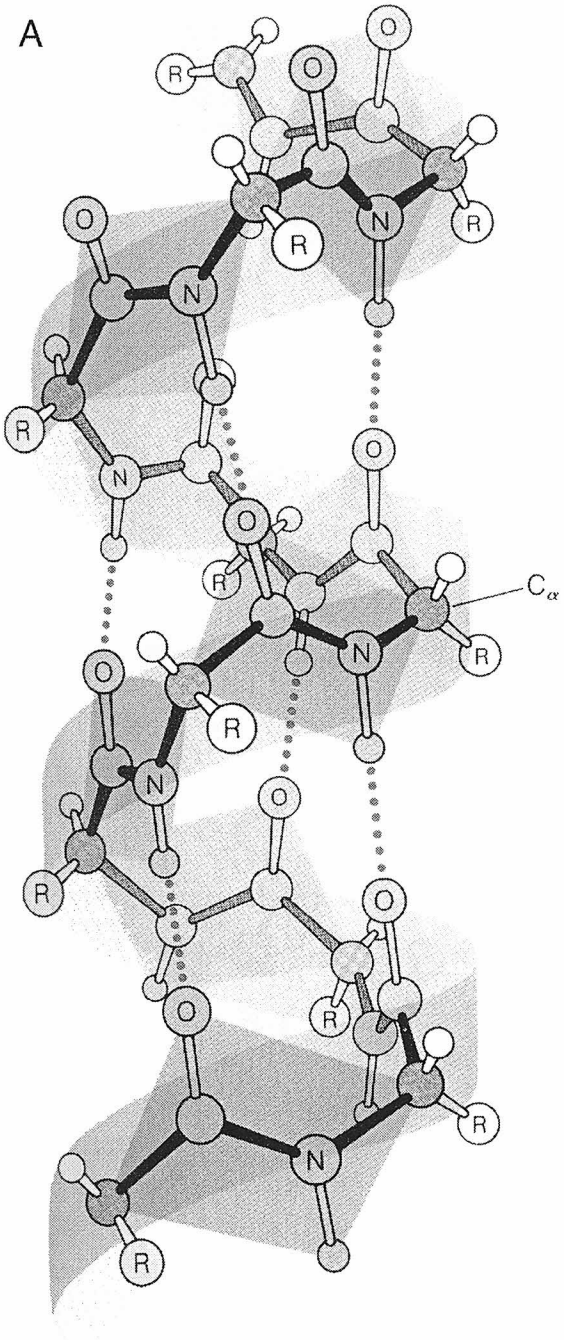
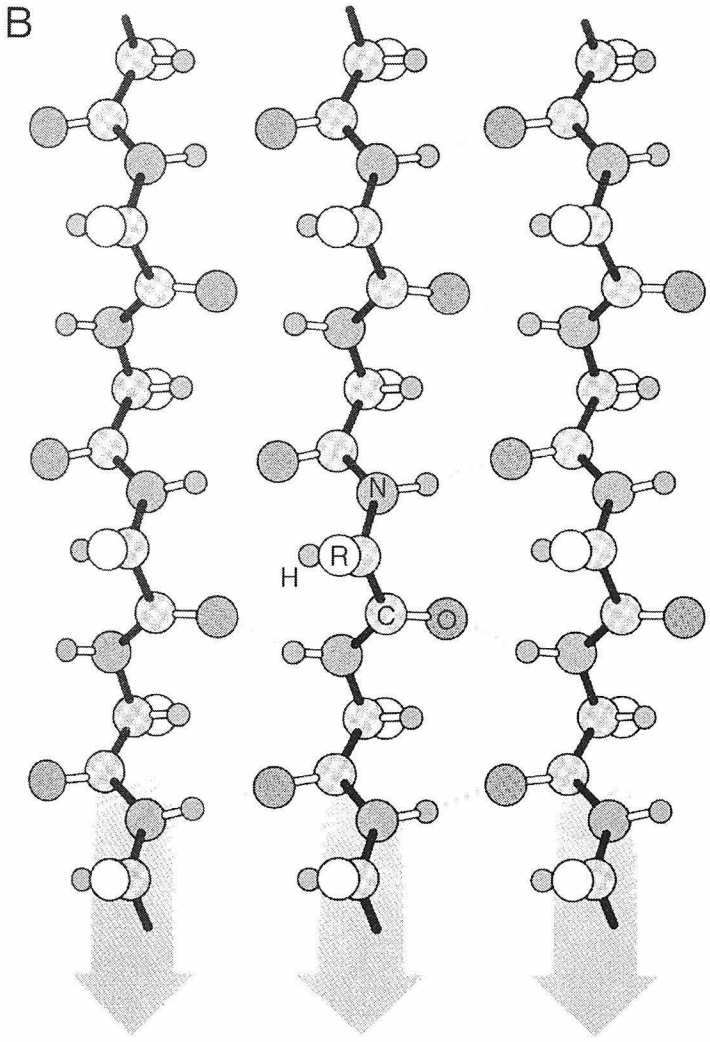

H
R
Side chain
H
O
N
C_α
ψ
C_α
ϕ
C
C
N
O
H
C_α

Figure 1-3. The major types of secondary structure. a) The α-helix. b) The parallel β-sheet, in which neighboring β-strands go in the same direction. c) The anti-parallel β-sheet, in which neighboring β-strands go in opposite directions. Figures copyright Irving Geis.


A
O
R
N
C_α


B
N
H
R
C
O

C

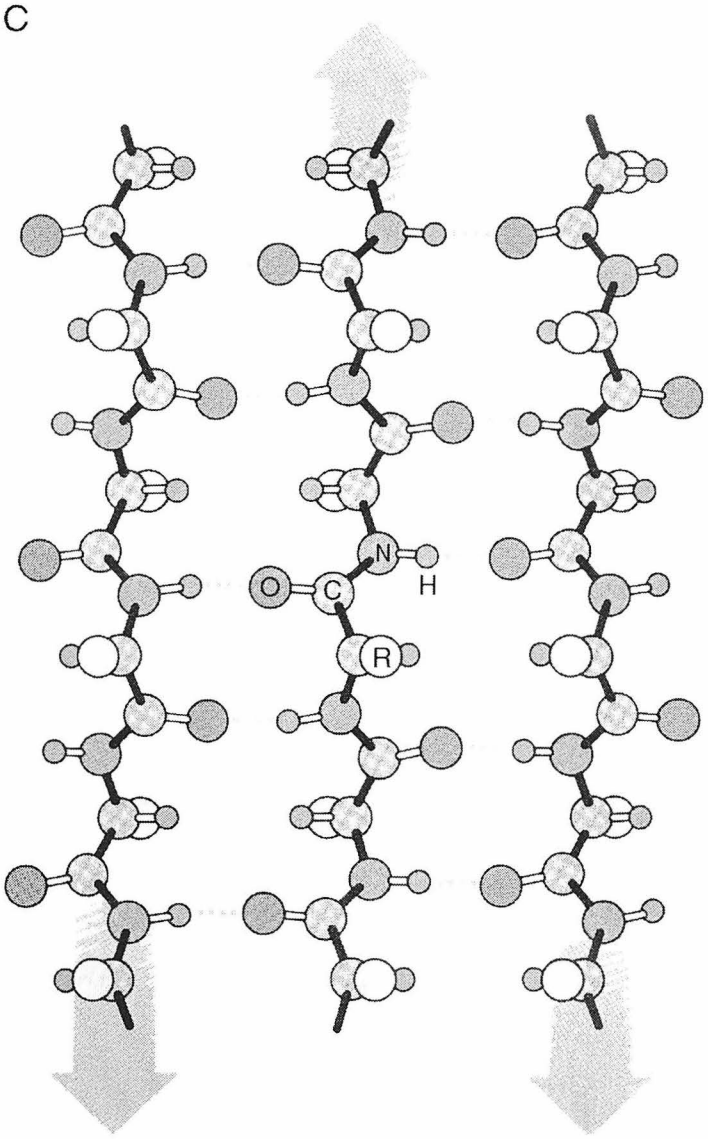
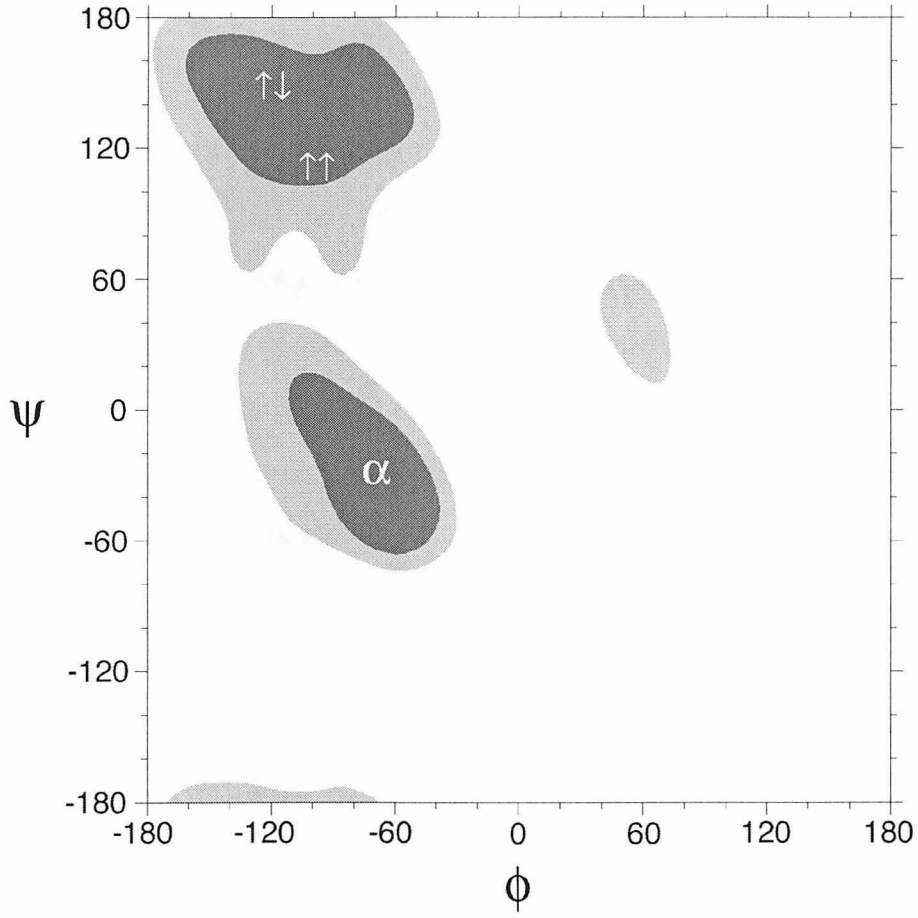


Figure 1-4. A "Ramachandran" plot showing the sterically allowed regions of (ϕ, ψ) space. The broad features of this plot are valid for all amino acids except glycine, which has more flexibility, and proline, which is more restricted. Specific regions on this plot are associated with the α-helix, the parallel β-sheet and the anti-parallel β-sheet, as shown. 80% of (ϕ, ψ) angles from crystallographically-determined structures lie within the dark regions of the plot, 95% lie within the dark and medium regions, and 98% lie within all the shaded regions.


180
120
60
0
-60
-120
-180
ψ
-180
-120
-60
0
60
120
180
ϕ
↑↓
↑↑
α

Chapter 2.

Computational Protein Design

The text of this chapter is partially adapted from the publication

Street A.G. and Mayo S.L. (1999) Structure 7, R105-R109.

Abstract

A "protein design cycle," involving cycling between theory and experiment, has led to recent advances in rational protein design. In particular a reductionist approach, in which protein positions are classified by their local environments, has aided development of an appropriate energy expression. Here we discuss the computational principles and practicalities of the protein design cycle, including energy minimization, backbone flexibility and negative design issues.

Introduction

There are many reasons to pursue the goal of protein design. In medicine and industry, the ability to precisely engineer protein hormones and enzymes to perform existing functions under a wider range of conditions, or to perform entirely new functions, has tremendous potential. Furthermore, in the case of rational protein design, the obtained knowledge would likely be linked to a more complete understanding of the forces underlying protein folding, enabling more rapid interpretation of the wealth of genomic information being amassed. Advances in protein design may also make possible the construction of a range of other self-organizing macromolecules.

Although some steps have been taken towards rationally designing functional enzymes (Wilson et al., 1991), such a goal lies some distance away. Currently, attention is focused on redesigning portions of proteins to insert particular motifs, increase stability or modify function. Examples include the engineering of metal binding centers, reviewed recently by Hellinga (Hellinga, 1998b), and the introduction of disulfide bonds (Pabo & Suchanek, 1986; Matsumura & Matthews, 1991; Yan & Erickson, 1994). Theoretical work in the context of lattice models has also led to important insights. This work has been recently reviewed (Dill et al., 1995; Shakhnovich, 1998).

Attempts to design entire proteins *de novo* have been increasingly successful over the past decade. Early design efforts typically led to poorly characterizable states or molten globules, instead of a single target fold (Betz et al., 1993). Other difficulties became apparent when a designed α-helical dimer (O'Neil & DeGrado, 1990) was shown to actually form a trimer (Lovejoy et al., 1993). This and subsequent studies relied on largely qualitative examination of the target molecule (Bryson et al., 1995), making generalization to other targets difficult.

This review focuses on the advances made in computational approaches to protein design. In particular, we examine those atomistic approaches which involve cycling between experiment and theory in a "protein design cycle."

Energy Expression

Atomistic protein design requires an energy expression or force-field to rank the desirability of each amino acid sequence for a particular backbone structure. Over the last decade, elements of a suitable energy expression for atomistic protein design have been suggested and explored. To avoid over-

fitting and to focus on only the most important contributors, the energy expression should contain as few terms as possible while maintaining predictive power. Communication between theory and experiment is required to determine which energy terms to include, and the relative importance of the included terms. In a protein design cycle, an energy expression is used to generate sequences which are subsequently made in the laboratory. Alterations and additions to the energy expression are then considered which improve the correlation between the computed and experimentally determined properties of the sequences. The improved energy expression is then used to generate new sequences, completing the cycle.

Energy Minimization

In order to experimentally test the energy expression, the minimum energy sequence on the target backbone must be determined. In the simplest implementation, the energy of every possible sequence is calculated using the energy expression, and the lowest energy sequence is reported. The size of most problems of interest renders this exhaustive approach impractical. Ignoring the possibility of multiple conformations of each amino acid, allowing the 20 naturally occurring amino acids at every position of a 100 amino acid protein yields 10^{130} possible sequence solutions. Clearly, ingenious energy minimization techniques are necessary.

Published search algorithms including self-consistent mean-field approaches (Lee, 1994; Vasquez, 1995; Koehl & Delarue, 1996), Monte Carlo techniques (Lee & Levitt, 1991; Hellinga & Richards, 1994), neural networks (Kono & Doi, 1996) and genetic algorithms (Desjarlais & Handel, 1995; Pedersen & Moult, 1996) share the advantage of being able to sample large combinatorial space but the disadvantage of not being guaranteed to find the

global optimal solution. By contrast dead-end elimination, and branch and terminate (discussed in more detail below) are search algorithms whose final solution is guaranteed to be the global optimum, but which require the discretization of side chain conformations into rotamers (Janin et al., 1978; Ponder & Richards, 1987). Such requirements will be discussed below. Search algorithms have been recently reviewed (Desjarlais & Clarke, 1998).

Dead-End Elimination

The dead-end elimination theorem was originally introduced (Desmet et al., 1992) to aid protein homology modeling, in which side chain identities are known and the adopted side chain conformations (or rotamers) are desired. Iteration of the theorem progressively eliminates rotamers which can be shown not to be part of the global minimum energy conformation (GMEC). Denoting positions on the protein backbone by i, j and specific rotamers at each position by i_c, j_c (where c is a position-specific index indicating the rotamer present), the energy E of a conformation can be written

$$E = E_{\text{template}} + \sum_i E(i_c) + \sum_i \sum_{j<i} E(i_c, j_c) \; . \tag{1}$$

Here E_{template} is the template (or backbone) self-energy, $E(i_c)$ is the energy of the rotamer i_c interacting with the template only, and $E(i_c,j_c)$ is the pairwise energy of interaction between rotamers i_c and j_c. The theorem states that, for a pair of rotamers r and t at the same position i (denoted i_r and i_t), if

$$E(i_r) + \sum_j \min_s E(i_r, j_s) > E(i_t) + \sum_j \max_s E(i_t, j_s); \quad i \neq j \tag{2}$$

then i_r is not in the GMEC. Conceptually, the theorem says that if the best possible energy a rotamer could achieve in its interactions with other rotamers is higher than the worst possible energy of another rotamer at the

same position, then it cannot be a member of the GMEC. This is illustrated in Figure 1a.

The method has since been improved substantially, so that it may be applied to the much larger problem of protein design, in which the number of rotamers allowed at each position may be an order of magnitude greater than in homology modeling. Typically an energy cutoff is applied to remove the worst rotamers from consideration before applying dead-end elimination (DeMaeyer et al., 1997). A less restrictive criterion (Goldstein, 1994) replaces (2) with

$$E(i_r) - E(i_t) + \sum_j \min_s \left(E(i_r, j_s) - E(i_t, j_s) \right) > 0; \quad i \neq j \ , \tag{3}$$

as demonstrated in Figure 1b. Goldstein also considered unifying two or more positions into one "super-residue." Critical to the ability of the method to find the GMEC is the extension of the theorem to eliminate pairs of rotamers (Desmet et al., 1994). Defining

$$\varepsilon([i_r, j_s]) = E(i_r) + E(j_s) + E(i_r, j_s) \tag{4}$$

and

$$\varepsilon([i_r, j_s], k_t) = E(i_r, k_t) + E(j_s, k_t) \ , \tag{5}$$

a rotamer pair $[i_r,j_s]$ is flagged if there exists another rotamer pair $[i_u,j_v]$ such that

$$\varepsilon([i_r, j_s]) - \varepsilon([i_u, j_v]) + \sum_k \min_t \left\{ \varepsilon([i_r, j_s], k_t) - \varepsilon([i_u, j_v], k_t) \right\} > 0; \quad i, j \neq k \ . \tag{6}$$

Flagged pairs are inconsistent with the GMEC and may be ignored in the "singles" summation of (3) as well as in future iterations of (6).

Application of (6) to every possible pair of rotamers involves a calculation which scales as the fourth power of the number of rotamers per

position, which significantly slows the search. By carefully employing quantities which can be calculated more quickly, it is possible to apply (6) only to pairs with a high likelihood of being flagged (Gordon & Mayo, 1998). Symmetry arguments further reduce the number of pairs which need to be examined by a factor of four.

Further enhancements can be derived from "conformational splitting" (N. Pierce, unpublished results), in which conformational space is partitioned by pulling, for a given position k, the interaction energies involving position k outside the summation in (3). The technique can be further extended to more than one such position.

Branch And Terminate

Branch and bound algorithms, from which the branch and terminate algorithm derives, have been applied to many problems of interest in structural biology in recent years (Gordon & Mayo, 1999). The search problem is arranged as a combinatorial tree, where each path through the tree corresponds to a solution to the problem. For protein design, each level of the tree corresponds to an amino acid position, and each node represents a particular rotamer at each position. The object is to find the one path through the tree which corresponds to the GMEC. Given a path down to a given level of the tree (i.e., with certain rotamers already chosen for some positions), a bounding energy is computable which is guaranteed to be lower than (or equal to) the lowest energy possible through the remainder of the tree. The algorithm keeps track of the lowest energy it has found so far, and exhaustively searches the combinatorial tree, in the process pruning away branches with higher bounding energies.

Two features of the branch and bound algorithm are apparent. First, the calculation of the bounding energy must balance stringency (so that as many branches are pruned as possible, resulting in faster execution) against the time it takes to compute (since the energy is calculated at each node, a complex bounding expression can significantly affect performance). A suitable balance can be found by recasting the energy (1) as

$$E = \sum_i \sum_{j \neq i} F(i_c, j_c) \tag{7}$$

where

$$F(i_c, j_c) = \frac{1}{2}\left(\frac{E(i_c) + E(j_c)}{p-1} + E(i_c, j_c) \right) \tag{8}$$

and p is the number of amino acid positions. Then, at a given level in the tree, all the rotamers above that level have been fixed (denote the set of such positions F), and the remaining rotamers are variable (V). One can then expand (7) into four terms, two of which are identical, to yield

$$E = \sum_{i \in \mathrm{F}} \sum_{\substack{j \in \mathrm{F} \\ j \neq i}} F(i_c, j_c) + 2 \sum_{i \in \mathrm{V}} \sum_{j \in \mathrm{F}} F(i_c, j_c) + \sum_{i \in \mathrm{V}} \sum_{\substack{j \in \mathrm{V} \\ j \neq i}} F(i_c, j_c) \ . \tag{9}$$

The most stringent bounding expression that can be derived from (9) is thus

$$E_{\mathrm{bound}} = \sum_{i \in \mathrm{F}} \sum_{\substack{j \in \mathrm{F} \\ j \neq i}} F(i_c, j_c) + \sum_{i \in \mathrm{V}} \min_r \left\{ 2 \sum_{j \in \mathrm{F}} F(i_r, j_c) + \sum_{\substack{j \in \mathrm{V} \\ j \neq i}} \min_s F(i_r, j_s) \right\} . \tag{10}$$

The second observation is that when the levels in the tree correspond to amino acid positions, their ordering is arbitrary. Yet the organization of the tree significantly impacts the performance of the algorithm – placing a branch at the lowest level of the tree that would be pruned if placed at the top level results in the same pruning step being repeated unnecessarily. This suggests a

pre-processing procedure in which each amino acid in turn is placed at the top level of the tree, and rotamers pruned from this position are discarded from the rest of the optimization. This procedure is called "termination," and may be repeated until no further rotamers are pruned. The branch and terminate algorithm applies termination at each level of the tree, resulting in a significant performance improvement over the original branch and bound algorithm (Gordon & Mayo, 1999).

The search algorithms described in these two sections can be used in concert. In particular, we have found that once application of dead-end elimination has reduced the number of rotamer conformations available by several (often over twenty) orders of magnitude, branch and terminate can frequently be used to find the GMEC more quickly than continuing with dead-end elimination. This opens up the possibility of applying computational protein design to ever larger systems.

Discretization of Side Chain Conformations

To place a reasonable limit on the complexity of the computation, the allowed side chain conformations are typically chosen from a library of discrete possibilities, known as rotamers. This discretization is necessary for some efficient search algorithms to be applicable – in particular, the dead-end elimination theorem.

Discretization of the side chain conformations increases the likelihood of "false negative" results. To be useful, atomistic protein design has only to output a subset of the sequences leading to the target fold, with simulation energies that correlate with their experimental stabilities. The simulation does not need to predict how well externally supplied sequences will fit the target fold. For example, the crystallographic structure of the Streptococcal

protein G B1 domain (GB1) (Gronenborn et al., 1991) shows Leu 7 in an unusual conformation which does not appear in standard rotamer libraries (Ponder & Richards, 1987). Therefore, an atomistic algorithm using such a library may not suggest Leu at position 7 in the top ranked sequences.

The effect of the size of the rotamer library has also been considered (DeMaeyer, et al., 1997; Tufféry et al., 1997); in general, the larger the better. However, if the library contains too many similar conformations of each amino acid, the energy landscape is flattened and energy minimization can be slow.

Residue Classification

A reductionist approach to protein design, in which subsets of a protein are designed independently, has proven fruitful. Computational attempts to design protein cores date back many years. More recently, there have been attempts to design surfaces and boundary positions as well.

The size of the design problem is reduced if only a subset of amino acid types need be considered in each of these three classes of residue positions. Protein cores are typically composed of hydrophobic amino acids, and protein surfaces are largely composed of hydrophilic amino acids, but the boundary residues must be selected from the full range of amino acids since these positions are observed to be both hydrophobic and hydrophilic. An automated way to classify residue positions is desirable, and a number of approaches have been described (Sun et al., 1995; Dahiyat & Mayo, 1997a).

The important components of the energy expression relevant to the core, surface and boundary will be discussed in the following sections.

The Core

Early attention on the protein design problem focused on the generally hydrophobic cores of proteins. It is believed that the folding process is driven principally by hydrophobic collapse of the polypeptide, implying that a well-designed hydrophobic core is crucial to the protein's structure and stability (Dill, 1990).

As might be expected, van der Waals forces (that is, packing constraints) are crucial when designing the protein core. Models in which packing constraints are the only element of the energy expression are able to predict the stabilities of core mutations with high accuracy, when polar substitutions are not allowed (Hellinga & Richards, 1994; Desjarlais & Handel, 1995; Dahiyat & Mayo, 1996; Dahiyat & Mayo, 1997b; Lazar et al., 1997). The importance of packing constraints can be determined by scaling the atomic van der Waals radii by a factor α. When α is varied to very high (>105%) or very low (<85%) values, implying too little or too much volume being packed into the available space, the resulting proteins exhibit unfolded or molten globule-like behavior (Dahiyat & Mayo, 1997b). This is not surprising. Too much volume clearly requires the backbone to shift to accommodate the excess (Baldwin et al., 1993). Too little volume would either leave cavities in the core, which have been shown to destabilize proteins (Lim & Sauer, 1989), or again force the backbone to shift to fill the cavity. When the protein backbone is significantly different from the model backbone, the model can no longer accurately predict the protein's stability, and there may cease to be a single stable folded state. The optimal value of α was found to be 90%, implying that a slight over-packing of hydrophobic residues in the core can actually stabilize a designed protein (Dahiyat & Mayo, 1997b). The benefit of using slightly diminished van der Waals radii can also be interpreted in

terms of accommodating some backbone and rotamer flexibility (discussed in a later section).

Consistent with the belief that the hydrophobic effect is a dominant cause of protein folding, the protein design cycle has been used to show that solvation effects also play an important role in the design of protein cores (Dahiyat & Mayo, 1996). The hydrophobic effect is usually approximated as an energy benefit proportional to the amount of solvent-accessible hydrophobic surface area that is buried upon folding (Eisenberg & McLachlan, 1986). A penalty for burying polar area may also be included. Calculation of solvation energies is complicated by the need to construct the energy expression as a sum of two-body interactions (Kurochkina & Lee, 1995; Street & Mayo, 1998).

An entropic term has been tested (Kono et al., 1998), which may improve correlation between predicted energy and biological activity (Hellinga & Richards, 1994). Such a term should in particular penalize methionine, whose loss of rotational freedom upon burial in a protein core can otherwise lead to destabilized proteins (Gassner et al., 1996).

The Surface

With the successful redesign of a range of protein cores, it is natural to consider the redesign of protein surfaces. Despite the incontrovertible role of the hydrophobic core in folding, the surface is also crucial to a protein's structure and stability.

The protein design cycle has been utilized to design surface sites, using as a starting point the energy expression determined from studies of protein cores. These studies showed the importance of electrostatics and hybridization-dependent hydrogen bonds (Dahiyat et al., 1997). In the case of α-helical surfaces, no further energy terms are necessary to achieve good

predictive ability. This is possibly because the side chains which are better hydrogen bond formers are also good α-helix formers, as quantified by α-helical propensity (Chakrabartty et al., 1994; Dahiyat, et al., 1997).

The above energy terms are not sufficient to design β-sheet surfaces (Hecht, 1994). It may be necessary additionally to directly bias the energy expression towards those side chains with good β-sheet propensities (see Chapter 3). This is physically justifiable because common energy expressions do not otherwise include side chain self-energies, which must at some level lead to propensities.

It is also possible that a main source of β-sheet stability is to be found elsewhere, for example in the hydrogen bonds that cause alignment with neighboring β-strands. In the case of anti-parallel β-strands, the turn joining the two strands plays an important role. Modifying the turn's component residues can seriously affect protein stability (Garrett et al., 1996; Ybe & Hecht, 1996; Blanco et al., 1998). In the case of non-continuous strands, it has been suggested that small clusters of hydrophobic area on the surface may help to set the register (Tisi & Evans, 1995). The hydrophobic effect may drive neighboring strands to align in such a way as to bury as much of the exposed hydrophobic area as possible, for example by covering it with long amphiphilic side chains. The role of hydrophobic exposure will be examined in Chapter 4.

The Boundary

Some residues cannot be easily classified as core or surface. Depending on the side chain orientation they can interact with either the protein's core or with the solvent. One example is Trp 43 of GB1 (Dahiyat & Mayo, 1997b), which is predicted by modeling to rotate out into the solvent when nearby

core residues are replaced with larger side chains. Such unfavorable behavior can be attenuated by a hydrophobic exposure penalty (Sun, et al., 1995; Dahiyat & Mayo, 1997b).

Recent work has shown that the design of boundary residues can lead to impressively enhanced stability (Malakauskas & Mayo, 1998). Just four boundary site mutations in the 56-residue GB1 improve the stability from 3.3 kcal/mol to 7.1 kcal/mol at 50 °C, converting a mesophilic protein into a hyperthermophilic protein.

Full de Novo Sequence Design

To date there exists only a single example of a complete sequence calculation in which the structure of the designed protein was experimentally shown to achieve the design target (Dahiyat & Mayo, 1997a). This calculation included one core position, 7 boundary positions and 18 surface positions, leading to a total of 10^{27} possible sequence solutions. The success of this design effort underscores the power of computational approaches.

