FORCE FIELD DEVELOPMENT IN PROTEIN DESIGN

Thesis by

Eric Stafford Zollars

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2006

(Defended May 25, 2006)

Acknowledgements

I have received a great deal of advice, encouragement, and assistance throughout my years at Caltech. I would first like to thank my advisor, Steve Mayo, for his guidance. I am continually amazed by his breadth and depth of knowledge concerning all things protein. Additionally, my committee members Niles Pierce, Pamela Bjorkman, and Doug Rees were excellent resources with whom I should have consulted more often.

My tenure in the Mayo lab saw a great deal of change: different buildings, different projects, and most importantly different people. I have seen nearly a complete turnover of postdocs and graduate students and the lab has always been a great environment in which to work and think. I would especially like to thank Kyle Lassila and Ben Allen not only for their intellectual contributions to my scientific development but also for their friendship and encouragement.

My family deserves credit for always supporting my academic goals and pushing me into science from an early age. It was certainly not easy to drive an hour each way to take a third grader to extra Russian language classes.

And finally, Jaime, my wife, for truly making me a better person. Her optimistic view on life gives me hope and her ambition gives me focus.

Abstract

Protein design requires the rapid evaluation of very large numbers of equations during the course of a calculation. These equations must represent the important contributors to protein stability in simple and accurate terms. Some physical phenomena are relatively easy to model such as van der Waals forces. Electrostatics and solvation in a protein environment are forces that are more difficult to adequately capture. Additionally, the balance of the terms used must be determined in order to design sequences that fold to stable, specific folds.

The electrostatic interactions within the protein and between the protein and solvent are important in both the stability and function of the protein. The effects of the protein-solvent interactions are evaluated using implicit models that consider the solvent as a bulk. These interactions are quantified using the Poisson-Boltzmann equation that must be solved using discrete numerical methods. We sought to avoid this performance hit by scaling a simpler model of electrostatics, Coulomb's law, to reproduce one aspect of the protein-solvent interaction: solvent screening. By dividing the Coulombic dielectric into two parts and scaling to correlate with the Poisson-Boltzmann results we significantly increased the strength of electrostatics in our force field that led to the design of a more stable engrailed homeodomain.

The second part of this work describes attempts to reparameterize our protein design force field. Many protein mutants have been expressed and biophysically characterized in the literature. We sought to use the measured stabilities of protein mutants in the literature to balance the terms in the force field. While we were able to produce a force field that could reproduce experimental energies, this force field led to unsatisfactory designed sequences. To more fully satisfy the unique conditions of a protein design force field we explored other

optimization techniques and found that the balance of the terms in the existing force field is nearly optimal.

Contents

1. Introduction

1.1. Background

Protein design is the process of creating new protein molecules with desired properties. Since the development of the molecular biological techniques that allowed for the easy mutation of protein sequences [1] proteins have been manipulated to explore sequence-stability relationships [2] as well as change their function. However, the complexity of the interactions within proteins makes the process of choosing positions to mutate a daunting task. While "rational" protein design can be accomplished by looking at a protein structure and deciding what mutations to make this becomes rapidly more difficult with an increasing number of mutations. The number of possible configurations increases exponentially with an increasing number of mutated positions. Assuming only the 20 natural amino acids at each of 50 positions (a very small protein) the number of possible combinations is 10^{65}. Thus, for all except the smallest of problems (1–2 interacting mutations) design-by-eye will be too difficult.

Methods designed to approach the complexity of large protein designs can be roughly grouped into evolutionary and computational methods. Evolutionary methods use the natural processes of mutation and recombination combined with an appropriate selection to progressively achieve the desired properties. Neither the structure of the molecule of interest or the explicit mechanism of the reaction considered must be known. Success of these methods include the stabilization of enzymes towards temperature or solvent [3], reversal of enantioselectivity [4], and creation of new function in enzymes [5]. Two main limitations exist for the evolutionary methods. One is that an experimental selection for the property of interest must exist so that each round of mutation/recombination can be enriched for the

property of interest. Additionally, only a small amount of all possible sequence combinations can be explored experimentally.

Computational protein design is the focus of the Mayo lab. We use computational methods to select a protein sequence of interest among the massive number of possible combinations. Thus, it is only limited by the amount of processing power available to be devoted to the problem. Computational protein design demands a great understanding of the systems to be designed.

1.2. History

The concept of protein design was first proposed by Eric Drexler in 1981 [6] as an extension of Richard Feynman's famous paper "There's Plenty of Room at the Bottom." While Feynman predicted the field currently known as nanotechnology, Drexler presented protein design as a way to construct atomic-scale machines. His vision of protein design leapfrogs Feynman's idea of small machines building smaller machines by using the already existing cellular machinery of protein synthesis. Especially prescient was his claim that attempts at protein design can occur before achievement of a full understanding of the nature of protein folding.

> Engineers (in contrast to scientists) need not seek to understand all proteins but only enough to produce useful systems in a reasonable number of attempts. An engineer designing a protein that has 1000 amino acids may choose among some 10^{1300} different amino acid sequences. It might be that only one in 10^9 (or even 10^{700}) randomly selected sequences would yield a predictable conformation, yet this tiny fraction represents a vast number of proteins. Through use of strategically placed charged groups, polar groups, disulphide bonds, hydrogen bonds, and hydrophobic groups, the engineer should be able to design proteins that not only fold predictably to a stable structure (sometimes) but that serve a planned function as well. Even a low success rate will lead to an accumulation of successful designs. Thus, the difficulties encountered in predicting the conformations of natural proteins do not seem insurmountable obstacles to protein engineering.

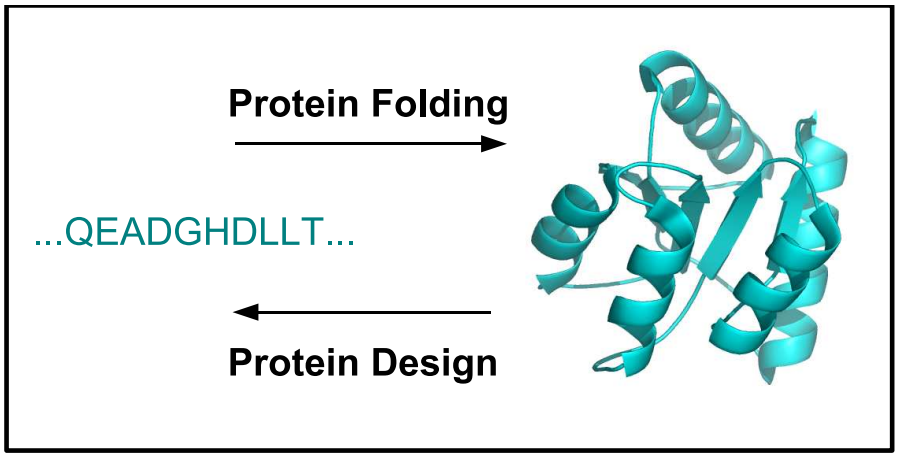


Figure 1.1. The problem of predicting a protein structure from just a sequence of amino acids is protein folding (left to right). The inverse, protein design (right to left), begins with a protein backbone structure and predicts the sequence of amino acids that would stabilize that structure.

The definition of protein design as the inverse of the protein folding problem is credited to Carl Pabo [7] (figure 1.1). By starting with a fixed protein backbone from a crystal structure it becomes unnecessary to predict the structure of any sequence of amino acids. Instead, the problem becomes determining the sequence of amino acids that are compatible with a given structure. A method that incorporated the idea of choosing sequences for a fixed protein backbone was presented by Ponder and Richards in 1987 [8]. They did not try to find the best sequence for a structure but attempted to list compatible sequences by using simple structural considerations. A number of advances were developed in this paper, most importantly the use of a rotamer library (see chapter 2) composed of 67 rotamers.

A further advance for protein design methodology originated in the related field of sidechain placement. The sidechain placement problem involves finding the correct orientation of sidechains on a known protein backbone. This can be considered a necessary component of the protein folding problem or the simplest possible protein design. While many orders of magnitude simpler than protein design there still exists a significant amount of combinatorial complexity in sidechain placement. Lee and Subbiah [9] applied a simulated annealing optimization algorithm [10, 11] to avoid the necessity of a systematic search of all possible combinations of sidechain configurations. Since this algorithm does not need to evaluate every possible solution, larger problems were able to be solved more quickly. The applica-

tion of simulated annealing to the sidechain placement problem was a significant advance as it allowed for the solution of the large combinatorial problem within a reasonable amount of time. An extension of this work led to the calculation of the energies of mutation [12] of a large number of experimentally determined mutations in lambda repressor [13]. The set of mutations were simple hydrophobic-to-hydrophobic mutations in the protein core; thus a very simple energy function was successful in qualitatively predicting the energetic effects of mutation. This showed that it was possible to predict the stability of mutations computationally, opening the door for more complex protein designs.

The complexity and heterogeneity of most protein structures makes the design of sequences of amino acids that fold to a given structure difficult. However, the existence of a simple model protein–the helix bundle–led to some early successes in protein design [14, 15]. These designs relied on the simple pattern of the helix bundle to make basic rules for design. More complicated designs soon appeared. Hellinga, Caradonna, and Richards developed procedures for designing a metal binding site in a protein [16]. Hurley et al. designed new cores for lysozyme [17] using an iterative approach similar to Ponder and Richards [8]. The first automated, *de novo* design of a complex protein core was Desjarlais and Handel [18] who designed a sequence of amino acids that repacked the entire core of the phage 434 cro protein and determined that it was well folded. Further developments included the design of a coiled-coil with some backbone flexibility [19], a heterotrimeric coiled-coil that incorporated specific electrostatics and negative design [20], and refinement of the balance of terms required for a specific, well-folded coiled-coil [21].

The first fully automated *de novo* design of an entire protein sequence was from the Mayo lab [22]. This achievement built on previous work in the Mayo lab. The Dead-End Elimination theorem, originally conceived as a rapid search algorithm for sidechain placement [23], was adapted for protein design and combined with a van der Waals potential and used to design sequences for a homodimeric coiled-coil fold [24]. Sequences designed by the van der Waals potential were poorly correlated with the experimental stabilities, but the introduction of a surface-area burial term led to increased correlation between calculated and experimental stabilities. Further work on the homodimeric coiled-coil included design of the surface positions using a hydrogen-bond potential, a penalty for burial of non-hydrogen-

bonded polar hydrogens, and a helix propensity term [25]. Quantitative conclusions on the utility of these various terms was not possible since nearly all designed sequences were of greater stability than the wildtype sequence but this increased stability demonstrated the importance of surface positions for stability. The introduction of a scaled van der Waals term [26] was designed to compensate for the lack of flexibility in the protein backbone and rotamers. Further work led to the redesign of the Streptococcal protein G beta1 domain yielding a hyperthermophilic protein with a melting temperature higher than 100 ^{O}C [27].

The ability to design sequences of amino acids that could reliably fold to the desired structure led to an interest in designing sequences that would also have specific functions such as binding or catalysis. While grafting metal-binding sites into proteins [16, 28] had some early success and antibodies can be generated with catalytic activity [29], converting an otherwise inactive protein into an active enzyme is a daunting task. Bolon and Mayo designed nucleophilic hydrolysis activity into thioredoxin [30]. The activity, while low, demonstrated kinetics that were suggestive of an enzymatic mechanism. Impressive work followed from the Hellinga lab with designed binding proteins that functioned as biosensors [31] and metabolic enzymes [32]. An exciting combination of techniques from protein design and protein folding led to the design of a sequence that folded to a structure previously not observed in nature [33]. These successes show that protein design continues to aid in the solution of more complicated problems and points to a future as one of the major tools in science and biotechnology.

1.3. Force field

1.3.1. Folding

The computational protein design software written and used in the Mayo lab is ORBIT (Optimization of Rotamers By Iterative Techniques). A principle component of ORBIT is the force field. A force field in the context of protein design is a collection of mathematical terms that together lead to a total calculated energy for a protein or part of a protein.

$$E_{Total} = E_{vanderWaals} + E_{nonpolar} + E_{hydrogenbonds} + \bullet\bullet\bullet$$

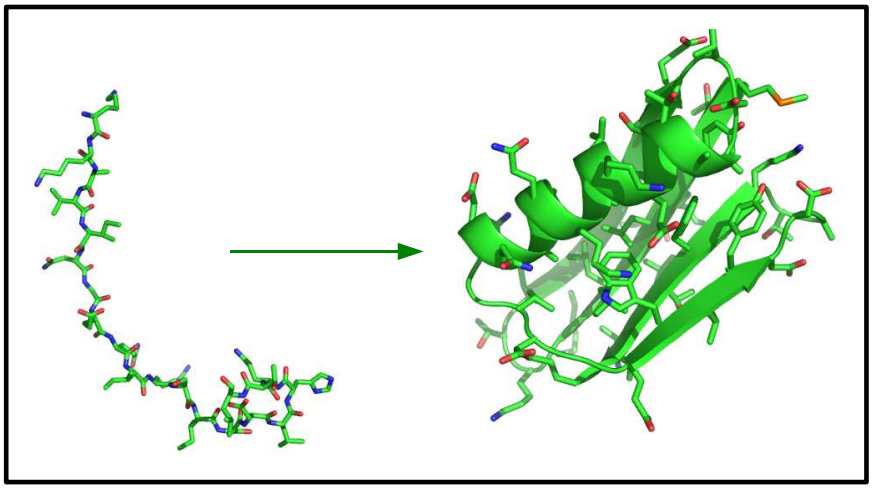


Figure 1.2. Protein folding. In reality many different conformations of the unfolded state collapse to give the folded conformation. The stability of the folded protein relative to the unfolded ensemble is ΔG_f.

In protein design, interactions are stabilizing or destabilizing relative to the folded state of the protein. When a protein folds it collapses from an ensemble of noncompact states to the native state that is significantly more compact (figure 1.2). The process of folding excludes many water molecules and the nonpolar amino acids bury themselves in the core of the protein away from the water environment. The stability of a protein is measured as a free energy of folding, ΔG_f, and thus includes both enthalpic and entropic components. Folded proteins are only marginally stable [34] and this marginal stability results from large opposing forces. The hydrophobic effect is thought to be the primary driving force of folding [35] with the loss of conformational entropy the opposing unfolding force. It is a delicate balancing act to appropriately represent these forces and accurately calculate the energy of a protein.

1.3.2. Pairwise calculations

Even with a complete understanding of the physical interactions within proteins the ability to represent these interactions mathematically would remain a problem. In addition, the massive combinatorial complexity of computational protein design precludes the calculation of large numbers of multibody interactions. The ideal, and potentially most accurate, calculation of energy would simultaneously include all of the atoms of the protein including

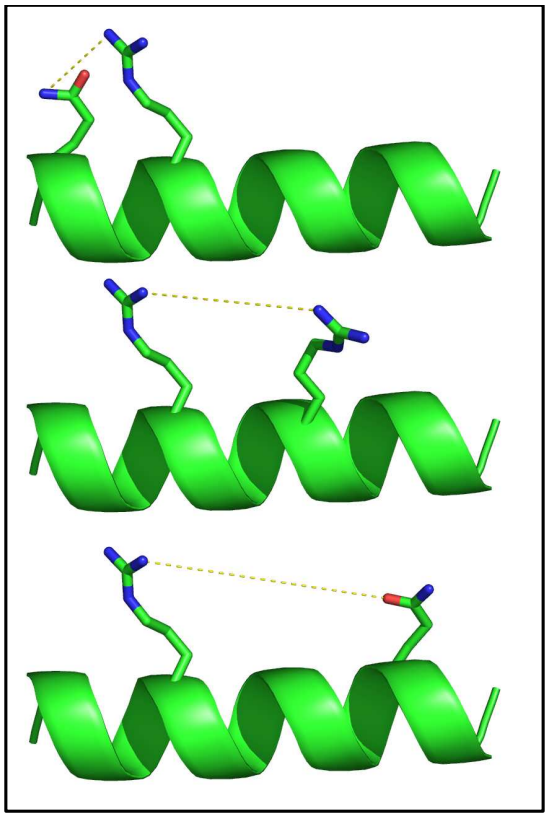


Figure 1.3. Pairwise interactions. Every interaction between pairs is calculated, ignoring the other residues present.

surrounding waters. Since this is computationally intractable, computational protein design must make simplifications and assumptions. The main simplification currently is the use of pairwise interaction energies (figure 1.3). The interaction energies between all pairs of residues are calculated (in the absence of all other designed positions) and summed to yield the overall energy of the protein. The validity and accuracy of this simplification have been evaluated and questioned [36, 37] but it remains a necessity at this point.

1.3.3. Surface area solvation

The hydrophobic amino acids tend to segregate to the protein core where they are not exposed to the water environment. Experiments with small molecules have shown that molecules with more hydrophobic surface area have more energetic benefit to being buried away from water. The surface area measured is the solvent accessible surface area (ASA)[38].

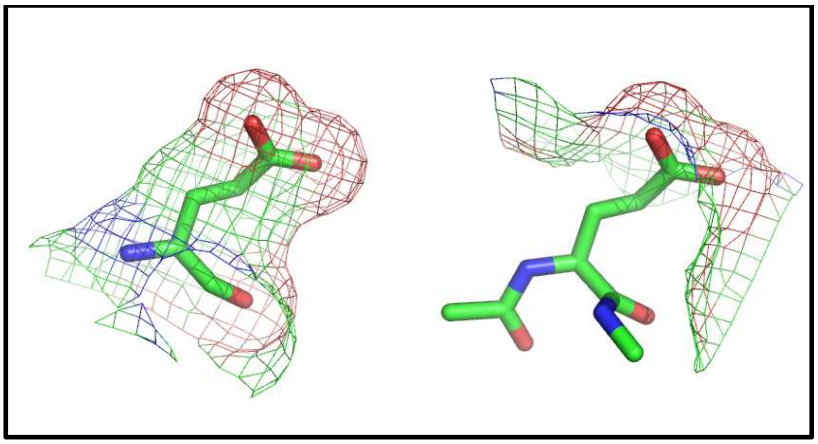


Figure 1.4. The mesh indicates the solvent accessible surface area of a hypothetical unfolded reference state (left) and the sidechain in the presence of the rest of the protein (right). The reduction in the mesh in the folded state is because some of the sidechain is no longer accessible to the solvent due to the presence of the rest of the protein.

atom	value
C	16 ± 2
N/O	-6 ± 4
O^-	-24 ± 10
N^+	-50 ± 9
S	21 ± 10

Table 1.1. Atomic solvation parameters (ASP) in cal mol^{-1} $Å^{-2}$ determined by Eisenberg and McLachlan [39]. O^- and N^+ are the charged versions of these atoms.

ASA is measured by rolling a sphere with the radius of a water molecule (1.4 Å) over the surface of the protein (figure 1.4). Quantitating the hydrophobicity of small molecules relies on measuring the free energy of transfer (ΔG_{trns}) of the small molecule from water to a more nonpolar solvent. By combining these two concepts Eisenberg and McLachlan arrived at an atomistic description of solvation in proteins with only five terms (table 1.1).

ORBIT surface area solvation reduces this to two terms, nonpolar (σ_{np}, 26 cal mol^{-1} $Å^{-2}$) and polar (σ_p, -100 cal mol^{-1} $Å^{-2}$) [36]. Due to the pairwise nature of ORBIT an additional parameter is introduced to reduce the overcounting of surface areas in the protein core [36]. The area of interest is the amount buried, i.e., the difference in area between the unfolded and folded states (figure 1.4). Other work [40] has shown that penalizing hydrophobic surface

area on the surface of proteins ($A_{np,e}$) leads to better designed sequences, leading to

$$E_{solvation} = -(\kappa + 1)\,\sigma_{np}\Delta A_{np,b} + \kappa\sigma_{np}A_{np,e} + \sigma_p\Delta A_{p,b}. \tag{1.1}$$

A represents surface area, either buried nonpolar (np, b), exposed nonpolar (np, e), or polar burial (p, b). κ is a term balancing the nonpolar burial and exposure terms.

The use of a single parameter such as ASP to describe the correlation between surface area and solvation energy assumes a linear relationship. This is strictly true only for small, linear hydrophobic molecules [41]. The correlation falls apart with polar atoms as polar solvation has more directional character. Even the interpretation of free energies of transfer become difficult with polar molecules [42]. This is because any degree of polarity of the solvent can lead to very different results when trying to obtain relative solvation scales (such as table 4.2). Such difficulties lead to more complex methods to treat polar solvation (chapter 3).

1.3.4. Hydrogen bonding

Hydrogen bonding has been known to be a significant contributor to protein structure and stability before a protein structure was ever visualized [43]. The hydrogen bond term of the force field takes into account the unique directional nature of this bonding interaction. In order to represent the geometric dependence of the hydrogen bond, angle terms are included in the energy term based on the hybridization of the donor and acceptor atoms [25].

$$E_{hbond} = D_0\left[5\left(\frac{R_0}{R}\right)^{12} - 6\left(\frac{R_0}{R}\right)^{10}\right] * F(\theta, \phi, \varphi). \tag{1.2}$$

Here R is the donor to acceptor distance, R_0 is an average hydrogen bond distance (2.8 Å), and D_0 is a well-depth term as in the van der Waals interaction. θ and ϕ represent donor-hydrogen-acceptor and hydrogen-acceptor-base angles respectively. φ is used to define the relationship between the two planes containing the donor and acceptor when both are sp^2 hybridized. This function is only used when the potential donating and accepting atoms are more than 2.6 Å and less than 3.2 Å apart, and $\phi - 109.5 < 90$ for $sp^3$to sp^3or $\phi > 90$ for $sp^3$to sp^2.

The angular functions are

$$F = \cos^2\theta\cos^2(\phi - 109.5) \quad \text{sp}^3\text{donor to sp}^3\text{acceptor}$$
$$F = \cos^2\theta\cos^2(\phi) \quad \text{sp}^3\text{donor to sp}^2\text{acceptor}$$
$$F = \cos^4\theta \quad \text{sp}^2\text{donor to sp}^3\text{acceptor}$$
$$F = \cos^2\theta\cos^2(\max[\phi,\varphi]) \quad \text{sp}^2\text{donor to sp}^2\text{acceptor}$$

1.3.5. van der Waals

The van der Waals forces are the weak forces that result from dispersion of electron clouds when non-bonded atoms are within interacting distance. The function used here is a Lennard-Jones 12-6 potential as implemented in the Dreiding force field [44].

$$E_{vdw} = D_0\left[\left(\frac{\alpha R_0}{R}\right)^{12} - 2\left(\frac{\alpha R_0}{R}\right)^{6}\right] \tag{1.3}$$

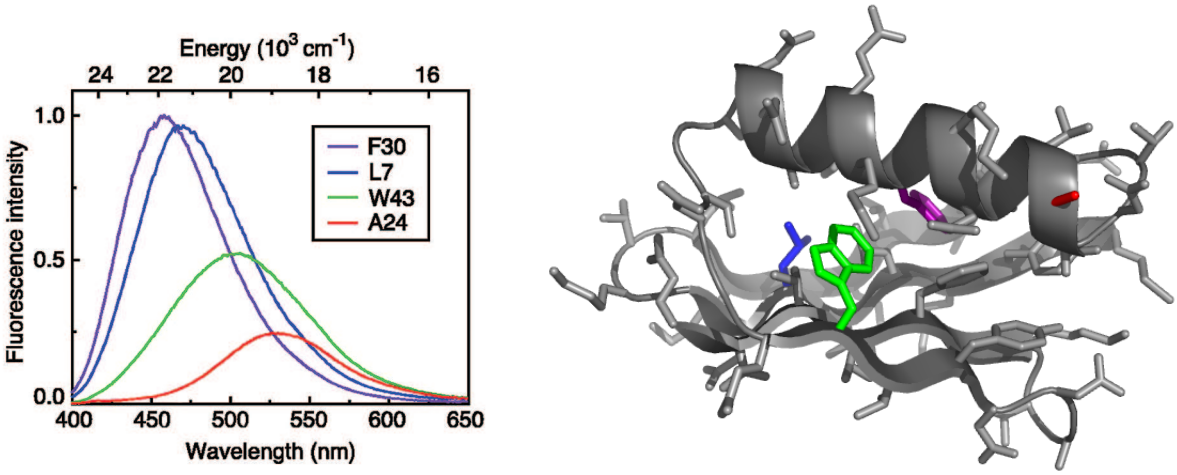
The distance between the atoms is R, R_0 is the average of the radii of the two atoms, and D_0 is an energetic weighting term that is based on the average well depth between the two atoms. α is a factor included in ORBIT to reduce the influence of steric overlapping, which, as it's twelfth power dependence demonstrates, is dominant at short distances. Since protein design necessitates the use of a fixed backbone, this factor, α, allows for some "flexibility" in sidechain packing that would normally be accomplished by minor movements in the backbone.

1.3.6. Coulombic electrostatics

Interactions between charges in the force field are represented with a Coulombic term with a distance dependent dielectric (εr).

$$E_{elec} = \frac{qq'}{(\varepsilon r)r} \tag{1.4}$$

Both charges and partial charges (i.e., the charges resulting from dipoles) are calculated using equation 1.4. Typically a value of 40 has been used for ε.



(a) (b)

Figure 1.5. (a) Different regions of polarizability in protein G as determined by Cohen et al. [46]. (b) Location of residues in protein G.