Backbone

Most atomistic protein design efforts require a fixed backbone. The calculation is performed under the assumption that the target backbone is precisely the backbone that will be achieved by the computed sequence. Fortunately, alterations in the backbone do not necessarily lead to large changes in the accessible sequence space (Su & Mayo, 1997). In one study, a 2 Å root mean square deviation (r.m.s.d.) in the backbone led to only a 0.5 Å r.m.s.d. in predicted side chain conformations (Tufféry, et al., 1997). Backbone flexibility can be modeled by using a softer van der Waals potential – in other words, giving the modeled atoms a fuzzy edge. This effect can be obtained by

using reduced atomic radii, which has been shown to improve the stability of designed proteins (Dahiyat & Mayo, 1997b).

Protein backbone movements may be incorporated if the backbone is parameterizable (Harbury et al., 1995; Su & Mayo, 1997), although to keep the calculation tractable, the number of side chain rotamer combinations may be limited. A coiled-coil with right-handed superhelical twist, whose backbone was necessarily designed *de novo*, has recently been reported (Harbury et al., 1998), where 216 amino acid sequences were considered.

Negative Design

The importance of negative design is the subject of much discussion. Recent work by Hellinga (Hellinga, 1998a) highlights the importance of this issue in computational protein design. The inverse-folding design method determines the sequence of amino acids whose energy is lowest when threaded onto the target backbone. It is conceivable that in some cases the computed sequence may actually prefer to fold to a different target structure, and that a sequence with a slightly higher computed energy would fold to the desired target (Figure 2). Unfortunately, knowledge of which structure will be adopted by the computed sequence requires a solution to the protein folding problem. Lattice models consisting of only two amino acid types can, however, be used to perform both sequence design and fold prediction. In this context, proposals to include non-thermodynamic potential functions aimed at addressing negative design issues have been developed (Shakhnovich & Gutin, 1993; Deutsch & Kurosky, 1996; Chiu & Goldstein, 1998), and are discussed in more detail in Chapter 6. The hydrophobic exposure penalty is one example of negative design that improves predictive power (Sun, et al., 1995; Dahiyat & Mayo, 1997b). Despite the power of lattice model simulations,

it has been suggested that the design procedure may be qualitatively different in such binary patterned systems (Micheletti et al., 1998).

Conclusions

The design of proteins which fold to a specified target backbone structure is becoming possible. Future advances are likely to follow from a tight coupling of experimental and computational work in a protein design cycle, with the near future revealing ever larger protein sequences being designed *de novo*. Discovering the forces critical to the determination of backbone conformation and their coupling to sequence selection is the major challenge in solving the "complete" protein design problem. A general ability to design specific protein structures will pave the way toward the goal of rationally designing novel functional molecules.

References

Baldwin EP, Hajiseyedjavadi O, Baase WA, Matthews BW. 1993. The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme. *Science 262*: 1715-1718.

Betz SF, Raleigh DP, DeGrado WF. 1993. De novo protein design: from molten globules to native-like states. *Curr Opin Struct Biol 3*: 601-610.

Blanco F, Ramirez-Alvarado M, Serrano L. 1998. Formation and stability of beta-hairpin structures in polypeptides. *Curr Opin Struct Biol 8*: 107-111.

Bryson JW, Betz SF, Lu HS, Suich DJ, Zhou HX, O'Neil KT, DeGrado WF. 1995. Protein design: a heirarchic approach. *Science 270*: 935-941.

Chakrabartty A, Kortemme T, Baldwin RL. 1994. Helix propensities of the amino-acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *Protein Sci 3*: 843-852.

Chiu TL, Goldstein RA. 1998. Optimizing potentials for the inverse protein folding problem. *Protein Eng 11*: 749-752.

Dahiyat BI, Gordon DB, Mayo SL. 1997. Automated design of the surface positions of protein helices. *Protein Sci 6*: 1333-1337.

Dahiyat BI, Mayo SL. 1996. Protein design automation. *Protein Sci 5*: 895-903.

Dahiyat BI, Mayo SL. 1997a. De novo protein design: fully automated sequence selection. *Science 278*: 82-87.

Dahiyat BI, Mayo SL. 1997b. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci USA 94*: 10172-10177.

DeMaeyer M, Desmet J, Lasters I. 1997. All in one: a highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Fold Des 2*: 53-66.

Desjarlais JR, Clarke ND. 1998. Computer search algorithms in protein modification and design. *Curr Opin Struct Biol 8*: 471-475.

Desjarlais JR, Handel TM. 1995. De novo design of the hydrophobic cores of proteins. *Protein Sci 4*: 2006-2018.

Desmet J, De Maeyer M, Hazes B, Lasters I. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature 356*: 539-542.

Desmet J, De Maeyer M, Hazes B, Lasters I. 1994. The dead-end elimination theorem: a new approach to the side-chain packing problem. In: Merz K, Jr, Le Grand S, eds. *The protein folding problem and tertiary structure prediction*. Boston: Birkhauser. pp 307-337.

Deutsch JM, Kurosky T. 1996. New algorithm for protein design. *Phys Rev Lett 76*: 323-326.

Dill KA. 1990. Dominant forces in protein folding. *Biochem 29*: 7133-7155.

Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD, Chan HS. 1995. Principles of protein folding - a perspective from simple exact models. *Protein Sci 4*: 561-602.

Eisenberg D, McLachlan AD. 1986. Solvation energy in protein folding and binding. *Nature 319*: 199-203.

Garrett JB, Mullins LS, Raushel FM. 1996. Are turns required for the folding of ribonuclease T1? *Protein Sci 5*: 204-211.

Gassner NC, Baase WA, Matthews BW. 1996. A test of the "jigsaw puzzle" model for protein folding by multiple methionine substitutions within the core of T4 lysozyme. *Proc Natl Acad Sci USA 93*: 12155-12158.

Goldstein RF. 1994. Efficient rotamer elimination applied to protein side-chains and related spin-glasses. *Biophys J 66*: 1335-1340.

Gordon DB, Mayo SL. 1998. Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J Comp Chem 19*: 1505-1514.

Gordon DB, Mayo SL. 1999. Branch and terminate: a combinatorial optimization algorithm for protein design. *Structure* in press.

Gronenborn AM, Filpula DR, Essign NZ, Achari A, Whitlow M, Wingfield PT, Clore GM. 1991. A novel, highly stable fold of the immunoglobin binding domain of streptococcal protein G. *Science 253*: 657-661.

Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. 1998. High-resolution protein design with backbone freedom. *Science 282*: 1462-1467.

Harbury PB, Tidor B, Kim PS. 1995. Repacking protein cores with backbone freedom: structure prediction for coiled coils. *Proc Natl Acad Sci, USA 92*: 8408-8412.

Hecht MH. 1994. De novo design of β-sheet proteins. *Proc Natl Acad Sci USA 91*: 8729-8730.

Hellinga HW. 1998a. Construction of a blue copper analogue through iterative rational protein design cycles demonstrates principles of molecular recognition in metal center formation. *J Am Chem Soc 120*: 10055-10066.

Hellinga HW. 1998b. The construction of metal centers in proteins by rational design. *Fold Des 3*: R1-R8.

Hellinga HW, Richards FM. 1994. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc Natl Acad Sci, USA 91*: 5803-5807.

Janin J, Wodak S, Levitt M, Maigret B. 1978. Conformation of amino acid side-chains in proteins. *J Mol Biol 125*: 357-386.

Koehl P, Delarue M. 1996. Mean-field minimization methods for biological macromolecules. *Curr Opin Struct Biol 6*: 222-226.

Kono H, Doi J. 1996. A new method for side-chain conformation prediction using a Hopfield network and reproduced rotamers. *J Comp Chem 17*: 1667-1683.

Kono H, Nishiyama M, Tanokura M, Doi J. 1998. Designing the hydrophobic core of Thermus flavus malate dehydrogenase based on side-chain packing. *Protein Eng 11*: 47-52.

Kurochkina N, Lee B. 1995. Hydrophobic potential by pairwise surface area sum. *Protein Eng 8*: 437-442.

Lazar GA, Desjarlais JR, Handel TM. 1997. De novo design of the hydrophobic core of ubiquitin. *Protein Sci 6*: 1167-1178.

Lee C. 1994. Predicting protein mutant energetics by self-consistent ensemble optimization. *J Mol Biol 236*: 918-939.

Lee C, Levitt M. 1991. Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core. *Nature 352*: 448-451.

Lim WA, Sauer RT. 1989. Alternative packing arrangements in the hydrophobic core of λ repressor. *Nature 339*: 31-36.

Lovejoy B, Choe S, Cascio D, McRorie DK, DeGrado WF, Eisenberg D. 1993. Crystal structure of a synthetic triple-stranded α-helical bundle. *Science 259*: 1288-1293.

Malakauskas SM, Mayo SL. 1998. Design, structure and stability of a hyperthermophilic protein variant. *Nature Struct Biol 5*: 470-475.

Matsumura M, Matthews BW. 1991. Stabilization of funcational proteins by introduction of multiple disulfide bonds. *Methods Enzymol 202*: 336-356.

Micheletti C, Seno F, Maritan A, Banavar JR. 1998. Design of proteins with hydrophobic and polar amino acids. *Proteins 32*: 80-87.

O'Neil KT, DeGrado WF. 1990. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science 250*: 646-651.

Pabo CO, Suchanek EG. 1986. Computer-aided model-building strategies for protein design. *Biochem 25*: 5987-5991.

Pedersen JT, Moult J. 1996. Genetic algorithms for protein structure prediction. *Curr Opin Struct Biol 6*: 227-231.

Ponder JW, Richards FM. 1987. Tertiary templates for proteins - use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol 193*: 775-791.

Shakhnovich EI. 1998. Protein design: a perspective from simple tractable models. *Fold Des 3*: R45-R58.

Shakhnovich EI, Gutin AM. 1993. Engineering of stable and fast-folding sequences of model proteins. *Proc Natl Acad Sci USA 90*: 7195-7199.

Street AG, Mayo SL. 1998. Pairwise calculation of protein solvent accessible surface areas. *Fold Des 3*: 253-258.

Su A, Mayo SL. 1997. Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci 6*: 1701-1707.

Sun SJ, Brem R, Chan HS, Dill KA. 1995. Designing amino acid sequences to fold with good hydrophobic cores. *Protein Eng 8*: 1205-1213.

Tisi LC, Evans PA. 1995. Conserved structural features on protein surfaces: small exterior hydrophobic clusters. *J Mol Biol 249*: 251-258.

Tufféry P, Etchebest C, Hazout S. 1997. Prediction of protein side chain conformations: a study on the influence of backbone accuracy on conformation stability in the rotamer space. *Protein Eng 10*: 361-372.

Vasquez M. 1995. An evaluation of discrete and continuum search techniques for conformational-analysis of side-chains in proteins. *Biopolymers 36*: 53-70.

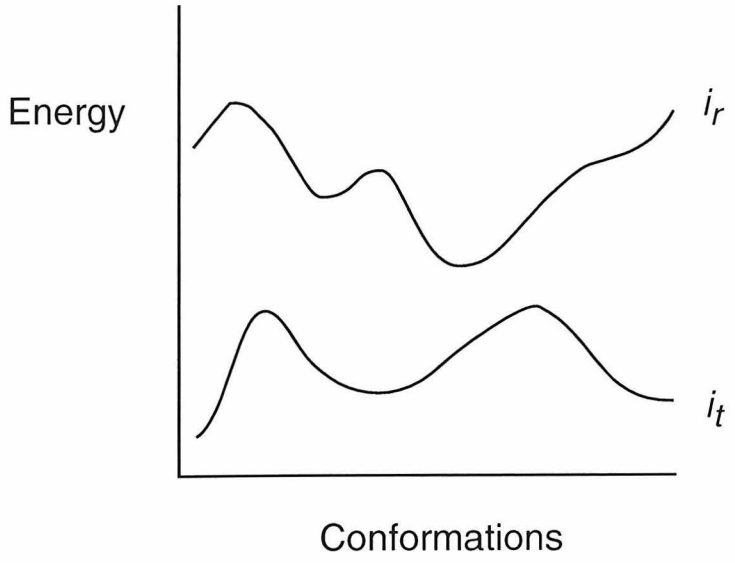
Wilson C, Mace JE, Agard DA. 1991. Computational method for the design of enzymes with altered substrate specificity. *J Mol Biol 220*: 495-506.

Yan YB, Erickson BW. 1994. Engineering of betabellin 14D: disulfide-induced folding of a β-sheet protein. *Protein Sci 3*: 1069-1073.

Ybe JA, Hecht MH. 1996. Sequence replacements in the central β-turn of plastocyanin. *Protein Sci 5*: 814-824.

Figure 2-1. The elimination of dead-ending rotamers. a) Criterion (2) eliminates a rotamer i_r if all conformations containing it have energies higher than all conformations containing some other rotamer i_t. b) Criteron (3) eliminates a rotamer i_r if, for every conformation containing it, replacing i_r with a rotamer i_t lowers the energy.

A



B

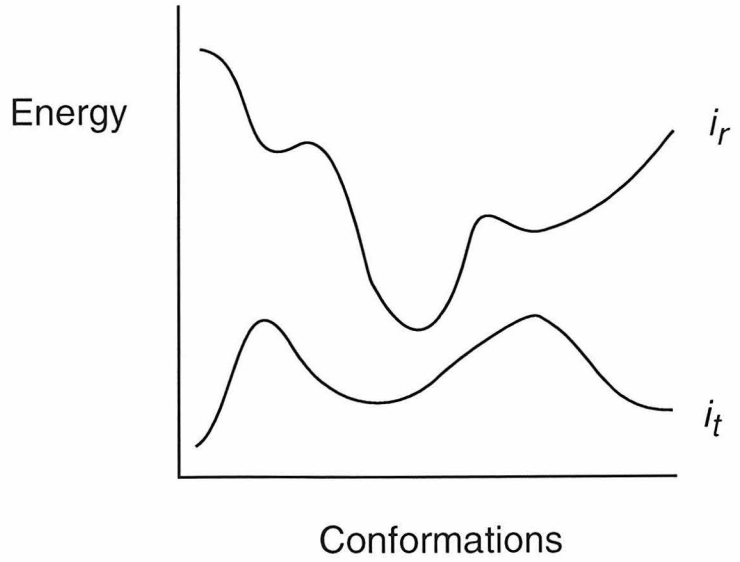
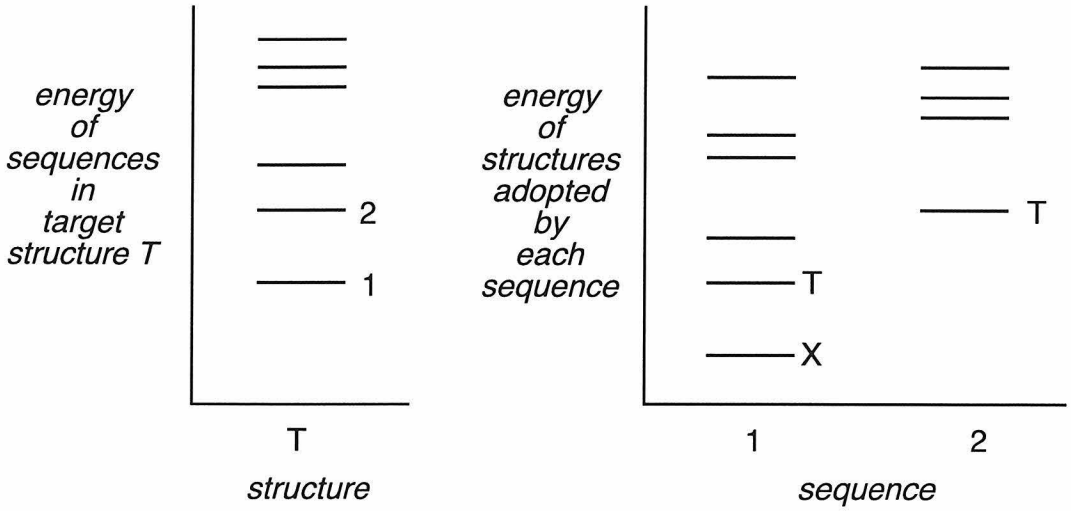


Figure 2-2. The role of negative design. Using a thermodynamic energy expression, a protein design algorithm computes that sequence 1 is the lowest energy sequence when threaded onto the target structure *T*. The ground state structure of sequence 1, however, is an alternative structure *X*. In this case, the design algorithm would ideally return sequence 2, the lowest energy sequence with ground state structure *T*.


energy of sequences in target structure T
2
1
T
structure
energy of structures adopted by each sequence
T
X
T
1
2
sequence

Chapter 3.

Understanding the Origin of Intrinsic β-Sheet Propensities

The text of this chapter is partially adapted from the publication

Street A.G. and Mayo S.L. (1999) Proceedings of the National Academy of Sciences U.S.A. 96, 9074-9076.

Abstract

The intrinsic secondary structure-forming propensities of the naturally occurring amino acids have been measured both experimentally in host-guest studies and by statistical examination of the protein structure databank. There has been significant progress in understanding the origins of intrinsic α-helical propensities but a unifying theme for understanding intrinsic β-sheet propensities has remained elusive. To this end we modeled dipeptides using a van der Waals energy function and derived Ramachandran plots for each of the amino acids. These data were used to determine the entropy and Helmholtz free energy of placing each amino acid in the β-sheet region of ϕ–ψ space. We quantitatively demonstrate that the dominant cause of intrinsic β-sheet propensity is the avoidance of steric clashes between an amino acid side chain and its local backbone. Standard implementations of electrostatic and solvation effects are seen to be less important.

Introduction

Understanding the relationship between a sequence of amino acids and its folded three-dimensional structure is of paramount importance for

protein design and protein folding studies. Conceptually, the relationship can be simplified by considering the formation of secondary and tertiary structure separately. One may then independently consider what forces drive the formation of secondary structure, and how these structures then pack together to form tertiary structure. Our concern here is with the first of these questions.

Examination of the frequencies of occurrence of the naturally occurring amino acids in α-helices or β-sheets of proteins of known structure led to the early recognition that amino acids exhibit differing propensities to form secondary structure (Chou & Fasman, 1974). The existence of stable helical peptides then enabled relatively unambiguous experimental determination of α-helical propensities (Lyu et al., 1990; O'Neil & DeGrado, 1990; Padmanabhan et al., 1990; Rohl et al., 1996), which agree with statistical studies of the protein structure database (Muñoz & Serrano, 1994). Together, these studies quantify the concept of "α-helical propensity," but do not elucidate the physical-chemical basis of propensities. Clarification of the physical-chemical basis of α-helix propensities awaited theoretical studies which compared distributions of side chain dihedral angles for each amino acid in a 9- or 11-residue α-helix and in a dipeptide standard state (Creamer & Rose, 1992; Creamer & Rose, 1994). These studies supported the view that the α-helical propensities of hydrophobic amino acids result from the loss of side chain entropy on folding. Thus alanine has the best α-helical propensity, since it loses no side chain entropy when its backbone is constrained to a helical conformation. Other studies have utilized molecular dynamics simulations using an elaborate energy expression (Hermans et al., 1992).

Because β-sheets do not appear to fold in isolation, experimental determination of β-sheet propensities has been more difficult than for α-

helices. A model protein with a suitable host site is required, and different choices yield different propensity scales (Kim & Berg, 1993; Minor & Kim, 1994b; Minor & Kim, 1994a; Smith et al., 1994; Luo et al., 1999). The preference for a certain amino acid to be in a β-sheet is therefore a more complicated issue than for α-helices, depending also on the structural context of the amino acid in the β-sheet. A statistical survey of the protein structure database nevertheless correlates well with an average of the experimental scales, supporting the idea that intrinsic β-sheet propensities do play an important role in determining a protein's stability (Muñoz & Serrano, 1994).

Correlation has been observed between one experimental β-sheet propensity scale and the ability of a side chain to sterically interfere with the formation of hydrogen bonds between its neighboring peptide group and solvent molecules (Bai & Englander, 1994). Electrostatic screening has also been proposed as an important factor (Avbelj & Moult, 1995). Other work has modeled equilibrium constants for secondary structure formation using a complex energy function (Finkelstein & Ptitsyn, 1977), which was extended to model β-sheet propensities (Finkelstein, 1995). There has also been related work modeling NMR coupling constants (Smith et al., 1996; Penkett et al., 1997). However, no concise theoretical description that fully explains the β-sheet propensities of the naturally occurring amino acids has yet emerged.

Our approach is to construct an ensemble of self-avoiding states of a dipeptide chain by fixing the bond angles and lengths and allowing the dihedral angles (ϕ, ψ and the χ's) to vary randomly over a uniform distribution. The resulting ensemble of structures represents the denatured state of the peptide. Assuming a microcanonical ensemble, the entropy change on occupying β-space is

$$\Delta S = k_B \ln \frac{W_\beta}{W} \quad (1)$$

where k_B is the Boltzmann constant, W is the number of members in the entire ensemble, and W_β is the number of members in β-space (i.e., those members with appropriate ϕ and ψ's, as defined in Methods). Comparison of ΔS calculated in this way (Table 1) with the experimentally observed β-sheet propensities is shown in Figure 1a. In order to average out as much as possible the context effects in individual experimental studies, we compare our results here with the average of the normalized available experimental data (Luo, et al., 1999). Excluding the amino acids Pro, Gly and Asn (discussed below), the correlation coefficient R is 0.92.

With the inclusion of an additional parameter to calibrate the calculated energies, this analysis can be furthered by assigning an energy ε_i to each self-avoiding chain i. The partition function over a canonical ensemble is

$$Q = \sum_i \exp(-\beta \varepsilon_i) \quad (2)$$

where $\beta = 1/k_B T$, T is the temperature, and where the summation is over all chains i in the ensemble. The change in Helmholtz free energy on folding into a β-sheet is then

$$\Delta A = -k_B T \ln \frac{Q_\beta}{Q} \quad (3)$$

where Q is the partition function for the entire ensemble and Q_β is the partition function for the β-space ensemble. However, the assigned energies ε_i may need to be scaled in order to correspond to experimental energies. This can be achieved by appropriately selecting a value of β. In order for the range of ΔA's to reproduce the experimental range of the ΔG's (for central strands

the experimental scales each range over approximately 2.5 kcal/mol excluding Gly and Pro), we select $1/\beta$ to be 9 kcal/mol. Comparison of ΔA calculated in this way with the experimentally observed β-sheet propensities is shown in Figure 1b, with $R = 0.95$.

It is conceivable that forces other than the van der Waals force may also play important roles in determining β-sheet propensity. The canonical ensemble formalism provides a convenient framework to explore this possibility, since the energies ε_i of each chain may include terms other than just the van der Waals energy. We therefore considered additional energy terms proportional to the amount of exposed (or, mathematically equivalently, buried) hydrophobic surface area, and electrostatic energies. No combination of these terms improved the correlation beyond that in Figure 1b. Electrostatic and solvation effects, in their standard implementations, are thus less important in determining β-sheet propensity.

Discussion

Our results reproduce the marked high preference in β-sheets for the β-branched amino acids Ile, Val and Thr, as well as the aromatic amino acids Phe and Tyr, and the marked low preference for Ala and Asp. Gly and Pro are excluded due to the imprecise determination of their experimental propensities. The only amino acid which lies significantly off the line of best fit in the figures is Asn. We note that, sterically, Asn and Asp have very similar side chains, so the calculated energies for the two are expected to be similar despite the wide experimental difference between their propensities. However, including surface area or charge terms in our energy expression does not improve the position of Asn. One possible explanation for Asn's better than expected experimental propensity is that hydrogen bonding may

play a greater role in determining the β-sheet propensity of Asn than for the other amino acids (Baker & Hubbard, 1984; Srinivasan et al., 1994).

One important implication of this work is that inherent β-sheet propensities can indeed be dissociated from context, as for α-helical propensities. In fact, the results of this study indicate that β-sheet propensity arises from even more local phenomena than α-helical propensity – namely, the steric interaction of an amino acid side chain with its local backbone. Thus, even in the absence of neighboring β-strands (Smith & Regan, 1995), the notion of β-sheet propensity remains valid. This agrees with studies in which a high correlation is seen between the statistically-derived preferences of amino acids in β-sheets and β-coils, where β-coils are defined to be residues in β-space but not in true β-sheets (Swindells et al., 1995). However, the existence of neighboring β-strands imposes additional contextual constraints – in particular, edge strands and central strands may present consistently different environments (Minor & Kim, 1994a). In contrast to the local nature of our description of β-sheet propensities, α-helical propensity is believed to arise from interactions between a side chain and the backbone of the neighboring turns (Creamer & Rose, 1992) (that is, from non-local interactions).

We have demonstrated that the dominant cause of intrinsic β-sheet propensity is the avoidance of steric clashes between an amino acid side chain and its local backbone. Standard implementations of electrostatic and solvation effects are less important. Our work shows, surprisingly, that the origins of β-sheet propensities may be more straightforward than those of α-helices.

Methods

We modeled each amino acid Xaa in a dipeptide environment, Ala-Xaa-Ala, with bond angles and lengths fixed (Brant & Flory, 1965). Each model peptide chain was created *de novo* using backbone and side chain dihedral angles chosen randomly from a uniform distribution. Chains were discarded if the DREIDING (Mayo et al., 1990) van der Waals energy of any atom exceeded a threshold of 2.5 kcal/mol; this threshold was chosen to best reproduce the standard Ramachandran plot for Ala (the results were not overly sensitive to changes in this value). The 1-4 van der Waals interaction energy was included except for intra-side chain contacts. Using chains which terminated at the Cα position on each flanking residue instead of full dipeptide chains did not significantly affect the results. All runs consisted of 10^5 successful chains, with relative standard errors of < 0.5%.

Our definition of β-space is based on the definition of Muñoz and Serrano (Muñoz & Serrano, 1994), bounded by the closed polygon with the following vertices in (ϕ, ψ) space: (-180, 180), (-54, 180), (-54, 90), (-144, 90), (-144, 108), (-162, 108), (-162, 126) and (-180, 126).

It is noted that the absolute propensities obtained depend quite sensitively on the N-Cα-Cβ bond angle, although the relative propensities do not. However, when this bond angle was allowed to vary according to a Gaussian distribution with mean 110° and standard deviation 2°, the reported correlations were not significantly affected.

Surface areas were calculated using the Connolly algorithm (Connolly, 1983), with a dot density of 10 $Å^{-2}$, a probe radius of zero and an add-on radius of 1.4 Å (Lee & Richards, 1971). Atoms that contribute to the hydrophobic surface area are carbon, sulfur, and hydrogen atoms attached to carbon and

sulfur. Trials were conducted using the side chain area only, and the side chain and backbone areas together.

Electrostatic energies were calculated using Gasteiger (Gasteiger & Marsili, 1980) or charge equilibration (Rappé & Goddard, 1991) point charges; neutral and charged versions of the side chains where appropriate were both tried, as were both $1/r$ and $1/r^2$ forms of the Coulomb potential. Trials were conducted using energies of the side chain only and alternatively of the full residue.

References

Avbelj F, Moult J. 1995. Role of electrostatic screening in determining protein main chain conformational preferences. *Biochem 34*: 755-764.

Bai Y, Englander W. 1994. Hydrogen bond strength and β-sheet propensities: the role of a side chain blocking effect. *Proteins 18*: 262-266.

Baker EN, Hubbard RE. 1984. Hydrogen bonding in globular proteins. *Prog Biophys Molec Biol 44*: 97-179.

Brant DA, Flory PJ. 1965. The configuration of random polypeptide chains. II. Theory. *J Am Chem Soc 87*: 2791-2800.

Chou PY, Fasman GD. 1974. Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. *Biochem 13*: 211-222.

Connolly ML. 1983. Solvent accessible surfaces of proteins and nucleic acids. *Science 221*: 709-713.

Creamer TP, Rose GD. 1992. Side-chain entropy opposes α-helix formation but rationalizes experimentally determined helix-forming propensities. *Proc Natl Acad Sci USA 89*: 5937-5941.

Creamer TP, Rose GD. 1994. α-helix-forming propensities in peptides and proteins. *Proteins 19*: 85-97.

Finkelstein AV. 1995. Predicted β-structure stability parameters under experimental test. *Prot Eng 8*: 207-209.

Finkelstein AV, Ptitsyn OB. 1977. Theory of protein molecule self-organization. I. thermodynamic parameters of local secondary structures in the unfolded protein chain. *Biopolymers 16*: 469-495.

Gasteiger J, Marsili M. 1980. Iterative partial equalization of orbital electronegativity — a rapid access to atomic charges. *Tetrahedron 36*: 3219-3228.

Hermans J, Anderson AG, Yun RH. 1992. Differential helix propensity of small apolar side-chains studied by molecular-dynamics simulations. *Biochem 31*: 5646-5653.

Kim CA, Berg JM. 1993. Thermodynamic β-sheet propensities measured using a zinc-finger host peptide. *Nature 362*: 267-270.

Lee B, Richards FM. 1971. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol 55*: 379-400.

Luo J-Y, Langen R, Olafson BD, Richards JH, Mayo SL. 1999. Both intrinsic secondary structure preference and context are determinants of β-sheet propensity (submitted).

Lyu PC, Liff MI, Marky LA, Kallenbach NR. 1990. Side-chain contributions to the stability of α-helical structure in peptides. *Science 250*: 669-673.

Mayo SL, Olafson BD, Goddard WA, III. 1990. Dreiding - a generic force-field for molecular simulations. *J Phys Chem 94*: 8897-8909.

Minor DL, Kim PS. 1994a. Context is a major determinant of β-sheet propensity. *Nature 371*: 264-267.

Minor DL, Kim PS. 1994b. Measurements of the β-sheet-forming propensities of amino acids. *Nature 367*: 660-663.

Muñoz V, Serrano L. 1994. Intrinsic secondary structure propensities of the amino acids, using statistical ϕ-ψ matrices: comparison with experimental scales. *Proteins 20*: 301-311.

O'Neil KT, DeGrado WF. 1990. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino-acids. *Science 250*: 646-651.

Padmanabhan S, Marqusee S, Ridgeway T, Laue TM, Baldwin RL. 1990. Relative helix-forming tendencies of nonpolar amino-acids. *Nature 344*: 268-270.

Penkett CJ, Redfield C, Dodd I, Hubbard J, McBay DL, Mossakowska DE, Smith RAG, Dobson CM, Smith LJ. 1997. NMR analysis of main-chain conformational preferences in an unfolded fibronectin-binding protein. *J Mol Biol 274*: 152-159.

Rappé AK, Goddard WA. 1991. Charge equilibration in molecular-dynamics simulations. *J Phys Chem 95*: 3358-3363.

Rohl CA, Chakrabartty A, Baldwin RL. 1996. Helix propagation and N-cap propensities of the amino acids measured in alanine-based peptides in 40 volume percent trifluoroethanol. *Prot Sci 5*: 2623-2637.

Smith CK, Regan L. 1995. Guidelines for protein design: the energetics of β sheet side chain interactions. *Science 270*: 980-982.

Smith CK, Withka JM, Regan L. 1994. A thermodynamic scale for the β-sheet forming tendencies of the amino acids. *Biochem 33*: 5510-5517.

Smith LJ, Bolin KA, Schwalbe H, MacArthur MW, Thornton JM, Dobson CM. 1996. Analysis of main chain torsion angles in proteins: prediction of NMR coupling constants for native and random coil conformations. *J Mol Biol 255*: 494-506.

Srinivasan N, Anuradha VS, Ramakrishnan C, Sowdhamini R, Balaram P. 1994. Conformational characteristics of asparaginyl residues in proteins. *Int J Pept Prot Res 44*: 112-122.

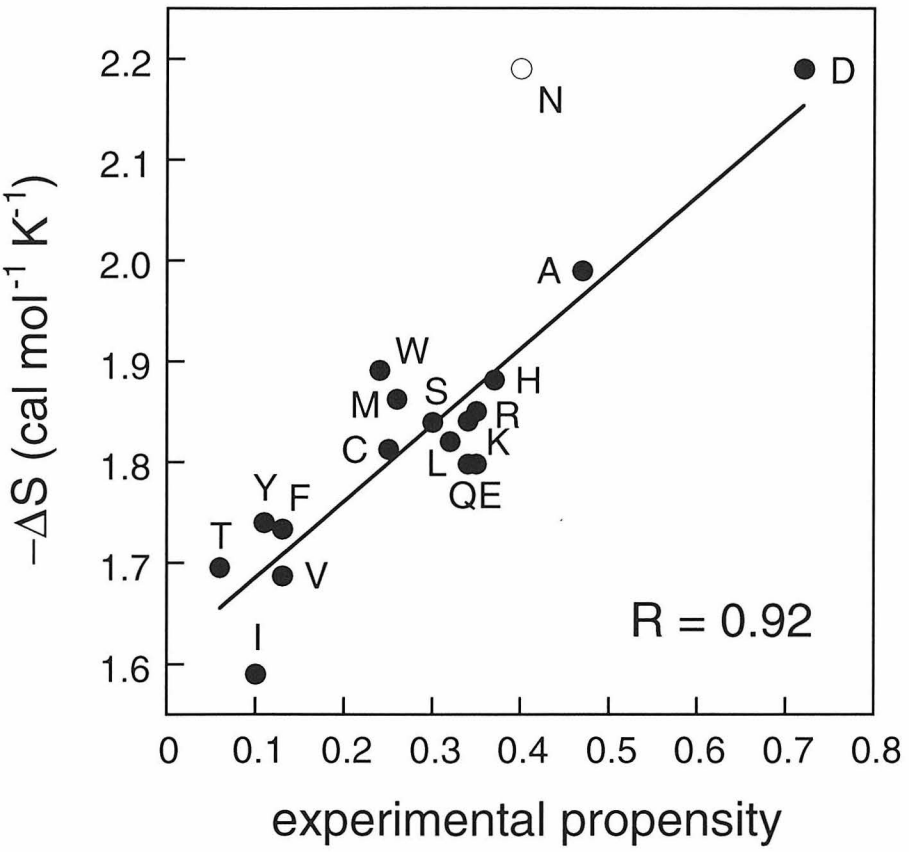
Swindells MB, MacArthur MW, Thornton JM. 1995. Intrinsic ϕ,ψ propensities of amino acids, derived from the coil regions of known structures. *Nature Struct Biol 2*: 596-603.

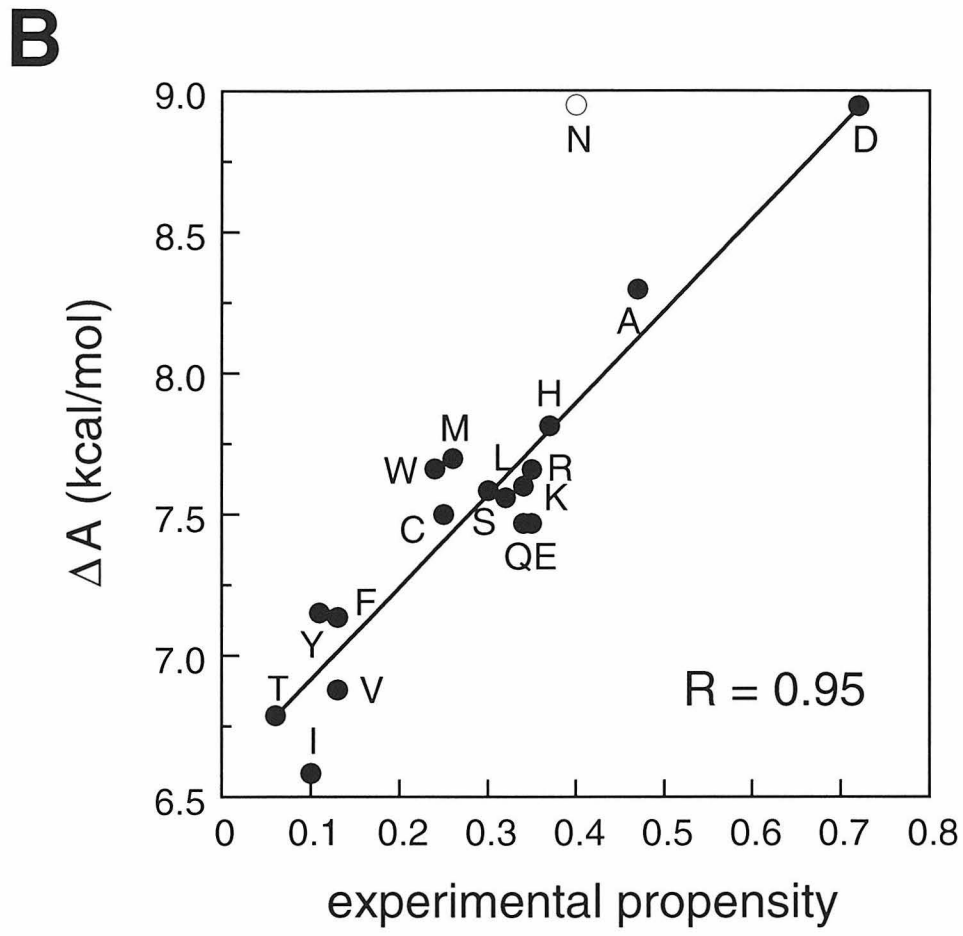
Table 3-1. Calculated change in entropy, ΔS, and Helmholtz free energy, ΔA, on folding into a β-sheet, and the average normalized experimental propensity (Luo, et al., 1999) of the naturally occurring amino acids. The average normalized experimental propensities are calculated from four published studies (Kim & Berg, 1993; Minor & Kim, 1994b; Minor & Kim, 1994a; Smith, et al., 1994) and a similar study in apo-azurin (Luo, et al., 1999). Each scale was normalized to range from zero to one, with Pro excluded, and averaged.

Amino acid	ΔS (cal mol^{-1} K^{-1})	ΔA (kcal mol^{-1})	Average normalized experimental propensity
I	-1.59	6.58	0.10
V	-1.69	6.88	0.13
T	-1.70	6.79	0.06
F	-1.73	7.14	0.13
Y	-1.74	7.15	0.11
E	-1.80	7.47	0.35
Q	-1.80	7.47	0.34
C	-1.81	7.50	0.25
L	-1.82	7.56	0.32
K	-1.84	7.60	0.34
S	-1.84	7.58	0.30
R	-1.85	7.66	0.35
M	-1.86	7.70	0.26
H	-1.88	7.81	0.37
W	-1.89	7.66	0.24
A	-1.99	8.30	0.47
D	-2.19	8.95	0.72
N	-2.19	8.95	0.40

Figure 3-1. Correlation between calculated and average normalized experimental β-sheet propensities (Luo, et al., 1999). All amino acids except Gly and Pro are shown. Asn, represented by the open circle, is discussed in the text. a) The negative of the entropy calculated using (1). b) Helmholtz free energy calculated using (2) and (3), and $1/\beta = 9$ kcal mol^{-1}.

A


2.2
2.1
2.0
1.9
1.8
1.7
1.6
−ΔS (cal mol-1 K-1)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
experimental propensity
N
D
A
W
H
M
S
R
K
C
L
QE
Y
F
T
V
I
R = 0.92


B
9.0
8.5
8.0
7.5
7.0
6.5
ΔA (kcal/mol)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
experimental propensity
N
D
A
H
M
W
L
R
K
C
S
QE
F
Y
T
V
I
R = 0.95

Chapter 4.

A Quantitative Model for Examining Hydrophobic Context Effects on β-Sheet Stability

Abstract

Previous work has shown that inclusion of a penalty for exposed hydrophobic solvent-accessible surface area can improve the designability of proteins. Here we demonstrate this experimentally. We mutate position 6 on the β-sheet surface of the β1 immunoglobulin-binding domain of streptococcal protein G to several different amino acids of similar β-sheet propensity. The melting temperatures of the mutant proteins are correlated with simulations using a previously published energy expression, with a modification that penalizes hydrophobic surface area exposure. We find that penalizing hydrophobic exposure at 1.6 times the strength at which hydrophobic burial is benefited increases the correlation between experiment and simulation to $R^2 = 0.97$. Our data support the claim that hydrophobic context is an important factor in the stability of β-sheets.

Introduction

Several quantitative methods have been proposed and tested which analyze the compatibility of possible sequences with a given protein fold (Hellinga et al., 1991; Hurley et al., 1992; Kono & Doi, 1994; Desjarlais & Handel, 1995; Harbury et al., 1995; Klemba et al., 1995; Nautiyal et al., 1995; Betz & Degrado, 1996; Dahiyat & Mayo, 1996). These algorithms calculate the spatial positioning and steric complementarity of side chains by explicitly

modeling the atoms of the considered sequence. Such techniques have typically focused on designing the cores of proteins where the van der Waals and sometimes hydrophobic solvation potentials are sufficient to yield reasonable results. Success has also been achieved by applying a hydrogen bonding potential to the design of α-helical surfaces (Dahiyat et al., 1997a). In addition, the design of an entire sequence for a small protein fold has been recently reported (Dahiyat & Mayo, 1997a). These potentials, however, are insufficient to design extensive β-sheet surfaces. We seek to extend the computational sequence selection approach to address β-sheet design with the goal of developing a complete *de novo* design algorithm.

The forces which are thought to be important in determining the formation and stability of β-sheet regions can be grouped into two categories. The first is the inherent propensity of the amino acids to form β-sheets, as determined by experimental host-guest site studies (Kim & Berg, 1993; Minor & Kim, 1994b; Smith et al., 1994) and by a statistical examination of the protein databank (Muñoz & Serrano, 1994). The second is the context of each site. Context encompasses both tertiary and hydrophobic effects. The tertiary, or structural, context of an amino acid is where it is physically located in the β-sheet; in particular, whether it is on an edge strand or a central strand (Minor & Kim, 1994a). We use hydrophobic context to mean the hydrophobic environment of a particular residue position, which can be quantitatively evaluated by measuring solvent-accessible hydrophobic surface areas (Otzen & Fersht, 1995). This study examines the role of hydrophobic context.

Our simulations implement a previously described benefit for the burial of hydrophobic surface area and a penalty for the burial of polar surface area (Dahiyat & Mayo, 1996; Street & Mayo, 1998). Improved correlation between calculated and experimentally determined stabilities is achieved

when the solvation potential is supplemented with a penalty for the exposure of hydrophobic surface area. An exposure penalty imposes a "reverse" hydrophobic effect (Pakula & Sauer, 1990).

Previous theoretical work (Sun et al., 1995) concluded that use of a hydrophobic exposure penalty for protein design leads to sequences with well separated native and denatured energies, and that a hydrophobic exposure penalty can be thought of as an example of negative design. Negative design is the process of choosing sequences such that structures other than the desired target structure are disfavored (Hecht et al., 1990). An excessive amount of exposed hydrophobic area on the surface of a protein can lead to a lack of specificity in folding, and therefore can lead to structural heterogeneity in the native state.

We chose as our model system the β1 immunoglobulin-binding domain of streptococcal protein G (GB1) (Gronenborn et al., 1991). A solution structure (Gronenborn, et al., 1991) and several crystal structures (Gallagher et al., 1994) are available to provide backbone templates for the side chain selection process. Its small size, 56 residues, makes computations feasible. GB1 contains no disulfide bonds and does not require a cofactor or metal ion to fold. We consider residue position 6, a surface position in the middle of a central β-strand in which the wildtype Ile is approximately 70% buried (Figure 1). The relatively high burial of Ile6 is achieved by van der Waals contacts with Lys4, Glu15, Thr51 and Thr53.

A series of mutations to amino acids of similar β-sheet propensity were made at position 6, and the resulting proteins' stabilities measured. The stability data were then used to determine the optimal strength of the hydrophobic exposure penalty.

Results and Discussion

Ile6 of GB1 was mutated to Thr, Val, Tyr and Phe. Because these amino acids have similar β-sheet forming propensities, stability differences between the respective proteins should reflect context specific effects. Mutations to amino acids of markedly different β-sheet propensity were not considered since this would have introduced a further variable into the experiment. Circular dichroism (CD) determined melting temperatures and ΔΔG's are listed in Table 1. We note the surprising result that, despite being polar and having high β-sheet forming propensity, Thr yields the lowest stability of the mutants considered, 10 °C below Ile, which supports the notion that hydrophobic context plays an important role at this site.

In order to quantitatively model the effect of the mutations, we used the dead-end elimination (DEE) theorem (Desmet et al., 1992; Gordon & Mayo, 1998) to find the minimum energy conformation of the 12 β-sheet surface residues shown in Figure 1. We introduced an additional energy term into our previously published energy expression (Dahiyat & Mayo, 1996; Dahiyat, et al., 1997a; Street & Mayo, 1998) in order to penalize solvent-exposed hydrophobic surface area,

$$E_{\text{penalty}} = \kappa\sigma_{\text{np}} A_{\text{np}}^{\text{exposed}}$$

where $A_{\text{np}}^{\text{exposed}}$ is the amount of solvent-exposed hydrophobic surface area and σ_{np} is an atomic solvation parameter for hydrophobic surface area burial. The dimensionless scale factor κ, whose value is to be determined, sets the strength of this term relative to the strength at which burial of hydrophobic surface area is benefited. Note that the fraction of a residue's hydrophobic area

that must be buried for a zero net surface area contribution to the energy is $\kappa / (\kappa+1)$.

When the simulation only includes terms for van der Waals, electrostatic and hydrogen-bonding, the calculated energies do not correlate with the experimentally observed melting temperatures (Figure 2a). The correlation rises to $R^2 = 0.72$ (Figure 2b) when, in addition to the above potential energy terms, hydrophobic burial is benefited and polar burial is penalized, but with $\kappa = 0$. The value of κ was determined by evaluating cross-validated correlation coefficients (Figure 3). When $\kappa = 1.6$, the correlation between the calculated energies and experimentally determined melting temperatures is $R^2 = 0.97$ (Figure 2c).

The high correlation between calculated energies and experimentally determined melting temperatures led us to examine the possibility of including amino acids with diminished β-sheet propensities at position 6. We did not expect these mutations to fit the previous model, but were interested in whether they would lie in the region where an adjustment to the simulation energies to penalize them for their lower propensities could account for any discrepancies. We made two such substitutions: Arg and Trp. The resulting molecules' melting temperatures (Table 1) are much higher than would be expected based purely on β-sheet propensities, but lower than expected based purely on context (by 3.2 and 1.1 kcal/mol, respectively, corresponding well to the magnitude of their β-sheet propensities). Thus, an additional energy term which explicitly penalizes lower β-sheet propensities could be used to model amino acids not considered in this study.

We have found that mutating a single β-sheet surface site on GB1 leads to changes in melting temperature as large as 10 °C, even when only amino acids of similar β-sheet propensity are considered. We have shown that a

solvation potential term that penalizes the exposure of solvent-accessible hydrophobic surface area can capture the context effects that are believed to be important in β-sheet formation. The improved potential function should prove useful for computational protein design efforts.

Materials and Methods

Modeling

The initial GB1 structure was taken from PDB entry 1pga (Bernstein et al., 1977; Gallagher, et al., 1994). The program BIOGRAF (Molecular Simulations Incorporated, San Diego, California) was used to generate explicit hydrogens on the structure which was then conjugate gradient minimized for 50 steps using the DREIDING force field (Mayo et al., 1990). All atoms except those of the side chains of residues 4, 6, 8, 13, 15, 17, 42, 44, 46, 51, 53, and 55 were held fixed for subsequent DEE calculations.

A Lennard-Jones 6-12 potential was used for van der Waals interactions with atomic radii scaled by 90% (Dahiyat & Mayo, 1997b). The Lee and Richards definition of solvent-accessible surface area (Lee & Richards, 1971) was used, areas being calculated with the Connolly algorithm (Connolly, 1983). Buried and exposed areas were calculated as previously described (Street & Mayo, 1998). We include a hydrogen-bonding and electrostatics potential (Dahiyat, et al., 1997a).

DEE optimization followed previously published methods (Gordon & Mayo, 1998). Calculations were performed on a 12 processor R10000-based Silicon Graphics Power Challenge.

Cross-validated R^2 values were calculated by removing each point (x_i, y_i) in turn and using least squares to predict its location (x_i, z_i) based on the remaining points. The cross-validated R^2 is then

$$\text{cross - validated } R^2 = \frac{\sum_i (y_i - \langle y \rangle)^2 - \sum_i (y_i - z_i)^2}{\sum_i (y_i - \langle y \rangle)^2}$$

where $\langle y \rangle$ is the average of the y_i's.

The simulation results depend, of course, on the precise rotamer library and solvation parameters selected. As in our previous work (Dahiyat, et al., 1997a), a backbone-dependent rotamer library was used (Dunbrack & Karplus, 1993). The results reported here use a library in which χ_1 angle values of all rotamers were expanded ±1 standard deviation about the mean value (known as the "e1" library), and solvation parameters that benefit hydrophobic burial by σ_{np} = 26 cal mol^{-1} Å^{-2} and penalize polar burial by σ_p = 100 cal mol^{-1} Å^{-2} (Street & Mayo, 1998). To increase the speed of larger calculations, previous work (Dahiyat et al., 1997b) has utilized the "a2h1p0" rotamer library, in which the χ_1 and χ_2 angles of aromatic side chain rotamers are expanded, the χ_1 angles of hydrophobic side chain rotamers are expanded, and only the mean χ_1 angles of polar side chain rotamers are used. An alternative solvation potential, in which hydrophobic burial is benefited by 48 cal mol^{-1} Å^{-2} and only polar hydrogen burial is penalized (when not engaged in hydrogen bonding), was also used (Dahiyat, et al., 1997a). The optimal value of κ found in this study is 1.6 for both of these energy functions. However, other combinations led to other values (Figure 4). This is to be expected, since the hydrophobic exposure penalty is itself a solvation parameter and so depends on the values of the other solvation parameters. We have reported the results for the most commonly used solvation parameters.

Mutagenesis and protein purification

A synthetic GB1 gene (Minor & Kim, 1994b) was cloned into a pET11a vector (Novagen), C-terminally His-tagged and used as the template for PCR mutagenesis. The correctness of the constructs was confirmed by DNA sequencing. The expression and purification of the various proteins followed published procedures.

Circular dichroism

CD spectra were measured on an Aviv 62DS spectrometer at pH 5.5, in 50 mM phosphate and 25 μM protein. A 1 mm path length cell was used. The temperature was controlled by a thermoelectric unit. Thermal melts were performed at 218 nm using 2 °C temperature steps with an averaging time of 10 s and an equilibration time of 90 s. The melting temperatures (T_m) were derived by evaluating the maximum of a $d\theta_{218}/dT$ versus T plot. T_m's were reproducible to within 0.5 °C. Protein concentrations were determined by UV spectrophotometry. ΔΔG's were calculated (Becktel & Schellman, 1987) using ΔH_m = 61 kcal/mol (Alexander et al., 1992), resulting in errors of ±0.08 kcal/mol. In the case of a 6-fold mutant of GB1 with a ΔT_m of 8 °C, the Becktel-Schellman method holds (Alexander et al., 1992). Guanidinium-induced denaturation was also used, with results in line with the melting temperature results but with greater experimental uncertainties (±0.2 kcal/mol), as determined from multiple measurements using an auto-titrator. The correlation between stabilities derived from guanidinium denaturation and from melting temperatures is shown in Figure 5.

References

Alexander P, Fahnestock S, Lee T, Orban J, Bryan P. 1992. Thermodynamic analysis of the folding of the streptococcal protein G IgG-binding domains β1 and β2 - Why small proteins tend to have high denaturation temperatures. *Biochem 31*: 3597-3603.

Becktel WJ, Schellman JA. 1987. Protein stability curves. *Biopolymers 26*: 1859-1877.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EE, Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The protein data bank: a computer-based archival file for macromolecular structures. *J Mol Biol 112*: 535-542.

Betz SF, Degrado WF. 1996. Controlling topology and native-like behavior of de novo-designed peptides - design and characterization of antiparallel 4-stranded coiled coils. *Biochem 35*: 6955-6962.

Connolly ML. 1983. Solvent accessible surfaces of proteins and nucleic acids. *Science 221*: 709-713.

Dahiyat BI, Gordon DB, Mayo SL. 1997a. Automated design of the surface positions of protein helices. *Prot Sci 6*: 1333-1337.

Dahiyat BI, Mayo SL. 1996. Protein design automation. *Prot Sci 5*: 895-903.

Dahiyat BI, Mayo SL. 1997a. De novo protein design: fully automated sequence selection. *Science 278*: 82-87.

Dahiyat BI, Mayo SL. 1997b. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci USA 94*: 10172-10177.

Dahiyat BI, Sarisky CA, Mayo SL. 1997b. *De novo* protein design: towards fully automated sequence selection. *J Mol Biol 273*: 789-796.

Desjarlais JR, Handel TM. 1995. De novo design of the hydrophobic cores of proteins. *Prot Sci 4*: 2006-2018.

Desmet J, De Maeyer M, Hazes B, Lasters I. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature 356*: 539-542.

Dunbrack RL, Karplus M. 1993. Backbone dependent rotamer library for proteins - an application to side-chain prediction. *J Mol Biol 230*: 543-574.

Gallagher T, Alexander P, Bryan P, Gilliland GL. 1994. Two crystal structures of the β1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. *Biochem 33*: 4721-4729.

Gordon DB, Mayo SL. 1998. Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J Comp Chem 19*: 1505-1514.

Gronenborn AM, Filpula DR, Essign NZ, Achari A, Whitlow M, Wingfield PT, Clore GM. 1991. A novel, highly stable fold of the immunoglobin binding domain of streptococcal protein G. *Science 253*: 657-661.

Harbury PB, Tidor B, Kim PS. 1995. Repacking protein cores with backbone freedom: structure prediction for coiled coils. *Proc Natl Acad Sci USA 92*: 8408-8412.

Hecht MH, Richardson JS, Richardson DC, Ogden RC. 1990. De novo design, expression, and characterization of Felix: a four-helix bundle protein of native-like sequence. *Science 249*: 884-891.

Hellinga HW, Caradonna JP, Richards FM. 1991. Construction of new ligand-binding sites in proteins of known structure 2. Grafting of buried transition-metal binding site into E. coli thioredoxin. *J Mol Biol 222*: 787-803.

Hurley JH, Baase WA, Matthews BW. 1992. Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme. *J Mol Biol 224*: 1143-1154.

Kim CA, Berg JM. 1993. Thermodynamic β-sheet propensities measured using a zinc-finger host peptide. *Nature 362*: 267-270.

Klemba M, Gardner KH, Marino S, Clarke ND, Regan L. 1995. Novel metal-binding proteins by design. *Nature Struct Biol 2*: 368-373.

Kono H, Doi J. 1994. Energy minimization method using automata network for sequence and side-chain conformation prediction from given backbone geometry. *Proteins 19*: 244-255.

Lee B, Richards FM. 1971. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol 55*: 379-400.

Mayo SL, Olafson BD, Goddard WA, III. 1990. Dreiding - a generic force-field for molecular simulations. *J Phys Chem 94*: 8897-8909.

Minor DL, Kim PS. 1994a. Context is a major determinant of β-sheet propensity. *Nature 371*: 264-267.

Minor DL, Kim PS. 1994b. Measurements of the β-sheet-forming propensities of amino acids. *Nature 367*: 660-663.

Muñoz V, Serrano L. 1994. Intrinsic secondary structure propensities of the amino acids, using statistical ϕ-ψ matrices: comparison with experimental scales. *Proteins 20*: 301-311.

Nautiyal S, Woolfson DN, King DS, Alber T. 1995. A designed heterotrimeric coiled coil. *Biochem 34*: 11645-11651.

Otzen DE, Fersht AR. 1995. Side-chain determinants of β-sheet stability. *Biochem 34*: 5718-5724.

Pakula AA, Sauer RT. 1990. Reverse hydrophobic effects relieved by amino-acid substitutions at a protein surface. *Nature 344*: 363-364.

Smith CK, Withka JM, Regan L. 1994. A thermodynamic scale for the β-sheet forming tendencies of the amino acids. *Biochem 33*: 5510-5517.

Street AG, Mayo SL. 1998. Pairwise calculation of protein solvent accessible surface areas. *Folding and Design 3*: 253-258.

Sun S, Brem R, Chan HS, Dill KA. 1995. Designing amino acid sequences to fold with good hydrophobic cores. *Prot Eng 8*: 1205-1213.

Table 4-1. Experimentally determined melting temperatures and free energy differences of the proteins used in this study.

Mutant	T_m (°C)	ΔΔG (kcal/mol)
Ile6	86.4	0
Val6	85.0	-0.24
Tyr6	80.2	-1.06
Phe6	79.8	-1.13
Thr6	75.7	-1.83
Arg6	82.7	-0.63
Trp6	76.5	-1.69

Figure 4-1. Schematic diagram of GB1 showing the 12 β-sheet surface residues considered in this study.