1.3.7. Occlusion

Proteins are heterogeneous polymers with irregular, collapsed native states. This heterogeneity can be observed in multiple contexts including hydrophobicity, polarizabilty (figure 1.5), and hydrogen bonding [45]. The most obvious characteristic of proteins is that amino acids tend to be more nonpolar the farther from the surface they occur. Surface area solvation is one attempt to capture this as an amino acid that has 100% buried surface area is likely to be distant from the surface. Another measure is density, as proteins also are more densely packed in the core than at the surface [47]. One way to measure this, is occlusion (also called contact-based, excluded volume) [48]. Since protein structures usually do not explicitly and accurately show the surrounding water molecules, atoms at the surfaces of proteins appear to be surrounded by nothing. Thus, an atom that is away from the surface is likely to be surrounded by more atoms and more occluded (figure 1.6).

The advantage of calculating occlusion is that it is very rapid, significantly faster than calculating surface area. At the simplest level, the number of atoms within a fixed distance can be counted. This does not take into consideration the size of the nearby atoms or the

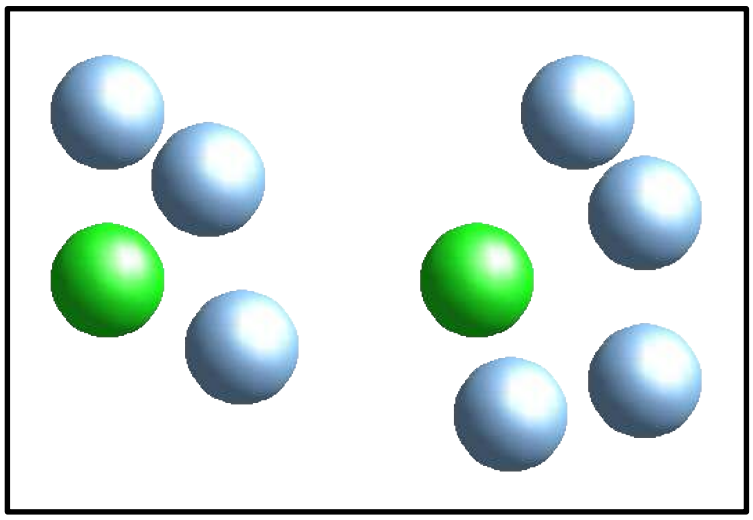


Figure 1.6. Representation of occlusion. The green atom on the right is more occluded by its neighbors than the green atom on the left.

distance from them. Including volume and distance into a description of occlusion gives

$$occ(i) = \sum V_j env(r). \tag{1.5}$$

This states that for an atom, i, sum over the volumes of nearby atoms, j, weighted by a envelope function that considers distance, r. A linear step function was used as the envelope function by Holm and Sander [49] to approximate the first solvent shell around an atom. A Gaussian form is used in Lazaridis and Karplus [50] because it leads to approximately 85% of the solvation free energy coming from the first solvent shell, in rough agreement with theory.

1.3.8. Secondary structure propensity

Proteins are primarily composed of combinations of the secondary structure elements alpha-helices and beta-strands. (Beta-strands are usually observed in multiples, leading to the tertiary structure of a beta-sheet). These secondary structure elements are the result of the conformations of the phi (ϕ) and psi (ψ) angles (figure 1.7). Before a single protein crystal structure existed Ramachandran hypothesized the existence of favorable combinations of

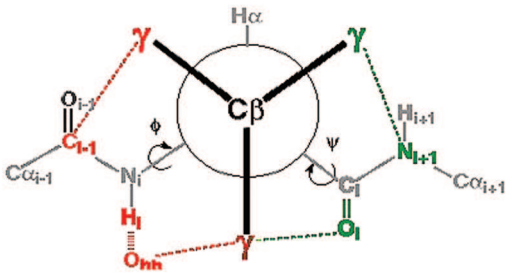


Figure 1.7. Representation of ϕ and ψ angles in a Newmann projection. Also shown are the potential clashes with the backbone that limit conformational space. Adapted from http://dunbrack.fccc.edu.

phi and psi based on sterics [51]. Regions of alpha and beta space are indicated in the Ramachandran plot (figure 1.8). These regions are the most favorable for all amino acids except glycine because of the absence of a β-carbon.

Ramachandran space can be further divided to reveal specific amino acid preferences. Certain residues are more likely to be found when the protein backbone is in specific conformations. In the simplest case this is purely due to sterics, such as when the backbone is in an alpha helical conformation beta-branched amino acids are unlikely to find a low energy conformation. Other factors are involved including solvation and dipole-dipole interactions [52, 53]. The use of this information in protein design benefits residues in the backbone environments where they are most energetically favored. The simplest method to obtain this information is to look at the occurrence of individual amino acids in specific regions of Ramachandran space by querying the Protein Data Bank [54]. The statistics that result need to converted to an energy term for use in ORBIT.

Conversion of a probabilistic term to an energy term requires using Boltzmann's equation

$$probability(y) \sim \exp\left(-\Delta G\left(y\right)\right) \tag{1.6}$$

and the assumption of an equilibrium distribution. Shortle [55] discussed this at length but in summary: if the use of equation (1.6) is predictive for a set of observables than

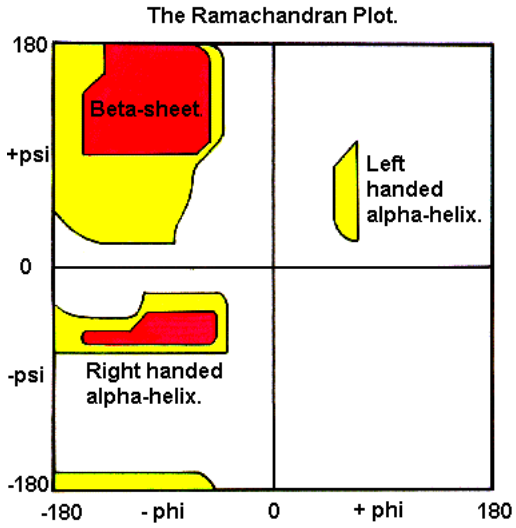


Figure 1.8. Ramachandran plot showing the favorable regions of ϕ, ψ space associated with the alpha helix and beta-sheet (in red). Yellow regions are more strained. Adapted from http://www.cryst.bbk.ac.uk.

the Boltzmann distribution applies. While this leads to interesting conclusions about the equilibrium properties of various protein structural features it is also useful in identifying the appropriate probability terms for analysis. If the observable of interest is the relative frequency of occurrence of a specific amino acid, x, in a defined region of phi-psi space, y, then

$$probability(x,y) = \frac{number\,of\,aa(x)\,in\,phipsi(y)}{number\,of\,aa(x)}. \tag{1.7}$$

But this probability term fails in predictions with equation (1.6) [56]. The probabilistic term to use is a propensity not a probability. A probability is the fractional occurrence of an event and is between 0 and 1. Propensity is a ratio of probabilities and a propensity of less than 1 is a less likely event and a probability greater than 1 is more likely event. The probability of occurrence of alanine is an illustrative example. A large percentage of all Ramachandran space is alpha helical and alanine occurs frequently in alpha helical space, leading to a pronounced favoritism to the helical region.

$$propensity(x,y) = \frac{number\,of\,aa(x)\,in\,phipsi(y)/number\,of\,aa(x)}{number\,of\,aa(ALL)\,in\,phipsi(y)/number\,of\,aa(ALL)}. \tag{1.8}$$

What this means is that propensity shows the likelihood of observing a particular amino acid in a specific phi-psi region over an average amino acid in that same region. For the alanine example, propensity gives the increased likelihood of observing alanine *relative* to other amino acids at any given position. Therefore, any use of statistical occurrence in the force field must be a measured propensity.

2. Rotamer Libraries

2.1. Introduction

2.1.1. Rotamers

Two simplifications used in most computational protein design programs currently are the uses of a fixed protein backbone and sidechain rotamers. Rotamer is the term used to describe the rotational isomers of protein sidechains (figure 2.1). Protein crystal structures show that sidechains are primarily in the low-energy conformations expected from simple physical chemistry. The first collection of these common rotamers was derived from only 19 protein structures [57]. Not until the exponential growth of the number of solved crystal structures in the 1990s could reliable data be gathered on the conformations of all sidechains. Importantly, a correlation between the backbone and sidechain dihedral angles was detected [58]. For some amino acids the observed minimum dihedral angles are different depending on the conformation of the protein backbone. A collection of rotamer conformations relative to backbone dihedral angles is a backbone-dependent rotamer library. A backbone-independent rotamer library does not consider the influence of the backbone dihedral angles on the sidechain conformation.

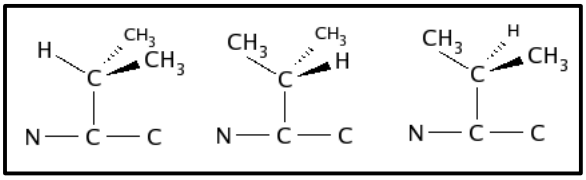


Figure 2.1. Three different rotamers for the amino acid valine.

The Mayo lab protein design software ORBIT (Optimization of Rotamers By Iterative Techniques) uses either a backbone-dependent (bbdep) or backbone-independent (bbind) rotamer library. Rotamers are fixed representations of inherently flexible sidechains. To more fully model this flexibility additional rotamers are added to the library by expanding around chi1 (e1 libraries) or both chi1 and chi2 (e2 libraries). Previous designs in the lab were based on rotamer libraries from Dunbrack and Karplus [59]. As new research into sidechain conformational preferences proceeded we wanted to incorporate these results and test them in our computational design procedure. We included the new rotamer libraries into ORBIT and designed new sequences for the core of protein G.

2.1.2. Dunbrack and Cohen

Even with the exponential growth in the number of structures in the PDB [54] there exist regions of Ramachandran space that are poorly represented. Most protein backbone dihedral angles (ϕ, ψ) are in alpha-helical or beta-sheet regions (figure 1.8). However, important functional regions of proteins occur frequently with backbone dihedral angles that are significantly outside of the two main regions. The work of Dunbrack and Cohen [60] attempted to address this by applying Bayesian statistics to more accurately predict the distributions of poorly represented areas of conformational space. Bayesian statistics allows for assumed *a priori* distributions to influence the data and potentially arrive at more reasonable *posterior* distributions in situations with sparse data. While qualitatively not much different from the previous Dunbrack rotamer library [58], there are quantitative differences as well as better statistics for use in determining frequency of various conformations and the standard deviations of the measured angles.

2.1.3. Lovell et al.

The rotamer library of Lovell et al. [61] emphasized quality over quantity. The use of higher resolution structures to determine rotamers was previously noted to lead to tighter conformer distributions (lower standard deviations) [8] but Lovell et al. extended this to include only highly resolved residues within high resolution structures. In order to exclude conformations with steric clash explicit hydrogens were included and sidechains with high van der Waals

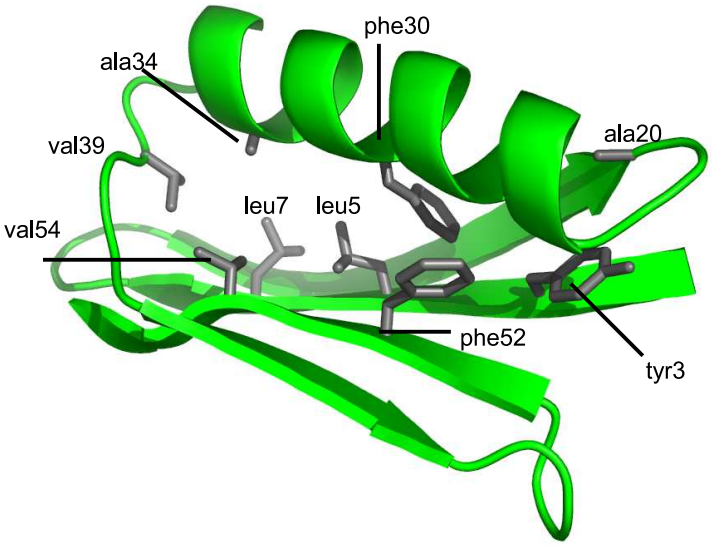


Figure 2.2. The core of wildtype protein G. The residues shown in grey are the positions allowed to change identity in the design.

overlap were eliminated. Additionally, the rotamers were defined based on modes of the conformer distributions rather than mean. This prevented the division of distributions into multiple peaks by incorrect binning. The resulting rotamer library was a minimal set of conformations (only 167) that covered 94.5% of observed sidechain conformations.

2.2. Results

One of the common systems used in studying protein properties is the β1-domain of protein G. It is a small, single domain, two-state folder and naturally relatively stable with a melting temperature of 86 ^{o}C. The core of the protein (as defined by any sidechain with less than 10% surface area exposed) is composed of ten hydrophobic residues: tyr3, leu5, leu7, ala20, ala26, phe30, ala34, val39, phe52, and val54 (figure 2.2). At each designed position the amino acids alanine, leucine, valine, isoleucine, phenylalanine, tyrosine, and tryptophan were considered. The force field used included van der Waals, surface area solvation, hydrogen bonding, and Coulombic electrostatics. The optimal sequences for each design were determined with HERO [62], and are labeled based on which rotamer library was used.

Positions 20, 26, and 34 remain alanine in all sequences as these positions are very spatially

Name	3	5	7	20	26	30	34	39	52	54	ORBIT E	T_m	#C
wildtype	F	L	L	A	A	F	A	V	F	V	–	86[b]	38
D02_bbdep_e0	F	L	I	A	A	F	A	V	F	V	-116.6	88[a]	38
D02_bbdep_e1	F	L	I	A	A	F	A	V	F	V	-122.4	88[a]	38
D02_bbdep_e2	**F**	**L**	**I**	**A**	**A**	**F**	**A**	**L**	**F**	I	**-128.8**	**90.5**	**40**
D02_bbind_e0	F	L	V	A	A	F	A	V	F	V	-108.3	84[c]	37
D02_bbind_e1	F	L	I	A	A	F	A	V	F	V	-118.6	88[a]	38
D02_bbind_e2	F	L	I	A	A	F	A	I	F	V	-131.3	91[d]	39
D96_bbdep_e0	**F**	**L**	**V**	**A**	**A**	**A**	**A**	**V**	**F**	**V**	**-107.3**	**57.2**	**31**
D96_bbdep_e1	F	I	I	A	A	A	A	V	F	V	-110.2	-	32
D96_bbdep_e2	F	L	I	A	A	F	A	V	F	V	-129.8	88[a]	38
D96_bbind_e0	**F**	**L**	**V**	**A**	**A**	**A**	**A**	**V**	**F**	**V**	**-98.3**	**57.2**	**31**
D96_bbind_e1	**F**	**L**	**I**	**A**	**A**	**A**	**A**	**V**	**F**	**V**	**-109.7**	**88.2**	**32**
D96_bbind_e2	F	L	I	A	A	F	A	I	F	V	-128.8	91[d]	39
Rich_e0_more	F	L	I	A	A	F	A	V	F	V	-115.5	88[a]	38
Rich_e1_less	F	L	V	A	A	F	A	V	F	V	-119.1	84[c]	37
Rich_e1_more	F	L	I	A	A	F	A	V	F	V	-126.3	88[a]	38
Rich_e2_less	**F**	**L**	**I**	**A**	**A**	**F**	**A**	**I**	**F**	**A**	**-122.7**	**91.3**	**37**
[a] Value from mutant IVV, page 3-39 in [63]													
[b] Value from page 3-39 in [63]													
[c] Value from mutant VVV, page 3-39 in [63]													
[d] From [26]													

Table 2.1. Rotamer library mutants. ORBIT E is the energy calculated by ORBIT for the designed sequence. T_m is the melting temperature of the protein. #C is the number of carbon atoms in all the designed positions (a rough estimate of packing density, since all residues are all carbon and buried). BOLD rows are sequences first created in this study.

restricted (position 20 by nearby backbone atoms, position 26 by the large amino acid at position 3, and position 34 points directly into the core including the fixed tryptophan at position 43). Position 3 is a tyrosine in wildtype but a phenylalanine in all mutants. Since position 3 is classified as a buried position, the surface area solvation penalizes the hydroxyl group on the tyrosine leading to selection of phenylalanine. Position 52 remains a phenylalanine in all sequences because the force field favors large, nonpolar amino acids in the core and phenylalanine 52 is the largest nonpolar residue possible at that position.

The remaining positions (5, 7, 30, 39, 54) do exhibit variation based on the rotamer library used. A known issue with the wildtype core of protein G is that Leu7 is in a non-rotameric conformation ($\chi_1 = -55^o$, $\chi_2 = 106^o$), whereas the canonical rotamers in this area of Ramachandran space would be $\chi_1 = 178^o$, $\chi_2 = 65^o$ 52% of the time and $\chi_1 = -60^o$, $\chi_2 = 177^o$ at 39% frequency [60]. Because the χ_2 angle of Leu7 is non-rotameric no designed sequence maintains a Leu at this position and attempt to fill space with an Ile, or less favorably a Val.

It is useful to analyze the trends within the series (e0, e1, e2). The ORBIT calculated energies drop as the number of rotamers increase. In this study the e2 expansion has a large effect on the "packability" of leucines and isoleucines (phenylalanine, tyrosine, and tryptophan have bulky rings and valine only has a χ_1 dihedral). The e2 libraries have more leucine and isoleucine conformations to fit into the spatially constrained core. Rich_e2_less appears to be an exception, but it is more appropriate to compare it with Rich_e1_less as both have low probability rotamers excluded that Rich_e1_more includes (see section 2.4). The calculated energy does not include any internal energies of the individual rotamers; if the canonical rotamer is assumed to be the lowest energy conformation then the expansions around χ_1 and χ_2 in the e1 and e2 libraries would increase the energy of the rotamer. If an internal energy was included the drop in energy would be less as the size of the rotamer library increased.

The number of carbon atoms packed into the core also increases with the size of the rotamer library. The D02_bbdep series relieves the strain in Leu7 with an Ile in both e0 and e1. The D02_bbdep_e2 library packs in two additional carbon atoms with Val39Leu and Val54Ile (figure 2.3). The D02_bbind series is similar except that e0 is unable to fit

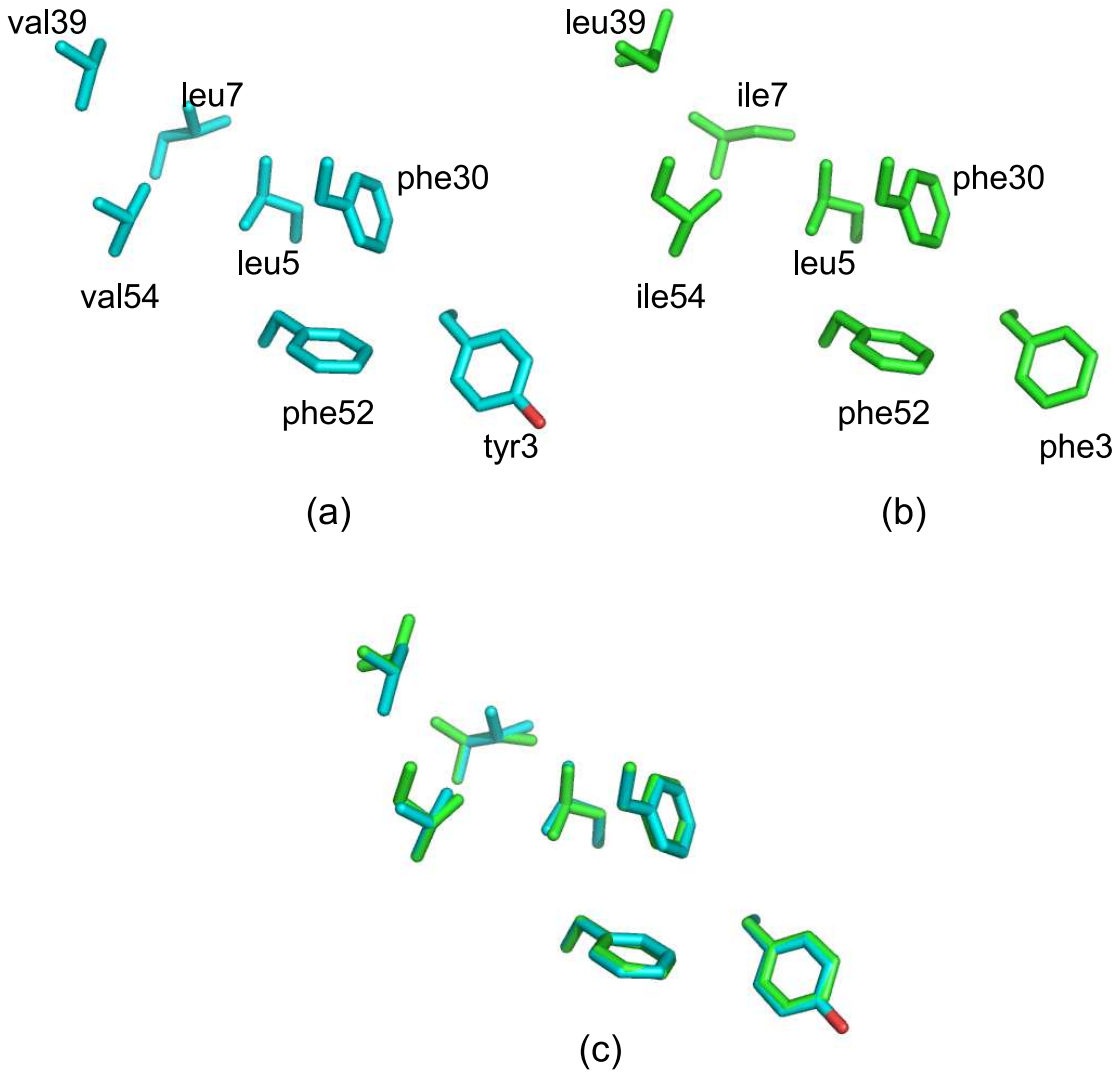


Figure 2.3. (a) Wildtype protein G showing non-alanine designed positions. (b) D02_bbdep_e2. (c) Superposition of mutations on wildtype. It is apparent that the conformation of Ile7 attempts to fill the same space as the strained Leu7.

an Ile at position 7 and must lose a carbon with val7 and e2 cannot fit another Ile at position 54 keeping val54. The isoleucine at position 54 with D02_bbdep_e2 may be a slight overpacking since the sequence with val54 is slightly more stable.

The older Dunbrack rotamer libraries (D96_) perform significantly worse with both the bbdep and bbind e0 libraries unable to place a phenylalanine at position 30. The core is considerably underpacked leading to a significantly destabilized protein (T_m= 57^{o}C). The difference between D96_ and D02_ is that the D96_ libraries have only one χ_2 dihedral per χ_1 dihedral (three total rotamers in e0) while the D02_ libraries have two χ_2 dihedrals per χ_1 dihedral (six total rotamers in e0). The additional dihedral value allows D02_ e0 and e1 libraries to fit a phenylalanine at position 30 when D96_ e0 and e1 libraries cannot. The D96_bbdep_e2 and D96_bbind_e2 libraries are able to fit a phenylalanine at position 30 because the expansions around χ_2 very closely approximate the wildtype dihedral.

The Richardson libraries performed quite well given their small size (167 rotamers in Rich_e0_more compared to 355 in DK96_bbind_e0). The Rich_e1_less library has some low frequency (>1%) rotamers removed that leads to removal of 2 of the 7 isoleucines and 1 of the 5 leucines. Thus Rich_e1_less cannot place an isoleucine at position 7 and suffers a stability loss with a valine. Compared to D02_bbdep_e2, Rich_e2_less places an isoleucine at position 39 but an alanine at 54 and is very stable (T_m = 91.3^{o}C). The previously most stable protein G core sequence has a valine at 54 [26] (or an isoleucine with D02_bbdep_e2). Therefore, some relief of overpacking may be occurring or position 54 is not greatly influencing the core stability (figure 2.4).

2.3. Discussion

With the continuing growth in the number of protein structures deposited to the PDB, larger and more detailed analyses of the properties of proteins are possible [54]. More information about protein sidechain conformations furthers the study of rotamers and their uses in homology modeling, fold prediction, and protein design [64]. We explored here two of the newer rotamer libraries and evaluated their performance within our protein design framework.

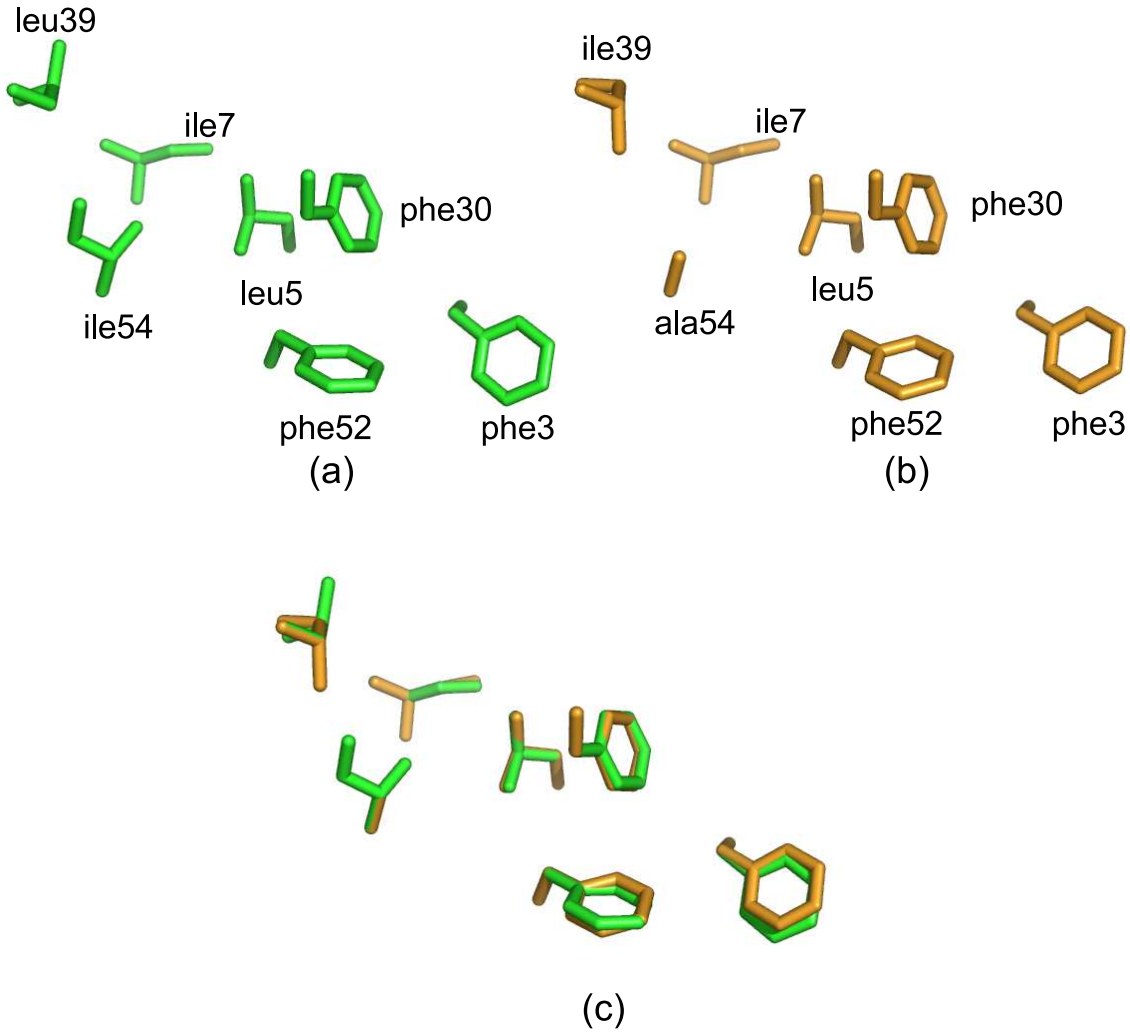


Figure 2.4. (a) D02_bbdep_e2. (b) Conformations of sidechains in Rich_e2_less showing the different conformation of Ile39 and the resulting Ala54. (c) Superposition of D02_bbdep_e2 and Rich_e2_less, two very stable sequences.

New core designs on protein G with a number of different rotamer libraries leads to new sequences, two (D02_bbdep_e2, Rich_e2_less) that are approximately as stable as the most stable protein G core previously known. The e2 libraries appear to be necessary to achieve high stability sequences. This implies that the expansion around the dihedral is a non-energetic contribution that merely accounts for fixed backbone and inflexible representations of sidechains. The inability of D96_bbdep_e0 to pack a phenylalanine into the core at position 30 demonstrates the importance of correct binning as the D96 libraries included two of the modes of a phenylalanine into one averaged bin. The value of accuracy and resolution in rotamer libraries is important for the small e0 libraries as well as the larger, expanded libraries to appropriately cover dihedral space.

2.4. Methods

Oligonucleotide mutatgenesis was carried out on existing lab stocks using inverse PCR. Nucleotide primers were ordered from Caltech. All DNA sequences were confirmed by Caltech Sequencing Facility. Recombinant proteins were expressed in BL21(DE3) *E. coli* cells (Stratagene) at 37 ^{O}C. Proteins were extracted from the cells using freeze-thaw [65]. Proteins were purified by HPLC using a reverse-phase C8 prep column (Zorbax) and linear acetonitrile-water gradient containing 0.1% (v/v) trifluoroacetic acid. Protein masses were determined by matrix-assisted time-of-flight and were as expected.

Circular dichroism (CD) data were obtained on an Aviv 62A DS spectropolarimeter with a thermoelectric cell holder and an autotitrator. All experiments were conducted at pH 5.5 in 50 mM sodium phosphate. Far UV wavelength scans were run at 1 and 99 ^{O}C to determine reversibility. Thermal denaturation data were at 218nm were collected every 1 ^{O}C from 1 to 99 ^{O}C, equilibrating for 90 seconds with 30 seconds of averaging (figures 2.5, 2.6, 2.7).

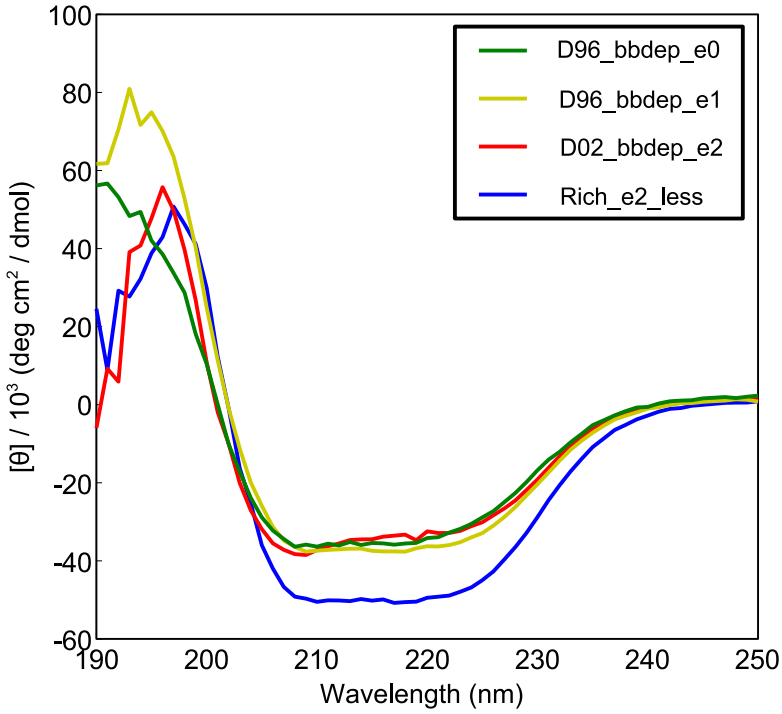


Figure 2.5. Far UV wavelength scans of new designs.

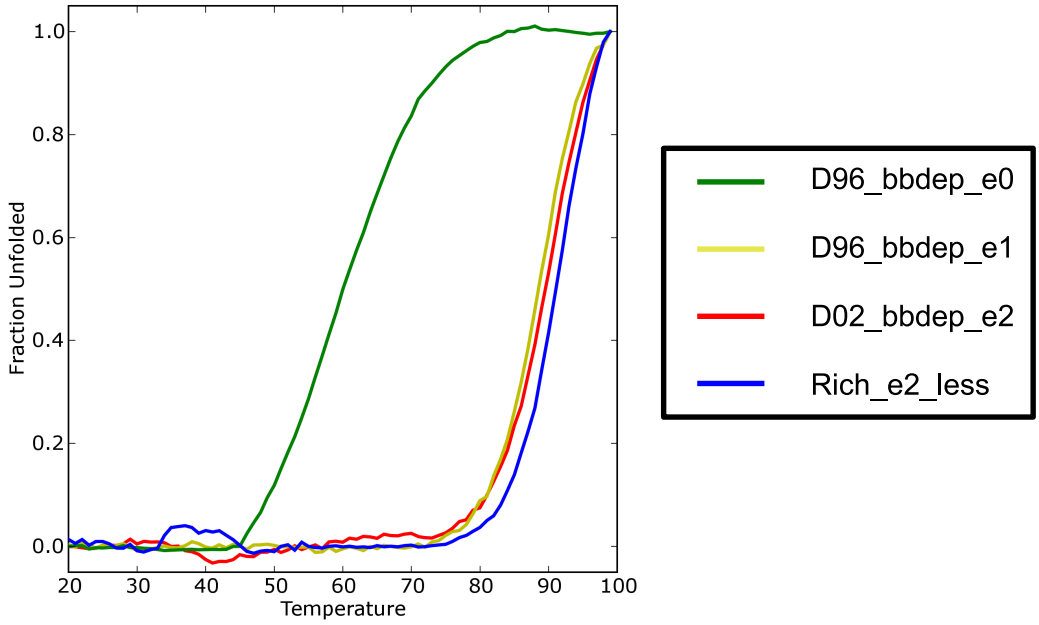


Figure 2.6. Temperature denaturations of new designs.

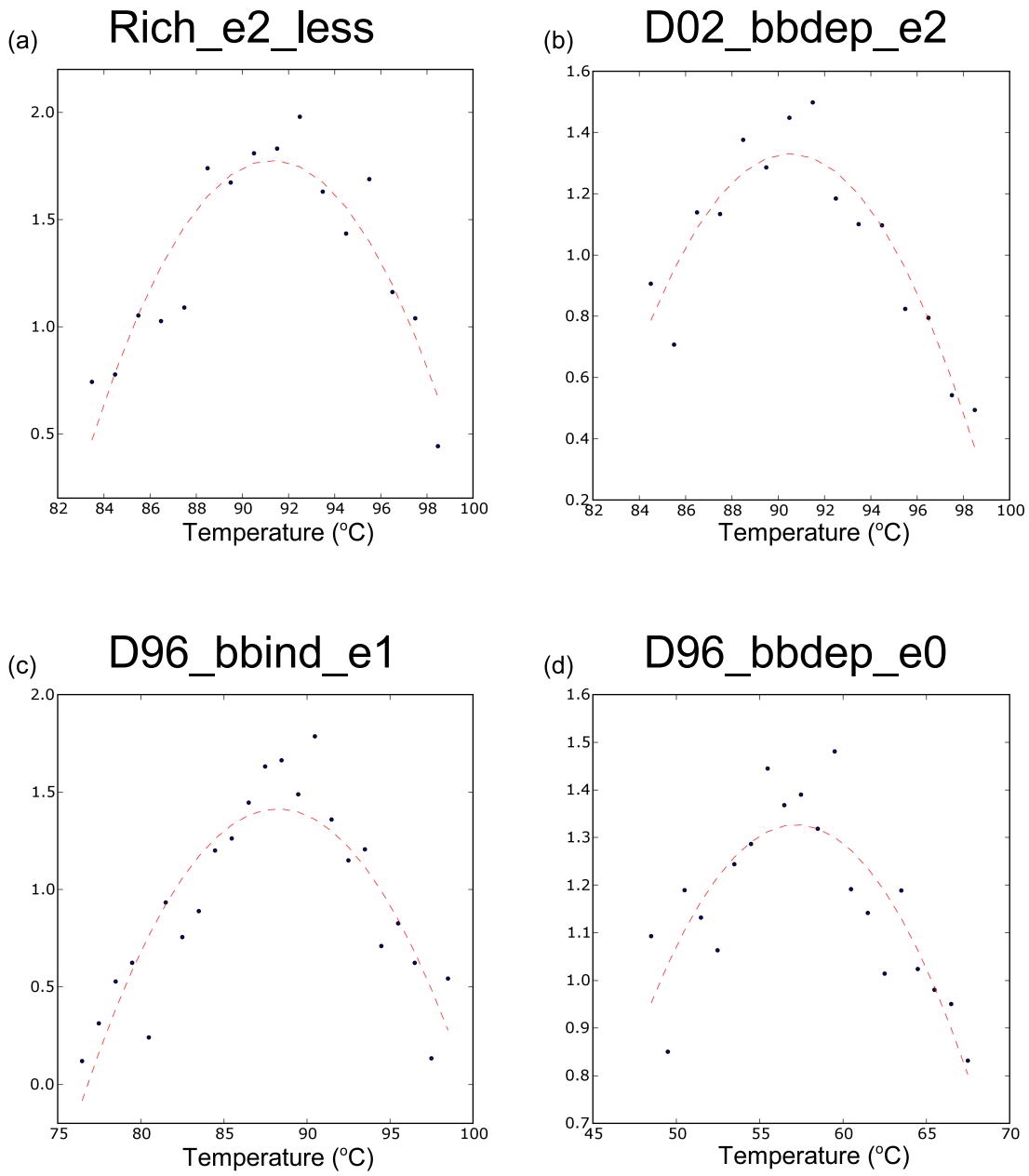


Figure 2.7. The temperature denaturation experiments for D02_bbdep_e2, Rich_e2_less, and D96_bbdep_e0 did not have appropriate posttransition baselines to use the standard fitting equation [66] and the inflection point has been used to compare stabilities. Plotted are the first derivatives of the thermal denaturation curves around the inflection points. A simple quadratic equation is fit to the first derivative data to arrive at the inflection point.

3. Electrostatics in Protein Design

This work was published in Protein Science 2006.

3.1. Introduction

The energy functions used in protein design must be rapidly evaluable due to the large size of protein design calculations. However, the physical interactions these energy functions are developed to represent are highly complex. Approximations introduced to increase the speed of calculations must also capture the intricate balance of the stabilizing and destabilizing interactions that lead to the observed marginal stability of proteins [34]. Electrostatic interactions contribute significantly to protein stability and function. Not only must intraprotein interactions be considered (hydrogen bonding, salt bridges, etc.), but effects due to the aqueous environment such as polar solvation and solvent screening need to be evaluated as well. It is computationally intractable to consider all individual water molecules surrounding a protein during a protein design calculation. Continuum approaches that consider the solvent at a macroscopic level using various numerical solutions of the Poisson-Boltzmann equation [67] have been used to predict sidechain pKas, the electrostatic component of binding, and other biologically important processes. However, a full Poisson-Boltzmann calculation is far too time consuming to be used at each step of a protein design calculation. Various methods attempting to reproduce the accuracy of Poisson-Boltzmann calculations within the restrictions of protein design include the adaptation of a solvent exclusion method [50] in designing a novel protein fold [33], a modified Tanford-Kirkwood approach [68] to design specific protein-protein interactions [69], use of a Born method in a new protein design force field [70], and a highly parameterized set of simple terms [71] in designing enzymatic activity

onto a previously catalytically inactive scaffold [32]. Work in this lab led to the development of a two-body decomposable implementation of a Poisson-Boltzmann calculation useful in protein design [72].

Previous design studies have shown both the importance of electrostatics and the need to improve the electrostatic component [73] of our protein design algorithm, ORBIT [24, 22]. Local interactions were shown to be underrepresented and hydrogen bonding was overrepresented relative to long range Coulombic interactions. Here we show a comparison of ORBIT electrostatic energies and those calculated using the finite difference Poisson-Boltzmann implementation in DelPhi [74] allowed a parametrization of the simple Coulombic equation term used in ORBIT. By scaling the dielectric value it is possible to approximate the energies calculated using the more accurate Poisson-Boltzmann method. Local interactions (sidechain-backbone) and longer range interactions (sidechain-sidechain) are parameterized separately. The polar solvation model used in this study penalizes the burial of non-hydrogen-bonded, non-backbone polar hydrogens.

3.2. Results

Electrostatic effects are studied in the background of the engrailed homeodomain, a small (51 amino acid) protein with three alpha helices. The wildtype sequence is not optimized for stability with seven positive charges distributed across a small amount of surface. The protein sequence resulting from a design calculation with the unoptimized electrostatic terms of ORBIT is NC0 [73]. While eliminating the charge excess, this designed protein was shown to have incorporated a number of unfavorable electrostatic interactions relative to wildtype: a reduced number of N-capping interactions and an increased number of potentially destabilizing interactions with the helix dipole. NC0 has a stability slightly greater than wildtype and is used as the baseline for ORBIT's electrostatic performance in this study. An updated rotamer library that was shown to lead to similar designed sequences as a previous library was used in the study reported here. The sequence designed with this new library but with the unoptimized electrostatic term is NC0_new. This protein was designed as a control for the new electrostatic term.

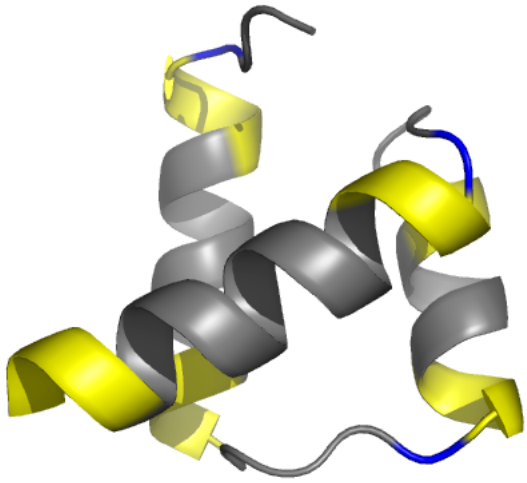


Figure 3.1. Engrailed homeodomain. Positions in yellow are considered to be interacting with the helix dipole. Positions in blue are N-capping positions.

The analysis of a small set of proteins suggested that a distance dependent dielectric of 5.1r for sidechain-backbone interactions and 7.1r for sidechain-sidechain interactions more closely predicts the energy calculated by DelPhi (see Methods). Previous work in this lab has used a distance-dependent dielectric of 40r. The new dielectric leads to a 7.8-fold increase in the strength of electrostatic interactions in the sidechain-backbone case and a 5.6-fold increase in the sidechain-sidechain case. Thus, while the importance of electrostatics is increased significantly in the design calculations, it is further increased in the sidechain-backbone case in order to address the concerns raised in Marshall et al. [73]. The design calculation using the optimized dielectric values and penalties based on the number of buried polar hydrogens is Dielec_H. Circular dichroism wavelength scans indicated that the designed proteins were well folded and alpha-helical. Thermal denaturation studies were carried out on NC0, NC0_new, and Dielec_H (see figure 3.2). All proteins unfolded completely and reversibly. For comparison, the thermal denaturation curve of wildtype engrailed is also included.

The dependence of protein design on the rotamer library used is shown clearly by the difference in sequence between NC0 and NC0_new (Fig 1). While the rotamer libraries

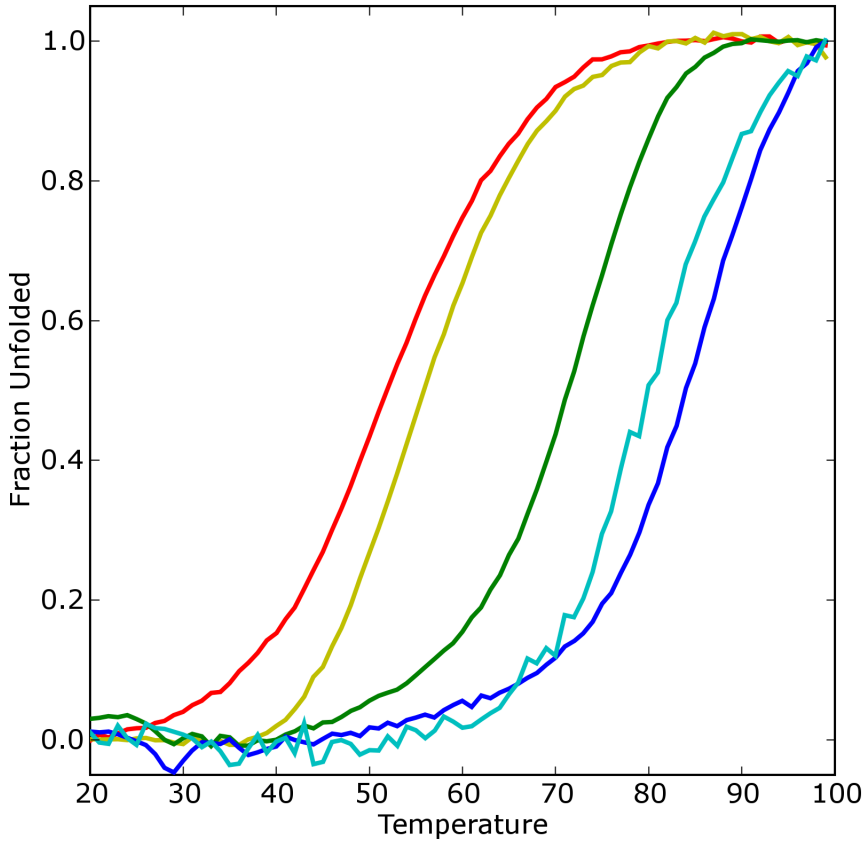


Figure 3.2. Thermal denaturation of engrailed homeodomain mutants. Red is wildtype, gold is NC0, green is NC0_new, cyan is NC3_Ncap, and blue is Dielec_H.

	Desolvation[a]	Sidechain/backbone[b]	Sidechain/sidechain[c]
WT	7.9	-8.5	-1.6
NC0	17.3	-8.1	-34.7
NC0_new	16.2	-10.5	-31.8
Dielec_H	14.1	-9.3	-32.7
NC3_Ncap	14.2	-14.9	-27.7
All energies are in kcal/mol			
[a] Sum of the sidechain desolvation energies			
[b] Sidechain/backbone screened Coulombic energy			
[c] Sidechain/sidechain screened Coulombic energy			

Table 3.1. Decomposition of DelPhi calculated electrostatics into sidechain desolvation, sidechain/backbone screened Coulombic energy and sidechain/sidechain screened Coulombic energy.

used have very similar numbers of rotamers and the calculations were otherwise identical, there are nine mutations between NC0 and NC0_new. The surface of a protein is much less constrained by sterics than the protein core, allowing a much greater choice of rotamers. A slight difference at one position leading to the choice of a different amino acid could propagate other changes across the protein surface during the design. The ORBIT calculated energies of these proteins are similar, but NC0_new is shown to have an unfolding transition temperature 21° C higher. By examining the electrostatic character of these sequences it is clear that NC0_new has both more beneficial N-capping interactions and less detrimental interactions with the helix dipole. There is some debate as to the importance of the helix dipole, especially at solvent-exposed positions [75, 76]. DelPhi analysis of the hypothetical structures of these sequences suggests that NC0_new experiences less of a desolvation penalty and better sidechain-backbone interactions than NC0 (table 3.1). Care must be taken with interpretation of these data due to the hypothetical nature of the structures and the strong conformation dependence of Poisson-Boltzmann calculations [77]. Thus while NC0_new and NC0 were not designed to be significantly different, the differences that are observed both in sequence and measured stability can be explained, at least qualitatively, by electrostatic differences.

The sequences of Dielec_H and NC0_new can now be compared to determine the effects of simply modifying the electrostatic term to include lower dielectrics. Dielec_H is a very

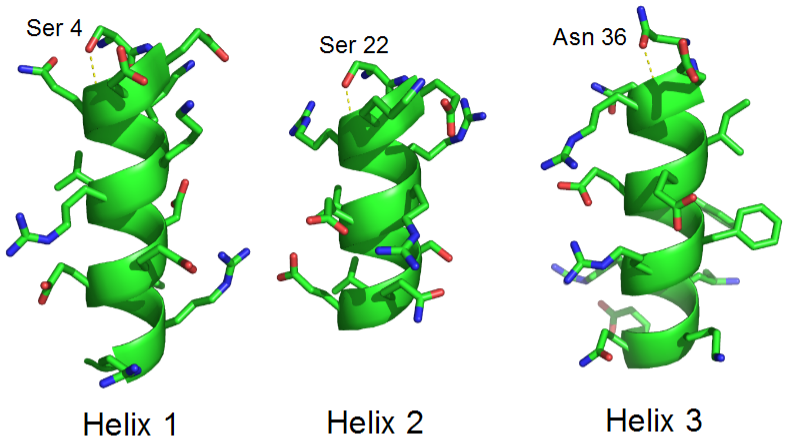


Figure 3.3. Three helices of engrailed homeodomain showing that ORBIT does predict all N-capping positions to be forming hydrogen bonds with the backbone.

stable protein, approximately as stable as NC3_Ncap [73] that was designed by preventing detrimental interactions with helix dipoles and forcing N-capping interactions by restricting amino acid composition. Just by increasing the strength of sidechain-backbone electrostatic interactions relative to sidechain-sidechain interactions appears to recover all three of the N-capping interactions in engrailed (figure 3.3). DelPhi calculations show that Dielec_H has more favorable desolvation and sidechain-sidechain energies than both NC0_new and NC3_Ncap (table 3.1). However, these values are conformation-dependent and the calculated difference in sidechain-backbone energy between Dielec_H and NC3_Ncap is dependent primarily on the interactions of three glutamates that would likely assume more than one conformation in solution.

3.3. Discussion

Protein design requires the evaluation of a large number of functions for a complete calculation. These functions need to both be rapid and accurate. Unfortunately, the complexity of the protein energy surface does not lend itself to a simple representation. While a full Poisson-Boltzmann calculation at each step would lead to a more accurate view of the electrostatic environment of the protein, implementing such a procedure remains computationally intractable. The need for approximate functions necessitates the evaluation of both

their accuracy and usefulness. A parameterization of the Coulombic term in ORBIT using the Poisson-Boltzmann equation implemented in DelPhi leads to an increase in the weight of electrostatics in the ORBIT force field as well as separate dielectric values for sidechain-backbone and sidechain-sidechain interactions. While it is difficult to conclusively state the exact interactions that lead to difference in stability, in this work we show experimentally that simple modifications to the Coulombic term lead to a designed protein that is stabilized by electrostatic interactions as well as recovering helix N-capping interactions, a stabilizing feature of natural protein sequences.