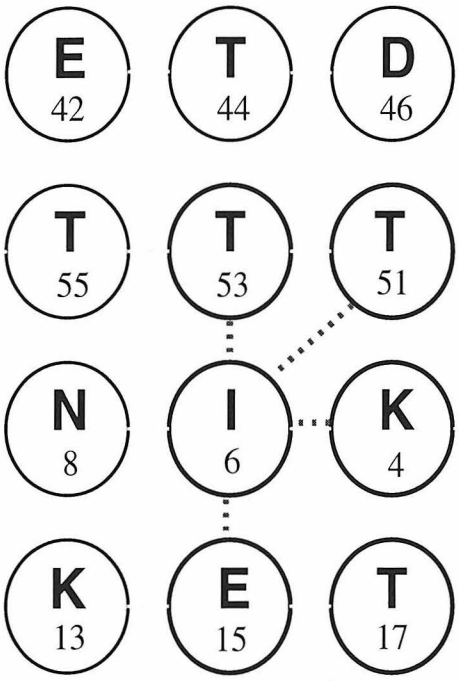
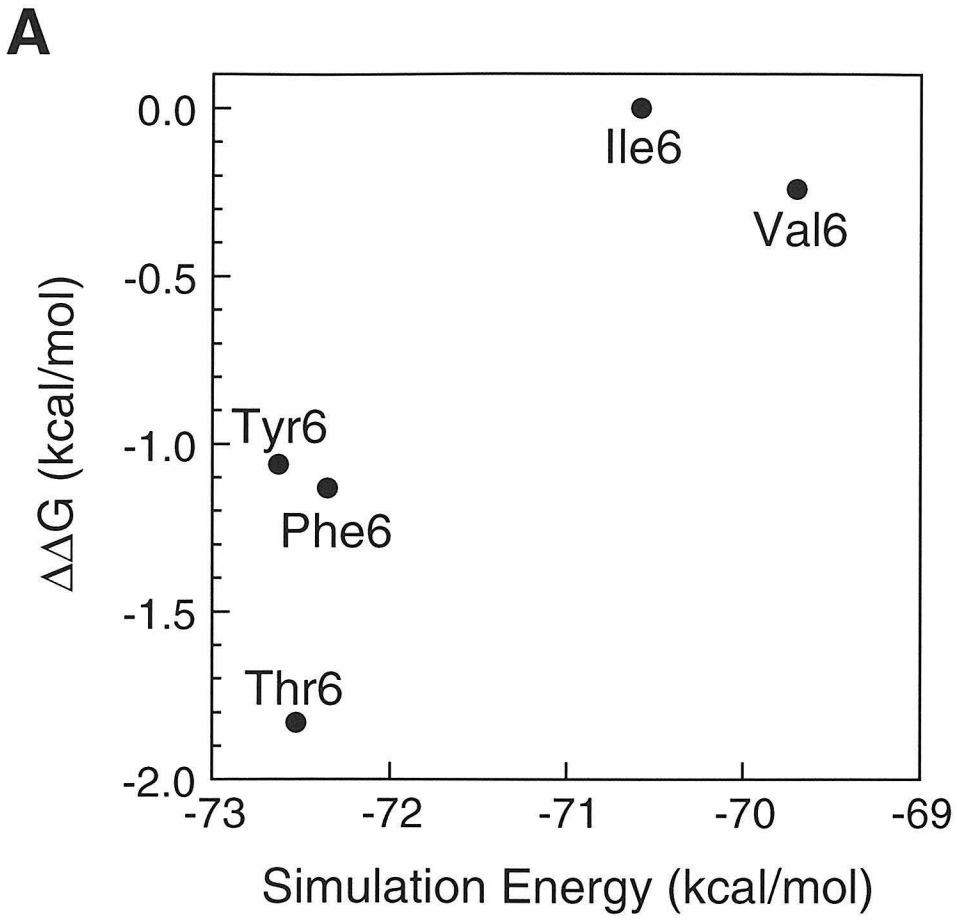
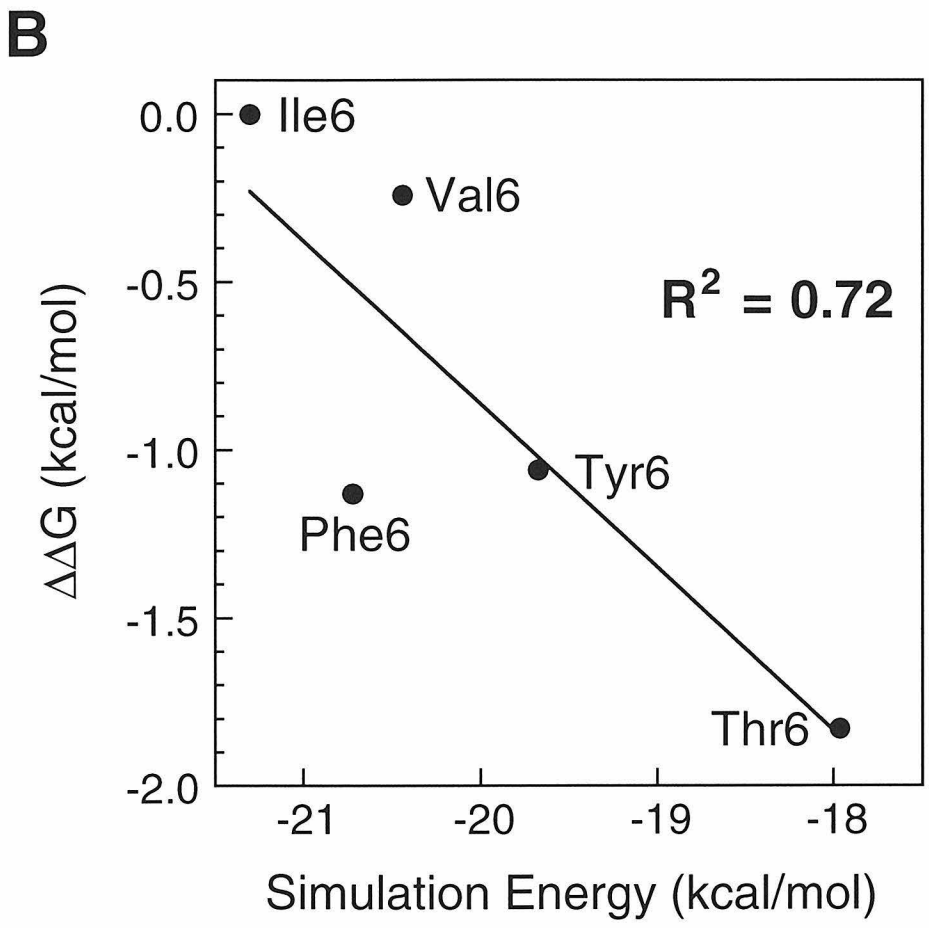

E
42
T
44
D
46
T
55
T
53
T
51
N
8
I
6
K
4
K
13
E
15
T
17

Figure 4-2. Correlation between energies calculated with different potential functions and the experimentally observed stabilities of the proteins used in this study. a) Potential function containing only van der Waals, electrostatics and hydrogen-bonding terms; b) Potential function that includes additional terms that benefit hydrophobic surface area burial and penalize polar surface area burial, but without the additional hydrophobic exposure penalty (i.e., $\kappa = 0$); c) Potential function including all terms with $\kappa = 1.6$.


A
0.0
-0.5
-1.0
-1.5
-2.0
ΔΔG (kcal/mol)
Ile6
Val6
Tyr6
Phe6
Thr6
-73
-72
-71
-70
-69
Simulation Energy (kcal/mol)


B
0.0
-0.5
-1.0
-1.5
-2.0
ΔΔG (kcal/mol)
Ile6
Val6
R2 = 0.72
Tyr6
Phe6
Thr6
-21
-20
-19
-18
Simulation Energy (kcal/mol)

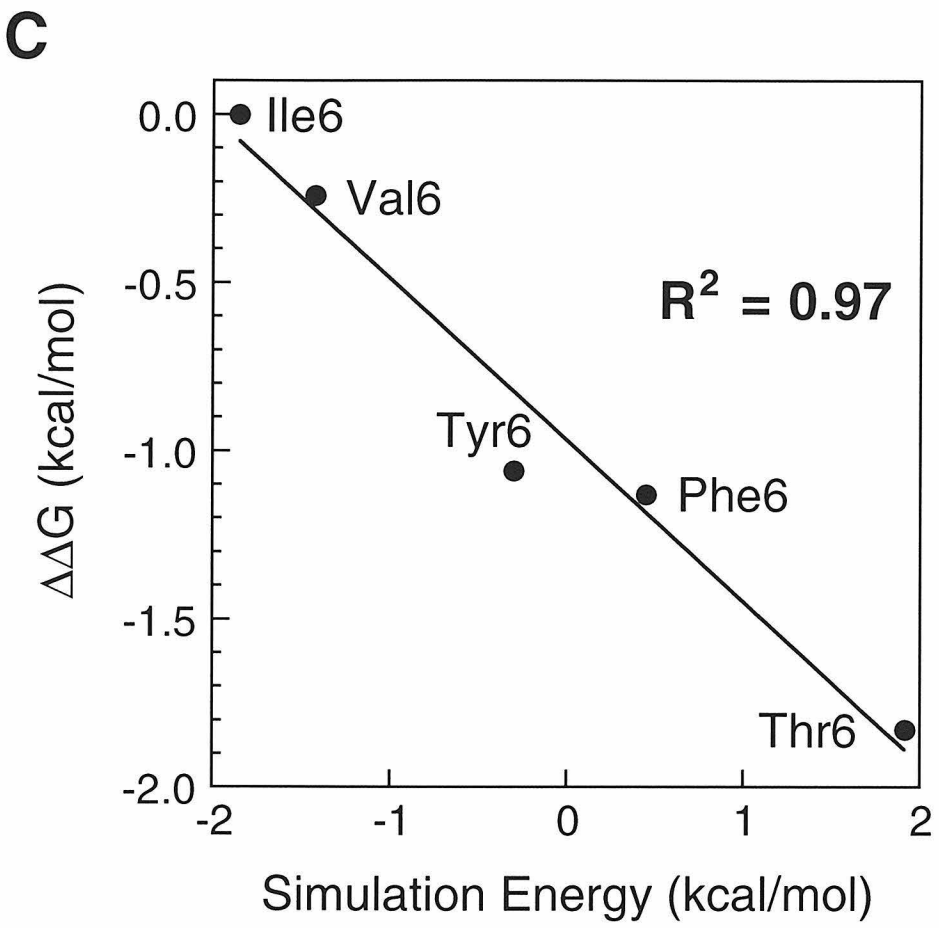

C
0.0
-0.5
-1.0
-1.5
-2.0
ΔΔG (kcal/mol)
Ile6
Val6
Tyr6
Phe6
Thr6
R2 = 0.97
-2
-1
0
1
2
Simulation Energy (kcal/mol)

Figure 4-3. Cross-validated correlation between stabilities and calculated energies of the five mutants as a function of the value of the exposed hydrophobic surface area penalty, κ.

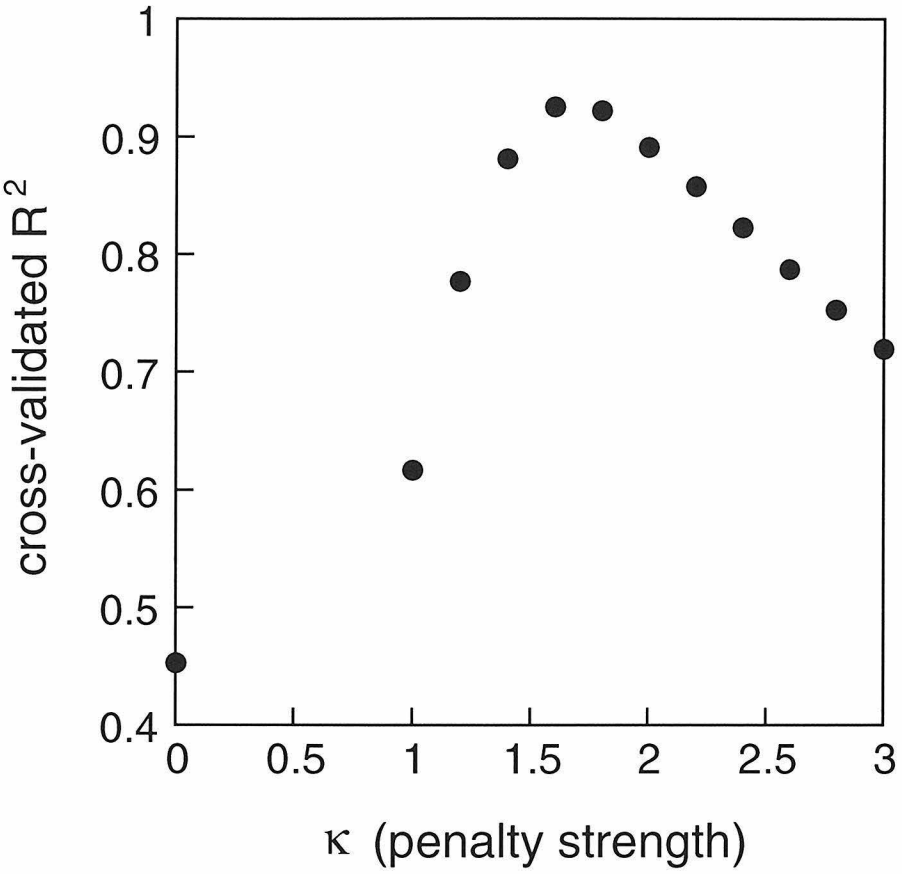

cross-validated R^2
1
0.9
0.8
0.7
0.6
0.5
0.4
0
0.5
1
1.5
2
2.5
3
κ (penalty strength)

Figure 4-4. Effect of rotamer library and solvation parameters on the correlation between stabilities and calculated energies of the five mutants, as a function of κ. Shown are the "e1" rotamer library (described in the text) with solvation potential A (σ_{np} = 26 cal mol^{-1} Å^{-2}, σ_p = 100 cal mol^{-1} Å^{-2}) (circles); "e1" with solvation potential B (σ_{np} = 48 cal mol^{-1} Å^{-2} and a polar hydrogen burial penalty of 2 kcal mol^{-1} per hydrogen) (squares); the "a2h1p0" library with solvation potential A (crosses) and with solvation potential B (diamonds); and the "e2" library with solvation potential A (plusses). In the "e2" library, the χ_1 and χ_2 angle values of all rotamers are expanded ± 1 standard deviation about the mean value. The correlation is greatest at κ = 1.6 for the most commonly used parameter sets but is less clear for other parameter sets.

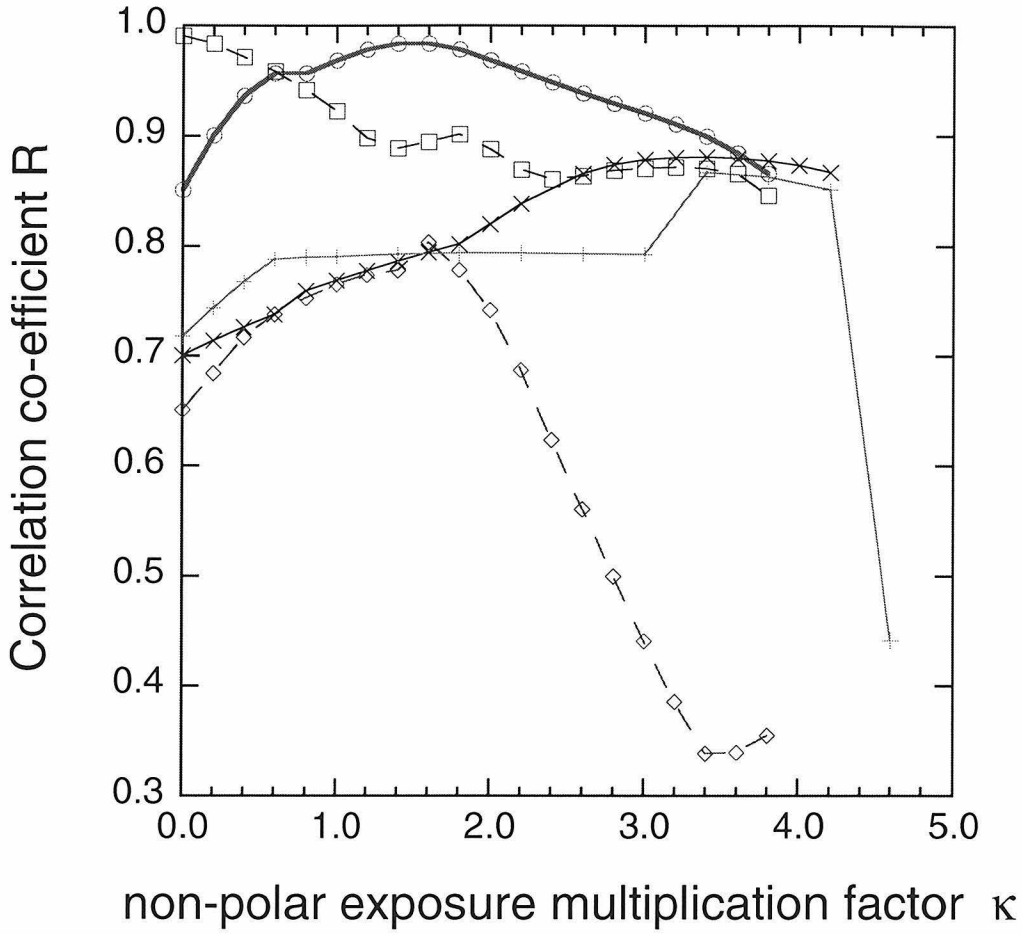
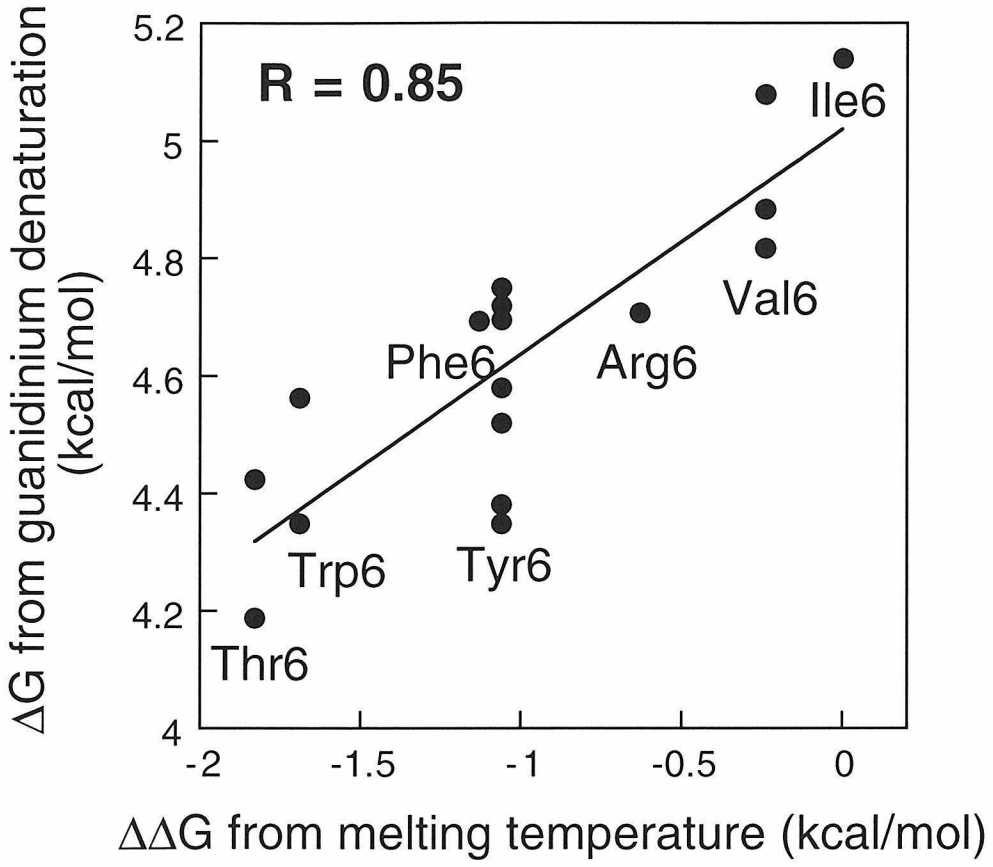

Correlation co-efficient R
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
1.0
2.0
3.0
4.0
5.0
non-polar exposure multiplication factor κ

Figure 4-5. Comparison of the stabilities obtained from proteins' melting temperatures via the Becktel-Schellman method, and the stabilities derived from guanidinium denaturation. Repeated temperature scans yielded very little change in the calculated melting temperature. Variation in stability resulting from repeated guanidinium denaturation experiments, however, was significant and is shown. In particular, the Tyr6 mutant was studied extensively.


R = 0.85
Ile6
Val6
Arg6
Phe6
Trp6
Tyr6
Thr6
5.2
5
4.8
4.6
4.4
4.2
4
-2
-1.5
-1
-0.5
0
ΔG from guanidinium denaturation (kcal/mol)
ΔΔG from melting temperature (kcal/mol)

Chapter 5.

Pairwise Calculation of Solvent-Accessible Surface Area

The text of this chapter is partially adapted from the publication

Street A.G. and Mayo S.L. (1998) Folding and Design 3, 253-258.

Abstract

Many algorithms for determining the energy state of a system depend for their tractability on the pairwise nature of an energy expression. Some energy terms, such as the standard implementation of the van der Waals potential, satisfy this criterion while others do not. One class of important potentials that is not pairwise involves benefits and penalties for burying hydrophobic and/or polar surface areas. It has been previously found that, in some cases, a pairwise approximation to these surface areas correlates with the true surface areas. We develop a pairwise expression with one scalable parameter that closely reproduces both the true buried and the true exposed solvent-accessible surface areas. We then refit our previously published coiled coil stability data (Dahiyat BI, Mayo SL, 1996, *Prot Sci* 5:895-903) to give solvation parameters of 26 cal/mol/Å^2 favoring hydrophobic burial and 100 cal/mol/Å^2 opposing polar burial.

Introduction

Many energy minimization schemes require an energy expression that depends exclusively on the superposition of two body interactions. Of particular interest to us is the dead-end elimination theorem (Desmet et al.,

1992) which allows at most two body interactions between amino acid sidechain rotamers and the protein backbone (or template) and between pairs of rotamers. Terms that depend on more than two bodies cannot be included. This leads to a general problem of accommodating surface area dependent terms in such energy expressions, since the buried and/or exposed surface areas of three or more interacting bodies cannot be calculated exactly as the sum of two body interactions.

The problem is exacerbated when calculating surface areas using the Lee and Richards definition of solvent-accessible surface area (Lee & Richards, 1971) where 1.4 Å is added to every atomic radius before calculation of the area. This increases the number of intersecting atoms and makes an accurate calculation of solvent-accessible surface areas by a two body method problematic (Figure 1a,b). As Figure 1b shows, a simple two body method to calculate exposed hydrophobic solvent-accessible surface areas correlates poorly with the true surface areas, and as such limits a simple two body method's utility in protein design calculations.

A two body approach has been considered in the context of increasing the speed of calculation of buried hydrophobic surface area for folding studies (Wodak & Janin, 1980; Kurochkina & Lee, 1995) where the areas of individual atoms or pseudo-atoms were calculated pairwise. These areas were either combined statistically (assuming randomly distributed atoms) or added and scaled, finding high correlation with the true Lee and Richards surface areas. The use of reduced van der Waal radii to compensate for pairwise over-counting has also been discussed (Hodes et al., 1979; Augspurger & Scheraga, 1996). Other (not necessarily pairwise) techniques for calculating surface areas have been recently reviewed (Connolly, 1996). Here we find empirically that by scaling only the portion of the expression for pairwise area that is subject to

over-counting, we can achieve excellent agreement with both the true buried and the true exposed solvent-accessible surface areas.

Results and Discussion

The pairwise calculation of surface areas used in this study differs in several key respects from that of our previous work (Dahiyat & Mayo, 1996). Here we include backbone atoms (N, HN, CA, HCA, C and O) in the calculation of surface areas. For each sidechain rotamer r at residue position i with a local tri-peptide backbone $t3$ ([CA, C, O]$_{i-1}$, [N, HN, CA, HCA, C, O]$_i$, [N, HN, CA]$_{i+1}$), we calculate $A^{o}_{i_r t3}$, the exposed area of the rotamer and its backbone in the presence of the local tri-peptide backbone, and $A_{i_r t}$, the exposed area of the rotamer and its backbone in the presence of the entire template t which is the protein backbone (Figure 2). The difference between $A^{o}_{i_r t3}$ and $A_{i_r t}$ is the total area buried by the template for a rotamer r at residue position i. For each pair of residue positions i and j and rotamers r and s on i and j, respectively, we calculate $A_{i_r j_s t}$, the exposed area of the rotamer pair in the presence of the entire template. The difference between $A_{i_r j_s t}$ and the sum of $A_{i_r t}$ and $A_{j_s t}$ is the area buried between residues i and j, excluding that area buried by the template. The pairwise approximation to the total buried surface area is

$$A^{\text{pairwise}}_{\text{buried}} = \sum_i \left(A^{o}_{i_r t3} - A_{i_r t} \right) + s \sum_{i<j} \left(A_{i_r t} + A_{j_s t} - A_{i_r j_s t} \right) \qquad (1)$$

As shown in Figure 2, the second sum in (1) over-counts the buried area. We have therefore multiplied the second sum by a scale factor s whose value is to be determined empirically. Expected values of s are discussed below.

Noting that the buried and exposed areas should add to the total area, $\sum_i A^{\circ}_{i_r t 3}$, the solvent-exposed surface area is

$$A^{\text{pairwise}}_{\text{exposed}} = \sum_i A_{i_r t} - s \sum_{i<j} \left(A_{i_r t} + A_{j_s t} - A_{i_r j_s t} \right) \qquad (2)$$

The first sum of (2) represents the total exposed area of each rotamer in the context of the protein template ignoring interactions with other rotamers. The second sum of (2) subtracts the buried areas between rotamers and is scaled by the same parameter s as in (1).

Some insight into the expected value of s can be gained from consideration of a close-packed face centered cubic lattice of spheres of radius r. When the radii are increased from r to R, the surface area on one sphere buried by a neighboring sphere is $2\pi R(R-r)$. We take r to be a carbon radius (1.95 Å), and R is 1.4 Å larger. Then, using

$$s = \frac{\text{true buried area}}{\text{pairwise buried area}}$$

and noting that each sphere has 12 neighbors, we have

$$s = \frac{4\pi R^2}{12 \times 2\pi R(R-r)}$$

This yields $s = 0.40$. We note that a close-packed face centered cubic lattice has a packing density of 74%, and that protein interiors have a similar packing density, although because many atoms are covalently bonded the close packing is exaggerated (Creighton, 1993; Richards & Lim, 1994). We therefore expect $s = 0.40$ to be a lower bound for real protein cores. For non-core residues, where the packing density is lower, we expect a somewhat larger value of s.

We classified residues from ten proteins ranging in size from 54 to 289 residues into core or non-core, as described in Materials and Methods (Table 1). The classification into core and non-core was made because core residues interact more strongly with one another than do non-core residues. This leads to greater over-counting of the buried surface area for core residues.

Considering the core and non-core cases separately, the value of s which most closely reproduced the true Lee and Richards surface areas was calculated for the ten proteins. The pairwise approximation very closely matches the true buried surface area (Figure 3a,b). It also performs very well for the exposed hydrophobic surface area of non-core residues (Figure 4b). The calculation of the exposed surface area of the entire core of a protein involves the difference of two large and nearly equal areas and is less accurate (Figure 4a); as will be shown, however, when there is a mixture of core and non-core residues, a high accuracy can still be achieved. These calculations indicate that for core residues s is 0.42 and for non-core residues s is 0.79.

To test whether the classification of residues into core and non-core was sufficient, we examined subsets of interacting residues in the core and non-core positions, and compared the true buried area of each subset with that calculated by (1) (using the above values of s). For both subsets of the core and of the non-core, the correlation remained high (R^2 = 1.00) indicating that no further classification is necessary (data not shown). (Subsets were generated as follows: given a seed residue, a subset of size two was generated by adding the closest residue; the next closest residue was added for a subset of size three, and this was repeated up to the size of the protein. Additional subsets were generated by selecting different seed residues.)

It remains to apply this approach to calculating the buried or exposed surface areas of an arbitrary selection of interacting core and non-core residues

in a protein. When a core residue and a non-core residue interact, we replace (1) with

$$A_{\text{buried}}^{\text{pairwise}} = \sum_i \left(A_{i_r t3}^{\circ} - A_{i_r t} \right) + \sum_{i<j} \left(s_i A_{i_r t} + s_j A_{j_s t} - s_{ij} A_{i_r j_s t} \right) \quad (3)$$

and (2) with

$$A_{\text{exposed}}^{\text{pairwise}} = \sum_i A_{i_r t} - \sum_{i<j} \left(s_i A_{i_r t} + s_j A_{j_s t} - s_{ij} A_{i_r j_s t} \right) \quad (4)$$

where s_i and s_j are the values of s appropriate for residues i and j, respectively, and s_{ij} takes on an intermediate value. Using subsets from the whole of 1pga, the optimal value of s_{ij} was found to be 0.74. This value was then shown to be appropriate for other test proteins (Figure 5a,b). The correlation shown in Figure 5b represents a substantial improvement over that shown in Figure 1b and demonstrates the utility of our approach.

In previous work we examined the ability of a simple van der Waals potential energy function to predict the thermal stability of a series of coiled coils (Dahiyat & Mayo, 1996). We noted a significant improvement in the correlation between calculated stabilities and experimentally measured stabilities when a hydrophobic burial benefit of $\sigma_{np} A_{\text{buried}}^{np}$ was included in the calculated energies, where σ_{np} is a hydrophobic solvation parameter whose value was determined to be 23 cal/mol/Å^2, and A_{buried}^{np} was the calculated buried hydrophobic area. The correlation between calculated energies and experimental melting temperatures was further improved by penalizing polar surface area burial by $\sigma_p A_{\text{buried}}^{p}$, where σ_p is a polar solvation parameter and A_{buried}^{p} was the calculated buried polar area. The best values of σ_{np} and σ_p were found to be 16 cal/mol/Å^2 and 86 cal/mol/Å^2, respectively, when both solvation terms were used together. In order to

benefit from the more accurate pairwise surface area method in protein design studies, it is necessary to update the values of σ_{np} and σ_p. We use (3) and the values of *s* described above. Residue 26 of the coiled coil used in the previous study was the only residue determined to be in the core. When only the hydrophobic burial benefit was considered, the best fit value of σ_{np} was determined to be 48 cal/mol/Å^2. When both the hydrophobic burial benefit and the polar burial penalty were considered together, the best fit values of σ_{np} and σ_p were determined to be 26 cal/mol/Å^2 and 100 cal/mol/Å^2, respectively (Figure 6).

By examining a test set of proteins of various sizes, we have determined that the true Lee and Richards buried and exposed surface areas can be approximated well as a superposition of two body interactions using (3) and (4), with values for the parameter *s* that depend on the structural context of each residue. For core residues *s* is 0.42, for non-core positions *s* is 0.79, and for interactions between core and non-core positions s_{ij} is 0.74.