3.4. Methods

The new dielectric values were obtained by performing electrostatic calculations on a small set of proteins and determining the value that would lead to the best fit. The protein structures downloaded from the PDB are: 1igh (β1 domain of protein G), 1rge (ribonuclease SA), 1rhe (Rhe VL), 1whi (L14 ribosomal protein), 1tta (transthyretin), 2rn2 (ribonuclease H), 3lzm (T4 lysozyme), and 1amm (gamma-B-crystalin). The DelPhi [74] calculations used a grid spacing of 2.0 grids per Å, an interior dielectric of 4.0, an exterior dielectric of 80.0, 0.050 M salt, and a probe radius of 1.4 Å. PARSE charges and radii were used [78]. In order to more directly compare with the terms in the ORBIT force field, the DelPhi results for both unfolded and folded states were separated into backbone and sidechain desolvation and screened Coulombic interactions. The description of the unfolded and folded states of the backbone and sidechains can be found in figures 1, 2, and 3 in [72]. The dielectric value used in the ORBIT Coulombic term is then scaled to more closely agree with the values calculated with DelPhi. Correlations coefficients of the fits between DelPhi electrostatic energies and scaled Coulombic energies (sidechain-sidechain and sidechain-backbone) are greater than 0.9 (figure 3.4). In order to facilitate comparison with previous work [73], electrostatic calculation in table 3.1 were performed with the dame DelPhi parameters as above with the exception of a probe radius of 0 Å.

The preparation of the engrailed homeodomain PDB [54] structure, 1enh, and the designed positions are the same surface positions as reported in Marshall et al. [73]. Residues

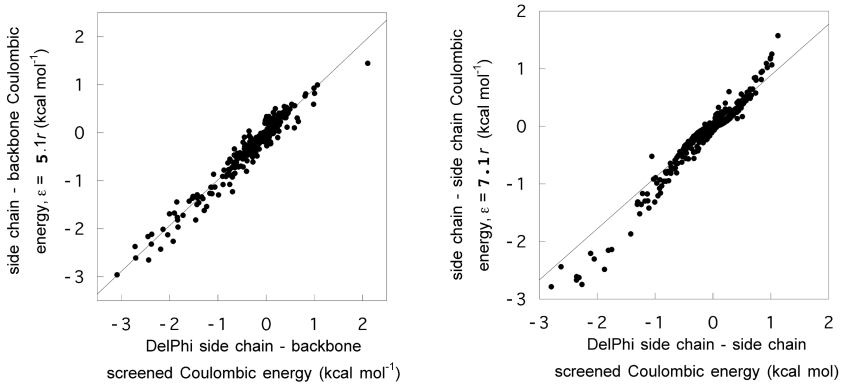


Figure 3.4. Parametrization of ORBIT Coulombic dielectric with data from DelPhi. Left: sidechain-backbone values. Right: sidechain-sidechain values.

allowed at the designed positions were Ala, Ser, Thr, Asp, Asn, His, Glu, Gln, Lys, and Arg. Rotamers are derived from the rotamer library of Dunbrack and Karplus [59] with expansion of one standard deviation about angles chi1 and chi2 of aliphatic residues, expansion of one standard deviation around chi1 of hydrophobic residues, and no expansion of polar residue dihedral angles. The force field in ORBIT contains van der Waals, Coulombic, hydrogen bond, and solvation terms [79]. The hydrogen bond term is geometry and hybridization dependent, as described [25]. The polar hydrogen burial term is calculated as 2.0 kcal/mol for each non-backbone, non-hydrogen bonded buried polar hydrogen, as described [25]. Sequence optimization was performed using DEE [23, 80, 81] or HERO [62]. The one best sequence for each design was expressed and purified for biophysical analysis.

Genes for the engrailed variants were prepared by recursive PCR [82] and cloned into pET-11a (Novagen). Wildtype engrailed expresses poorly and was cloned into a plasmid that had been engineered to include N-terminal His-tags and a ubiquitin domain with a ubiquitin-specific cleavage site [83]. DNA sequencing confirmed the identity of all variants. Proteins were expressed in BL21(DE3) *E. coli* cells (Stratagene) and isolated with freeze-thaw [65] or sonication. Proteins were purified by HPLC as in Marshall et al. [73] or nickel exchange columns (Qiagen). Cleavage of the protein of interest from the fusion domain occurred by use of the protease UCH-L3 (Boston Biochem) at 37 ^{o}C for 1 to 4 hours. Proteins were confirmed with MALDI-TOF mass spectrometry. Temperature denaturation

circular dichroism was carried out as described [73].

4. Computational Tuning of the Force Field

4.1. Introduction

The accuracy of the force field used in protein design is of fundamental importance. The ability to design proteins with interesting properties relies on the force field to correctly represent the physical interactions within the protein. The great difficulty is that these physical interactions are not perfectly understood at a quantitative level. In some cases, even a qualitative understanding eludes consensus. This is demonstrated in the debate surrounding the importance of electrostatics for protein stability with one camp suggesting these interactions are destabilizing [84, 85] while another maintaining that they are the primary stabilizing force in proteins [86, 87].

There may be multiple ways to express the energy terms algorithmically and a method must be developed to determine the best combination for the force field. As ORBIT was developed in the Mayo lab from the simplest coiled-coil designs to engineering new enzymatic functions, energy terms were added in a stepwise manner that allowed experimental validation at each step. This is the "Design Cycle," theoretical work is translated to computational designs that are validated by building the proteins. Knowledge from the experimental phase can then be fed back into the cycle. The cycle is mainly limited at the experimental validation step. First, even the simplest of proteins will take time and effort to successfully express, purify, and biophysically characterize. Second, the force field ideally would be tested on many proteins to ensure transferability between systems. Third, with more complex terms more data would have to be collected from more proteins. A method

is needed to increase the speed of development of new and different force field terms.

Since the techniques for site-specific mutation of proteins became commonplace [1], protein biophysicists have perturbed proteins and measured the effects. The increase in protein crystal structures is equally important as structural changes (or lack thereof) of mutations could be explicitly shown rather than assumed. A variety of techniques and systems are represented in the literature. In general, the free energy of folding (ΔG_f) of the wildtype (unmutated) protein is compared to the free energy of folding of the mutated protein. The difference between these values ($\Delta\Delta G_f$) is attributed to the effects of the mutation.

$$\Delta\Delta G_f = \Delta G_f mutant - \Delta G_f wildtype$$

This obviously leads to selection bias in the data because the mutations were chosen by investigators to explore a particular interaction and only those mutations that lead to folded proteins are reported. However, for the most part those interactions that were considered interesting to the experimental biophysicists will be important for the ORBIT force field to represent accurately. The balance between van der Waals and nonpolar solvation, the importance of electrostatics, and hydrogen bonds were all explored and are represented in the literature.

Previous work has attempted to explain observed free energies of mutation through structural analysis or calculations. Brian Matthews pioneered this work with detailed structural and thermodynamic studies of lysozyme [88, 2]. An early calculation-based study was also one of the founding studies for protein design [12]. In that study a series of mutants previously designed and characterized by the Sauer lab [13] was explored with a Monte Carlo sidechain placement algorithm that considered only van der Waals energy. The analysis was dramatically successful, predicting the measured energy of nearly all the mutants with a very good correlation. However, it must be emphasized that the predicted mutations were the simplest possible. All were in one system and all were from one hydrophobic residue to another hydrophobic residue.

Another more recent study by the Serrano lab attempted to predict the energies of single mutants across a large set of different proteins [89]. By using an existing database,

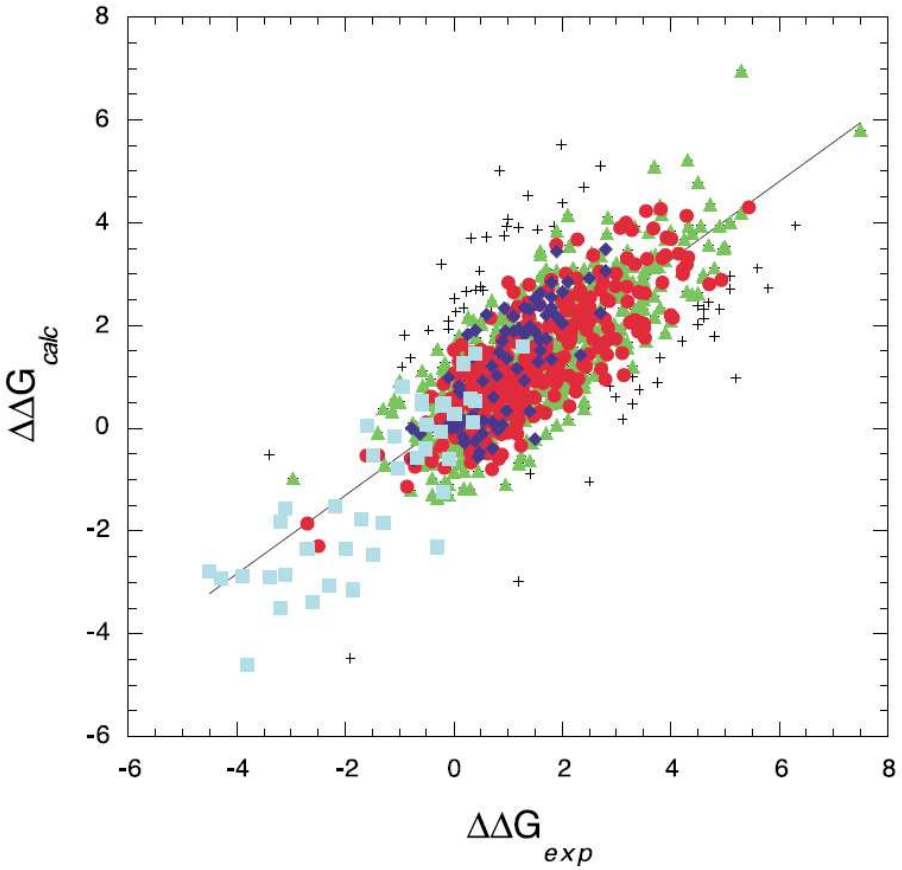


Figure 4.1. Results of the FOLDEF prediction of free energies of mutation [89]. A correlation coefficient R = 0.83 is arrived at by excluding 5% of the data as outliers (crosses). Red circles were in the training dataset, green triangles are the the test set, light blue squares are specific, disruptive lysozyme mutants, and dark blue diamonds are protein-protein complexes.

PROTHERM [90], Serrano evaluated their force field on over 1000 mutants and performed quite well (Fig 4.1). Only the wildtype structures were explicitly used; the mutant structures were built from the wildtype using WHATIF [91]. Because the mutant structures were hypothetical the mutations were restricted to those that could be predicted with some degree of confidence. Only single deletion mutants (loss of groups from a sidechain *e.g.* Leu→Val, or anything to Ala) or single substitution mutants (exchange of one atom for another *e.g.* T⇌V) were considered. What must be noted here is that non-disruptive single mutations are relatively easily to predict; as can be seen from figure 4.1 nearly all are destabilizing ($\Delta\Delta G_f > 0$) and the stabilizing mutants show systematic deviation from the correlation line. Thus while the correlation is overall good the predictive ability of the method for

multiple mutations would significantly degrade from accumulation of error.

The work here attempted to use the collected biophysical data in the literature to parameterize the ORBIT force field. As the ORBIT force field is based on physical terms (rather than statistical) it would ideally reproduce experimental physical results. It is shown here that producing similar results to experimental data not a sufficient condition for an accurate ORBIT force field.

4.2. Methods

4.2.1. Database development

A number of constraints designed to minimize error in the dataset were designed from the outset. Previous studies, most notably Guerois et al.[89], required only the structure of the wildtype protein. The structures of the mutant proteins were predicted from the wildtype. Mutations can introduce structural effects into the rest of the protein (see Notes in table at the end of chapter 4) that can lead to unpredicted results. We minimized this error by requiring that X-ray crystal structures exist for both the wild type and mutant sequences. The crystal structures were required to be greater that 2.0 Å resolution to reduce the incidence of structural errors such as misfitting and multiple conformations [61]. An initial round of analysis suggested further refinements including verifying that the wildtype and mutant structures have the same sequence away from the site of mutation, that both structures have the same number of missing residues (preferably zero), and that the two structures are the best possible comparison crystallographically (same space group, packing, etc.) The protein must be a known two-state folder (for accurate thermodynamic measurements), monomeric, with no intrinsic ligands. The available thermodynamic data must be taken in a pH range from 5 to 7.5 in low salt conditions. (Unfortunately, this does remove a significant amount of data as much early protein thermodynamic data was taken at pH 2). PROTHERM was used as an initial set, but as the searching ability of the hosted database was not useful, the database as of August 23, 2003 was copied directly from the site to allow searching and manipulation of the data locally. The data in PROTHERM was rarely complete and often self-contradictory leading to a primary reference search. In many cases the free energy of

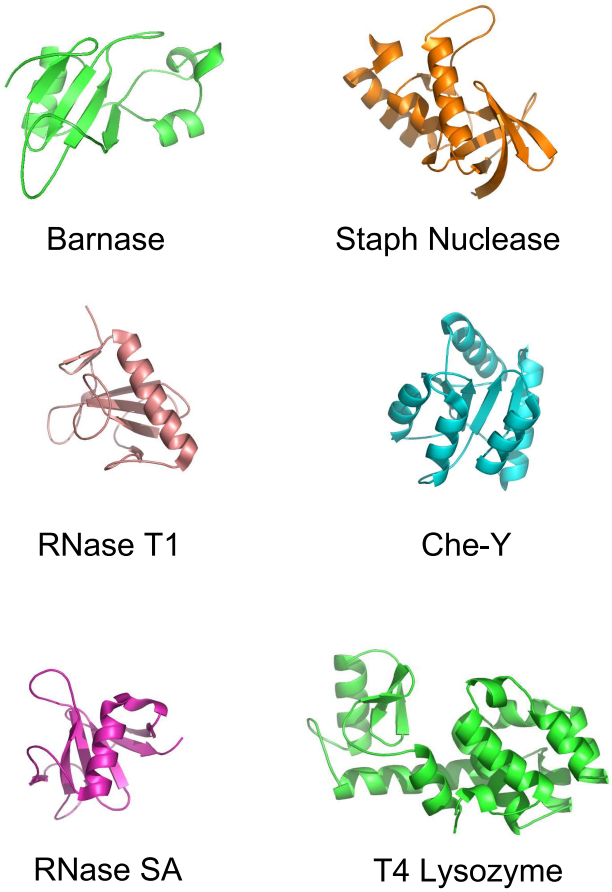


Figure 4.2. Structures in the database. Characteristically small, monomeric, two-state folders.

unfolding of the wildtype was slightly different when measured by different labs or different methods. Since the value of interest for this study is the difference between the free energy of unfolding of the wildtype and mutant, priority was given to data measured in the same lab by the same method, then different lab by same method. The collected data is in the table at the end of the chapter. There are 269 mutants across 18 different proteins. Of all the mutants in the dataset, 54% are in T4 lysozyme.

4.2.2. Optimization

As the ORBIT force field is composed of physical terms it ideally should predict differences in measured physical properties. The collection of $\Delta\Delta G$ points is therefore an excellent parametrization and test set. Since the force field is a linear sum of terms the simplest optimization for parametrization is an ordinary least squares (OLS) method.

$$\begin{aligned}
\Delta\Delta G_{f1} &= B_1 \bullet E_{vdW1} + B_2 \bullet E_{elec1} + B_3 \bullet E_{hbond1} + B_4 \bullet E_{NP1} + B_5 \bullet E_{Pol1} \\
\Delta\Delta G_{f2} &= B_1 \bullet E_{vdW2} + B_2 \bullet E_{elec2} + B_3 \bullet E_{hbond2} + B_4 \bullet E_{NP2} + B_5 \bullet E_{Pol2} \\
\Delta\Delta G_{f3} &= B_1 \bullet E_{vdW3} + B_2 \bullet E_{elec3} + B_3 \bullet E_{hbond3} + B_4 \bullet E_{NP3} + B_5 \bullet E_{Pol3}
\end{aligned}$$

Each line represents the energy of one mutation: $\Delta\Delta G_{f1}$ on the left is the measured free energy of mutation between the wildtype and mutant and on the right the equation $B_1 \bullet E_{vdW1} + B_2 \bullet E_{elec1} + B_3 \bullet E_{hbond1} + B_4 \bullet E_{NP1} + B_5 \bullet E_{Pol1}$ is the ORBIT-calculated energy difference for the mutation. The least squares optimization will find the set of parameters $[B_1, B_2, B_3, B_4, B_5]$that lead to the best "fit" between the measured values, $\Delta\Delta G_f$, and the ORBIT-calculated energies. The parameters that lead to the lowest sum of squared error is the optimum solution for the set of linear equations. The standard deviation for a parameterization, σ, is used as a measure of fit, with a lower value indicating a better fit. While this optimization method guarantees the best answer for the given set of equations, it is impossible to determine whether the force field terms included are optimal. The solution from the optimization, $[B_1, B_2, B_3, B_4, B_5]$, is the set of parameters for the ORBIT force field that best predicts the measured $\Delta\Delta G_f$. A variety of tools were used for the mathematics in this project including Octave (*www.octave.org*), Matlab (*www.mathworks.com*), scipy (*www.scipy.org*), and the BVLS bounded least squares solver (*http://lib.stat.cmu.edu/general/bvls*, [92]).

4.2.3. Force field terms

4.2.3.1. Standard ORBIT

For a more complete discussion of the force field terms used in this study also see the Introduction, section 1.3.

The ORBIT suite of protein design programs have evolved from the very simple function of designing positions in coiled-coils [24, 25] to designing catalytic activity into a previously inactive protein [30]. The force field has changed as well with the addition and development of terms. The most obvious first rule is that no atomic overlap will occur in a protein. This and dispersion effects are taken into account with a simple 12-6 van der Waals potential. Another well-known trait of proteins is that hydrophobic residues tend to be buried in the core of the protein away from the aqueous solvent. The van der Waals term alone was sufficient to create well-folded coiled-coil designs [24]. A coiled-coil is a very simple design because the residues occur in a set pattern, i.e. it is exactly known where a hydrophobic group should occur. Thus, even though only a van der Waals term was used to design the coiled-coil the algorithm was only choosing among other hydrophobic groups at the hydrophobic positions. The agreement between the calculated energies of the designed molecules and the experimentally measured molecules was improved by the addition of hydrophobic burial potential. The benefit for burial of hydrophobic residues was correlated to surface area; the larger the residue buried the more beneficial energy that choice would receive.

Continuing with the design of of coiled-coils but including the surface positions polar residues need consideration. Just as hydrophobic residues primarily exist in the buried core, polar groups exist on the exposed surface and importantly the burial of polar atoms in a nonpolar region is destabilizing. Dahiyat et al. [25] included a penalty for the burial of polar hydrogens not otherwise participating in a hydrogen bond and a term to benefit to formation of beneficial hydrogen bonds. An additional term included in this design was a statistical term representing helical propensity (taking into consideration that due to a variety of factors some amino acids are found more frequently in alpha helices). The molecules designed with these terms were significantly more stable than wildtype, showing

for the first time that optimizing the polar surface can lead to large stabilizations of a protein. To further deal with electrostatics, a simple Coulombic term was added to the force field [79]. With a high, distance-dependent dielectric (40r) the Coulombic term is primarily functional in minimizing close range destabilizing interactions. A surface area based term for dealing with polar solvation was developed as an alternative to the polar hydrogen burial term.

Previous implementations in ORBIT have included helix-propensity scales and simple beta-sheet propensity scales [93, 94]. These have broadly defined Ramachandran space into relatively large regions. However, high resolution treatment of Ramachandran space leads to increased accuracy of a secondary structure propensity term [56].

4.3. Least squares parameterization

4.3.1. Core hydrophobic-to-hydrophobic mutations

The first parametrization attempted was the simplest. The first successful designs using ORBIT used only a van der Waals term for mutations between hydrophobic amino acids, thus we looked at performance on hydrophobic-to-hydrophobic mutants in the core of the protein. Initially we compared the correspondence between the standard ORBIT force field with the measured experimental values (see figure 4.3). While there is clustering in the region of the figure where ORBIT predicts correctly the destabilizing mutations the large number and large values of the outliers contribute to an overall negative correlation between the experimental and predicted energies. A simple parametrization leads to a much improved correlation in this case (figure 4.4). The new parameters that result from the OLS minimization are shown in table 4.1. In excellent agreement with expectation the dielectric is lowered and the penalty for the burial of polar surface area is lowered. The hydrogen bond term is reduced to a nearly insignificant amount, which with the lowered dielectric value is in agreement with previous work [73].

A simple analysis of the points shows clustering of various types of mutations (figure 4.5). Figure 4.5(a) shows a cluster of mutations between Leu, Ile, and Val. These are small mutations that differ by at most one methylene group and the reparameterized ORBIT is

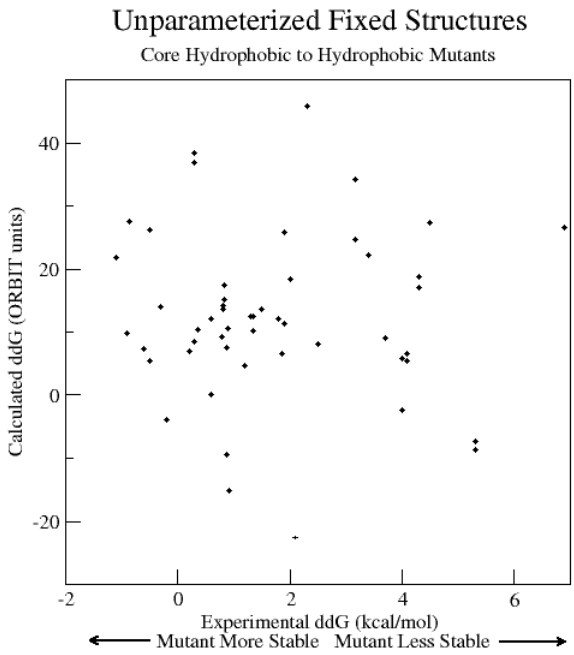


Figure 4.3. Unparameterized ORBIT force field compared with experimental data for protein core hydrophobic to hydrophobic mutations. 'Fixed' structures indicates that the mutant and wildtype structures have been chosen with regard to crystal packing, low alignment RMSD, and identical lengths.

	Old	New
van der Waals	1	1
Dielectric	40	12.5
Hydrogen bond	8	0.18
Non-polar burial benefit	0.026	0.026
Polar burial penalty	0.1	0.016

Table 4.1. Parameters obtained from hydrophobic-to-hydrophobic core mutants. Parameters are compared by normalizing the value of the van der Waals term to 1.

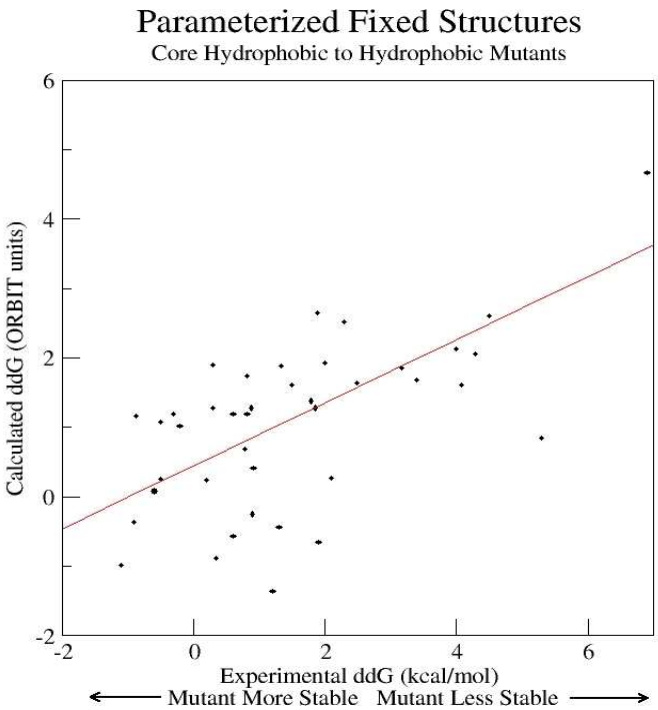
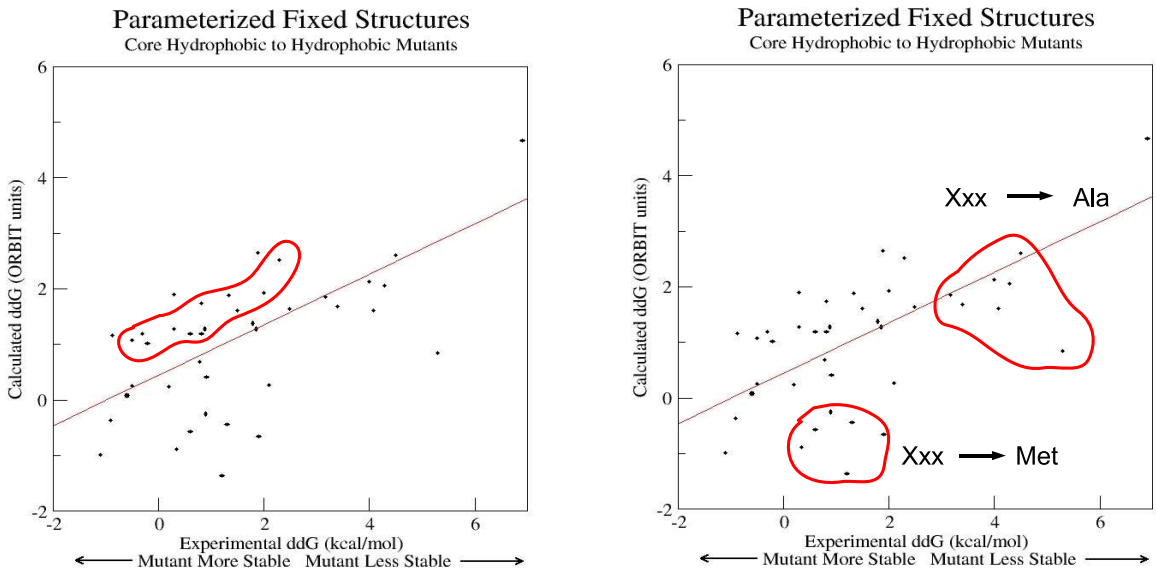


Figure 4.4. Parametrized hydrophobic to hydrophobic core mutants. Correlation coefficient, R = 0.61.

able to correctly predict the magnitude of the destabilizing mutations. The clusters in figure 4.5(b) show mutations to methionine and alanine from any other residue. ORBIT predicts both groups to be more stabilizing than experimentally observed. In the methionine case this could be due to an entropic effect as these are all mutations that occur in the core of proteins. When methionine is conformationally restricted by burial it suffers a large entropic penalty from a large loss in configurational freedom. Entropy loss for the various sidechains have been estimated in a number of studies [95, 96]. Using sidechain entropy values calculated by a self-consistent mean field approach in [96] the change in entropy due to the mutation is shown in figure 4.6. Again this shows some clustering, with the Xxx → Met mutations showing the expected entropic penalty and the Xxx → Ala mutations showing mostly beneficial entropic changes (not all, as there are some multiple mutant groups included in this set, with one mutation Xxx → Ala and another mutation). However, adding an entropic term to the standard ORBIT force field

$$\Delta\Delta G_f = B_1 * E_{vdw} + B_2 * E_{elec} + B_3 * E_{hbond} + B_4 * NP_{bur} + B_5 * P_{bur} + B_6 * ddS \quad (4.1)$$



(a) Mutations between Leu, Ile, and Val.

(b) Specific mutations.

Figure 4.5. Clustering of mutations.

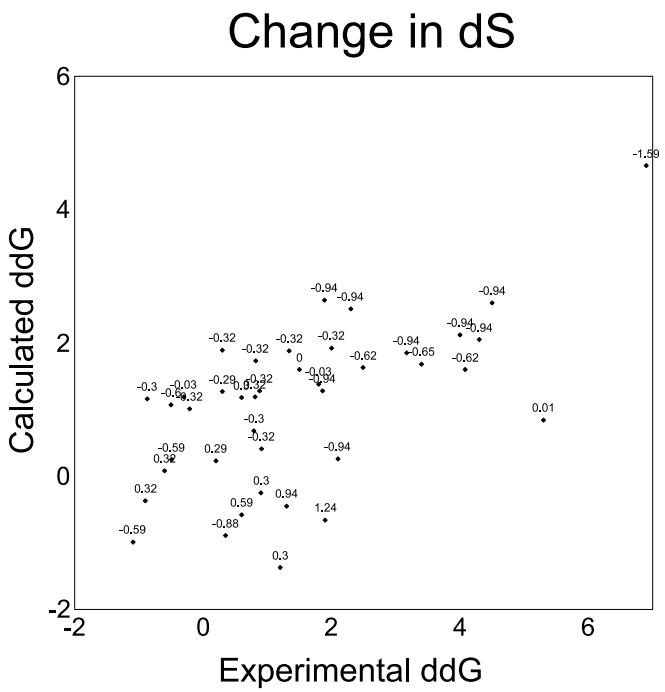


Figure 4.6. Change in entropy due to sidechain mutation.

and reparameterizing does not lead to any increased correlation between calculated ORBIT values and the experimental measurements.

While it seems reasonable to conclude that an entropic penalty for methionine burial should be included in the force field (see A) the deviation of the cluster of Xxx → Ala mutations requires further explanation. These mutations lead to the creation of cavities in the core of proteins which are experimentally known to destabilize proteins. There is debate as to whether this is a solvation effect due to loss of hydrophobic surface area or a packing effect due to loss of beneficial van der Waals interactions [2, 97]. The ORBIT nonpolar benefit term gives a uniform 26 cal mol^{-1} per $Å^2$ of nonpolar surface area buried in the folded state. This linearity of the surface area to energy relationship is doubtful [41]. To determine whether a more detailed treatment of nonpolar solvation in ORBIT would increase accuracy a measure of the relative hydrophobicities of the amino acids is needed. As hydrophobicity is just the preference of an amino acid for a nonpolar solvent over water, energies of transfer of the different amino acids from water to a hydrophobic medium are

Amino acid	ΔG_{trns}	Polarity
Arg	1.37	charged
Lys	1.35	charged
Asp	1.05	charged
Glu	0.87	charged
Asn	0.82	polar
Gln	0.30	polar
Ser	0.05	polar
Gly	0	polar
His	-0.18	polar
Thr	-0.35	polar
Ala	-0.42	nonpolar
Pro	-0.98	nonpolar
Tyr	-1.31	nonpolar
Val	-1.66	nonpolar
Met	-1.68	nonpolar
Cys	-1.34	nonpolar
Leu	-2.32	nonpolar
Phe	-2.44	nonpolar
Ile	-2.46	nonpolar
Trp	-3.07	nonpolar

Table 4.2. Free energies of transfer of the amino acids from water to octanol in kcal/mol. Adapted from [98].

used to quantify this preference. The core of a protein is not completely nonpolar and thus an accepted representation of the character of the protein core is octanol (but see [42] for an interesting discussion on this). The free energies of transfer of the amino acids from water to octanol are shown in table 4.2. All the mutants in this core hydrophobic-to-hydrophobic dataset are 100% buried, thus the full free energy of transfer is applicable (i.e. a complete transfer from aqueous environment to protein environment occurs). A comparison between the experimental free energy of transfer and the ORBIT nonpolar solvation term is shown in figure 4.7. While exhibiting the correct trend, the data shows a great deal of spread. Replacing the solvation terms in the ORBIT force field with the free energy of transfer,

$$\Delta\Delta G_f = B_1 * E_{vdw} + B_2 * E_{elec} + B_3 * E_{hbond} + B_4 * \Delta G_{trns} \tag{4.2}$$

does improve the correlation between ORBIT calculated energies and the experimental en-

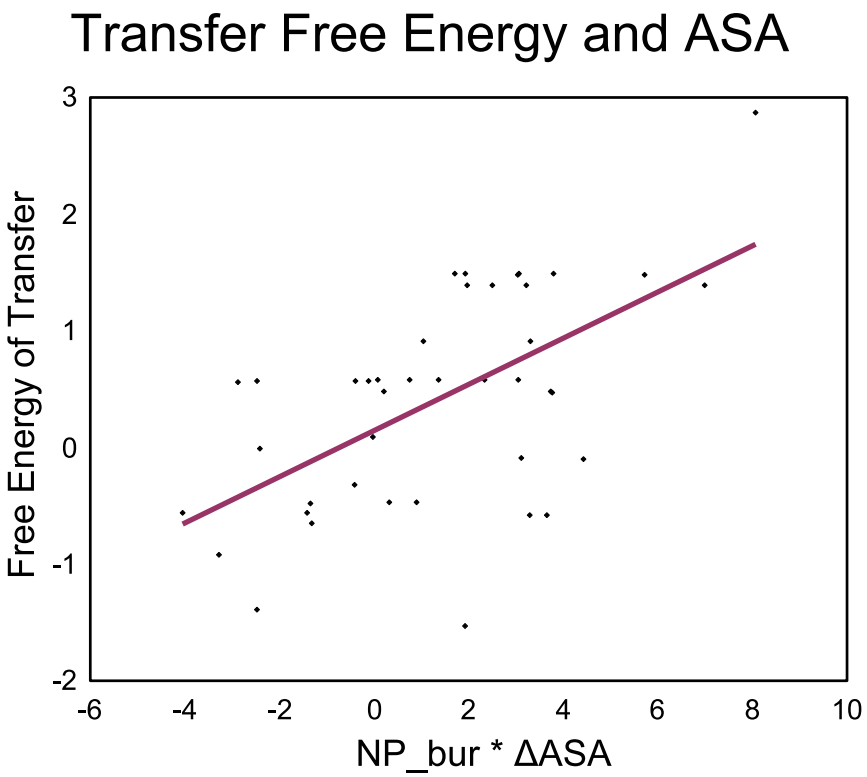


Figure 4.7. Comparison of free energy of transfer of the amino acids and non-polar benefit energy calculated by ORBIT. NP_bur is the ORBIT value for nonpolar benefit energy of 26 cal mol^{-1} per $Å^2$.

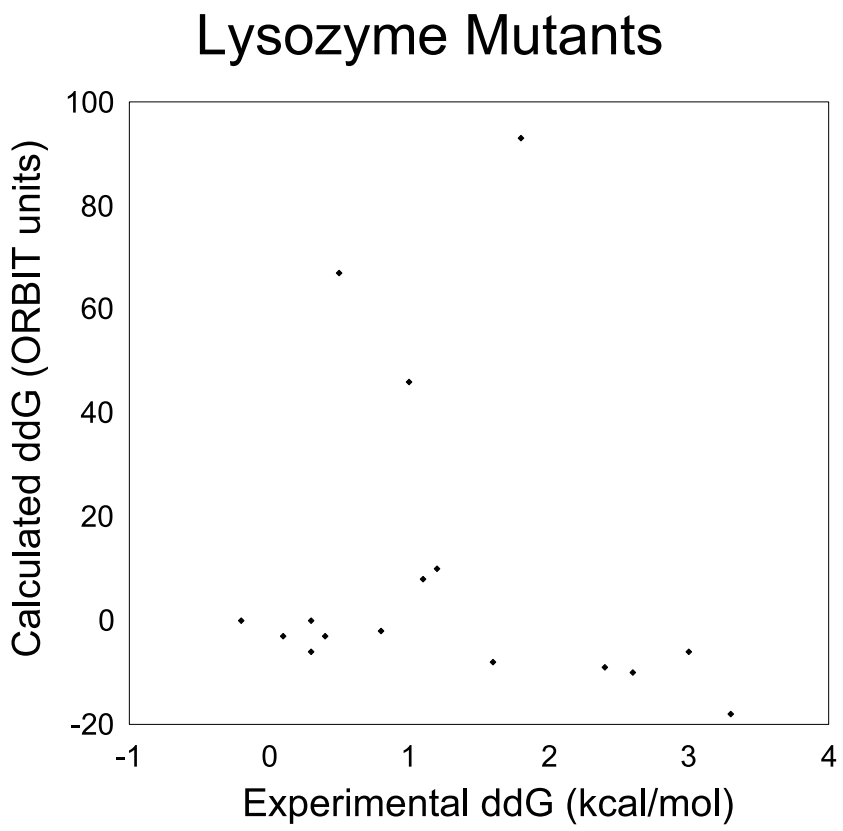


Figure 4.8. Comparison of calculated and measured energies in ORBIT redesigns of lysozyme core.

ergies. The measure of error is reduced 15% relative to the standard ORBIT force field.

A previous study (A, [99]) produced a number of new designs of the T4 lysozyme core. None were more stable than wildtype and the calculated energies of the sequences differed dramatically from the measured values once the sequences were expressed and characterized (figure 4.8). Calculating the energies of the sequences with the newly reparameterized ORBIT force field (table 4.1) leads to a closer approximation of the calculated and experimental values (figure 4.9). This indicates that the reparameterized force field is better predicting measured stabilities.

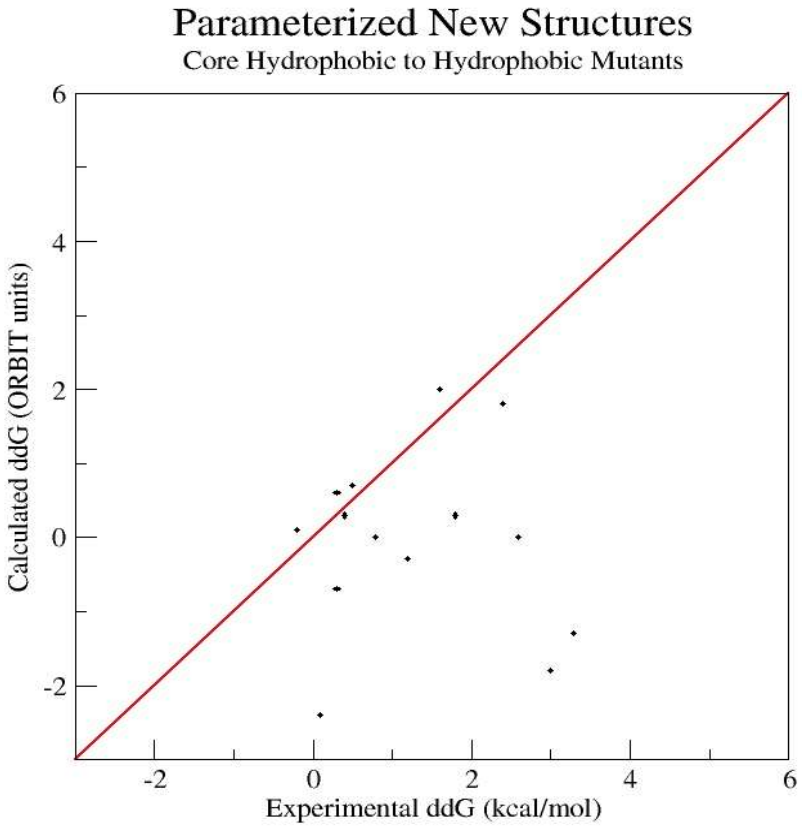


Figure 4.9. Lysozyme designs energies calculated with a reparameterized force field.

Term	Old	New
van der Waals	1	1
Dielectric	40r	3.7r
Hydrogen bond	8.0	0.31
NP_bur	0.026	0.116
Pol_bur	0.100	-0.004

Table 4.3. New parameters (normalized to 1 for van der Waals) for the standard ORBIT force field to better predict $\Delta\Delta G$ values of the entire dataset. The strength of electrostatics is increased by nearly a factor of 10, hydrogen bond energy is reduced to 0.31 kcal/mol, the benefit energy for nonpolar surface area is 4.5 times larger, and the penalty against polar burial is nearly eliminated.

4.3.2. Full dataset

4.3.2.1. Standard ORBIT force field

When all possible mutations in the database were evaluated the performance of the optimization significantly degraded. The data are obviously much more complicated with mutations that change polarity and hydrogen bonding patterns, large changes in sequence, and variability in protein environment. The initial unparameterized ORBIT force field performance on the dataset of mutants is shown in figure 4.10. A slight positive correlation exists between calculated and experimental $\Delta\Delta G$ for the destabilizing mutants (those to the right of 0 on the x-axis) but obviously the predictive value of the standard ORBIT force field is poor. An OLS optimization leads to an improvement in correlation (figure 4.11) but still far from an ideal behavior that would approach the trendline in figure 4.11. Nevertheless, the parameters from the optimization are interesting, dramatically changing the balance of terms in the force field (table 4.3). A negative weight on the polar burial penalty (P_bur) effectively turns this term into a very small benefit for the burial of polar surface area. This is partially due to mutations in the database that explored the benefit of burying polar groups in proteins [87] (e.g. the isosteric mutation val→thr). Effectively the OLS optimization cannot penalize the burial of polar surface area in general because there are cases where burial of polar surface area is the reason for a favorable energy of mutation. The relative contribution of the force field energy terms to the calculated energy is shown in table 4.4. The parameterization changes the benefit energy from relying primarily on van

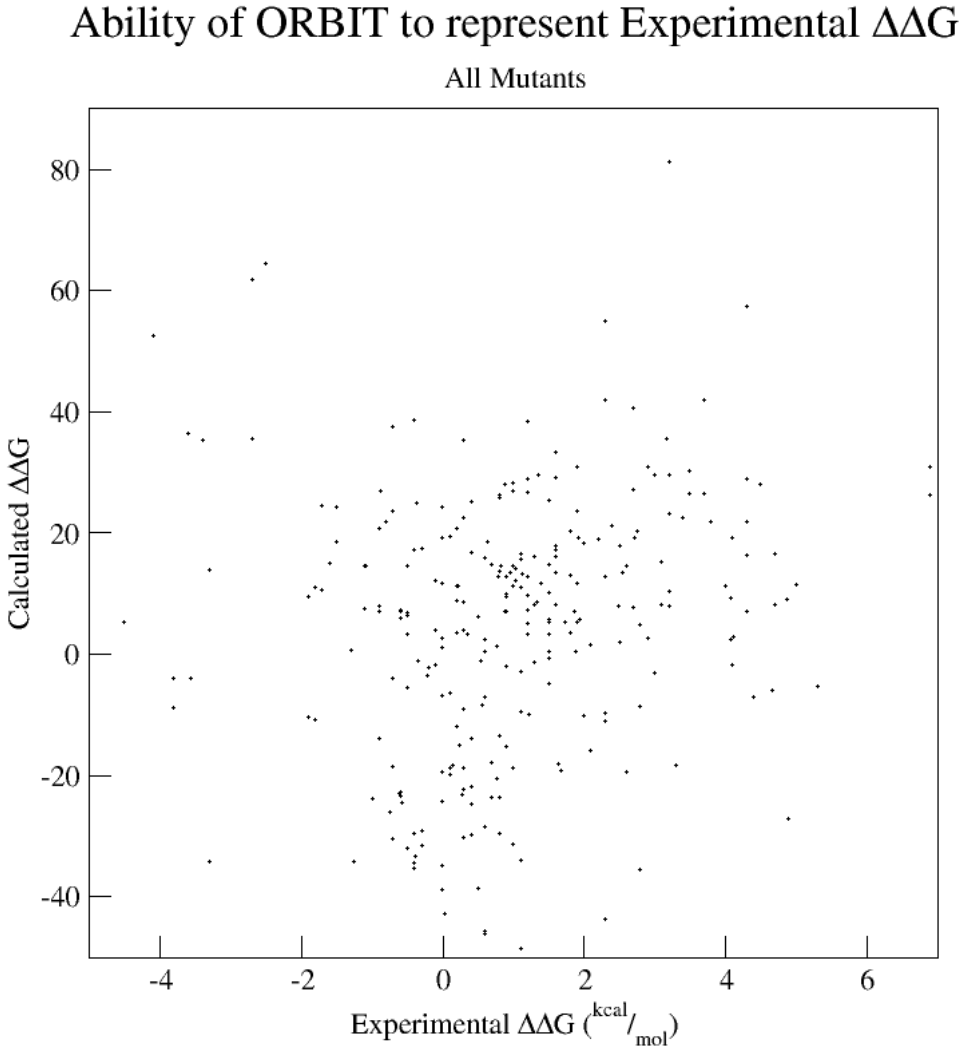


Figure 4.10. All mutations with unparameterized ORBIT force field.

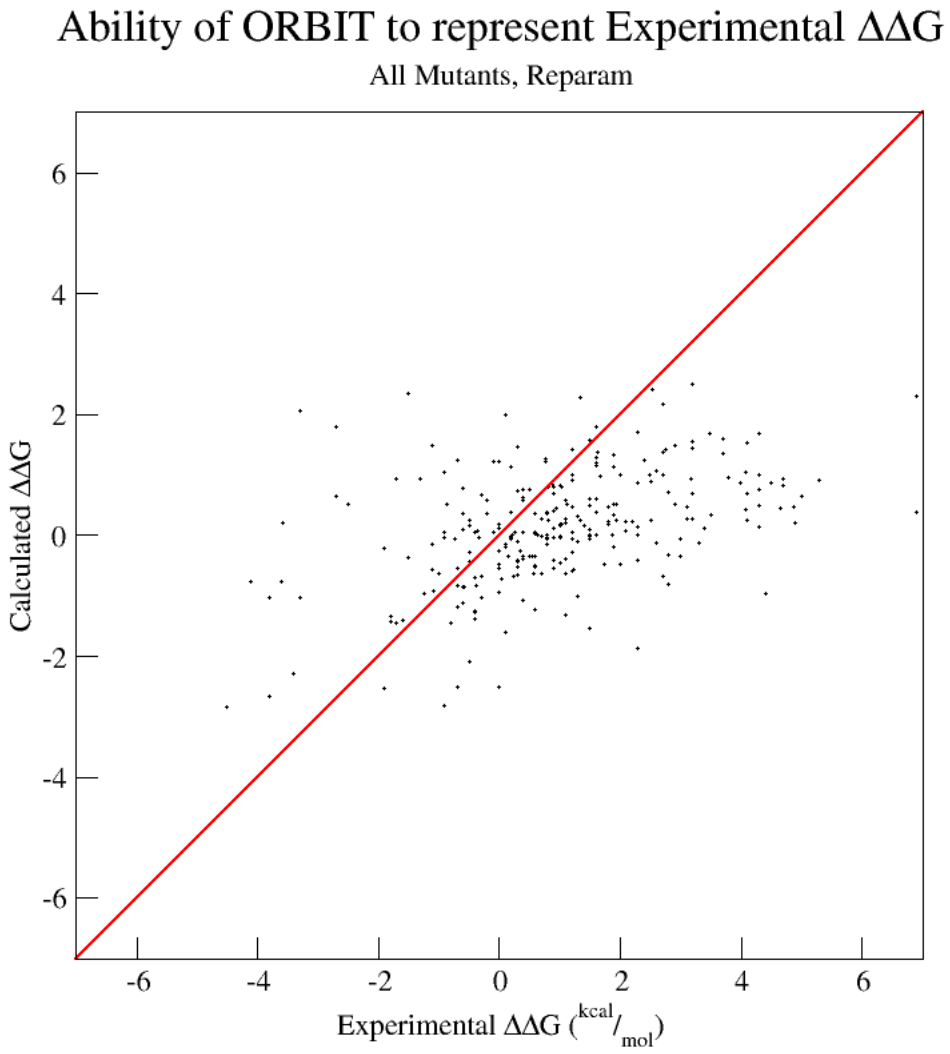


Figure 4.11. Parametrized ORBIT force field on full dataset. $\sigma = 3.6$. The line shown is *not* a correlation line, it is present only to guide the eye.

	Benefit Energy				Penalty Energy
	van der Waals	Electrostatics	Hydrogen bond	NP_burial	Polar burial
Standard	49%	2%	16%	33%	100%
Parameterized	21%	8%	1%	65%	-3%[a]
[a] In the parameterized case polar surface area is a 3% contributor to the benefit energy					

Table 4.4. The contribution of the force field terms to the calculated energy. In the standard ORBIT force field the polar burial term penalizes the burial of polar surface area and is nearly the same magnitude as the total benefit energy from the sum of van der Waals, electrostatics, hydrogen bonding and nonpolar burial. No penalty energy exists in the parameterized force field and the relative contribution of van der Waals and nonpolar burial is switched.

der Waals energy to a primary emphasis on nonpolar burial. This is notable because the van der Waals energy is beneficial when specific interactions between residues in the protein are optimal, whereas nonpolar burial is a very non-specific term.

4.3.2.2. Free energies of transfer

As with the set of core hydrophobic-to-hydrophobic mutations, free energies of transfer of the amino acids (table 4.2) were used as a representation of the solvation of the amino acids. Again, this is to determine whether a more detailed treatment of solvation (one value for each amino acid instead of just one value for hydrophobic surface area and one value for polar surface area) leads to an improvement in prediction accuracy. The set of mutations in the full database includes a variety of environments and many of the residues are less than fully buried. Therefore, these positions should not receive the full free energy of transfer value since they are not fully transferred from one environment to another. In this implementation the amount of solvation energy a residue receives is proportional to how buried the residue is in the folded protein. (This neglects differences in exposure in the unfolded state.) If a residue is 100% buried it will receive the full value from table 4.2. Full residues are evaluated, not individual atoms, thus a lysine residue that is 40% buried will be penalized 0.54 kcal/mol even if the terminal charged group is nearly fully exposed. In practice this is rarely observed: the terminal group for lysine usually has only a slightly smaller fraction buried than the residue as a whole. The force field with the free energy of

Term	Standard	Transfer
van der Waals	1	1
Dielectric	40r	6.9r
Hydrogen bond	8.0	-0.17
Transfer	NA	7.8

Table 4.5. Normalized parameters. The lowered dielectric again increases the strength of the electrostatic term in the force field. The hydrogen bond term is reduced to nearly zero. The results of these parameters on the calculated energy distributions is seen in table 4.6.