Methods

We considered ten representative proteins whose Brookhaven Protein Databank codes (Bernstein et al., 1977) are listed in Table 1. The program BIOGRAF (Molecular Simulations Incorporated, San Diego, California) was used to generate explicit hydrogens on the structures which were then conjugate gradient minimized for 50 steps using the DREIDING force field (Mayo et al., 1990).

We classified residues as core or non-core using an algorithm that considered the direction of each sidechain's Cα–Cβ vector relative to a surface computed using only the template Cα atoms with a carbon radius of 1.95 Å, a probe radius of 8 Å and no add-on radius. A residue was classified as a core

position if both the distance from its Cα atom (along its Cα–Cβ vector) to the surface was greater than 5.0 Å and the distance from its Cβ atom to the nearest point on the surface was greater than 2.0 Å (Dahiyat & Mayo, 1997). The advantage of such an algorithm is that a knowledge of the amino acid type actually present at each residue position is not necessary.

Surface areas were calculated using the Connolly algorithm with a dot density of 10 $Å^{-2}$ (Connolly, 1983), using a probe radius of zero and an add-on radius of 1.4 Å (Lee & Richards, 1971) and atomic radii from the DREIDING forcefield (Mayo, et al., 1990). Atoms that contribute to the hydrophobic surface area are carbon, sulfur, and hydrogen atoms attached to carbon and sulfur.

Energy calculations and parameter optimizations for the coiled coil system were performed as previously described (Dahiyat & Mayo, 1996).

References

Augspurger JD, Scheraga HA. 1996. An efficient, differentiable hydration potential for peptides and proteins. *J Comp Chem 17*: 1549-1558.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EE, Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The protein data bank: a computer-based archival file for macromolecular structures. *J Mol Biol 112*: 535-542.

Connolly ML. 1983. Solvent accessible surfaces of proteins and nucleic acids. *Science 221*: 709-713.

Connolly ML. 1996. Molecular surfaces: a review. *Network Science 2*: `http://www.netsci.org/Issues/1996/Apr/articles.html`.

Creighton TE. 1993. *Proteins: Structure and Molecular Properties*. New York: W.H. Freeman and Co.

Dahiyat BI, Mayo SL. 1996. Protein design automation. *Prot Sci 5*: 895-903.

Dahiyat BI, Mayo SL. 1997. De novo protein design: fully automated sequence selection. *Science 278*: 82-87.

Desmet J, De Maeyer M, Hazes B, Lasters I. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature 356*: 539-542.

Hodes ZI, Némethy G, Scheraga HA. 1979. Model for the conformational analysis of hydrated peptides. Effect of hydration on the conformational stability of the terminally blocked residues of the 20 naturally occurring amino acids. *Biopolymers 18*: 1565-1610.

Kurochkina N, Lee B. 1995. Hydrophobic potential by pairwise surface area sum. *Prot Eng 8*: 437-442.

Lee B, Richards FM. 1971. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol 55*: 379-400.

Mayo SL, Olafson BD, Goddard WA, III. 1990. Dreiding - a generic force-field for molecular simulations. *J Phys Chem 94*: 8897-8909.

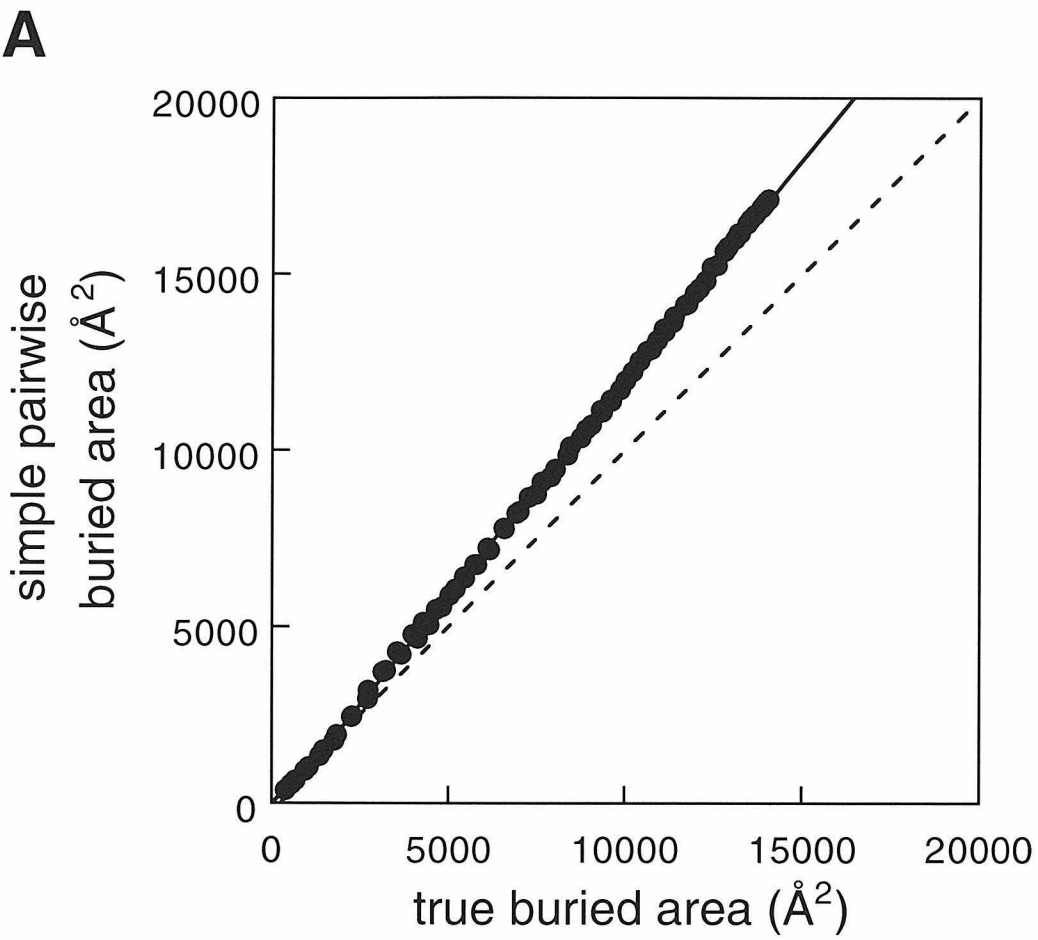
Richards FM, Lim WA. 1994. An analysis of packing in the protein folding problem. *Q Rev Biophys 26*: 423-498.

Wodak SJ, Janin J. 1980. Analytical approximation to the accessible surface area of proteins. *Proc Natl Acad Sci USA 77*: 1736-1740.

Table 5-1. Selected proteins, total number of residues and the number of residues in the core and non-core of each protein (Gly and Pro were not considered).

Brookhaven identifier	Total size	Core size	Non-core size
1enh	54	10	40
1pga	56	10	40
1ubi	76	16	50
1mol	94	19	61
1kpt	105	27	60
4azu-A	128	39	71
1gpr	158	39	89
1gcs	174	53	98
1edt	266	95	133
1pbn	289	96	143

Figure 5-1. Comparison of true solvent-accessible surface area and that calculated with the simplest pairwise technique (equations (1) and (2) with s = 1) for subsets of 1mol. a) Buried area. The line of best fit has slope 1.24 and a correlation coefficient R^2 = 1.00. Differences between calculated and true buried areas vary from 0 to 22%. b) Exposed hydrophobic area, with differences between calculated and true areas from 0 to 250% for small areas, converging to 100% for areas above 1000 Å^2. The line of best fit (not shown) has slope 0.00 and R^2 = 0.00. In each case, a dashed line of slope 1 is shown.


A
20000
15000
10000
5000
0
simple pairwise
buried area (Å2)
0
5000
10000
15000
20000
true buried area (Å2)

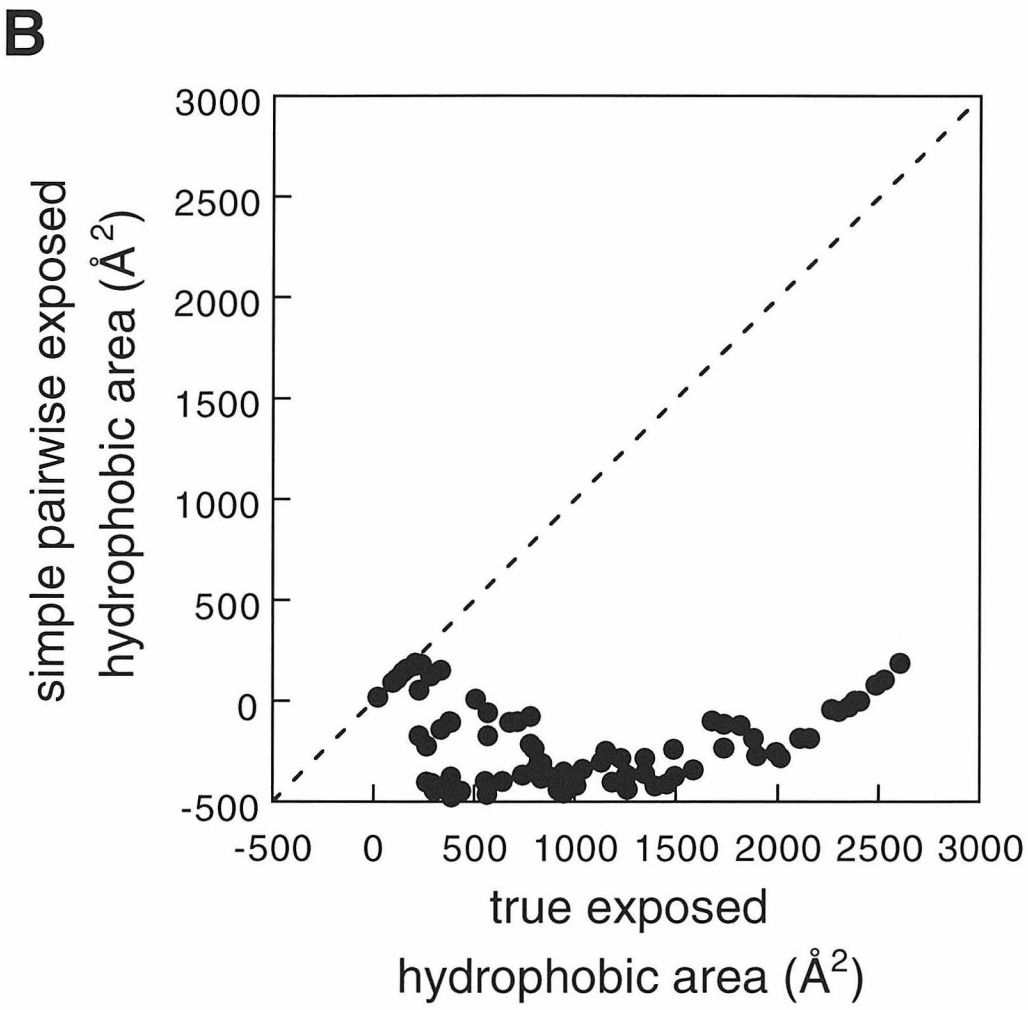
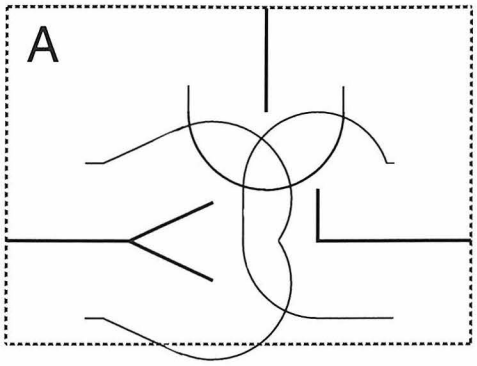
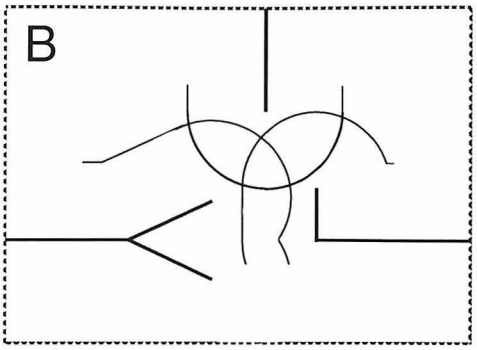
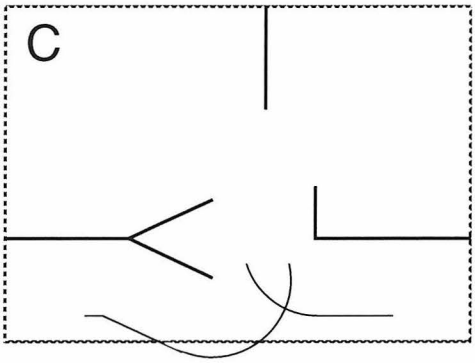
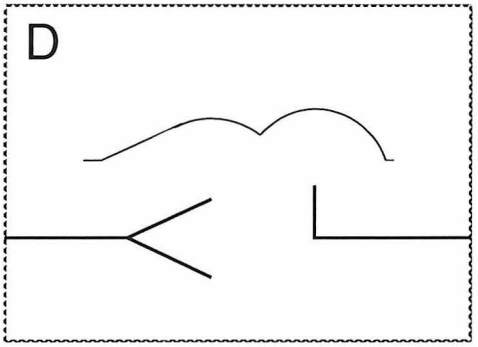

B
simple pairwise exposed hydrophobic area (Å2)
3000
2500
2000
1500
1000
500
0
-500
-500 0 500 1000 1500 2000 2500 3000
true exposed hydrophobic area (Å2)

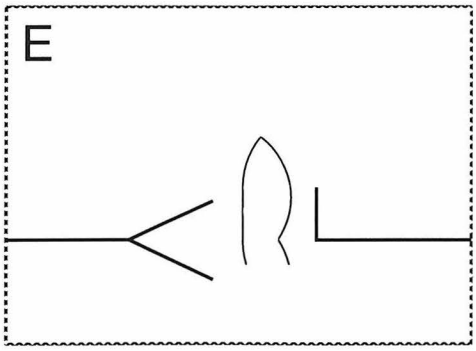
Figure 5-2. Areas involved in calculating the buried and exposed areas of equations (1) and (2). The dashed box is the protein template (i.e., the protein backbone), the heavy solid lines correspond to three rotamers at three different residue positions, and the lighter solid lines correspond to surface areas. a) $A^{\circ}_{i_r t3}$ for each rotamer. b) $A_{i_r t}$ for each rotamer; notice that the template has buried some area from the lower two rotamers. c) $\left(A^{\circ}_{i_r t3} - A_{i_r t}\right)$ summed over the three residues. The upper residue does not bury any area against the template except that buried in the tri-peptide state $A^{\circ}_{i_r t3}$. d) $A_{i_r j_s t}$ for one pair of rotamers. e) The area buried between rotamers, $\left(A_{i_r t} + A_{j_s t} - A_{i_r j_s t}\right)$, for the same pair of rotamers as in (d). f) The area buried between rotamers, $\left(A_{i_r t} + A_{j_s t} - A_{i_r j_s t}\right)$, summed over the three pairs of rotamers. The area intersected by all three rotamers (and only that area) is counted twice and is indicated by the double lines. The buried area calculated by (1) is the area buried by the template, represented in (c), plus s times the area buried between rotamers, represented in (f). The scaling factor s accounts for the over-counting shown by the double lines in (f). The exposed area calculated by (2) is the exposed area in the presence of the template, represented in (b), minus s times the area buried between rotamers, represented in (f).


A


B


C


D


E

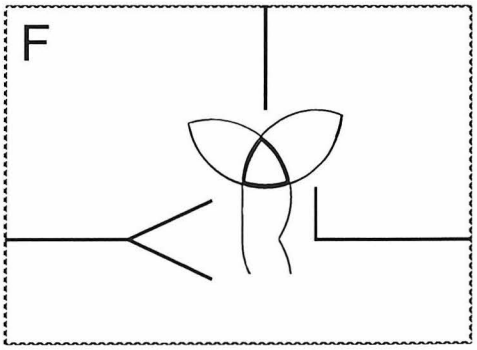
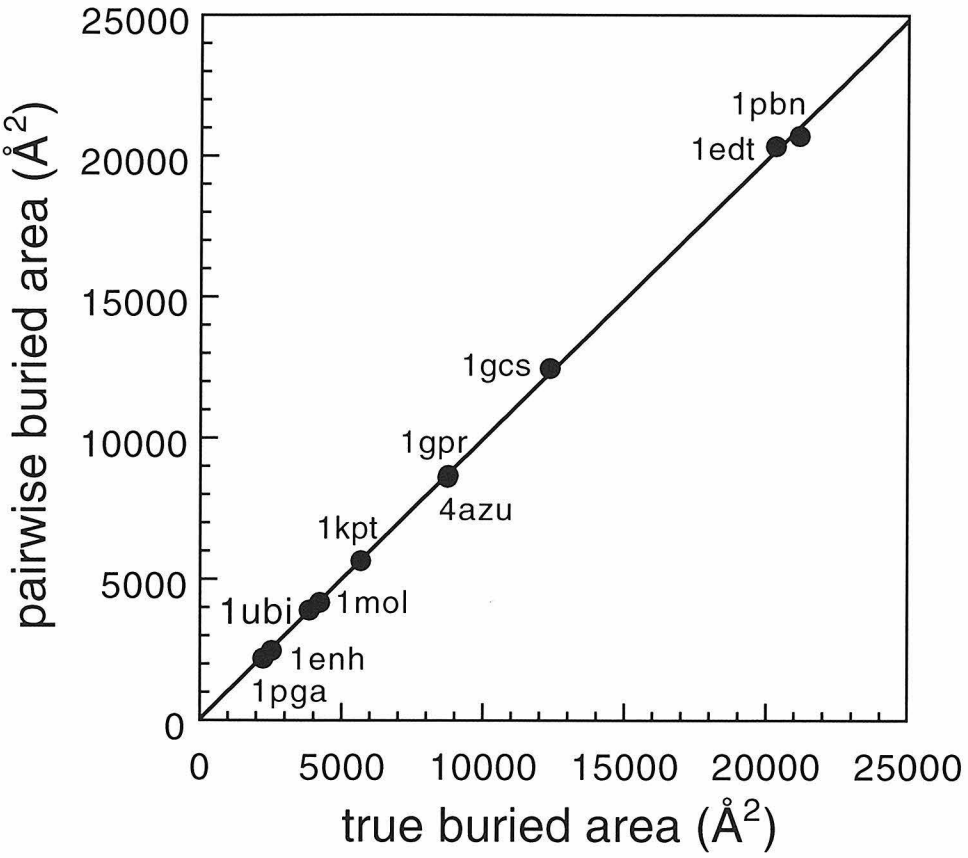

F

Figure 5-3. Comparison across ten proteins of the true buried surface area and the pairwise buried surface area calculated using (1). a) Core residues using s = 0.42. b) Non-core residues using s = 0.79. In each case the correlation coefficient R^2 = 1.00. The lines of best fit have slope 0.99 and 1.00 respectively, and differences between calculated and true buried areas are at most 2.5%.

A


25000
20000
15000
10000
5000
0
pairwise buried area (Å^2)
0
5000
10000
15000
20000
25000
true buried area (Å^2)
1pbn
1edt
1gcs
1gpr
4azu
1kpt
1mol
1ubi
1enh
1pga

B

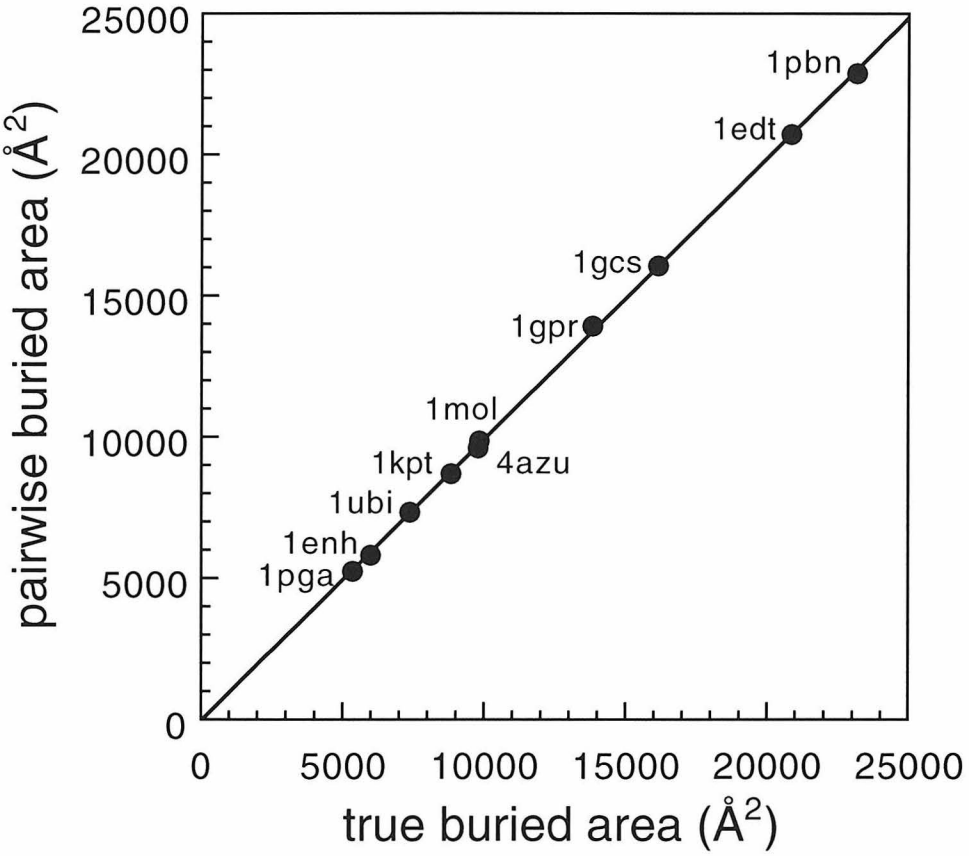
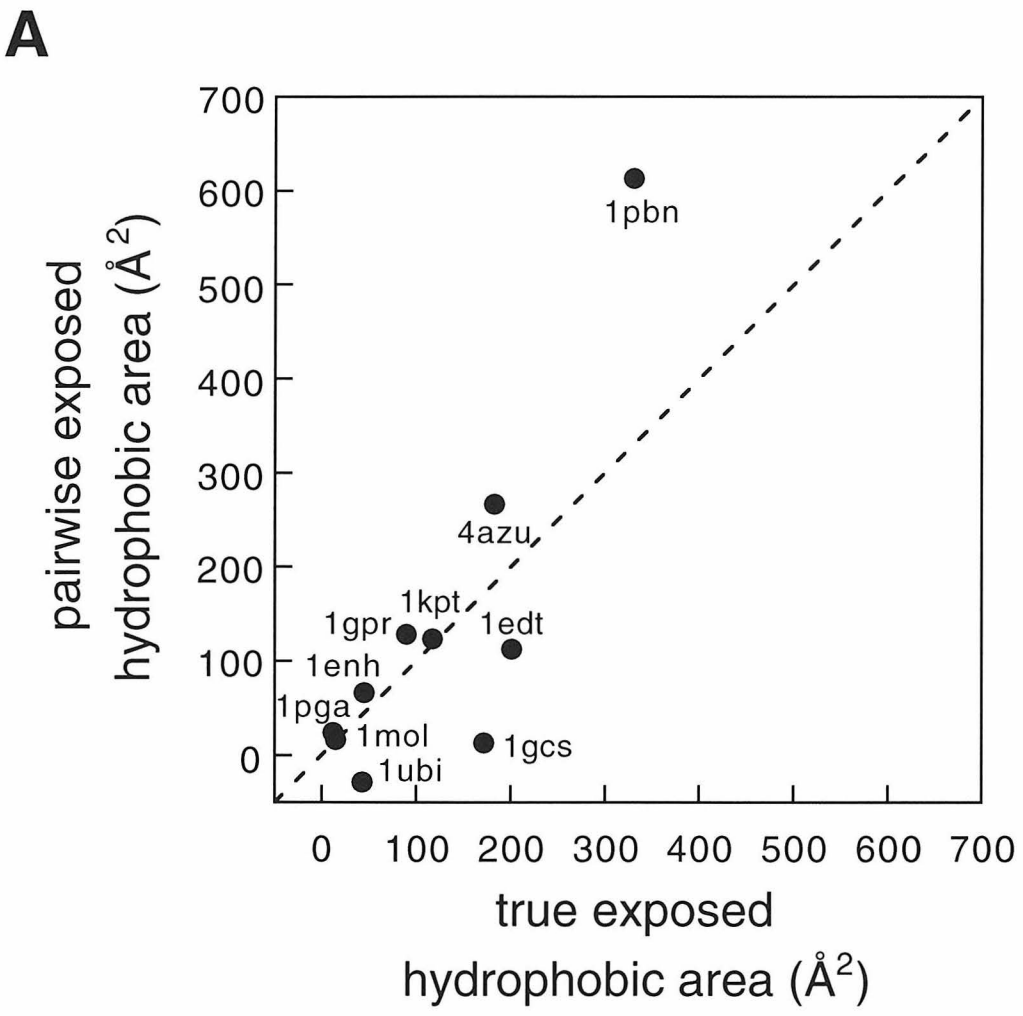

25000
20000
15000
10000
5000
0
pairwise buried area (Å^2)
0
5000
10000
15000
20000
25000
true buried area (Å^2)
1pbn
1edt
1gcs
1gpr
1mol
1kpt
4azu
1ubi
1enh
1pga

Figure 5-4. Comparison across ten proteins of the true exposed hydrophobic surface area and the pairwise exposed hydrophobic surface area calculated using (2). a) Core residues using $s = 0.42$, with $R^2 = 0.69$ (for reference, a dashed line of slope 1 is shown). The maximum difference between calculated and true exposed hydrophobic areas is 170%. b) Non-core residues using $s = 0.79$. The line of best fit has slope 1.02 and a correlation coefficient $R^2 = 1.00$. The maximum difference between calculated and true exposed hydrophobic areas is 5%.