	Benefit Energy				Penalty Energy
	van der Waals	Electrostatics	Hydrogen bond	NP_burial	Polar burial
standard	49%	2%	16%	33%	100%
transfer	47%	9%	<1%	44%	100%

Table 4.6. Distribution of energy between the terms when parameterized with free energy of transfer as the solvation term. Nonpolar burial benefit is increased relative to the standard ORBIT distribution but not nearly as much as the reparameterized standard ORBIT (table 4.4). The polar burial penalties in both cases are the only penalty terms but in the standard ORBIT energy distribution the polar burial term nearly offsets the entire benefit energy while the polar burial energy when using free energies of transfer is only 18% of the magnitude of the benefit energy.

transfer solvation term is

$$\Delta\Delta G_f = B_1 * E_{vdw} + B_2 * E_{elec} + B_3 * E_{hbond} + B_4 * \Delta G_{trns} * \%buried. \tag{4.3}$$

The result of the parameterization is figure 4.12. The ability of equation (4.3) to predict experimental $\Delta\Delta G_f$ is improved slightly relative to the standard ORBIT force field (σ is improved to 3.3 from 3.6). The parameters obtained are in table 4.5 with the energy distribution between the terms in table 4.6.

As with the reparameterization of the standard ORBIT force field the strength of the electrostatic term is increased (by lowering the dielectric value) and the hydrogen bond term is decreased (in this case to a negative value, but effectively zero). In this case, however, the polar burial penalty term is not reduced to zero. By parameterizing equation (4.3) with only one value for the free energies of transform, solvation as a whole is weighted in the force field, but each individual free energy of transfer value is maintained (i.e. tryptophan retains exactly 3.07 kcal/mol more nonpolar burial benefit energy than glycine, as in table

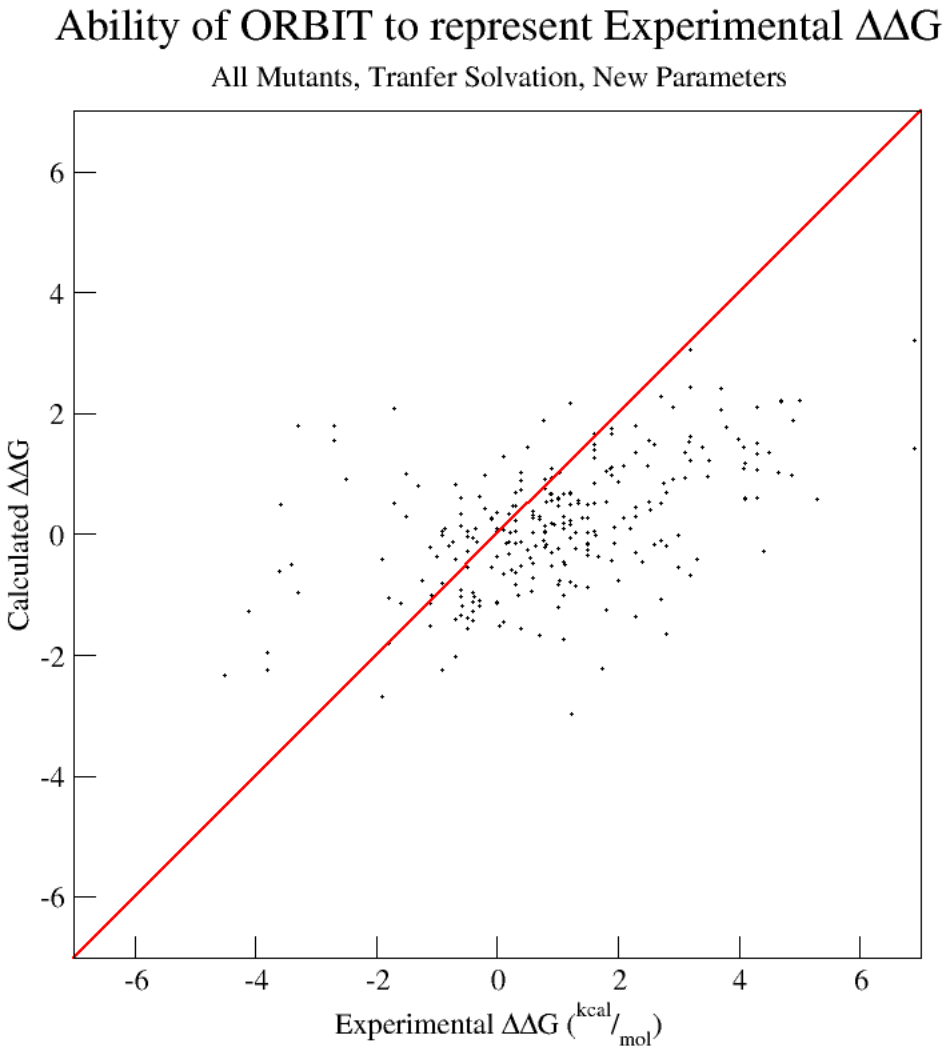


Figure 4.12. Values reparameterized using free energies of transfer to represent solvation. The standard deviation, $\sigma = 3.3$, a modest improvement over the standard ORBIT force field.

4.2). In effect, in order to get nonpolar burial benefit significantly weighted in equation 4.3 the polar burial penalty must "come along for the ride." This is also the reason why the nonpolar burial energy is a smaller proportion of the total benefit energy in this case than in the previous case when nonpolar burial benefit and polar burial penalty were parameterized separately – allowing the polar burial penalty to be nearly eliminated. While not reduced to zero, the polar burial penalty is still significantly smaller than in the standard ORBIT force field. The polar burial penalty in the standard ORBIT energy distribution is nearly equal and opposite the sum total of all the beneficial energy terms; with the free energy of transfer solvation term the polar burial penalty is roughly 18% of the magnitude of the total benefit energy.

4.3.2.3. Methionine penalty

Previous work [100, 63] has shown that a special consideration of methionine leads to improved agreement between calculated ORBIT energies and the measured stabilities of proteins. This is attributed to the significant entropic penalty methionine experiences when conformationally restricted in the folded protein. The core hydrophobic-to-hydrophobic set of mutations did not indicate that a general entropic term improved the correlation between ORBIT calculated energies and experimentally measured energies, but the full mutation dataset contains many more mutations involving methionine including one (1LWG) that has eight mutations to methionine. The protein 1LWG is also one of the most destabilized in the set at 4.9 kcal/mol.

This is a simple change to the force field – if a methionine is present that position receives a penalty, otherwise no difference.

$$\Delta\Delta G_f = B_1 * E_{vdw} + B_2 * E_{elec} + B_3 * E_{hbond} + B_4 * NP_{bur} + B_5 * P_{bur} + B_6 Met \quad (4.4)$$

The results of the parameterization are shown in figure 4.13, the new parameters in table 4.7, and the energy distribution in table 4.8. The parameterization of equation 4.4 leads to a significant improvement in the accuracy of the energy predictions (σ is 3.2, down from 3.6 for the force field without the methionine penalty). This large drop in error is largely

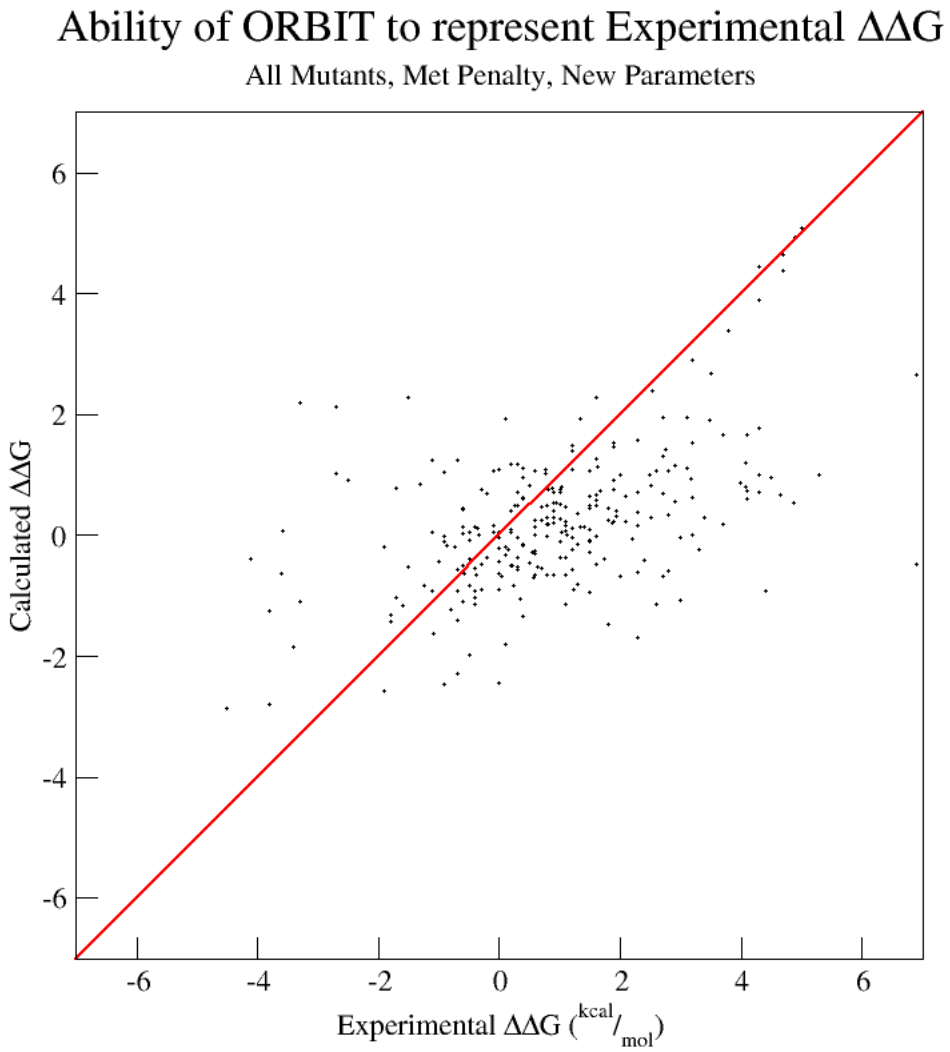


Figure 4.13. Result of reparamterizing the standard ORBIT force field with the addition of a methionine penalty. Due to the significant number of methionine mutations in the data set this leads to a significant improvement with the standard deviation, σ, decreasing from 3.6 to 3.2. A number of points with experimental $\Delta\Delta G$s greater than 4 kcal/mol (destabilizing) fall directly on the trendline now. These points are the multiple methionine mutants in the set.

Term	Standard	Plus Met Penalty
van der Waals	1	1
Dielectric	40r	6r
Hydrogen bond	8.0	-0.4
NP_bur	0.026	0.058
P_bur	0.100	0.011
Met Penalty	8	31.8

Table 4.7. The value for the methionine penalty used in previous work [100] was 8 kcal/mol. ORBIT units are relatively arbitrary, so a methionine penalty of 31.8 does not realistically imply a 31.8 kcal/mol penalty in reality (see Table 4.8 for the energy distribution). Unlike the standard ORBIT parameterization, the polar burial penalty is maintained at a reasonable value. The hydrogen bond term is again practically nullified.

	Benefit Energy			Penalty Energy	
	van der Waals	Electrostatics	NP_burial	Polar burial	Met Penalty
standard	49%	2%	33%	100%	0%
met penalty	36%	8%	55%	50%-100%	0%-50%
per residue	-0.26	tiny	-0.4	0.05	0.64

Table 4.8. Distribution of energy between the terms in the standard ORBIT force field and the with the addition of a methionine penalty. The hydrogen bonding column has been left out as it contributes very little in the new parameterization. In the protein with 13 methionines the total methionine penalty is nearly equal the polar burial penalty. The "per residue" row breaks down the energy by residue: the new parameters give -0.26 kcal/mol per residue for van der Waals, -0.4 kcal/mol per residue for non-polar burial, 0.05 kcal/mol per residue for polar burial, and 0.64 kcal/mol for each methionine. The entropic penalty for methionine is estimated to be 1.24 kcal/mol [96].

due to the number of multiple methionine mutants in the database that are significantly destabilizing. However, the parameters (table 4.7) and the energy distribution (table 4.8) show that the addition of the methionine penalty allows a more realistic value for the polar burial penalty than the parameterized force field without the methionine penalty (table 4.3), suggesting that the addition of this term does correctly model some behavior of the proteins. Additionally, while a methionine penalty of 31.8 may seem excessively large, this value is not in true kcal/mol. The parameterized energy distributions in table 4.8 show that on a per residue basis the methionine penalty is about 0.64 kcal/mol as compared with an average per residue value of the nonpolar burial penalty of -0.4 kcal/mol and an estimated value of the entropic penalty of methionine of 1.24 kcal/mol [96].

4.3.2.4. Lazaridis and Karplus (LK) solvation

The Lazaridis and Karplus [50] implicit solvent model for protein solvation leads to a large increase in detail in the description of the solvation model. The 17 atom types with three associated values (volume, ΔG^{ref}, ΔG^{free}) are parameterized within the model based on true thermodynamic data. The implementation of the LK model in ORBIT divides the atom types into a polar group and a nonpolar group that can weighted independently of the other. It is also possible to consider protein backbone ($b\_$) and sidechain/rotamer ($r\_$) atoms separately.

$$\Delta\Delta G_f = B_1 * E_{vdw} + B_2 * E_{elec} + B_3 * E_{hbond} + B_4 * b\_P_{bur} + B_5 * r\_NP_{bur} + B_6 * r\_P_{bur} \quad (4.5)$$

The results of the parameterization are shown in figure 4.14, the parameters in table 4.9, and the energy distribution in table 4.10. A large improvement is seen in the accuracy of the calculated energies as compared to the experimental energies and the standard deviation, σ, has dropped to 2.57. The correlation of the full dataset is measurable at R = 0.6, and if 5% of the data are excluded as "outliers" similar to Guerois et al. [89], the correlation coefficient improves to R = 0.71. The energy distribution in table 4.10 reveals that the backbone polar atoms contribute more to the total penalty energy than the sidechain polar atoms. The peptide backbone is primarily polar and a significant amount is buried by the surrounding

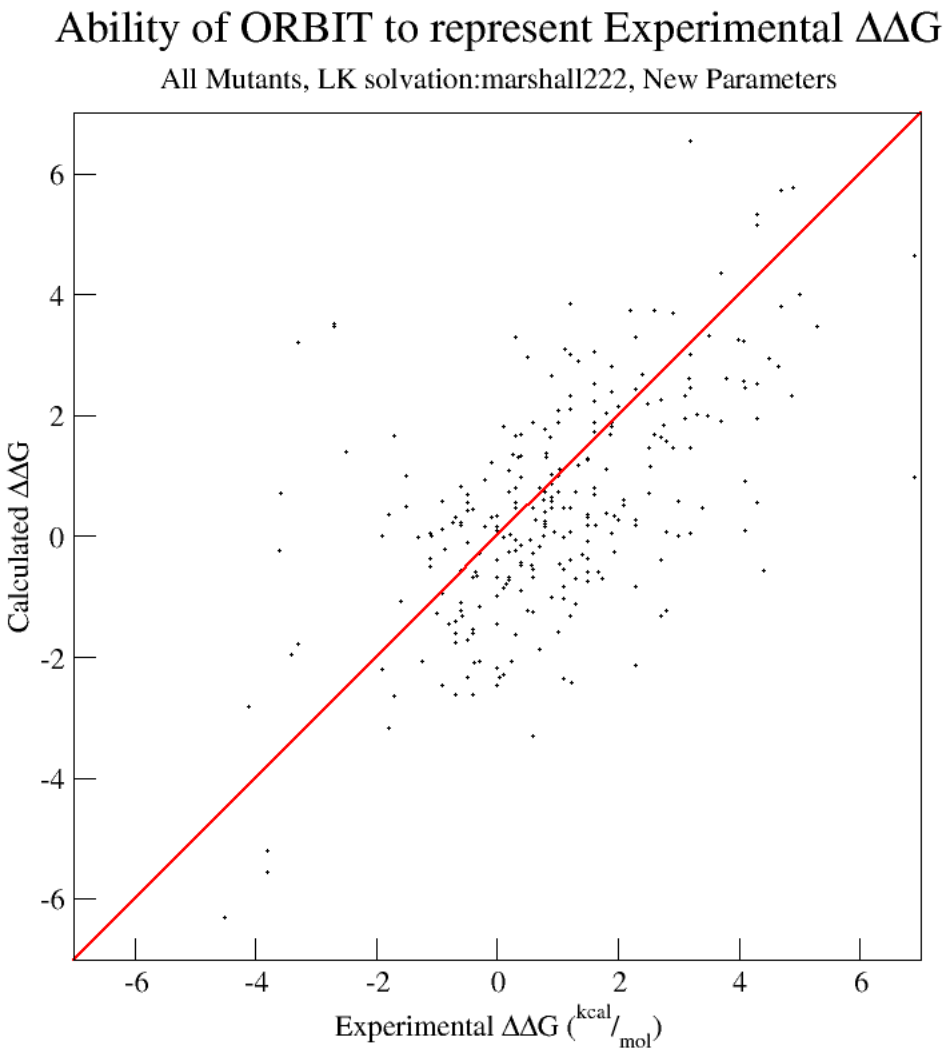


Figure 4.14. Noticeable improvement in correlation between calculated and experimental values. The standard deviation, σ, is reduced to 2.57. A line of best fit to the data (not shown) gives a correlation coefficient, R = 0.6. (If 5% of the data are removed at "outliers" as in Guerois et al. [89], the correlation coefficient improves to R = 0.71). LK parameter set used: marshall222.lkmodel.

Term	LK
van der Waals	1
Dielectric	7.9r
Hydrogen bond	-0.33
b_P_bur	15.7
r_NP_bur	14.3
r_P_bur	1.1

Table 4.9. Parameters obtained for LK model solvation. b_P_bur contains polar backbone atoms. r_NP_bur contains nonpolar rotamer atoms. r_P_bur contains polar rotamer atoms. Similar to previous parameterizations the polar burial penalty (for sidechains) is weighted less heavily. See table 4.10.

	Benefit Energy			Penalty Energy	
	van der Waals	Electrostatics	r_NP_bur	b_P_bur	r_P_bur
LK	30%	5%	63%	75%	25%

Table 4.10. Distribution of energy between the terms in the force field with LK solvation. Notable is the large fraction of the penalty energy that comes from burial of polar groups in the backbone. The total penalty energy is 54% of the magnitude of the total benefit energy. Hydrogen bond contribution is excluded as it is negligible.

sidechains. However, it is likely that the polar burial penalty on the rotamers (r_P_bur) is underweighted as seen in the previous surface area solvation parameterizations. The penalty energy from the two polar burial terms (b_P_bur and r_P_bur) is quite high at 54% of the magnitude of the total benefit bnergy, again from the high contribution of buried polar backbone atoms (that is not explicitly calculated in the surface area solvation methods.)

4.3.2.5. Failed to improve

Other terms and modifications were tried with the ORBIT force field that did not lead to an improvement in the ability of ORBIT to predict experimental $\Delta\Delta G$s. These include:

- Separation of Coulombic electrostatics into rotamer/rotamer and rotamer/template as in chapter 3 and Marshall et al. [73].
- Separation of hydrogen bonding into rotamer/rotamer and rotamer/template.
- Regional dielectrics based on RESCLASS classification. The regions of the protein are separated into core, boundary, and surface by RESCLASS and parameterized

separately to find three different dielectrics. It was expected to find a low dielectric value for the core, higher for the boundary, and highest for the surface. This did not occur.

- Different charge sets.
- Rotamer probability factors/self-energy. Rotamers that occur with a low probability are usually high-energy conformations. Introducing a probability term or a conformational self-energy term would account for this. The parameterization process used in this study is not the best way to test this type of term, as there is no choosing among rotamers.
- Secondary structure probability term as implemented in ORBIT (not a propensity term).
- Calculating 1,4-van der Waals interactions, using a repulsive-only van der Waals term, and not calculating van der Waals.

This does not prove that these terms would not be useful in protein design, just that they did not improve the correlation between calculated and experimental $\Delta\Delta G$s.

4.3.2.6. Distance restrictions

The difference in structure between the wildtype and mutant proteins in the database is very small. The Cα RMSD between them ranges from 0.06Å to 0.6Å. The energy calculations are quite sensitive to position however, and if any difference between structures is non-random this will influence the energy calculation. For example, the mutation Asp57Ala in protein 1E6M would be expected to lose electrostatic interactions, lose beneficial contacts, and experience a change in solvation energy (figure 4.15). The difference in calculated energies for the two structures would be expected to concentrate in the region surrounding the mutation. With the LK parameterization, this occurs to some degree (figure 4.16). The largest difference in calculated energy between the wildtype and mutant is at position 57 and results from loss of van der Waals and electrostatic interactions and a gain in favorable solvation energy from removing a buried polar hydroxyl group. However, at the 13 next

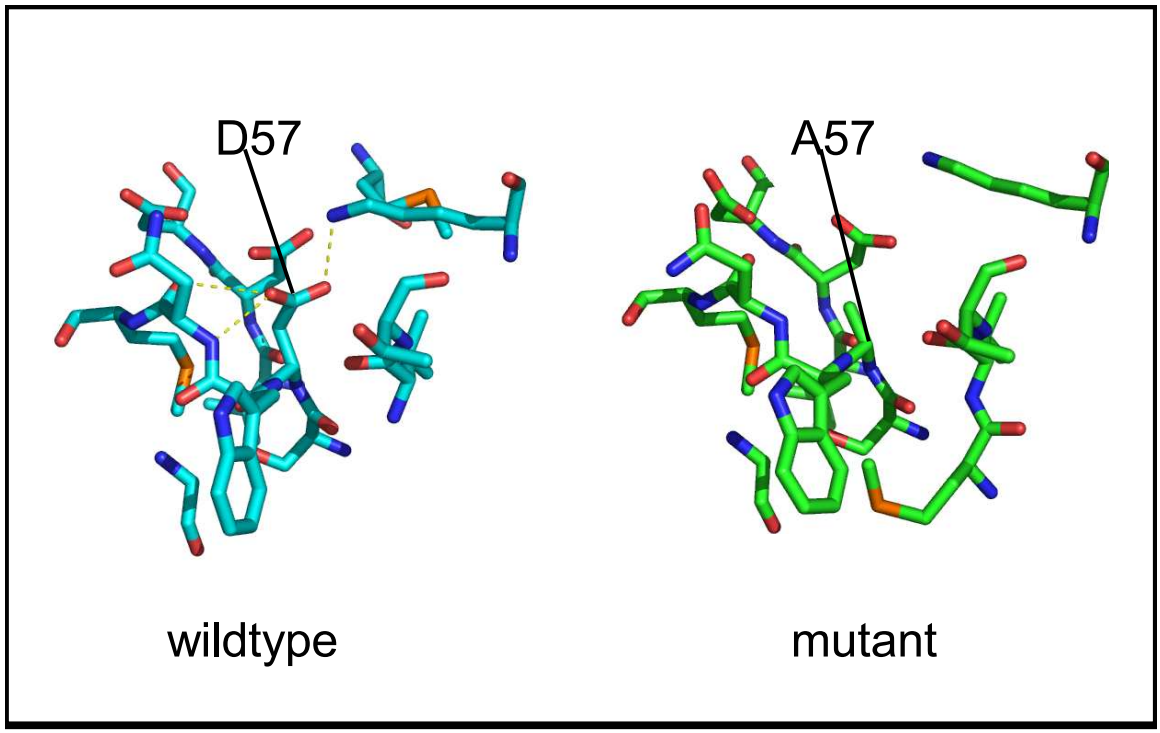


Figure 4.15. Loss of potential interactions in the partially buried D57A mutant in Che-Y.

pos	dvdw	dElec	dHbond	db_P_bur	dr_NP_bur	dr_P_bur	Total	abs(Total)
57	**0.1805**	**0.1648**	**-0.0221**	**-0.2198**	**-0.3475**	**-0.8004**	**-1.0444**	**1.0444**
75	0.0022	-0.2751	0.0162	-0.0413	-0.0315	-0.0884	-0.4179	0.4179
127	0.0721	0.0000	0.0000	-0.0474	0.3131	-0.0140	0.3238	0.3238
106	**0.2827**	**-0.0066**	**0.0000**	**-0.6015**	**0.1723**	**-0.1382**	**-0.2915**	**0.2915**
126	0.0043	0.1231	0.0084	0.0168	0.0297	0.0763	0.2585	0.2585
67	0.0793	0.0578	-0.0154	-0.2039	0.0551	-0.2309	-0.2581	0.2581
78	-0.0560	-0.0369	0.0000	0.0095	-0.2010	0.0558	-0.2286	0.2286
3	0.1111	0.0050	0.0164	-0.1909	0.0471	-0.2091	-0.2204	0.2204
26	-0.0257	0.1080	-0.0049	0.0322	0.0131	0.0947	0.2174	0.2174
22	0.0192	0.1454	0.0060	0.0055	0.0182	0.0118	0.2061	0.2061
64	**0.0390**	**-0.0893**	**-0.0092**	**-0.0217**	**0.0195**	**-0.1308**	**-0.1926**	**0.1926**
123	0.0247	0.0000	0.0000	0.0013	0.1613	0.0034	0.1908	0.1908
84	0.0458	0.0000	0.0000	-0.0338	0.1759	-0.0042	0.1837	0.1837
7	0.0121	-0.0350	0.0006	0.1126	0.0299	0.0566	0.1769	0.1769

57	**Point Mutant**
75	Non-local, different backbone conformation in loop
127	Non-local, leucine in a different conformation
106	**Local, tyrosine 106 in different conformation**
126	Non-local, different backbone conformation in loop
67	Non-local, glutamate in different conformation on surface
78	Non-local, methionine in different conformation
3	Non-local, N-terminal disorder
26	Non-local, lysine in different conformation on surface
22	Non-local, arginine in different conformation on surface
64	**Local, aspartate 64**
123	Non-local, isoleucine 123
84	Non-local
7	Non-local, lysine in different conformation on surface

Figure 4.16. Distribution of energy values around the site of mutation, position 57. These are difference measurements, thus dvdw is the difference in van der Waals energy at the position between the wildtype and mutant. Negative values are more favorable interactions for the mutant. D57A shows the largest difference in calculated energies between the structures, as expected. 11/13 of the next largest calculated energy differences are non-local and potentially spurious results. db_P_bur, dr_NP_bur, and dr_P_bur are the differences in calculated energies of backbone polar burial, rotamer non-polar burial, and rotamer polar burial.

(a) **LK Parameterization**

	4Å	6Å	8Å
dielectric	4.5r	8.7r	4.8r
hydrogen bond	0.2	-0.03	0.04
NP_bur	6.1	6.3	7.4
P_bur	0.9	1.0	1.0
	σ=2.01	**σ=2.0**	**σ=2.06**

(b) **Only Surface Mutations**

	4Å	6Å	8Å
dielectric	2.1r	2.9r	1.0r
hydrogen bond	3.9	0.5	3.1
NP_bur	20	12	44
P_bur	0.9	0.5	-1.4
	σ=1.34	**σ=1.42**	**σ=1.47**

(c) **Core/Boundary Mutations**

	4Å	6Å	8Å
dielectric	5.5r	6.2r	8.2r
hydrogen bond	0.2	0.6	0.8
NP_bur	5.1	5.4	5.8
P_bur	1.0	1.1	1.1
	σ=2.27	**σ=2.13**	**σ=2.06**

Figure 4.17. LK parameterization results when only considering the energy of interactions within 4Å, 6Å, or 8Å. (a) Parameterization using all points in the database, similar to section 4.3.2.4 but with a distance restriction. (b) Parameterization using only mutations that occur at the surface of the protein. (c) Parameterization using only mutations that occur in the core or boundary regions of the protein. P_bur includes b_Pol_bur and r_Pol_bur for a general polar burial term.

largest calculated energy differences only two would likely be involved in energy differences resulting from the mutation. In order to explore potentially complicating non-local effects the effect on distance from mutation site was evaluated.

The results of the distance studies with the LK solvation model are shown in figure 4.17. When compared with the unrestricted full dataset parameterization in section 4.3.2.4 the strength of the nonpolar burial benefit is reduced relative to the other terms (leads to nonpolar burial energy of the same magnitude of the van der Waals energy). The 6Å restriction has a slightly smaller measure of error than the 4Å. The mutant and wildtype

structures have the most variation unrelated to the region of mutation on the surfaces. Increased flexibility of surface sidechains and occasionally crystal packing artifacts leads to different conformations of sidechains that are due more to chance than true energetic reasons. The parameterization of mutations that occur on the surface of the protein shows large variability between the different distance restraints. This suggests that the conformational noise of the surface is influencing the parameters. (The low standard deviations (σ) are do both to the lower number of points in the set and the generally lower effect that mutations have on the surface of proteins.)

The core/boundary mutations in figure 4.17c lead to parameterizations that are relatively stable between the different distance sets. Interestingly, in the 4Å set the balance between van der Waals and nonpolar burial is the most favored to van der Waals in the entire study. When conformational noise is significantly reduced by only looking at the nearest neighbors, the specific benefit term (van der Waals) is favored over the non-specific benefit term (nonpolar burial).

4.3.2.7. Design tests

The goal of the entire parameterization study is to find parameters and terms for ORBIT that improve the stability of designed sequences. To this end a number of designs using parameters from the least squares optimizations were performed (figure 4.18).

These designs are all performed on engrailed homeodomain (figure 3.1). Initial design attempts immediately showed the limitations of the parameterizations. A full design that allows any amino acid at every position performs miserably (not shown). This is known behavior of the standard ORBIT force field with the default experimentally tested parameters. This is also the reason for such negative design terms as the nonpolar exposure penalty that attempts to maintain more natural sequences by reducing the number of hydrophobic groups on the surface [26]. But such negative design terms are non-thermodynamic and are not included in an optimization to thermodynamic measurements.

The results shown in figure 4.18 are split into three sections with the first box (NC0, NC3_Ncap, and dielec_H) containing known results form previous surface designs of engrailed. The second box (SA_default, SA_surf, LK_surf_control, LK_surf) show surface

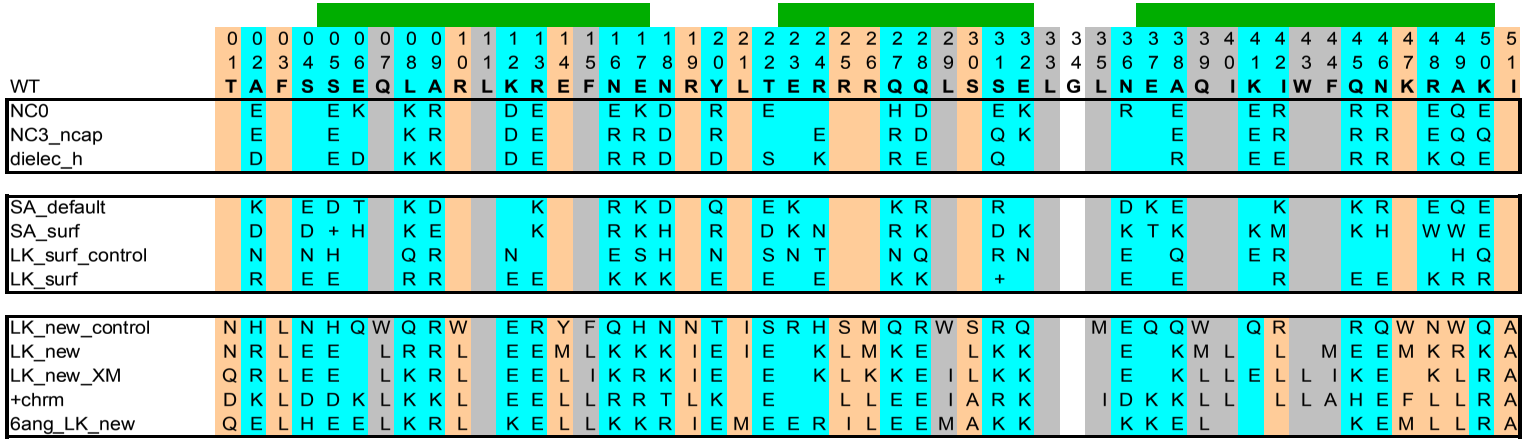


figure 4.18 Design results of engrailed homeodomain using new parameters. The column color coding indicates the position in the protein: grey is core, orange is boundary, and cyan is surface. The green bars across the top indicate the positions of the helices. The first set of boxed sequences shows sequences of the surface of engrailed homeodomain previously designed and experimentally verified (NC0, NC3_Ncap, and dielec_h).The second box includes surface redesigns of engrailed using standard ORBIT surface area solvation (including the hydrophobic exposure penalty). The surface designs use a new rotamer library e2QERK0 (all residues are expanded about chi1 and chi2 except Q, E, R, and K, which are not expanded). Parse charge sets and marshall222.lkmodel internal parameters are used. The SA_surf parameters used are dielectric = 3.7r, hbond = 0.31, NP_bur = 0.116, and P_bur = -0.04 from Table 4.3. LK_surf_control used NP_bur = 1.0, P_bur = 1.0, rest standard ORBIT. The other LK sequences used reparameterized values: NP_bur = 6.3, P_bur = 1.0, hbond = 0.2, and dielectric = 4.6r.

designs using parameters obtained from the least squared minimizations. The last box shows the results form restricted full designs: designs that allow only hydrophobic residues in the core and only polar and charged residues on the surface. Boundary positions are allowed to choose residues from either group.

The standard ORBIT force field (surface area solvation, nonpolar exposure penalty) used in the surface design is SA_default. It is weighted towards large charged amino acids (lys, glu, arg) because they can have more favorable contacts (van der Waals) as well as bury more hydrophobic surface area. The reparameterized surface area force field that in weighted heavier towards nonpolar burial shows similar results. The default LK (LK_surf_control) results are dramatically different with a large number of asparagines on the surface. This is due to a significantly different distribution of energy between the terms (figure 4.10) that reduces the nonpolar burial benefit allowing relatively more favorable electrostatic interactions of which asparagine can take advantage. The reparameterized LK, LK_surf, heavily weights nonpolar burial and results in a completely unnatural surface composed nearly entirely of arginine, glutamate, and lysine.

Full designs, with sequence restrictions on core, boundary, and surface are shown in the last box. These are to more fully explore the effects of the new LK parameters on the rest of the protein. The nearly exclusive use of van der Waals in the standard LK force field leads to a number of tryptophans in the core due to their large size alone. The parameterizations of LK force fields all show similar surface behaviors as the surface only design (glutamate, arginine, lysine) and the cores are decently packed with primarily leucine and isoleucine. The addition of a methionine penalty term (LK_new_XM) removes a buried methionine, and using charmm19 [50] internal LK paramters instead of marshall222 leads to a slightly more varied surface (a few aspartates included). However, the true effect of the new parameters is observed with the boundary positions (orange in figure 4.18) that can choose any residue (polar or nonpolar). Nearly all are nonpolar due to the emphasis on nonpolar burial in the reparameterized force field models.

4.4. FOLDX implementation

In order to compare our results of prediction of free energies of mutation with published work of a similar nature the FOLDX force field of Serrano and co-workers [89] was implemented as a module in ORBIT. In many ways the FOLDX force field is simpler than the ORBIT force field and requires only the calculation of energies of the existing sidechains of a protein. No design is required and only the single body terms from SETUP were required to implement FOLDX. Their approach is simple and they make the dubious assumption that terms and parameters derived from fitting experimental energies of mutation have any application in protein design. This detracts from their claim that the results will be useful in future protein design efforts.

4.4.1. The FOLDX force field

4.4.1.1. Occlusion based solvation and van der Waals

Occlusion is used to represent the environments of the protein and to scale the contribution of solvation and van der Waals to the overall calculated energy. The standard form [49, 48] of the occlusion equation is used

$$occ(i) = \sum V_j \frac{-d^2}{e^{2\sigma^2}}. \tag{4.6}$$

But a further development is made by defining a scaled fraction called the sfactor,

$$sfactor(i) = \frac{occ(i) - occmin(i)}{occmax(i) - occmin(i)} \tag{4.7}$$

which leads to a representation of atom burial in a protein as a fraction between 0 and 1. In the form of equation 4.7 an sfactor value of 1 indicates an atom that is fully buried, i.e. is maximally occluded. A maximally occluded atom has full van der Waals and full desolvation values.

4.4.1.2. Electrostatics

Electrostatic interactions are only calculated between charged residues, thus only aspartate Oδ1 and Oδ2, glutamate Oϵ1 and Oϵ2, and lysine Nζ are included. No hydrogens are used in the force field. The dielectric used in the Coulombic equation is scaled by the sfactor.

4.4.1.3. Hydrogen bonding

Backbone-backbone hydrogen bonds are calculated. All polar residues have the potential to hydrogen bond if less than 3.6 Å distance apart and satisfy crystallographic-base angle restraints. Polar-polar hydrogen bonds are given 1.3 kcal/benefit and polar/charge hydrogen bonds receive 1.4 kcal/mol benefit.

4.4.1.4. Entropy

Two separate entropy functions are used. One is basically a secondary structure term that is described as a penalty for fixing the backbone based on the sidechain identity. The other is the entropy loss of a sidechain upon protein folding. This is estimated by scaling the full entropy loss for a sidechain calculated in [95] by the sfactor.

4.4.1.5. Other terms

If atoms are too close in space a van der Waals clash score is used as a penalty. The backbone entropy is scaled down if the backbone dihedrals are in a loop region of Ramachandran space. Also used is a term to reduce solvation if water networks are predicted to be involved.

4.4.1.6. Parameterization

Parameterization of the FOLDX force field by Guerois et al. [89] began with the conservative hydrophobic-to-hydrophobic mutants in their already restricted set (see section 4.1) and optimized the van der Waals, hydrophobic solvation, and sidechain entropy terms. On the full set of mutations these parameters were held constant and the other terms in the force field were parameterized with a grid search. The distribution of energy in the parameterized FOLDX force field is in figure 4.19.

	Non-Polar		Polar, hbonded	
	Buried	Exposed	Buried	Exposed
Nonpolar solvation	48.4	35.2	15.9	15.7
Polar solvation	11.2	19.7	26.9	23.2
van der Waals	21.1	18.8	9.2	8.6
Hydrogen bond	-	-	28	25.7
Electrostatic	-	-	2.8	5.4
Main-chain entropy	7.1	17.9	3.3	4.3
Side-chain entropy	10.5	6.0	11.0	15.1

figure 4.19 Distribution of energy in the FOLDX force field, separated by polarity and burial.

4.4.2. Results

Results are summarized in figure 4.20. FOLDX does not perform any better than the LK parameterization that has a correlation coefficient of R = 0.71 on the full set (no outliers removed). FOLDX does not have any advantage in predicting free energies of mutation. Indeed, its performance is increased by using the ORBIT hydrogen bonding potential instead of the FOLDX hydrogen bond potential.

4.5. One-body wildtype optimization

4.5.1. Introduction

We have found that parameterization of a force field for protein design requires more than the ability to predict the energies of the sequences of existing proteins. The force field must also discriminate among all possible sequences. As discussed in Chiu and Goldstein [101] the best sequence energetically for a given structure may not fold to that structure and equivalently, an optimized chemical-physics potential may not be optimal for protein design. Using model systems they were able to show that conformations of sequences generated with an exact energy function did not lead to recovery of the same sequences on those conformations with the same energy function. Instead, a different energy function was required to design the appropriate sequence for the structures.

What is needed then is an optimization strategy that leads to selection of an optimum sequence with exclusion of other sequences. The exclusion of sequences is negative design,

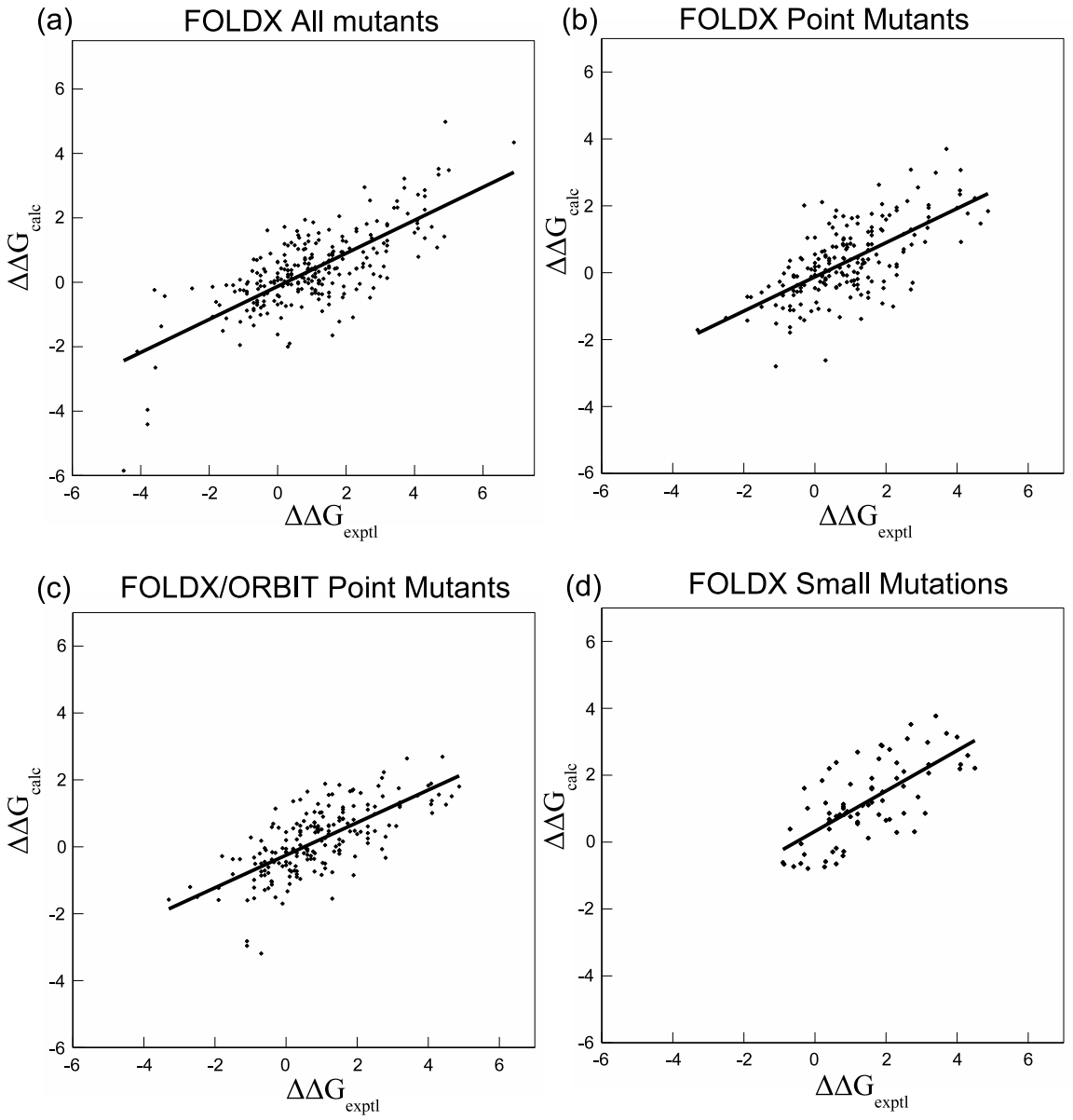


figure 4.20 Results of FOLDX force field on $\Delta\Delta G$. All plots have removed the 5% worst-fitting points as "outliers" to agree with Guerois et al. [89]. (a) FOLDX on all mutants in the data set, R = 0.74. (b) FOLDX on only the point mutants, R = 0.66. (c) FOLDX on point mutants with ORBIT hydrogen bonding potential, R = 0.7. (d) FOLDX (with ORBIT hydrogen bond) on small conservative mutants as in Guerois et al., R = 0.73.

as the goal is not to design *for* a specific sequence but *against* all others. Dill and colleagues [40] found that a combination of positive and negative design led to the selection of the optimal sequences in model systems. The penalty against exposure of hydrophobic surface area in ORBIT is based on a similar term in Dill's work. But to optimize for one sequence against all others it is necessary to know the identity of that one sequence. Early work with model systems [102, 103, 101], could discover the optimal sequence by enumeration of all possible combinations. Non-optimal sequences could then be generated.

The goal of protein design is to generate the optimal sequence for a given structure. Thus, we want to optimize the protein design strategy for an answer that is unknowable beforehand. A first approximation is given by Kuhlman and Baker [104]: the wildtype sequence for any protein is "close to optimal" for that structure. Many sequences were generated for a variety of backbones and 51% of all core positions and 27% of positions overall were identical to wildtype. While their argument is inherently circular (a force field optimized to give wildtype sequences is used to get statistics on wildtype recovery) there is experimental justification for the claim. Designs of protein cores consistently maintain significant identity with wildtype [26, 13] including the inability to design more stable cores of T4 lysozyme [100]. While there are cases where proteins have been mutated to greater stability [27, 105, 73], in general it is probably true that the most stable sequence for a given structure will be similar to the wildtype sequence.

If most natural protein sequences are close to optimal for their structures than those most likely to be closest to optimal are those proteins that are stable at high temperatures. Proteins are not selected by evolution to be absolutely stable. Function, folding kinetics, and neutral genetic drift all factor into the sequence choice for a protein that may interfere with optimal stability. Thermophiles have little room for error in sequence as any destabilization may lead to denaturation and destruction of the protein. Nearly all proteins have a temperature of maximum stability near room temperature. Thermophiles tend to be more stable at room temperature with a wider range of temperature where they maintain a folded structure [106] (figure 4.21). A collection of thermophilic proteins would be an excellent set to use as optimum sequences. Single domain, two-state thermophilic proteins of known structure are shown in table 4.11. Optimizing the force field to recover the sequences of the

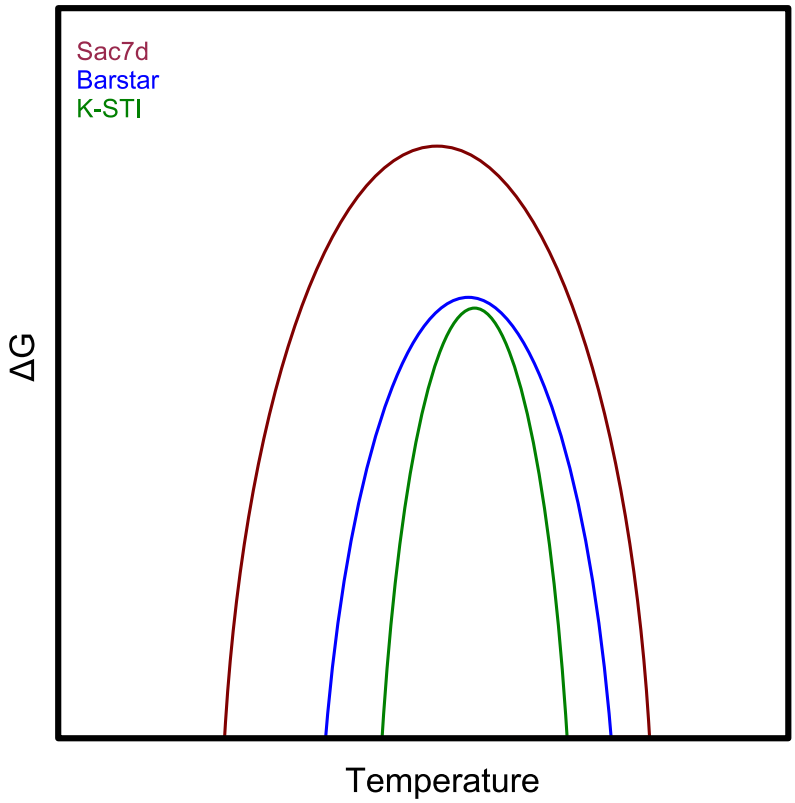


figure 4.21 Protein stability curves. The maximum point of the curves is the maximum stability of the protein. Where the curves cross the x-axis is the point of zero energy, or denaturation. Proteins that denature at higher temperatures have broader curves. Adapted from [107].

Name	Melting Temp	Size	PDB
Sac7d	90.7	66	1SAP
BsHPr	73.4	87	2HID
Barstar	72.7	90	1A19
GDH Domain II	69.6	150	1B26
ADA2h	77	80	1AYE
BcCsp	86	70	1C9O
Ubiquitin	70	72	1UBQ
Protein G	81	50	1PGA

Table 4.11. Thermophilic proteins.

thermophiles over all other possible sequences will lead to a force field optimized for protein design.

4.5.2. Implementation

A procedure similar to the method described in Kuhlman and Baker [104] was implemented as a module in ORBIT. Similar to the FOLDX force field described in section 4.4, only one-body terms need to be calculated. At each position the amino acid is mutated to all possible amino acids and the energy evaluated. By adjusting the parameters in the force field we attempt to maximize the function

$$Q = \frac{1}{N} \sum^{N} \frac{\exp\left(-E\left(aa_{wt}\right)\right)}{\sum_i \exp\left(-E\left(aa_i\right)\right)} \tag{4.8}$$

where $E\left(aa_{wt}\right)$ is the energy of the wildtype residue at the position, $E\left(aa_i\right)$ is the energy of each amino acid, N is the number of amino acids in the protein. If the wildtype sequence is highly favored at every position Q will approach 1. The parameters in the force field are evaluated with a grid search.

4.5.3. Results

The results of the one-body optimizations are shown in figure 4.22. For the most part the resulting parameters are in agreement with the standard ORBIT parameters. Thousands of different combinations of parameters are tested with the grid search for each set of terms and many combinations lead to roughly the same value of Q (figure 4.23). Wildtype recovery is approximately 50% for the highest scoring values during optimization. figure 4.22c includes additional terms including an occlusion-based entropy similar to the sidechain entropy term in FOLDX (section 4.4.1.4), a rotamer probability term to reduce high energy conformations, and a secondary structure propensity value. These additional terms have relatively small contributions to the overall energy relative to the van der Waals and solvation terms.