A
700
600
500
400
300
200
100
0
pairwise exposed
hydrophobic area (Å2)
1pbn
4azu
1kpt
1gpr
1edt
1enh
1pga
1mol
1gcs
1ubi
0
100
200
300
400
500
600
700
true exposed
hydrophobic area (Å2)

B

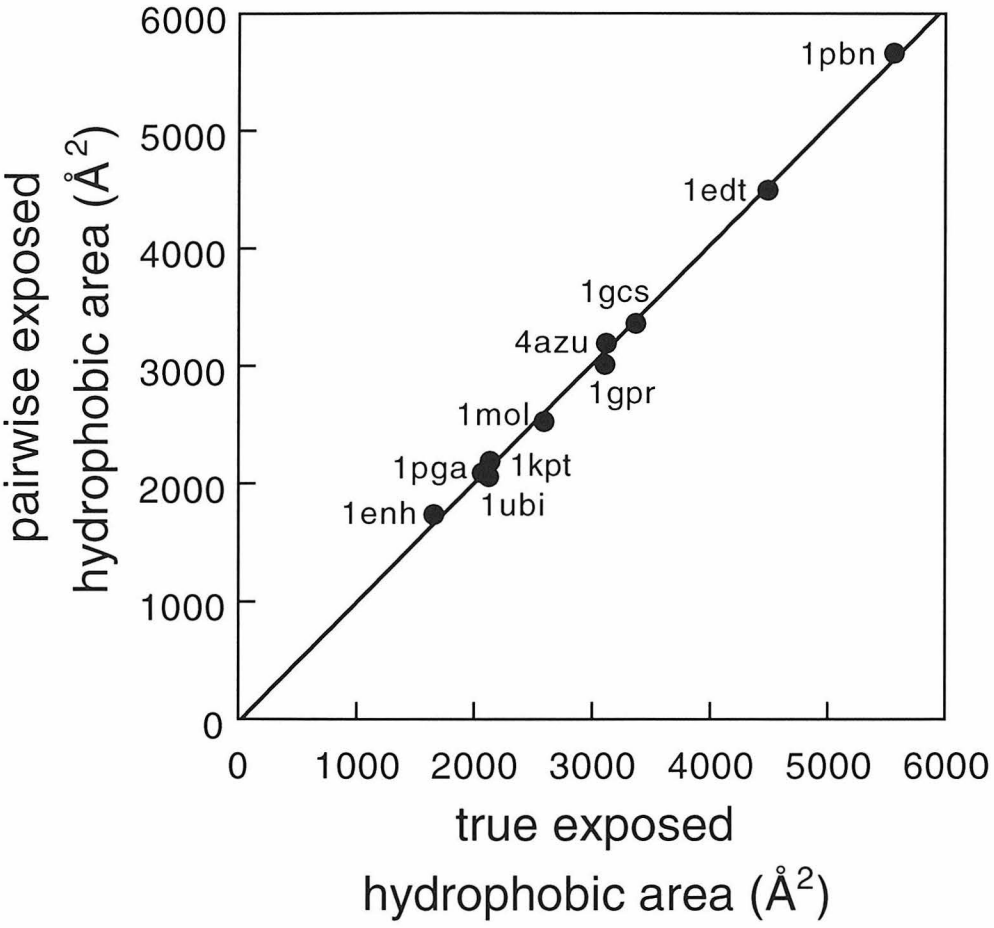
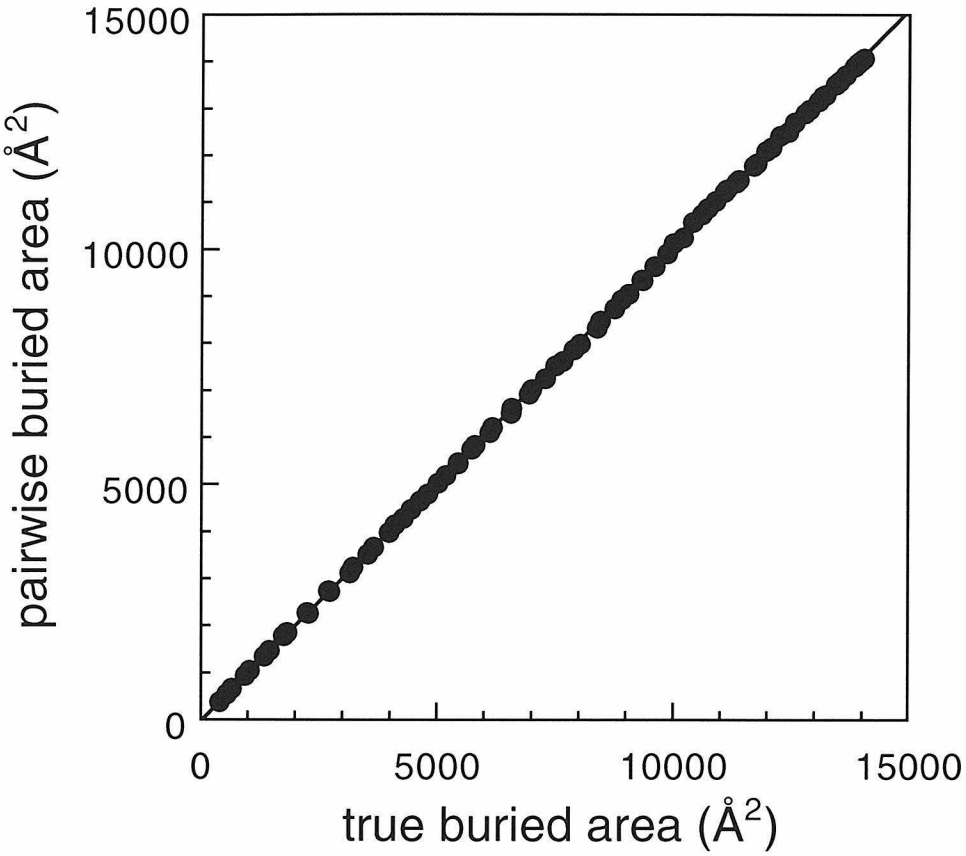

6000
5000
4000
3000
2000
1000
0
0
1000
2000
3000
4000
5000
6000
pairwise exposed hydrophobic area (Å^2)
true exposed hydrophobic area (Å^2)
1pbn
1edt
1gcs
4azu
1gpr
1mol
1pga
1kpt
1ubi
1enh

Figure 5-5. Comparison of true surface area and that calculated with (3) and (4) for subsets of 1mol using $s_{ij} = 0.74$. The subsets are the same as in Figure 1. a) Buried area. The line of best fit has slope 1.01, a correlation coefficient R^2 = 1.00, and a maximum difference between calculated and true buried area of 2%. b) Exposed hydrophobic area. The line of best fit has slope 1.05 and a correlation coefficient R^2 = 1.00, with differences between calculated and true areas from 0 to 30% for small areas, converging to 5% for areas above 1000 Å^2. These percent differences represent approximately an order of magnitude improvement over Figure 1.

A


15000
10000
5000
0
pairwise buried area (Å^2)
0
5000
10000
15000
true buried area (Å^2)

B

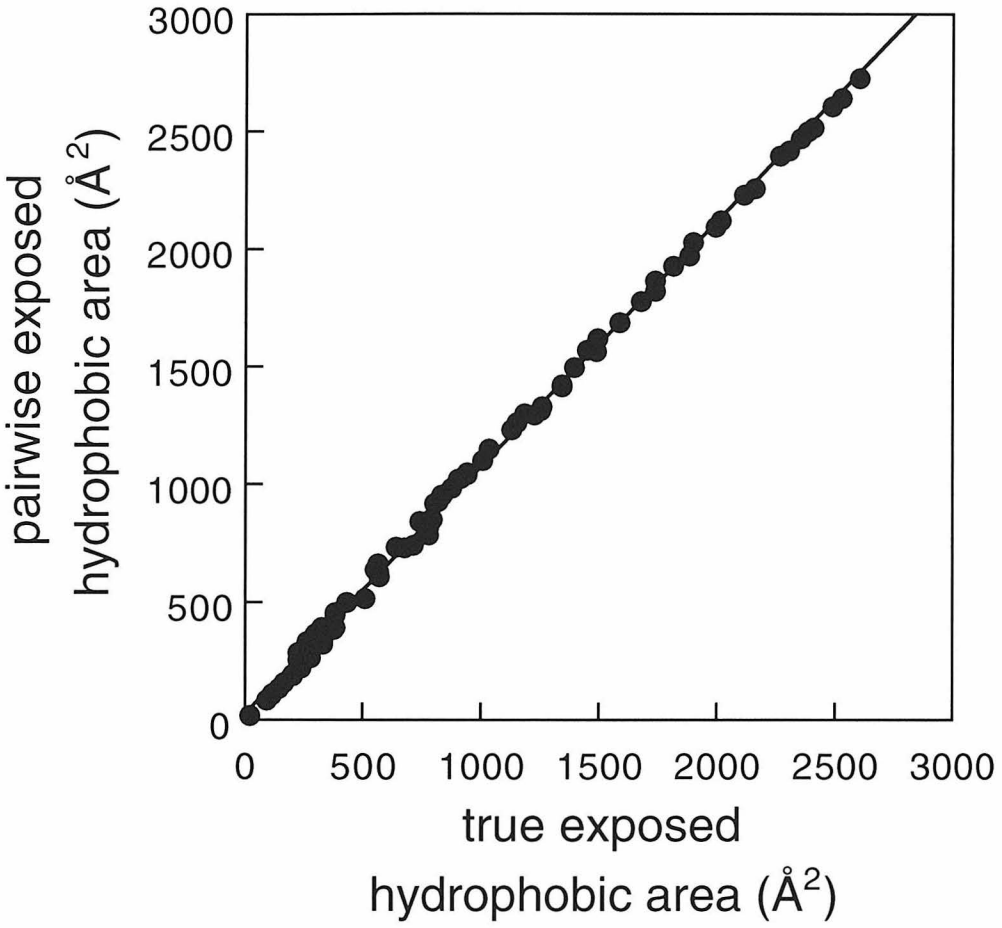
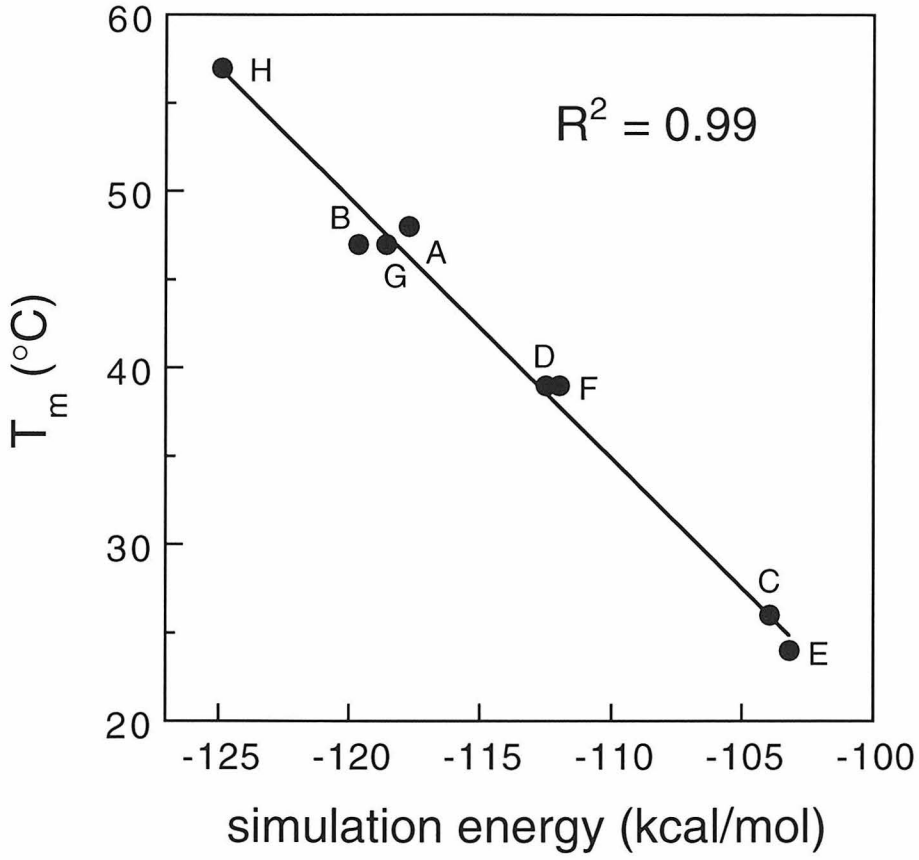

3000
2500
2000
1500
1000
500
0
pairwise exposed
hydrophobic area (Å^2)
0
500
1000
1500
2000
2500
3000
true exposed
hydrophobic area (Å^2)

Figure 5-6. Correlation between calculated and measured stability for designed coiled coils using buried surface areas calculated using (3) (compare to Figure 5b of (Dahiyat & Mayo, 1996)). Solvation parameter values are 26 cal/mol/Å^2 favoring hydrophobic burial and 100 cal/mol/Å^2 opposing polar burial. The labels A through H correspond to proteins PDA-3A through PDA-3H of (Dahiyat & Mayo, 1996).


60
50
40
30
20
T_m (°C)
H
B
A
G
D
F
C
E
R^2 = 0.99
-125
-120
-115
-110
-105
-100
simulation energy (kcal/mol)

Chapter 6.

Designing Real Protein β-Sheet Surfaces by *Z*-Score Optimization

The text of this chapter is partially adapted from the manuscript

Street A.G., Datta D., Gordon, D.B. and Mayo S.L. (to be submitted)

Abstract

Studies of lattice models of proteins have suggested that the appropriate energy expression for protein design may include non-thermodynamic terms in order to accommodate negative design concerns. A method has been developed to improve protein design in lattice model studies where enumeration of all possible sequences, and their ground state structures, is possible. The method maximizes a quantity known as the "Z-score," which compares the lowest energy sequence whose ground state structure is the target structure to an ensemble of random sequences. Here we show that, in certain circumstances, the technique can be applied to real proteins. The energy expression is then optimized using the assumption that the wildtype sequence is a low energy sequence (and its ground state is known to be the target structure). The new energy expression is used to design the β-sheet surfaces of two real proteins. We find experimentally that the resulting proteins are stable and well folded, and in one case is even more thermostable than the wildtype.

Introduction

Much effort in the field of computational protein design is directed towards developing a potential function to rank the compatibility of amino acid rotamer sequences with a target structure (Gordon et al., 1999). In a "protein design cycle" (Dahiyat & Mayo, 1996; Street & Mayo, 1999), the potential function is developed by cycling between experiment and simulation, so that the computational potential ideally approaches nature's "true" potential. This technique has had some remarkable recent successes (Dahiyat & Mayo, 1997a; Malakauskas & Mayo, 1998).

The approach nevertheless rests on a controversial assumption. Rotamer sequences are threaded onto the target structure, and the sequence with the lowest energy (as determined by the potential function) is reported as the best sequence for that structure. It is conceivable, though, that in some circumstances this sequence will not adopt the desired ground state structure. An extreme example is provided by imagining that the true potential function is one that only benefits hydrophobic contacts (and hydrophobic-polar and polar-polar interactions contribute zero energy) (Lau & Dill, 1989). Then, for any target structure, an all-hydrophobic sequence must be one of the best sequences. This sequence, of course, is not likely to fold specifically to the target structure – some polar residues ought to be included to characterize the surface of the molecule. Overcoming this problem involves introducing non-thermodynamic considerations to the design procedure, collectively known as "negative design" (Hellinga, 1997).

There are a number of schemes proposed to implement negative design, often specifically to solve the problem of the example in the last paragraph (or variations on it based on the Ising model of ferromagnetism). Perhaps the simplest is to use a fixed sequence composition, that is, to hold

the total number of hydrophobic and polar residues constant (Shakhnovich & Gutin, 1993). Even with this constraint, however, designed sequences are frequently found to fold to alternative structures of lower energy than the target structure (Shakhnovich, 1994; Yue et al., 1995). Alternatively, instead of minimizing the potential function, it is possible to choose a sequence to maximize the occupation probability of the target structure (Micheletti et al., 1998b; Seno et al., 1998).

Other approaches employed in lattice model studies involve adding non-thermodynamic terms to the potential function. One method is to introduce a "clamping potential" to force the molecule into the target structure, and then to minimize the difference between the clamping potential and the "true" potential (Kurosky & Deutsch, 1995; Deutsch & Kurosky, 1996). Another approach involves the addition of a penalty for exposing hydrophobic surface area (Sun et al., 1995).

Negative design is thus clearly important, at least in lattice model studies with simple potential functions and a limited set of amino acids (Crippen, 1996; Micheletti et al., 1998a). For real proteins and more physical potential functions, negative design can be necessary to guarantee the correct multimeric state of designed proteins (Harbury et al., 1993). A penalty for exposing hydrophobic surface area has also been shown to improve the designability of real proteins (Dahiyat & Mayo, 1997b; Malakauskas & Mayo, 1998).

In this chapter we take yet another approach to determining the optimal potential function for protein design, in which we maximize the energy gap between a low energy sequence known to fold to the target structure, and the average energy of an ensemble of random sequences threaded onto a target structure (Chiu & Goldstein, 1998). In a cubic 3x3x3

lattice simulation, the desired "true" potential can be selected manually and the protein folding problem can be solved. Thus a sequence *S*, whose ground state structure is the target structure, can be determined and its energy calculated. If the distribution of energies of the random sequences is assumed to be Gaussian, the success of the test potential for protein design is measured by the energy gap between the mean of the distribution and the energy of sequence *S*, normalized by the standard deviation of the distribution (Figure 1). This quantity is known as the *Z*-score of the sequence *S* on the target structure. The test potential is then adjusted to maximize the *Z*-score.

Chiu and Goldstein applied the method to a 3x3x3 lattice model, using statistically-derived pair potentials (Miyazawa & Jernigan, 1985) as the "true" potential. They found that the potential generated by maximizing the *Z*-score across many structures led to significantly better success at solving the protein design problem than the true potential. Here we show that the technique does not transfer readily to real proteins in their entirety. Nevertheless, we show that the technique can be applied to certain subsections of proteins. In particular we use it to design the β-sheet surfaces of the β1 immunoglobulin-binding domain of streptococcal protein G (GB1) and of a variant of poplar apoplastocyanin with the metal binding site removed (PCV).

The *Z*-Score Applied to Real Proteins

One of the key assumptions of the lattice model method of Chiu and Goldstein (Chiu & Goldstein, 1998) is that the energies of random sequences threaded onto the target structure form a Gaussian distribution. It would be surprising if this assumption were to hold for real proteins. In particular, one would expect that placing random amino acid side chains in the core of a protein would typically lead to unresolvable steric clashes, especially since the

modeled backbone of the target structure is held rigid. Indeed, Figure 2a shows the distribution of potential energies of random sequences threaded onto the core of GB1. The distribution is clearly not Gaussian, with most sequences yielding enormous energies. A Gaussian distribution may be achievable by using a statistically derived pair potential instead of an atomistic van der Waals potential, but designs using pair potentials have not yielded uniquely characterizable folded states (Isogai et al., 1999).

When only surface residues are considered, the situation is improved. For α-helix and β-sheet surface residues of GB1, the distribution of energies of random sequences is close to Gaussian, as shown in Figures 2b and 2c, respectively. Thus it appears that on the surface, even randomly selected amino acids are always able to find suitable rotamers that avoid severe steric interference. The *Z*-score analysis may therefore provide some insight into the appropriate potential function for α-helix and β-sheet surface design, provided one can find an appropriate sequence with which to calculate the *Z*-score. In lattice models, one knows the true potential function and can exhaustively search all conformations to solve the protein folding problem (Shakhnovich & Gutin, 1993). Hence the *Z*-score of a structure could be calculated using the lowest-energy sequence whose ground state is the target structure.

In contrast, in the lattice model study of Chiu and Goldstein (Chiu & Goldstein, 1998), the *Z*-score is actually calculated without knowledge of this lowest-energy sequence. One thousand 27-residue random amino acid sequences are constructed, which are found to correspond to 992 unique ground state structures. Eight sequences are discarded to yield a one-to-one correspondence between structures and sequences. The *Z*-score is calculated for each sequence in its ground state structure, using the 992 sequences to

determine the energy distribution. The potential function is then modified to maximize an appropriately formed average of the *Z*-scores. Thus, the reference sequence used to calculate the *Z*-score is not necessarily the lowest-energy sequence whose ground state structure is the target structure, but instead an arbitrary sequence whose ground state structure is the target structure. Nevertheless, the resulting potential function is significantly better for protein design than the "true" potential.

In our application of the theory to real proteins, we therefore expect that any arbitrary sequence known to fold to a target structure will suffice for calculating the *Z*-score of that structure. Given an experimentally determined structure, we can thus use the protein's wildtype sequence to calculate its *Z*-score. In essence, the method then chooses the potential function which locates the protein's wildtype sequence as far as possible down the tail of the distribution of energies.

Since a number of successful computational redesigns of α-helical surfaces have been reported (Dahiyat et al., 1997; Morgan, 2000), we chose to examine the *Z*-score technique on the β-sheet surface, where there have been few successful computational protein design efforts. Negative design issues are also expected to play a larger role in β-sheet design (Hecht, 1994). Rather than maximizing the *Z*-score of a large number of structures, as a first step we consider just one structure, so that the resulting potential function is optimized for protein design on that structure. This method should increase the possibility of the technique being successful for at least the one selected structure. The resulting potential function may then be applied to other proteins to test its generality, or a new potential function may be calculated by considering more protein structures. In particular, we chose to apply the technique to the eight β-sheet surface residues of GB1 which are not involved

in stabilizing interactions with neighboring turns (Figure 3a), and to the seven β-sheet surface residues on one face of PCV (Figure 4a).

The computational potential function, E, included van der Waals interactions, E_{vdW} (Mayo et al., 1990; Dahiyat & Mayo, 1997b), electrostatics, E_{elec}, and a hydrogen bonding potential, E_{HB} (Dahiyat, et al., 1997), a bias for secondary structure propensity, E_{SS} (Dahiyat, et al., 1997), and solvation energies. The solvation energies were a benefit for burial of hydrophobic surface area, A_{np}^{buried}, a penalty for burial of polar surface area, A_{polar}^{buried}, and a penalty for exposure of hydrophobic surface area, $A_{np}^{exposed}$ (Street & Mayo, 1998), and a further penalty for polar hydrogen burial, E_{phb} (Dahiyat, et al., 1997).

$$E = vE_{vdW} - \sigma_{np}A_{np}^{buried} + \xi_{np}A_{np}^{exposed} + \sigma_{p}A_{polar}^{buried} + \frac{1}{\varepsilon}E_{elec} + DE_{HB} + PE_{phb} + E_{SS}(N) \quad (1)$$

The magnitude of the van der Waals interactions, v, was held fixed and the relative magnitudes of the other seven energy terms (σ_{np}, ξ_{np}, σ_{p}, ε, D, P, and N as shown, where E_{SS} is an exponential function of N) were allowed to vary individually until the Z-score was maximized.

Results and Discussion

The resulting potential functions are shown in Table 1. For GB1, the maximum Z-score is 2.6, i.e., the wildtype sequence is assigned an energy lower than 99.5% of all possible sequences. For PCV, the maximum Z-score is 2.2. Also shown in Table 1 is the potential function built up over many experiments using the protein design cycle, which has been successful in particular for core design and α-helix surface design (Street & Mayo, 1999). The Z-score optimized potential functions exhibit some interesting common

features. The hydrophobic burial benefit, which is the main embodiment of the hydrophobic effect (Wesson & Eisenberg, 1992), has disappeared. This reflects the relative lack of importance of hydrophobic burial on the surface of proteins (although there may be some role for small hydrophobic clusters on the surface of β-sheets (Tisi & Evans, 1995)). The other solvation parameters are broadly similar to the experimental potential function.

The most dramatic difference from the protein design cycle potential is the increased importance of electrostatic interactions. The value of the dielectric constant used in the protein design cycle is similar to that of water, and leads to electrostatic interactions being de-emphasized. This value was never experimentally tested, however. Although saltbridges are not encouraged, the hydrogen bonding potential from the protein design cycle is quite strong (an ideal hydrogen bond receives a benefit of 8.0 kcal/mol). The *Z*-score optimized dielectric constant is an order of magnitude smaller, closer to unity. This is justifiable because we are considering effects at the molecular level, where the assumptions behind the use of the dielectric constant break down. The screening effect of solvent is also approximated by using a distance attenuated Coulomb potential (Mayo, et al., 1990).

To determine if the *Z*-score technique may be useful, this potential function must be used for real protein design. We used a combination of dead-end elimination (Desmet et al., 1992; Gordon & Mayo, 1998) and branch and terminate (Gordon & Mayo, 1999) to find the lowest energy sequence for each β-sheet surface, using the new potential functions. (These minimization algorithms are guaranteed to produce the absolute lowest energy sequence, unlike stochastic algorithms such as Monte Carlo.)

The resulting GB1 variant, GB1-Z1, is a five-fold mutant of the wildtype protein. One can clearly see the impact of the electrostatic term in

the potential function. The modeled side chain configurations are shown in Figure 3, alongside those of the wildtype crystal structures (Gallagher et al., 1994). A cluster of threonines and an isoleucine have been replaced by cross-strand saltbridge networks, Asp42 to Arg55, and Arg6 to Glu53 to Lys44. The wildtype saltbridge formed by Lys4 and Glu15 is maintained. Such cross-strand saltbridges might be expected to contribute to β-sheet formation and stability, and surface networks of saltbridges are postulated to be a stabilizing factor in hyperthermophilic proteins (Elcock, 1998; de Bakker et al., 1999).

The resulting PCV variant, PCV-Z1, is a three-fold mutant of the wildtype protein. The modeled side chain configurations are shown in Figure 4, alongside those of the apoplastocyanin wildtype crystal structure (Garrett et al., 1984). Again, the impact of the electrostatic term is clear, with a saltbridge network formed by Glu18, Lys95, Lys97 and Glu79.

The designed proteins were made experimentally using standard molecular biology techniques and their properties measured. Their far UV circular dichroism spectra overlay those of the wildtype proteins. The melting temperature of GB1-Z1 was determined to be 71 °C (Figure 5). The melting temperature of GB1 is 86 °C. Although the designed protein is not as stable as the wildtype protein, it appears to fold to the correct structure. Although the literature contains many examples of alterations to the β-sheet surface of GB1, we know of no instances resulting in greater than wildtype stability. This is the first example of a well formed, many-stranded β-sheet designed through purely computational means.

The results for PCV-Z1 were even more impressive. The melting temperature of PCV-Z1 was determined to be 64 °C, compared to the melting temperature of PCV of 56 °C (Figure 6). The designed protein is thus even

more stable than the natural one. To our knowledge, this is the first time a natural protein's stability has been increased by redesigning its β-sheet surface.

Materials and Methods

Simulation

The core residues of GB1 are positions 3, 5, 7, 20, 26, 30, 34, 39, 52, and 54. The eight β-sheet surface positions of GB1 considered here are 4, 6, 15, 17, 42, 44, 53, and 55. The α-helix surface positions of GB1 are 24, 27, 28, 31, 32, 35, and 36. The seven β-sheet surface positions of PCV considered here are 18, 20, 79, 81, 93, 95, and 97. These follow from our residue classification algorithm (Dahiyat & Mayo, 1997a). The potential function used in Figure 1 is derived from the protein design cycle, shown in Table 1.

The *Z*-score maximization algorithm searched along each potential function basis vector (that is, varying the scale factor for each energy term in (1)) individually to maximize the *Z*-score. The search was initiated at the potential function derived from the protein design cycle, from the van der Waals potential alone, and from other random potentials, and always converged to the same result. Further, the ordering of the search through basis vectors had no effect on the result. It was found that this optimization algorithm was sufficient to find the maximum *Z*-score.

The *Z*-score was calculated using 4000 random sequences to determine the energy distribution of the potential function on the structure, resulting in an uncertainty in the *Z*-score of ±0.04. The random sequences were composed of the polar amino acids Ser, Thr, Asp, Asn, Glu, Gln, Lys and Arg, as well as the hydrophobic amino acids Ala, Val and Ile. The results were surprisingly robust to changes in the set of amino acids considered. In particular, the

results were not significantly different if Ala was removed from consideration, or if His, Met and Gly were included.

In contrast to the case in lattice models, real amino acids may adopt many different conformations, or rotamers. The energy of a given amino acid sequence on a structure is thus calculated by minimizing the energy across all possible rotamer configurations, using dead-end elimination. For this procedure a backbone-dependent rotamer library was used (Dunbrack & Karplus, 1993), in which the χ_1 angles of all hydrophobic amino acid rotamers were expanded ±1 standard deviation about the mean value (Dahiyat, et al., 1997).

Experimental

A synthetic GB1 gene (Minor & Kim, 1994) was cloned into a pET11a vector (Novagen) and used as the template for QuikChange mutagenesis (Qiagen). A synthetic PCV gene was constructed by recursive PCR (Prodromou & Pearl, 1992). The genes were confirmed by DNA sequencing. The expression and purification of the protein followed published procedures, and was verified by mass spectrometry. The 56-residue form of GB1 (with N-terminal methionine processed) and the 100-residue form of PCV (including the N-terminal methionine) were used. PCV was derived from wildtype poplar apoplastocyanin (Garrett, et al., 1984) by removing its metal binding site through the mutations His37 to Val and Cys84 to Ala. These mutations are in the core of the molecule and are not expected to interact with changes to the surface of the protein. The melting temperature of PCV was observed to be 56 °C compared to 51 °C for unmodified apoplastocyanin.

Far UV circular dichroism spectra were measured on an Aviv 62DS spectrometer. The spectra of GB1 and GB1-Z1 were measured at pH 5.5, in 50 mM phosphate and 50 μM protein, using a 1 mm path length, with thermal melts performed at 218 nm using 2 °C temperature steps with an averaging time of 30 s and an equilibration time of 2 min. A guanidinium denaturation of GB1-Z1 was also performed using an auto-titrator and a 10 minute equilibration time, yielding a stability of 3.6 kcal/mol at 1 °C (Santoro & Bolen, 1988). The spectra of PCV and PCV-Z1 were measured at pH 7.0, in 50 mM potassium phosphate, 0.5 M sodium sulfate, and 70 μM protein, with thermal melts performed at 210 nm. The melting temperatures were derived by evaluating the maximum of a $d\theta/dT$ versus T plot. Protein concentration was determined by UV spectrophotometry.

Conclusion

We report the first time a natural protein's stability has been increased by redesigning its β-sheet surface. Further, it is notable that we have in fact designed two stable protein β-sheet surfaces using different potential functions. Indeed, further application of the technique to other proteins suggests yet different potentials may be appropriate. This supports the belief that there may be alternative routes taken by nature to stabilize protein surfaces, and which may be taken in *de novo* design too (Cordes et al., 1996). Of course, one test of this proposal is to use the potential derived from one protein to design the β-sheet surface of another, and preliminary results in this regard appear promising. A further advantage of the approach outlined in this chapter is that it could lead to a faster turn-around time for protein design, since it optimizes the potential function with less frequent recourse to experiment.

References

Chiu TL, Goldstein RA. 1998. Optimizing potentials for the inverse protein folding problem. *Prot Eng 11*: 749-752.

Cordes MHJ, Davidson AR, Sauer RT. 1996. Sequence space, folding and protein design. *Curr Opin Struct Biol 6*: 3-10.

Crippen GM. 1996. Failures of inverse folding and threading with gapped alignment. *Proteins 26*: 167-171.

Dahiyat BI, Gordon DB, Mayo SL. 1997. Automated design of the surface positions of protein helices. *Prot Sci 6*: 1333-1337.

Dahiyat BI, Mayo SL. 1996. Protein design automation. *Prot Sci 5*: 895-903.

Dahiyat BI, Mayo SL. 1997a. De novo protein design: fully automated sequence selection. *Science 278*: 82-87.

Dahiyat BI, Mayo SL. 1997b. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci USA 94*: 10172-10177.

de Bakker PIW, Hunenberber PH, McCammon JA. 1999. Molecular dynamics simulations of the hyperthermophilic protein Sac7d from *Sulfolobus acidocaldarius*: contributions of salt bridges to thermostability. *J Mol Biol 285*: 1811-1830.

Desmet J, De Maeyer M, Hazes B, Lasters I. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature 356*: 539-542.

Deutsch JM, Kurosky T. 1996. New Algorithm for Protein Design. *Phys Rev Lett 76*: 323-326.

Dunbrack RL, Karplus M. 1993. Backbone dependent rotamer library for proteins - an application to side-chain prediction. *J Mol Biol 230*: 543-574.

Elcock AH. 1998. The stability of salt bridges at high temperatures: implications for hyperthermophilic proteins. *J Mol Biol 284*: 489-502.

Gallagher T, Alexander P, Bryan P, Gilliland GL. 1994. Two crystal structures of the β1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. *Biochem 33*: 4721-4729.

Garrett TPJ, Clingeleffer DJ, Guss JM, Rogers SJ, Freeman HC. 1984. The crystal structure of poplar apoplastocyanin at 1.8-A resolution - the geometry of the copper-binding site is created by the polypeptide. *J Biol Chem 259*: 2822-2825.

Gordon DB, Marshall SA, Mayo SL. 1999. Energy functions for protein design. *Curr Opin Struct Biol 9*: 509-513.

Gordon DB, Mayo SL. 1998. Radical performance enhancements for combinatorial optimization algorithms based on the dead-end elimination theorem. *J Comp Chem 19*: 1505-1514.

Gordon DB, Mayo SL. 1999. Branch and terminate: a combinatorial optimization algorithm for protein design. *Structure* in press.

Harbury PB, Zhang T, Kim PS, Alber T. 1993. A switch between 2-stranded, 3-stranded and 4-stranded coiled coils in gcn4 leucine-zipper mutants. *Science 262*: 1401-1407.

Hecht MH. 1994. De novo design of β-sheet proteins. *Proc Natl Acad Sci USA 91*: 8729-8730.

Hellinga HW. 1997. Rational protein design: combining theory and experiment. *Proc Natl Acad Sci USA 94*: 10015-10017.

Isogai Y, Ota M, Fujisawa T, Izuno H, Mukai M, Nakamura H, Iizuka T, Nishikawa K. 1999. Design and synthesis of a globin fold. *Biochem 38*: 7431-7443.

Kurosky T, Deutsch JM. 1995. Design of copolymeric materials. *J Phys A 27*: L387-L393.

Lau KF, Dill KA. 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules 22*: 3986-3997.

Malakauskas SM, Mayo SL. 1998. Design, structure and stability of a hyperthermophilic protein variant. *Nature Struct Biol 5*: 470-475.

Mayo SL, Olafson BD, Goddard WA, III. 1990. Dreiding - a generic force-field for molecular simulations. *J Phys Chem 94*: 8897-8909.

Micheletti C, Seno F, Maritan A, Banavar JR. 1998a. Design of proteins with hydrophobic and polar amino acids. *Proteins 32*: 80-87.

Micheletti C, Seno F, Maritan A, Banavar JR. 1998b. Protein design in a lattice model of hydrophobic and polar amino acids. *Phys Rev Lett 80*: 2237-2240.

Minor DL, Kim PS. 1994. Measurements of the β-sheet-forming propensities of amino acids. *Nature 367*: 660-663.

Miyazawa S, Jernigan RL. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules 18*: 534-552.

Morgan CS. 2000. Full sequence design of an alpha-helical protein and investigation of the importance of helix dipole and capping effects in helical protein design. *PhD thesis*. California Institute of Technology.

Prodromou C, Pearl LH. 1992. Recursive PCR: a novel technique for total gene synthesis. *Prot Eng 5*: 827-829.

Santoro MM, Bolen DW. 1988. Unfolding free energy changes determined by the linear extrapolation method. 1. Unfolding of

phenylmethanesulfonyl α-chymotrypsin using different denaturants. *Biochem 27*: 8063-8068.

Seno F, Micheletti C, Maritan A, Banavar JR. 1998. Variational approach to protein design and extraction of interaction potentials. *Phys Rev Lett 81*: 2172-2175.

Shakhnovich EI. 1994. Proteins with selected sequences fold into unique native conformations. *Phys Rev Lett 72*: 3907-3910.

Shakhnovich EI, Gutin AM. 1993. Engineering of stable and fast-folding sequences of model proteins. *Proc Natl Acad Sci USA 90*: 7195-7199.

Street AG, Mayo SL. 1998. Pairwise calculation of protein solvent accessible surface areas. *Folding and Design 3*: 253-258.

Street AG, Mayo SL. 1999. Computational protein design. *Structure 7*: R105-R109.

Sun S, Brem R, Chan HS, Dill KA. 1995. Designing amino acid sequences to fold with good hydrophobic cores. *Prot Eng 8*: 1205-1213.

Tisi LC, Evans PA. 1995. Conserved structural features on protein surfaces: small exterior hydrophobic clusters. *J Mol Biol 249*: 251-258.

Wesson L, Eisenberg D. 1992. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Prot Sci 1*: 227-235.

Yue K, Fiebig KM, Thomas PD, Chan HS, Shakhnovich EI, Dill KA. 1995. A test of lattice protein folding algorithms. *Proc Natl Acad Sci USA 92*: 325-329.

Table 6-1.

Potential functions determined through different methods. The energy terms considered are shown in (1). The van der Waals energy scale factor v was held fixed. A potential function has been developed using the protein design cycle (Street & Mayo, 1999), and has been successful for core and α-helix surface design in particular. The *Z*-score method applied to the β-sheet surface of PCV and of GB1 yield new potential functions. Also shown are the ranges over which each parameter may be changed while keeping the *Z*-score within 5% of its maximum (when the other parameters are kept fixed). The units of the solvation parameters are kcal/mol/Å^2.

Energy term	Design cycle	PCV	Range	GB1	Range
van der Waals v	1.0	1.0	n.a.	1.0	n.a.
np burial σ_{np}	0.05	0.0	0.0 – 0.01	0.0	0.0 – 0.02
np exposure ξ_{np}	0.05	0.10	0.04 – 0.16	0.06	0.02 – 0.08
polar burial σ_p	0.0	0.0	0.0 – 0.04	0.03	0.01 – 0.06
dielectric ε	40.0	4.0	2.0 – 6.0	4.0	2.0 – 6.0
H-bond D	8.0	1.0	1.0 – 8.0	6.0	1.0 – 8.0
polar H burial P	2.0	9.0	6.0 – 15.0	3.0	1.0 – 7.0
secondary structure bias N	n.a.	1.0	0.0 – 1.4	1.4	0.8 – 1.6

Figure 6-1. The assumed distribution of energies of sequences threaded onto the target structure. Sequence S_0 is the lowest energy sequence whose ground state structure is the target structure. Note that there may be sequences of lower energy which do not fold to the target structure. By altering the energy function non-thermodynamically, negative design seeks to move these sequences above S_0.

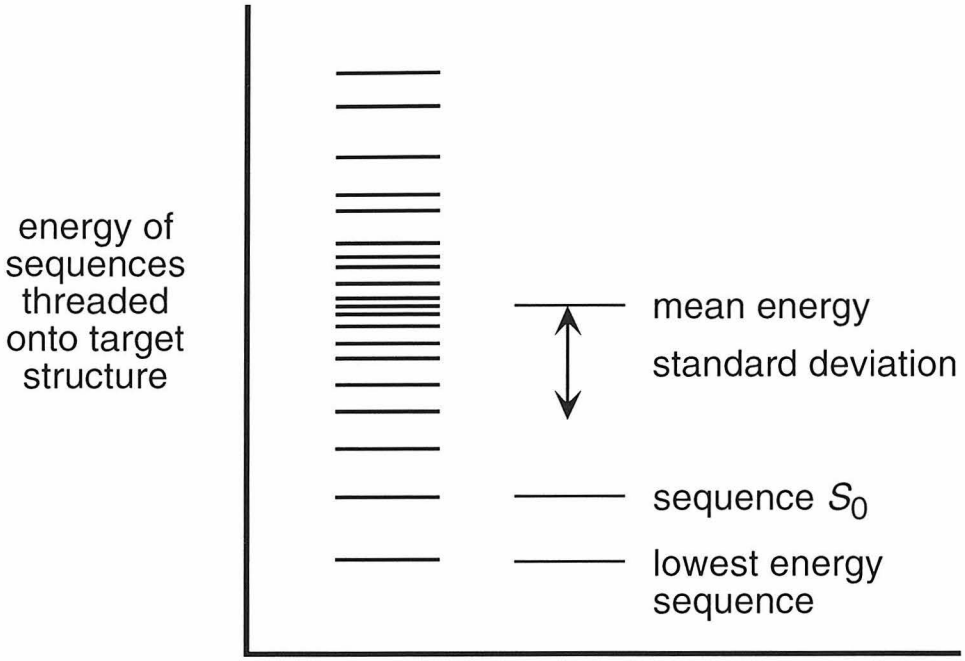
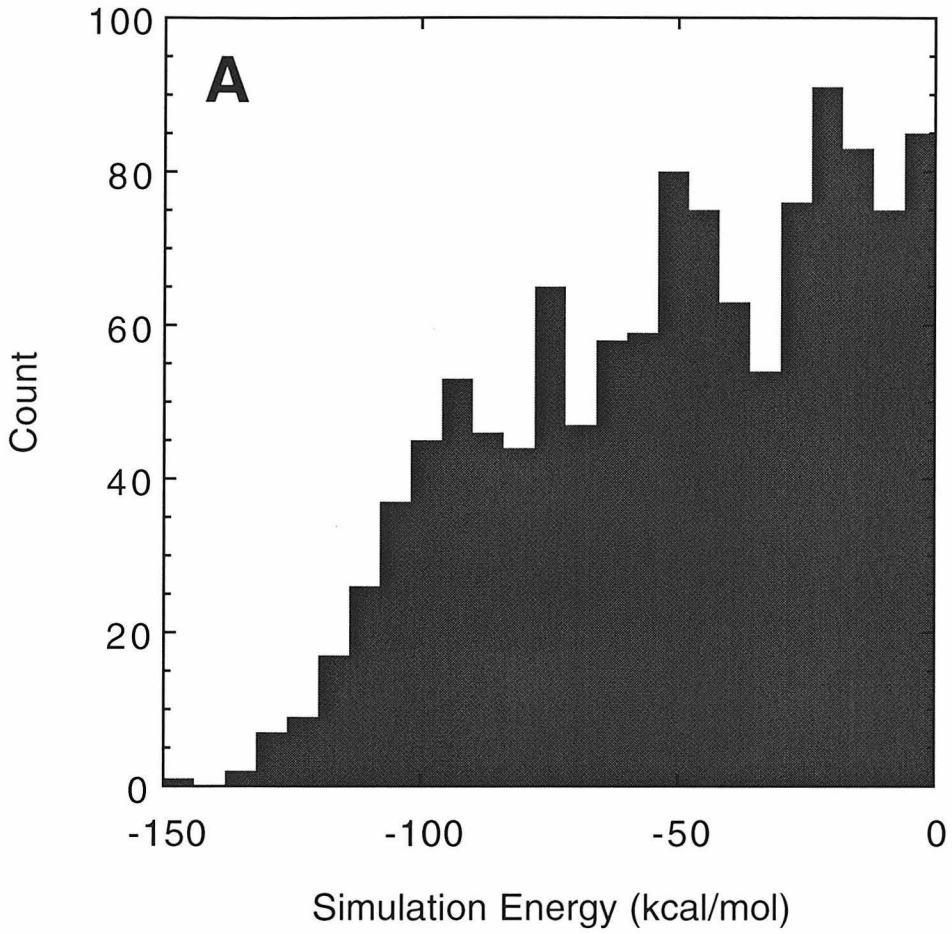
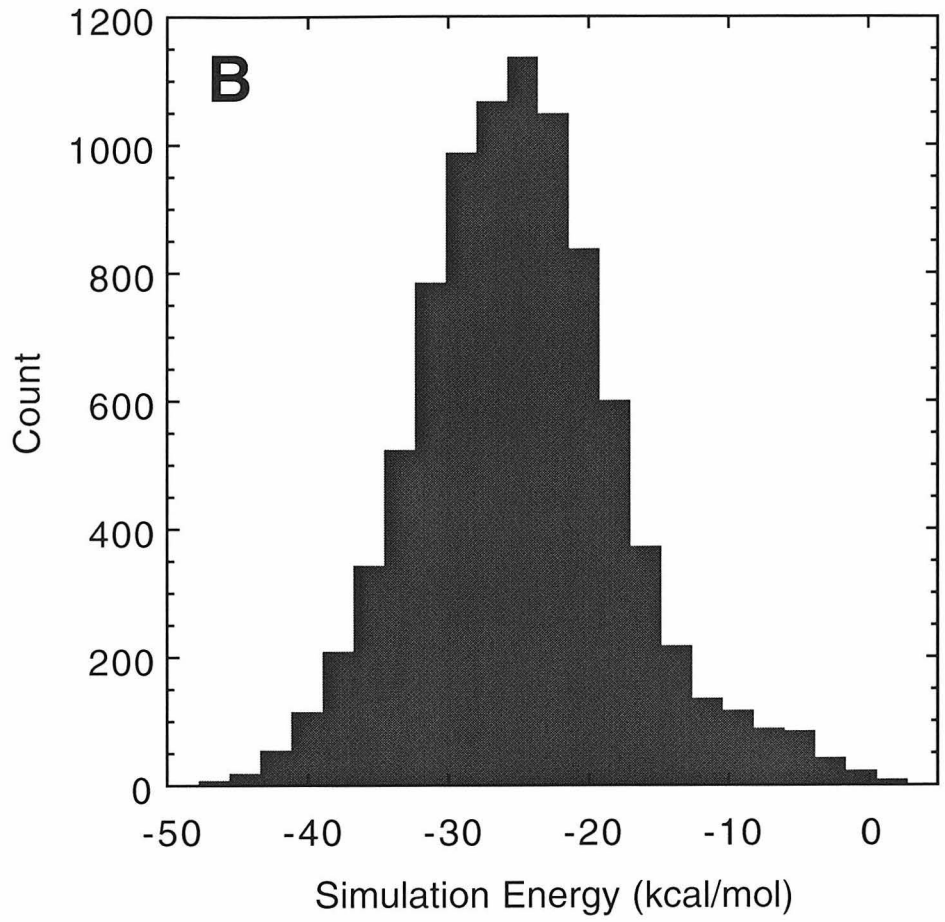

energy of sequences threaded onto target structure
mean energy
standard deviation
sequence S_0
lowest energy sequence

Figure 6-2. The actual distribution of energies of various subsets of the real protein GB1, using the potential function derived from the protein design cycle (Table 1). a) The core (only the 2.5% lowest-energy sequences are shown), b) the α-helix surface, and c) the β-sheet surface.


A
100
80
60
40
20
0
Count
-150
-100
-50
0
Simulation Energy (kcal/mol)


B
1200
1000
800
600
400
200
0
Count
-50
-40
-30
-20
-10
0
Simulation Energy (kcal/mol)

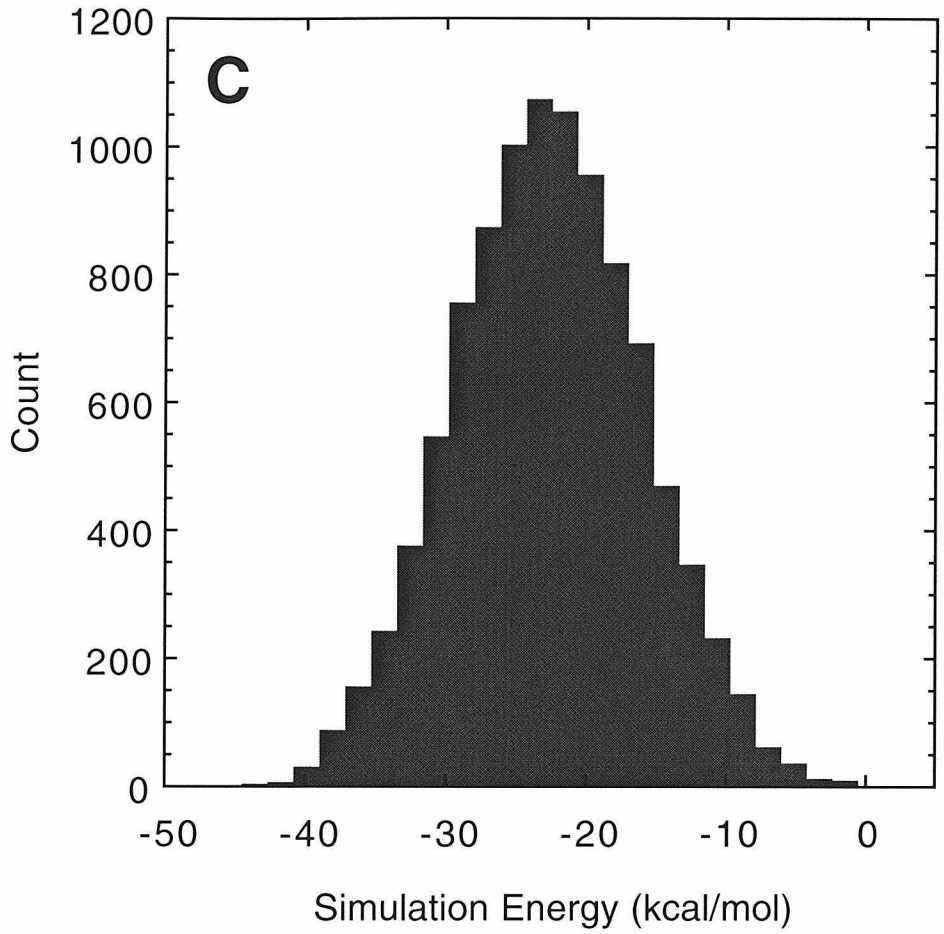
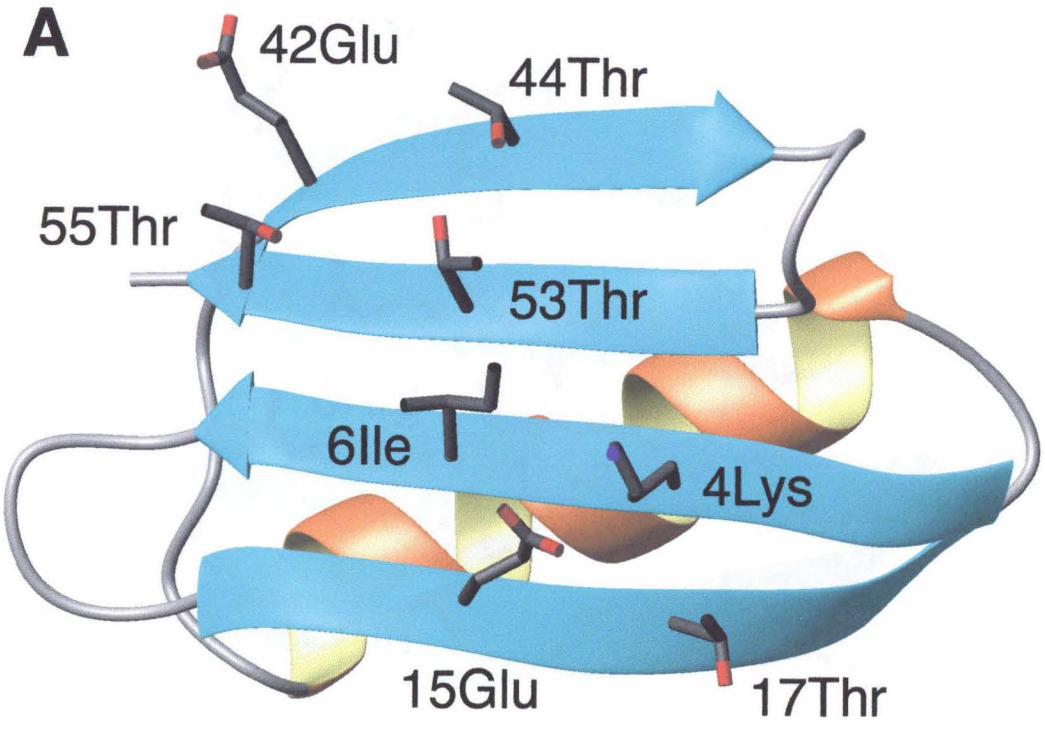

C
1200
1000
800
600
400
200
0
Count
-50
-40
-30
-20
-10
0
Simulation Energy (kcal/mol)

Figure 6-3. Views of the eight designed positions on the β-sheet surface of GB1. a) The crystographically-determined wildtype side chain orientations, and b) the orientations modeled using the *Z*-score-derived potential function.


A
42Glu
44Thr
55Thr
53Thr
6Ile
4Lys
15Glu
17Thr

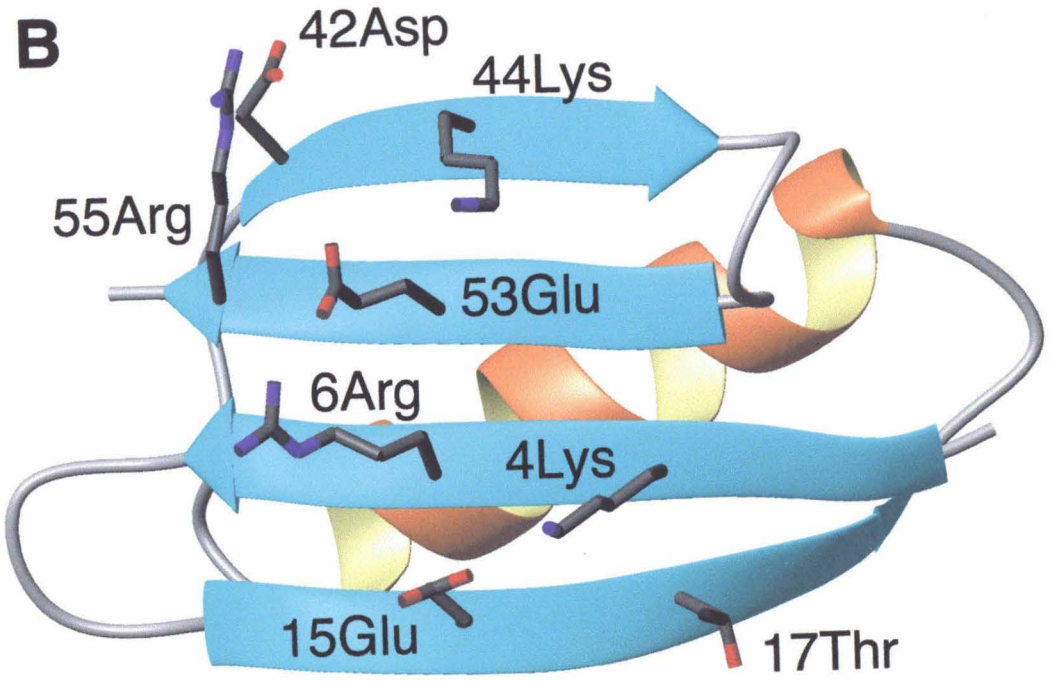
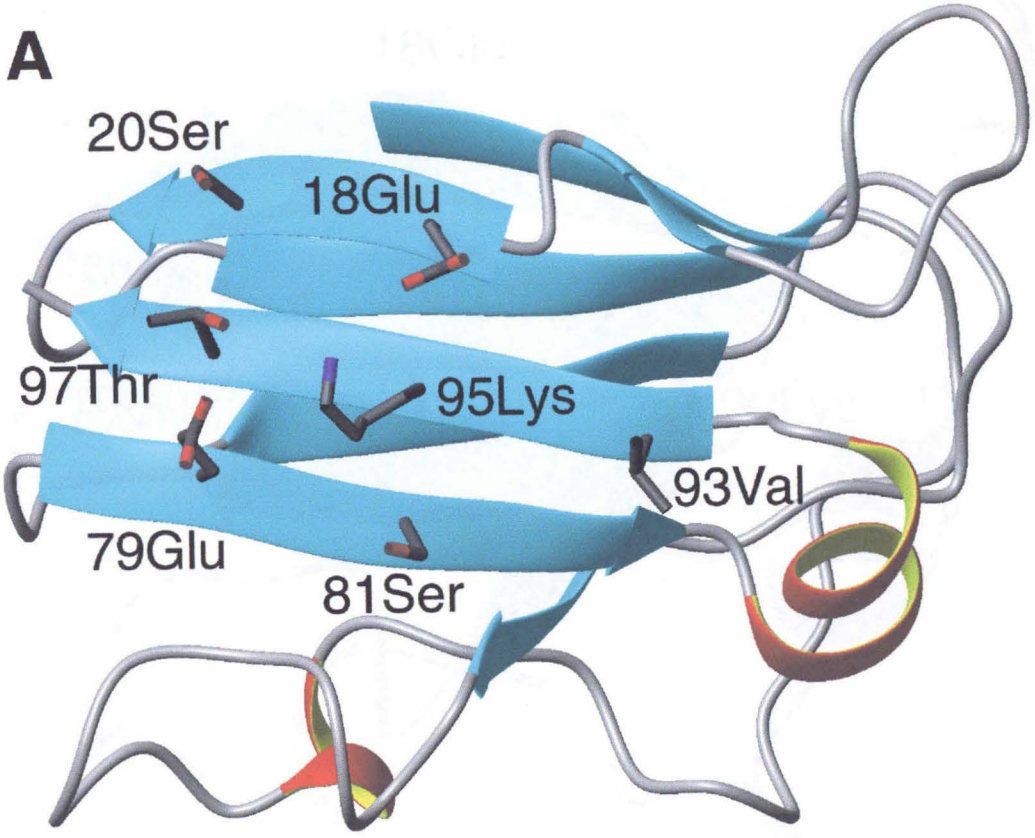

B
42Asp
44Lys
55Arg
53Glu
6Arg
4Lys
15Glu
17Thr

Figure 6-4. Views of the seven designed positions on the β-sheet surface of PCV. a) The crystographically-determined wildtype side chain orientations, and b) the orientations modeled using the *Z*-score-derived potential function.


A
20Ser
18Glu
97Thr
95Lys
93Val
79Glu
81Ser

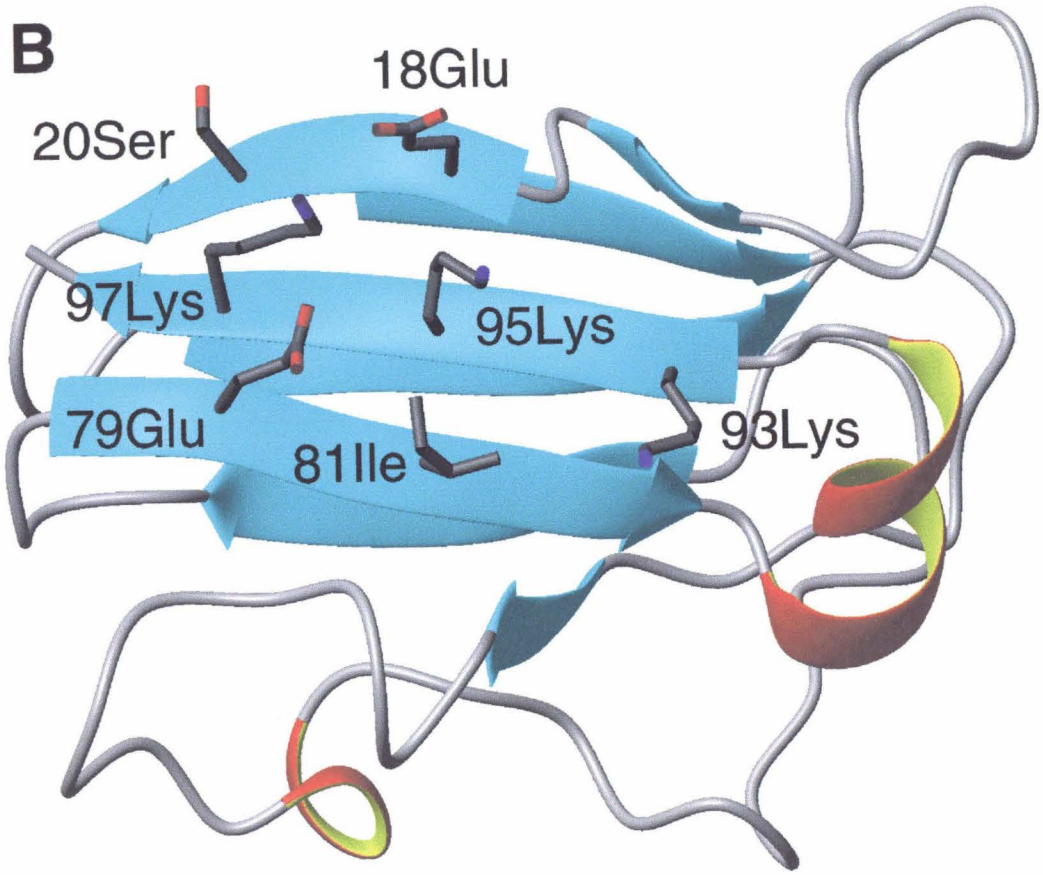

B
18Glu
20Ser
97Lys
95Lys
79Glu
81Ile
93Lys

Figure 6-5. Circular dichroism measurements of GB1 (open circles) and GB1-Z1 (solid circles) with temperature at 218 nm. Their melting temperatures are respectively 86 °C and 71 °C.

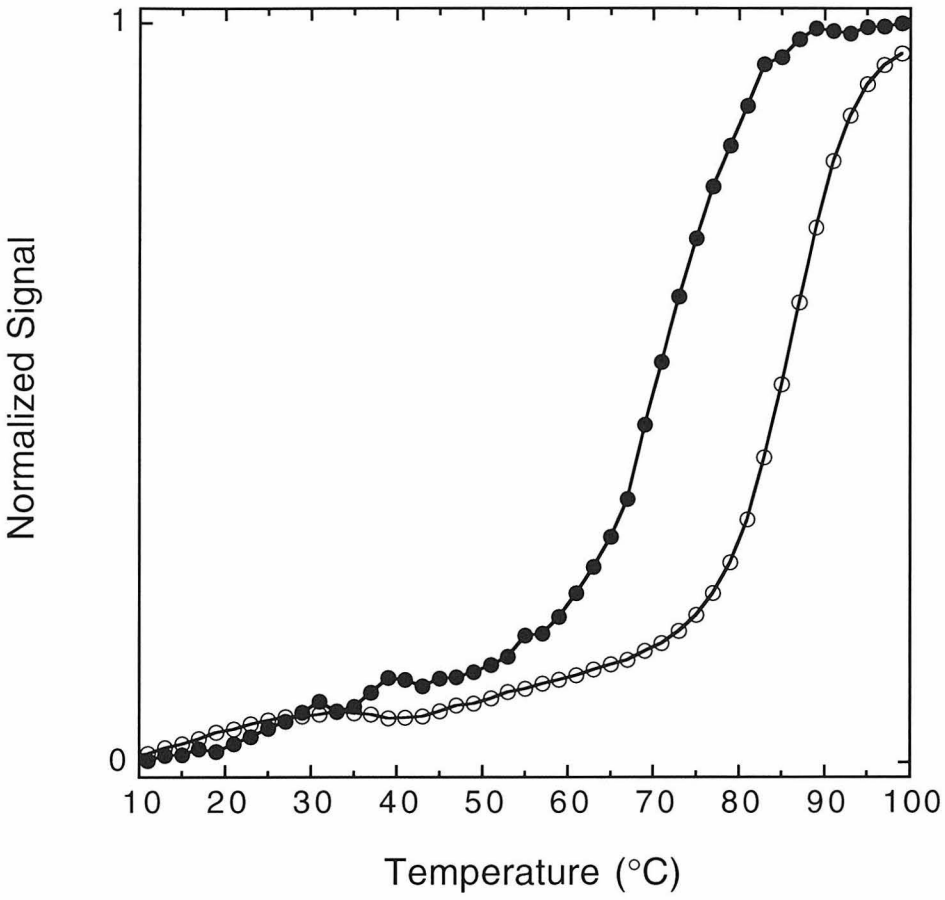
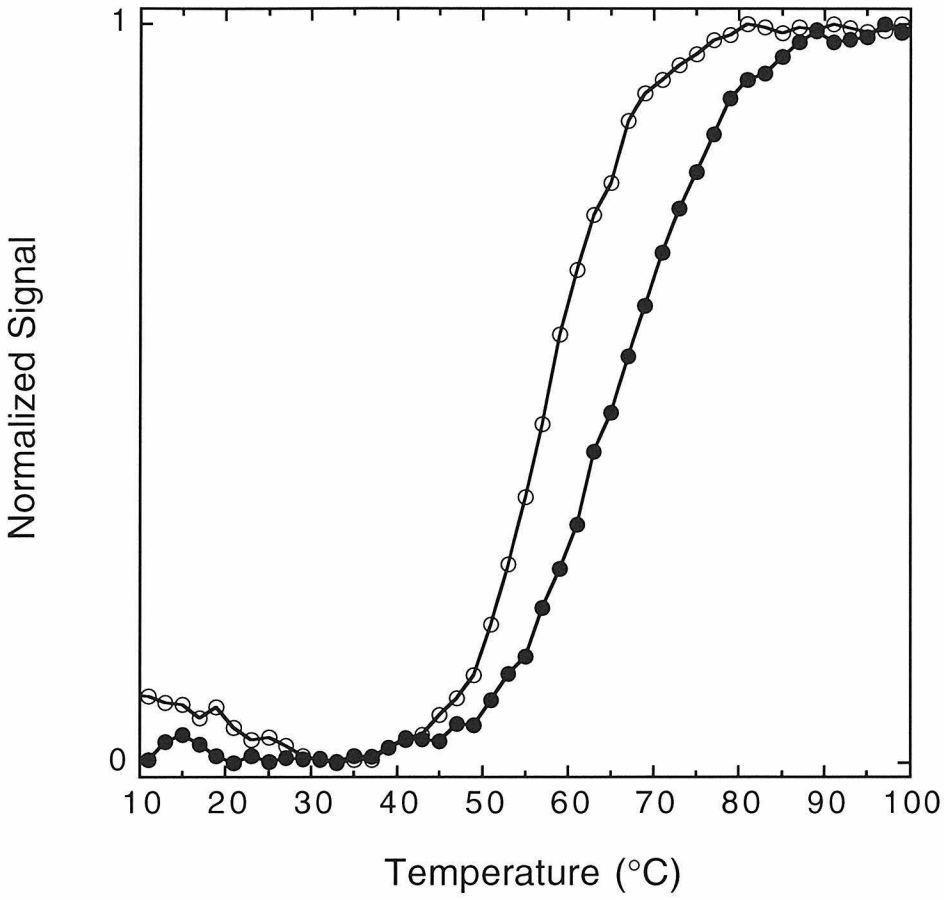

1
0
Normalized Signal
10
20
30
40
50
60
70
80
90
100
Temperature (°C)

Figure 6-6. Circular dichroism measurements of PCV (open circles) and PCV-Z1 (solid circles) with temperature at 210 nm. Their melting temperatures are respectively 56 °C and 64 °C.


1
0
Normalized Signal
10
20
30
40
50
60
70
80
90
100
Temperature (°C)

Appendix A.

Calculations for Rational Design of a Catalytic Antibody

Introduction

A significant goal of protein design is the design of novel enzymes. Although there are notable examples of rational design leading to fully functional proteins with improved thermostability (Malakauskas & Mayo, 1998), improving or redesigning protein function is more difficult. Here we discuss an attempt to use our protein design method to improve the binding strength of an enzyme.

We chose to study the catalytic antibody AZ-28 (Ulrich et al., 1997). Catalytic antibodies are artificial enzymes which can be created by injecting an animal with an antigen containing a transition state analog (that is, a stable molecule which resembles the hypothesized transition state in a chemical reaction – hapten **1** of Figure 1 in the case of AZ-28). Antibodies are created in the animal which bind to the antigen, and may be purified. These antibodies should then catalyze the desired reaction (substrate **2** to product **3** in this case) by forcing reactants into the transition state, and thus over the activation barrier to the reaction. The process can result in effective novel catalysts (Schultz & Lerner, 1995). It is our hope that by applying the techniques of rational protein design to this essentially evolutionary approach, further improvements in catalysis can be made.

There are two versions of AZ-28, since the immune system refines its response over time – the initial "germline" antibody is first on the scene, and later gives rise to "mature" antibodies with small numbers of mutations from

the germline that bind more efficiently to the antigen. One would expect that the mature antibody would be the more efficient catalyst, but in the case of AZ-28 the germline has a 35-fold greater rate enhancement over the mature antibody (a turnover rate, k_{cat}, of 0.80 min^{-1} versus 0.023 min^{-1} – for comparison, natural enzymes are usually 10^2 to 10^6 sec^{-1}). The reason for this requires an understanding of the structure of the antibody, which is available for the mature antibody only (Ulrich, et al., 1997), and shown in Figure 2.

The antibody AZ-28 catalyzes the oxy-Cope rearrangement of substrate **2** to product **3**. The reaction proceeds fastest when the central cyclohexyl ring is coplanar with the flanking phenyl substituents, so that π-orbital overlap is maximized. The structure of the mature antibody reveals, however, that the cyclohexyl ring is bound at right angles to the phenyl rings. An examination of the structure indicates that L34 Asn (position 34 of the light chain) is largely responsible for holding the cyclohexyl ring in its conformation. In the germline antibody, this residue is L34 Ser . This suggests that in the germline antibody, the cyclohexyl ring is free to rotate – hence its increased activity (Ulrich, et al., 1997).

The considerations above suggest a useful role for rational protein design. Computationally, one is not limited by the necessity to create stable analogs of the transition state. For example, we can require in our modeling that the phenyl and cyclohexyl rings remain coplanar – a requirement that is difficult to achieve experimentally.

Results and Discussion

In order to cause minimal disruption to the binding site, we modeled the hapten in its crystallographically determined structure, with the deepest

bound 5-phenyl ring rotated through 180°. At an 80° rotation, the π-orbital overlap between the 5-phenyl and cyclohexyl rings is greatest.

We identified those hydrophobic residues within 7.1 Å of the 5-phenyl ring, and additionally H35 Glu (a polar residue), which sterically clashes with the rotated 5-phenyl substituent. These are not positions which were mutated in the original analyses (Ulrich, et al., 1997). All positions were constrained to remain hydrophobic (except for H35 Glu). A detailed rotamer library was used, as discussed in Materials and Methods. A hydrophobic burial benefit was included, which has been shown to improve designability (Dahiyat & Mayo, 1996) and which should increase the amount of hydrophobic packing around the hapten.

The resulting sequences are shown in Table 1. There are some concerted changes from the wildtype sequence. In particular, H103 Trp and H37 Val are replaced with H103 Phe and H37 Ile for an improvement in simulation energy of 2.5 kcal/mol.

As discussed earlier for the wildtype sequence, the zero degree sequence cannot accommodate the rotated 5-phenyl. The +80° sequence, however, has a simulation energy of -262 kcal/mol when it is combined with the unrotated 5-phenyl. This is actually better than its energy with the 80°-rotated hapten, because there have been no mutations from small side chains to large side chains to hold the 5-phenyl in its new conformation. The +80° sequence may therefore prefer the hapten conformation found in the crystal structure, but should at least be flexible enough to allow the 5-phenyl to rotate. With a fixed backbone, this is the best result possible without leading to destabilizing steric clashes. It does point to the desirability, however, of including some backbone flexibility in future calculations.

Whether ignoring backbone flexibility in rational protein design can nevertheless lead to some improvements in catalysis must of course await experimental verification. The sequences in Table 1 were sent to the Schultz group in July 1998 and the resulting molecules' stabilities were not available at the time of writing.

Materials and Methods

The program BIOGRAF (Molecular Simulations Incorporated, San Diego, California) was used to generate explicit hydrogens on the structure of the mature antibody AZ-28 provided by the Schultz group. This was conjugate gradient minimized for 50 steps using the DREIDING force field (Mayo et al., 1990). All atoms except those of the side chains in question were held fixed for subsequent DEE calculations. Two crystallographically observed waters in the binding site were left in place during the minimization and subsequent calculations.

A Lennard-Jones 6-12 potential was used for van der Waals interactions with atomic radii scaled by 90% (Dahiyat & Mayo, 1997). The Lee and Richards definition of solvent-accessible surface area (Lee & Richards, 1971) was used, areas being calculated with the Connolly algorithm (Connolly, 1983). Buried areas were calculated as previously described (Street & Mayo, 1998). An atomic solvation parameter of $\sigma_{np} = 48$ cal/mol/Å^2 was used to favor hydrophobic burial (Street & Mayo, 1998). We include a hydrogen-bonding and electrostatics potential (Dahiyat et al., 1997).

As in our previous work (Dahiyat et al., 1997), a backbone-dependent rotamer library was used (Dunbrack & Karplus, 1993). The χ_1 and χ_2 angles of all rotamers were expanded ±1 standard deviation about the mean value. DEE optimization followed previously published methods (Dahiyat & Mayo, 1996).

Calculations were performed on a 12 processor R10000-based Silicon Graphics Power Challenge.

The resulting structure files, which were sent to Alex Varvak of the Schultz group, use our numbering scheme, which differs from the original numbering scheme because it increases from 1 monotonically, continuing where it left off from one chain to the next. The light chain therefore has the same numbering in both schemes, but the heavy chain is quite different, particularly because of residues such as 100a, 100b, 100c in the original. Both schemes are shown in Figure 3.

References

Connolly ML. 1983. Solvent accessible surfaces of proteins and nucleic acids. *Science 221*: 709-713.

Dahiyat BI, Gordon DB, Mayo SL. 1997. Automated design of the surface positions of protein helices. *Prot Sci 6*: 1333-1337.

Dahiyat BI, Mayo SL. 1996. Protein design automation. *Prot Sci 5*: 895-903.

Dahiyat BI, Mayo SL. 1997. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci USA 94*: 10172-10177.

Dunbrack RL, Karplus M. 1993. Backbone dependent rotamer library for proteins - an application to side-chain prediction. *J Mol Biol 230*: 543-574.

Lee B, Richards FM. 1971. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol 55*: 379-400.

Malakauskas SM, Mayo SL. 1998. Design, structure and stability of a hyperthermophilic protein variant. *Nature Struct Biol 5*: 470-475.

Mayo SL, Olafson BD, Goddard WA, III. 1990. Dreiding - a generic force-field for molecular simulations. *J Phys Chem 94*: 8897-8909.

Schultz PG, Lerner RA. 1995. From molecular diversity to catalysis: lessons from the immune system. *Science 269*: 1835-1842.

Street AG, Mayo SL. 1998. Pairwise calculation of protein solvent accessible surface areas. *Folding and Design 3*: 253-258.

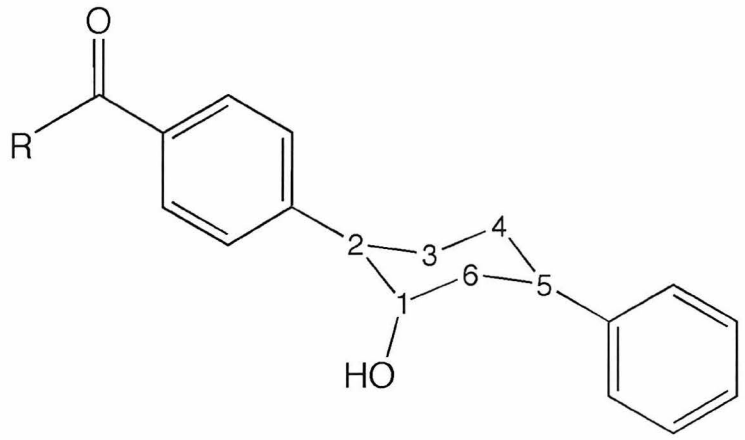
Ulrich HD, Mundorff E, Santarsiero BD, Driggers EM, Stevens RC, Schultz PG. 1997. The interplay between binding energy and catalysis in the evolution of a catalytic antibody. *Nature 389*: 271-274.

Table A-1. Designed sequences for the mature AZ-28 antibody. Different sequences are predicted for different orientations of the deepest bound 5-phenyl ring. Vertical bars indicate there is no change from wildtype. Simulation energies (kcal/mol) are also shown for each sequence. The wildtype energy (*) is not directly comparable, however, since it is calculated by applying the energy expression directly to the minimized crystal structure, rather than by forcing the wildtype side chains to their nearest available rotamers.

5-phenyl	L	L	L	L	L	H	H	H	H	H	H	H	Energy
rotation	36	89	91	96	98	35	37	45	47	91	93	103	
(wildtype)	F	L	Y	Y	F	E	V	L	W	Y	A	W	-308*
0	|	|	F	L	W	|	I	|	F	|	|	F	-267
+20	|	V	F	L	W	|	I	|	F	|	|	F	-266
+40	|	V	F	L	W	A	I	|	F	|	|	F	-260
+60	|	V	F	L	W	A	I	|	F	|	|	F	-259
+80	|	V	F	L	W	A	I	|	F	|	|	F	-250
-80	L	|	F	L	|	A	I	|	F	|	|	F	-237
-60	L	|	F	L	|	A	I	|	F	|	|	F	-251
-40	|	|	F	L	W	N	I	|	F	|	|	F	-261
-20	|	|	F	L	W	|	I	|	F	|	|	F	-265

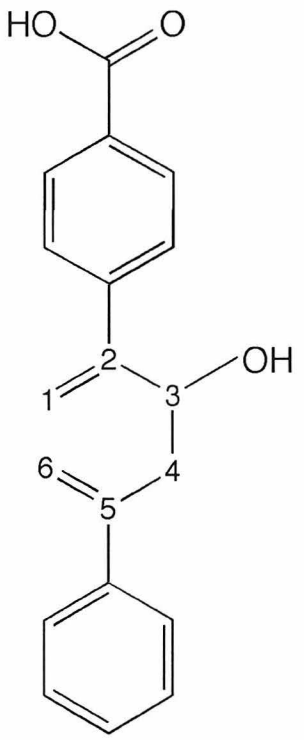
Figure A-1. Transition-state analog (**1**) and the reaction catalyzed by antibody AZ-28 (**2**, **3**). Figure adapted from (Ulrich, et al., 1997).

1



R = $NH(CH_2)_2O(CH_2)_2NH_2$

2



3

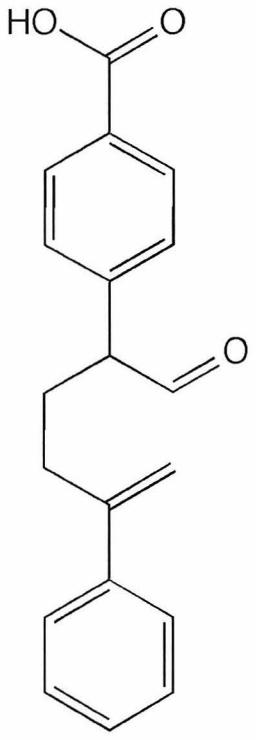


Figure A-2. The structure of the antigen-binding fragment (F_{ab}) of the AZ-28 mature antibody. The light chain is shown in yellow, the heavy chain in orange. The hapten **1** (excluding the group R) is shown in its binding site. The cyclohexyl ring is seen to be at right angles to the flanking phenyl rings.

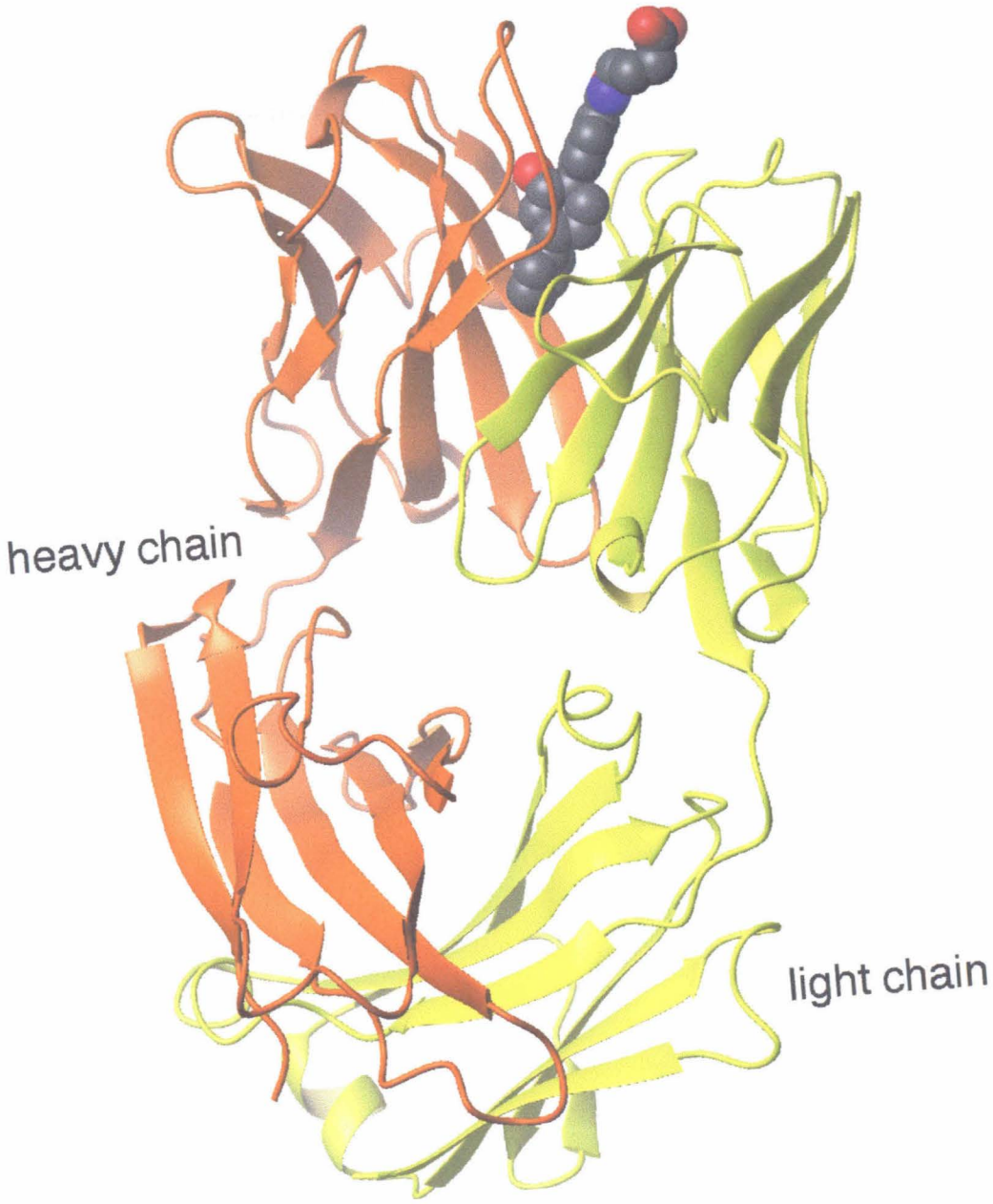
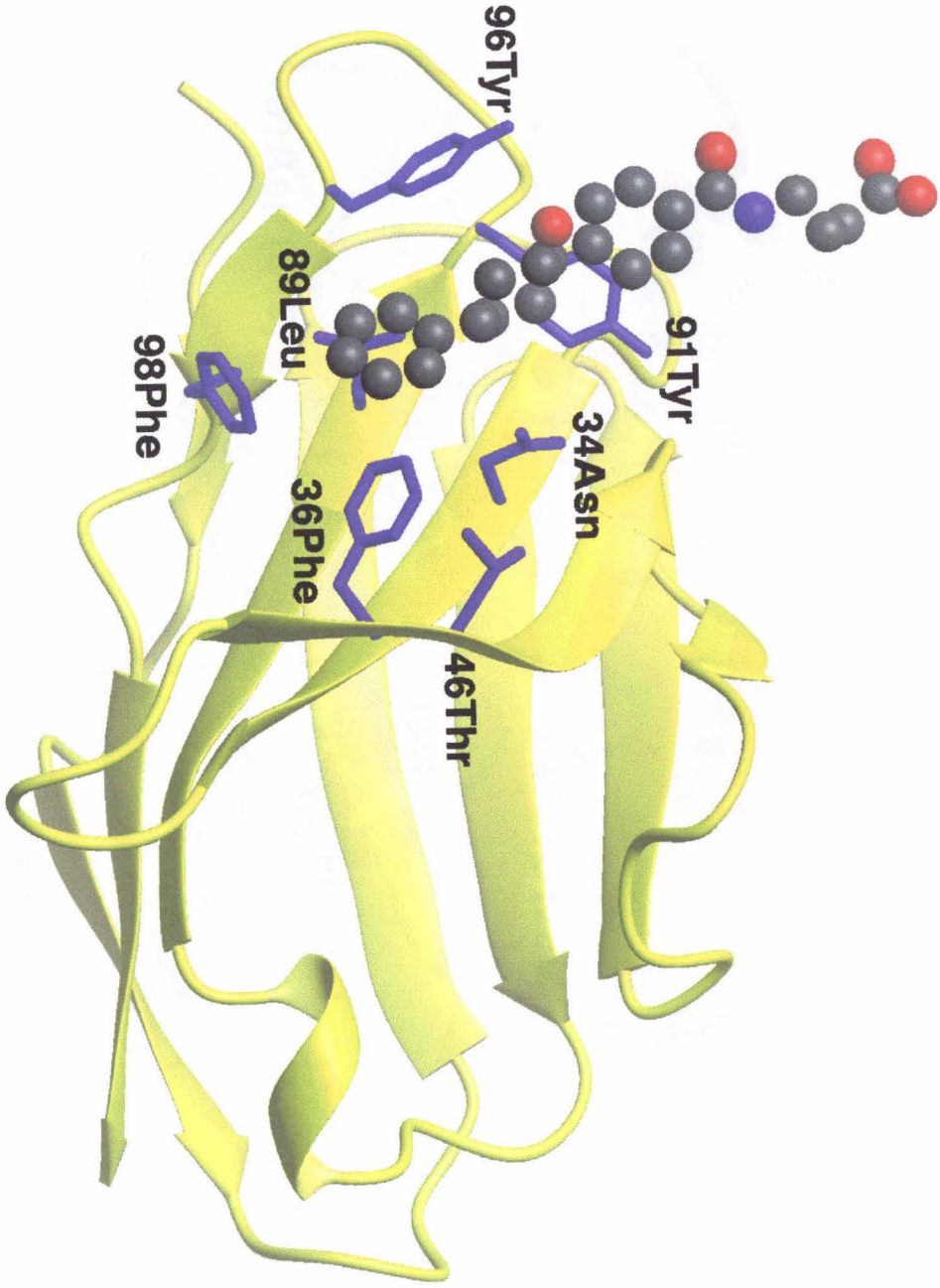
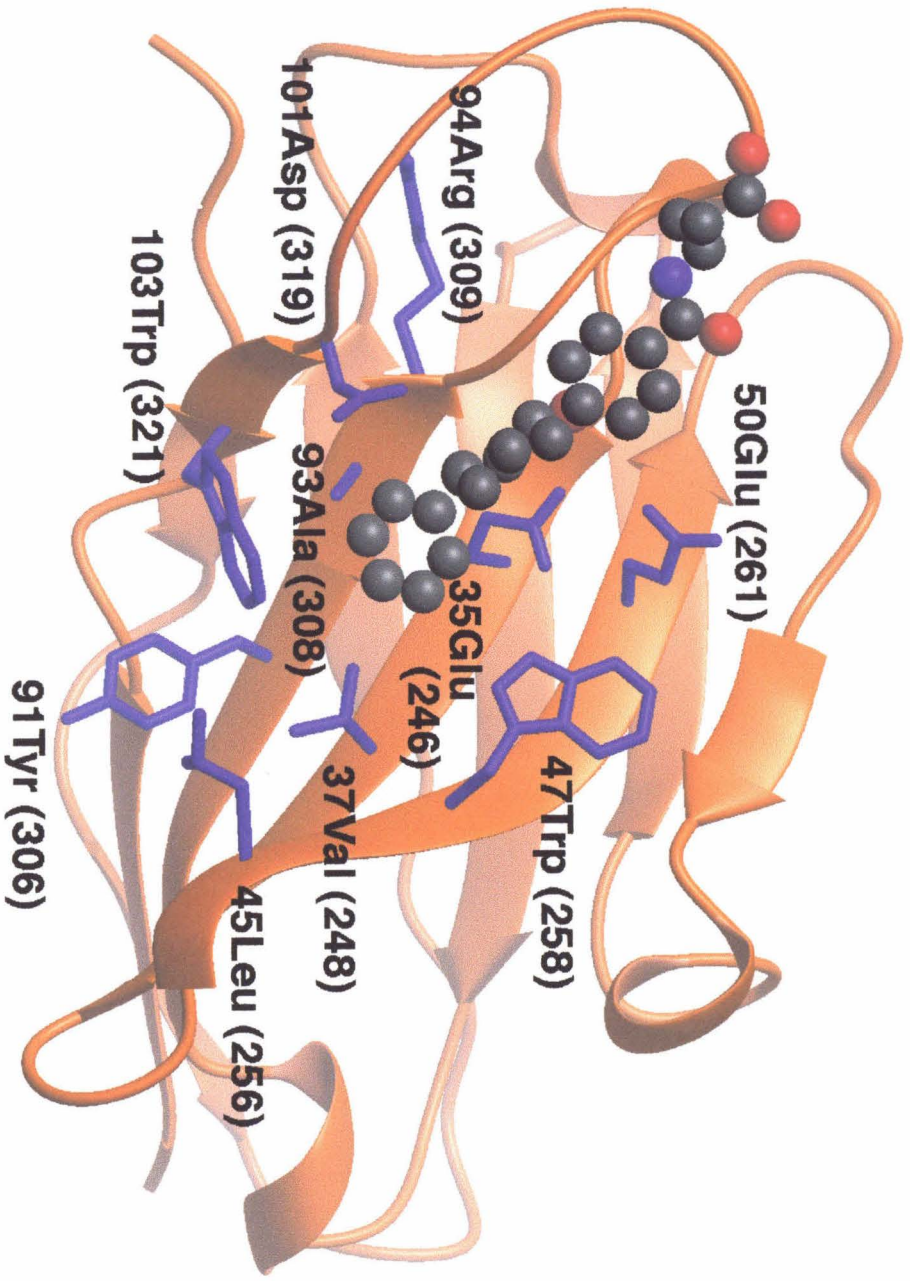

heavy chain
light chain

Figure A-3. Views of the AZ-28 mature antibody binding site. The displayed side-chains are those close to the 5-phenyl group, a) with the heavy chain removed from view, b) with the light chain removed from view. The original and our sequential numbering schemes are both shown.


96Tyr
89Leu
98Phe
91Tyr
34Asn
36Phe
46Thr


50Glu (261)
94Arg (309)
101Asp (319)
103Trp (321)
93Ala (308)
35Glu (246)
47Trp (258)
37Val (248)
45Leu (256)
91Tyr (306)