(a) Standard ORBIT

Term	Default	New	Range (10% Qmax)
van der Waals	-	-	-
dielectric	40r	2r	1 – 4
hydrogen bond	8	16	12 – 16
NP_burial	0.026	0.034	0.026 – 0.034
NP_exposure	1.6	1.7	1.2 – 1.7
Pol_burial	0.1	0.06	0.05 – 0.08

(b) ORBIT with LK

Term	Default	New	Range (10% Qmax)
van der Waals	-	-	-
dielectric	40r	2r	2 -6
hydrogen bond	8	12	8 -18
NP_burial	1	4.8	1.0 – 6.0
NP_exposure	1	0	0 – 2
Pol_burial	1	0.8	0.1 – 1.2

(c) Additional Terms

Term	Default	New	Range (10% Qmax)
van der Waals	-	-	-
dielectric	40r	5r	3 - 7
hydrogen bond	8	7	5 – 10
NP_burial	0.026	0.034	0.026 – 0.042
NP_exposure	1.6	1.7	1.2 – 1.7
Pol_burial	0.1	0.04	0.03 – 0.06
Entropy	-	4	3 – 6
Probability	-	1	1 – 2
Propensity	-	4	4 – 6

figure 4.22 Results of the one-body optimizations. (a) Optimization using the standard surface area solvation. The new parameters lead to Q = 0.37. (b) Optimization with LK solvation. The new parameters lead to Q = 0.34. (c) Standard ORBIT plus an occlusion-based entropy term, a rotamer probability term, and a secondary structure propensity term. Q is increased to 0.39.

Q	Elec	Hbond	NP_bur	NP_exp	Pol_bur
37.26	2	16	0.034	1.7	0.06
37.25	3	16	0.034	1.7	0.05
37.25	2	16	0.034	1.7	0.08
37.21	3	16	0.030	1.7	0.05
37.21	3	14	0.034	1.7	0.05
37.19	2	16	0.034	1.5	0.06
37.18	2	16	0.030	1.7	0.08
37.17	2	14	0.034	1.7	0.06
37.16	2	16	0.034	1.5	0.08
37.14	3	14	0.030	1.7	0.05
37.14	3	16	0.034	1.5	0.05
37.13	2	14	0.034	1.5	0.06
37.12	3	16	0.026	1.7	0.05
37.12	2	16	0.030	1.7	0.06
37.1	3	12	0.034	1.7	0.05
37.1	2	16	0.030	1.5	0.08
37.1	3	16	0.030	1.5	0.05
37.09	2	14	0.030	1.7	0.06
37.08	2	16	0.026	1.7	0.08

figure 4.23 Distribution of Q values with different force field parameters.

4.5.4. Discussion

The one-body optimizations suggest that the standard ORBIT parameters are certainly in the range of the optimal values for protein design. There may be a range of parameters that will be more accurate for specific designs. The electrostatic contribution is increased (lowered dielectric), the hydrogen bond term remains high, and the polar burial penalty is reduced. This behavior may be due to using a thermophilic protein set. Thermophilic proteins are observed to have higher numbers of charged and polar groups than their mesophilic homologs [108]. Introduction of new terms to the force field is shown to slightly improve the Q score and is easy to obtain first pass parameters for the terms with this method.

One-body optimization is very fast due to the lack of pairwise calculations. This is also a limitation. Each mutation in the optimization occurs in the background of all the surrounding residues, introducing potential bias to the calculation. As a test the optimization was allowed to choose between a rotamer from the library and the sidechain conformation from the crystal structure (not shown). The crystal structure sidechain conformation was

chosen frequently leading to increased wildtype recovery. This bias does not exist in a protein design calculation where only two designed positions at a time are available for energy calculations. This suggests the need for a more rigorous optimization.

4.6. Future Directions

Multiple position wildtype optimization has some history in the Mayo lab. Street and Mayo [109] extended model studies [102, 103] by using Z-score optimization on real proteins. A small number (7-8) of surface positions on the beta-sheets of two different proteins were populated with random rotamers. The Z-score is a measure of separation of the energy of the wildtype sequence from the average of the ensemble of random sequences. Parameters of the force field are altered to maximize the Z-score. This method is restricted to surfaces as it requires a well-distributed ensemble of random sequences that cannot exist in the confines of a protein core.

Ben Allen has combined wildtype recovery and multiple position design in his Computational Protein Design Suite (CPDS) (personal communication). By separating the energy matrix into separate components for each force field term it is possible to change the force field parameters without recalculating the pairs interactions each time. This is still a time consuming process as it is a small protein design calculation for each step. We have implemented a more detailed occlusion-based solvation method as well as using secondary structure propensity. With Ben's modifications to the FASTER algorithm this becomes a possible, if not rapid calculation. An initial large scale calculation has been completed by designing 47 clusters of 4 - 15 residues across seven of the thermophilic proteins in table 4.11 (1ubq, 1a19, 1aye, 1pga, 1c9o, 1azq, and 1sph). It remains to be seen the difference between one-body and two-body optimization methods.

Name	Mut PDB	WT PDB	Mutation	ddG	M1	T2	S3	Note
RNase P2	1B4O	1JIC	F31A	-3.7	T	[110]	[111]	-
RNase Ba	1BAN	1BNI	S91A	-1.93	U	[112]	-	-
RNase Ba	1BAO	1A2P	Y78F	-1.35	U	[112]	-	-
RNase Ba	1BNS	1BNI	T26A	-1.94	U	[112]	-	-
RNase Ba	1BRG	1BNI	F7L	-4.1	U	[113]	-	-
RNase Ba	1BRH	1A2P	L14A	-4.3	U	[112]	[114]	-
RNase Ba	1BRI	1BNI	I76A	-1.89	U	[112]	[114]	a
RNase Ba	1BRJ	1A2P	I88A	-4	U	[112]	[114]	-
RNase Ba	1BRK	1A2P	I96A	-3.17	U	[115]	[114]	-
RNase Ba	1BSA	1A2P	I51V	-1.8	U	[112]	[116]	b
RNase Ba	1BSB	1A2P	I76V	-0.82	U	[112]	[116]	c
RNase Ba	1BSC	1A2P	I88V	-1.34	U	[117]	[116]	d
RNase Ba	1BSD	1A2P	I96V	-0.88	U	[115]	[116]	-
RNase Ba	1BSE	1A2P	L89V	-0.3	U	[112]	[116]	e
RNase Ba	1B20	1B2X	R69S	-2.72	U	[118]	[119]	-
RNase Ba	1B21	1B2X	R69S;D93N	-3.49	U	[118]	[119]	-
RNase Ba	1B2Z	1B2X	D93N	-4.11	U	[118]	[119]	-

BPTI	1AAL	7PTI	C30V;C51A	0.5	T	[120]	-	f
Che-Y	1E6K	1JBE	D12A	2.5	U	[121]	-	-
Che-Y	1E6L	1JBE	D13A	2.7	U	[121]	-	-
Che-Y	1E6M	1JBE	D57A	3.3	U	[121]	-	-
Che-Y	1UDR	1JBE	K91D;K92A;I96K;A98L	1.8	U	[121]	-	-
Che-Y	1AB5	1JBE	F14N;V21T	2.7	U	[122]	-	g
Che-Y	1AB6	1JBE	F14N;V86T	1.7	U	[122]	-	h
IL-1B	1HIB	4I1B	T9G	-2.6	G	[123]	-	-
λrepressor	1LLI	1LMB	V36L;M40L;V47I	0.5	G	[124]	[125]	i
HPR	1OPD	1POH	S46D	1.5	U	[126]	-	-
RNase Sa	1BOX	1RGG	N39S	-2.3		[127]	-	-
RNase Sa	1UCI	1RGG	V2T	-0.9		[128]	-	-
RNase Sa	1UCJ	1RGG	V36T	-1.3		[128]	-	-
RNase Sa	1UCK	1RGG	V43T	-0.5		[128]	-	-
RNase Sa	1UCL	1RGG	V57T	-4.4		[128]	-	j
RNase Sa	1I8V	1RGG	Y80F	-1.5		[129]	-	-
RNase Sa	1I70	1RGG	Y86F	-0.3		[129]	-	-
RNase T1	1LRA	1RNT	E58A	-0.8	U	[130]	-	-

RNase T1	1RGC	1RLS	Q25K	0.9	U	[130]	-	-
RNase T1	1TRP	1RNT	Y45W;W59Y	-0.9	T	[131]	-	k
RNase T1	1TRQ	1RNT	W59Y	-0.9	T	[131]	-	l
RNase T1	7RNT	1RNT	Y45W	0.74	T	[131]	-	-
RNase T1	1RHL	1RNT	G23A	-1.1		[132]	-	-
RNase T1	1I2E	1I0V	V16A	-2.49	U	[133]	[133]	-
RNase T1	1G02	1I0V	V16S	-4.66	U	[133]	[133]	-
RNase T1	1I2G	1BVI	V16T	-3.2	U	[133]	[133]	-
RNase T1	1FYS	1I0V	V16C	-5.3	U	[133]	[133]	m
RNase T1	1FZU	1I0V	V78A	-4.08	U	[133]	[133]	-
RNase T1	1I3I	1I0V	V78T	-4.08	U	[133]	[133]	-
RNase T1	1I3F	1I0V	V89S	-4.87	U	[133]	[133]	-
RNase T1	3BIR	1RNT	H92N	-1.73	U	[134]	-	-
RNase T1	4RNT	1RNT	H92A	-1.04	U	[134]	-	-
RNase T1	5BIR	1RNT	H92Q	-0.57	U	[134]	-	-
RNase T1	2HOH	1BVI	N9A	-0.76	U	[135]	-	-
RNase T1	4HOH	1BVI	T93A	-0.7	U	[135]	-	-
RNase T1	5HOH	1BVI	N9A;T93A	-1.22	U	[135]	-	-

Staph Nuclease	1EY5	1EY0	T33V	0.4	G	[136]	-	-
Staph Nuclease	1EY6	1EY0	T41I	0.7	G	[136]	-	-
Staph Nuclease	1EY4	1EY0	S59A	0.5	G	[136]	-	-
Staph Nuclease	1EY7	1EY0	S128A	0.7	G	[136]	-	-
Staph Nuclease	1EY8	1EY0	P117G;H124L;S128A	3.4	G	[136]	-	-
Staph Nuclease	1EY9	1EY0	T41I;P117G;H124L;S128A	4.1	G	[136]	-	-
Staph Nuclease	1EYA	1EY0	T33V;T41I;P117G;H124L:S128A	3.8	G	[136]	-	-
Staph Nuclease	1EYC	1EY0	T41I;S59A;P117G;H124L;S128A	3.8	G	[136]	-	-
Staph Nuclease	1EZ6	1EY0	T33V;T41I;S59A;P117G;H124L;S128A	4.5	G	[136]	-	-
Staph Nuclease	1KAB	1STN	K116G	0	T	[137]	-	-
Staph Nuclease	1KAA	1STN	K116A	0		[138]	-	-
Staph Nuclease	1KDA	1STN	K116D	-0.11	G	[139]	-	-
Staph Nuclease	1KDB	1STN	K116E	0	G	[139]	-	-
Staph Nuclease	1KDC	1STN	K116N	-0.24	G	[139]	-	-
Staph Nuclease	1SNO	1STN	H124L	1.8	G	[140]	-	-
Staph Nuclease	1SNP	1STN	H124L;P117G	3	G	[140]	-	-
Staph Nuclease	1SYE	1STN	P117T	0	G	[141]	-	-
Staph Nuclease	1SYG	1STN	P117A	0.8	G	[141]	-	-

Staph Nuclease	1SYC	1STN	P117G	1.7	G	[141]	-	-
Gene V	1VQA	1VQB	V35A;I47L	-2.9	G	[142]	-	-
Gene V	1VQC	1VQB	V35I;I47F	-2	G	[142]	-	-
Gene V	1VQD	1VQB	V35I;I47L	-1.1	G	[142]	-	-
Gene V	1VQE	1VQB	V35I;I47M	-2.8	G	[142]	-	-
Gene V	1VQF	1VQB	V35I;I47V	-3	G	[142]	-	-
Gene V	1VQG	1VQB	I47L	-0.6	G	[142]	-	-
Gene V	1VQH	1VQB	I47M	-2.1	G	[142]	-	-
Gene V	1VQI	1VQB	I47V	-2.5	G	[142]	-	-
Gene V	1VQJ	1VQB	V35I	-0.6	G	[142]	-	-
Gene V	1AE1	1GVP	L32R	-1.6	G	[143]	-	-
Gene V	1AE3	1GVP	R82C	-1.5	G	[143]	-	-
Gene V	1GKH	1GVP	K69H	-1.3	G	[143]	-	-
CI-2	1COA	2CI2	I76V	0.21	G	[144]	[144]	n
CI-2	1YPA	2CI2	S31A;E33A;E34A	-1.67	G	[145]	-	-
CI-2	1YPB	2CI2	S31G;E33A;E34A	-1.63	G	[145]	-	-
CI-2	1YPC	2CI2	E33A;E34A	-0.76	G	[145]	-	-
T4 lysozyme	129L	1L63	A93T	0	T	[146]	-	-

T4 lysozyme	130L	1L63	T151S	0.4	T	[146]	-	-
T4 lysozyme	131L	1L63	T26S	0.6	T	[146]	-	-
T4 lysozyme	149L	2LZM	I3L	0.7	T	[147]	-	-
T4 lysozyme	172L	2LZM	I3C	-1.2	T	[148]	-	o
T4 lysozyme	173L	2LZM	K16E;R119E;K135E;K147E	-1	T	[149]	-	-
T4 lysozyme	189L	2LZM	I3L;S38D;A41V;A82P;N116D;V131A;N144D	3.57	T	[147]	-	p,q
T4 lysozyme	195L	1L63	A129L	-1.3	T	[150]	[150]	r
T4 lysozyme	196L	1L63	A129M	-1.9	T	[150]	[150]	s
T4 lysozyme	197L	1L63	A129M;F153A	-4.3	T	[150]	[150]	-
T4 lysozyme	198L	1L63	L121A;A129L	-1.1	T	[150]	[150]	t
T4 lysozyme	199L	1L63	L121A;A129M	-1	T	[150]	[150]	-
T4 lysozyme	1CTW	1L63	I78A	-1.2	T	[151]	-	u
T4 lysozyme	1CU0	1L63	I78M	-1.5	T	[152]	-	-
T4 lysozyme	1CU2	1L63	L84M	-1.9	T	[152]	-	-
T4 lysozyme	1CU3	1L63	V87M	-2.3	T	[152]	-	-
T4 lysozyme	1CU5	1L63	L91M	-0.8	T	[152]	-	-
T4 lysozyme	1CU6	1L63	L91A	-2.6	T	[151]	-	-
T4 lysozyme	1CUP	1L63	I100M	-1.6	T	[152]	-	-

T4 lysozyme	1CUQ	1L63	V103M	-1.2	T	[152]	-	-
T4 lysozyme	1CV0	1L63	F104M	-0.4	T	[152]	-	-
T4 lysozyme	1CV4	1L63	L118M	-0.7	T	[152]	-	-
T4 lysozyme	1CV3	1L63	L121M	-0.8	T	[152]	-	-
T4 lysozyme	1CV5	1L63	L133M	-0.4	T	[153]	-	-
T4 lysozyme	1CV6	1L63	V149M	-2.8	T	[152]	-	-
T4 lysozyme	1CVK	1L63	L118A	-3.2	T	[151]	-	-
T4 lysozyme	1D2W	1L63	I27M	-3.1	T	[151]	[154]	-
T4 lysozyme	1D2Y	1L63	I50M	-0.4	T	[151]	[154]	-
T4 lysozyme	1D3J	1L63	L66M	-1	T	[151]	[154]	-
T4 lysozyme	1KS3	1L63	L118M;L121M	-1.9	T	[153]	-	-
T4 lysozyme	1KW5	1L63	L84M;L91M;L99M	-3.1	T	[153]	-	-
T4 lysozyme	1KY0	1L63	L84M;L91M;L99M;F153M	-3.8	T	[153]	-	-
T4 lysozyme	1KW7	1L63	L84M;L91M;L99M;L133M	-3.5	T	[153]	-	-
T4 lysozyme	1D3M	1L63	L84M;L91M;L99M;L118M;L121M	-4.3	T	[153]	-	-
T4 lysozyme	1L0J	1L63	L84M;L91M;L99M;L118M;L121M;F153M	-4.7	T	[153]	-	-
T4 lysozyme	1KY1	1L63	L84M;L91M;L99M;L118M;L121M;L133M	-4.7	T	[153]	-	-
T4 lysozyme	1CX7	1L63	L84M;L91M;L99M;L118M;L121M;L133M;L153M	-5	T	[153]	-	-

T4 lysozyme	1L0K	1L63	L84M;L91M;L99M;V111M;L118M;L121M;L133M	-4.3	T	[153]	-	-
T4 lysozyme	1LWG	1LW9	L84M;V87M;L91M;L99M;V111M;L118M;L121M;L133M	-4.9	T	[153]	-	-
T4 lysozyme	1L00	2LZM	Q105A	-0.6	T	[155]	-	-
T4 lysozyme	1L10	2LZM	T157I	-1.2	T	[156]	-	-
T4 lysozyme	1L16	2LZM	G156D	-2.3	T	[157]	-	-
T4 lysozyme	1L17	2LZM	I3V	-0.4	T	[148]	-	-
T4 lysozyme	1L18	2LZM	I3Y	-2.3	T	[148]	-	-
T4 lysozyme	1L19	2LZM	S38D	0.6	T	[158]	-	-
T4 lysozyme	1L20	2LZM	N144D	0.41	T	[147]	-	-
T4 lysozyme	1L21	2LZM	N55G	-0.6	T	[159]	-	-
T4 lysozyme	1L22	2LZM	K124G	-0.1	T	[159]	-	-
T4 lysozyme	1L23	2LZM	G77A	0.4	T	[160]	-	-
T4 lysozyme	1L24	2LZM	A82P	0.57	T	[147]	-	-
T4 lysozyme	1L27	2LZM	P86D	-0.4	T	[161]	-	-
T4 lysozyme	1L29	2LZM	P86H	-1.5	T	[161]	-	-
T4 lysozyme	1L31	2LZM	P86R	-1.1	T	[161]	-	-
T4 lysozyme	1L33	2LZM	V131A	0.39	T	[147]	-	-
T4 lysozyme	1L34	2LZM	R96H	-2.8	T	[162]	-	-

T4 lysozyme	1L37	2LZM	T115E	0.3	T	[163]	-	-
T4 lysozyme	1L38	2LZM	Q123E	0.4	T	[163]	-	-
T4 lysozyme	1L40	1L63	N144E	0.5	T	[163]	-	-
T4 lysozyme	1L41	1L63	K83H;A112D	-1.5	T	[163]	-	-
T4 lysozyme	1L44	2LZM	R119E	-0.04	T	[149]	-	-
T4 lysozyme	1L45	2LZM	K135E	-1	T	[149]	-	-
T4 lysozyme	1L46	2LZM	K147E	-0.7	T	[149]	-	-
T4 lysozyme	1L47	2LZM	R154E	-1.1	T	[149]	-	-
T4 lysozyme	1L54	1L63	M102K	-6.9	T	[164]	-	-
T4 lysozyme	1L55	1L63	D92N	-1.4	T	[158]	-	-
T4 lysozyme	1L56	2LZM	K60P	0	T	[160]	-	-
T4 lysozyme	1L57	2LZM	N116D	0.6	T	[147]	-	-
T4 lysozyme	1L59	1L63	T109N	0.1	T	[158]	-	-
T4 lysozyme	1L60	2LZM	G113A	0.3	T	[160]	-	-
T4 lysozyme	1L61	1L63	S38N	0	T	[158]	-	-
T4 lysozyme	1L62	1L63	T108D	0.6	T	[158]	-	-
T4 lysozyme	1L63	2LZM	C54T;C97A	-0.6	T	[163]	-	-
T4 lysozyme	1L64	1L63	N40A;K43A;S44A;E45A;L46A;D47A;K48A	-2.54	T	[165]	-	-

T4 lysozyme	1L65	1L63	D47A	-0.95	T	[165]	-	-
T4 lysozyme	1L66	1L63	K43A	-1.03	T	[165]	-	-
T4 lysozyme	1L67	1L63	L46A	-1.86	T	[165]	-	v
T4 lysozyme	1L68	1L63	S44A	0.34	T	[165]	-	-
T4 lysozyme	1L76	1L63	D72P	-2.7	T	[166]	-	-
T4 lysozyme	1L77	1L63	M102L	-0.79	T	[117]	[117]	w
T4 lysozyme	1L79	1L63	L99F;V111I	-0.89	T	[117]	[117]	x
T4 lysozyme	1L80	1L63	L99F;M102L;V111I	-1.12	T	[117]	[117]	y
T4 lysozyme	1L81	1L63	L99F;M102L;F153L	-0.22	T	[117]	[117]	z
T4 lysozyme	1L82	1L63	L99F;M102L;V111I;F153L	-0.63	T	[117]	[117]	-
T4 lysozyme	1L85	1L63	F153A	-3.4	T	[150]	[167]	aa
T4 lysozyme	1L86	1L63	F153I	-0.2	T	[167]	[167]	ab
T4 lysozyme	1L87	1L63	F153L	0.3	T	[167]	[167]	ac
T4 lysozyme	1L88	1L63	F153M	-0.6	T	[150]	[167]	ad
T4 lysozyme	1L89	1L63	L99A;F153A	-6.9	T	[167]	[167]	ae
T4 lysozyme	1L90	1L63	L99A	-4.5	T	[167]	[167]	af
T4 lysozyme	1L91	1L63	L99F	-0.3	T	[167]	[167]	ag
T4 lysozyme	1L92	1L63	L99I	-1.5	T	[167]	[167]	ah

T4 lysozyme	1L93	1L63	L99M	-0.6	T	[167]	[167]	ai
T4 lysozyme	1L94	1L63	L99V	-2	T	[167]	[167]	aj
T4 lysozyme	1L95	1L63	F153V	-1.8	T	[167]	[167]	ak
T4 lysozyme	1L98	2LZM	Q105E	-1.1	T	[155]	-	-
T4 lysozyme	1L99	2LZM	Q105G	-1.5	T	[155]	-	-
T4 lysozyme	1LYE	1L63	T59V	-1.5	T	[168]	-	-
T4 lysozyme	1LYF	1L63	T59S	-0.2	T	[168]	-	-
T4 lysozyme	1LYG	1L63	T59N	-1.1	T	[168]	-	-
T4 lysozyme	1LYH	1L63	T59G	-1.6	T	[168]	-	-
T4 lysozyme	1LYI	1L63	T59D	-1.2	T	[168]	-	-
T4 lysozyme	1LYJ	1L63	T59A	-1.5	T	[168]	-	-
T4 lysozyme	1P2L	1LW9	V87I	-0.3	T	[99]	[99]	-
T4 lysozyme	1P2R	1LW9	I78V	-0.8	T	[99]	[99]	-
T4 lysozyme	1P36	1LW9	I100V	-0.4	T	[99]	[99]	-
T4 lysozyme	1P37	1LW9	V87I;I100V;V103I;M106I;V111A;M120Y;L133F;V149I;T152V	-2.6	T	[99]	[99]	al
T4 lysozyme	1P3N	1LW9	V87I;I100V;M102L;M106I;V111A;M120Y;L133F;V149I;T152V	-1.6	T	[99]	[99]	am
T4 lysozyme	1P46	1LW9	M106I	0.2	T	[99]	[99]	-
T4 lysozyme	1P64	1LW9	L133F	-0.3	T	[99]	[99]	-

T4 lysozyme	1P6Y	1LW9	M120Y	-0.1	T	[99]	[99]	-
T4 lysozyme	1P7S	1LW9	V103I	-0.5	T	[99]	[99]	-
T4 lysozyme	1PQD	1LW9	V87I;I100V;M102L;V103I;M106I;V111A;M120Y;L133F;V149I;T152V	-2.4	T	[99]	[99]	an
T4 lysozyme	1PQI	1LW9	I78V;V87M;M120Y;L133F;V149I;T152V	-3.3	T	[99]	[99]	-
T4 lysozyme	1PQJ	1LW9	V87I;I100V;M102L;V103I;M106I;M120Y;L133F;V149I;T152V	-1.8	T	[99]	[99]	ao
T4 lysozyme	1PQK	1LW9	I78V;L118I;M120Y;L133F;V149I;T152V	-3	T	[99]	[99]	ap
T4 lysozyme	1PQO	1LW9	L118I	-1.2	T	[99]	[99]	-
T4 lysozyme	1QT6	1L63	E11H	0.1	T	[169]	-	-
T4 lysozyme	1QT7	1L63	E11N	-0.1	T	[169]	-	-
T4 lysozyme	1TLA	2LZM	S117F	1.1	T	[169]	-	-
T4 lysozyme	1TOL	1L63	A146C	-1.5	T	[170]	-	-
T4 lysozyme	1QS5	1L63	A98L	-4.3	T	[171]	-	aq
T4 lysozyme	1QS9	1L63	A98V	-3.2	T	[171]	-	ar
T4 lysozyme	1QSB	1L63	A98C	-1	T	[171]	-	as
T4 lysozyme	1QSQ	1L63	M106A	-1.9	T	[151]	-	-
T4 lysozyme	1QTB	1L63	A42V	-2.7	T	[171]	-	at
T4 lysozyme	1QTC	1L63	A129F	-1.2	T	[171]	-	au
T4 lysozyme	1QTD	1L63	A129W	-2.2	T	[171]	-	av

T4 lysozyme	1QTH	1L63	A98M	-3.2	T	[171]	-	aw
T4 lysozyme	222L	1L63	M102A	-2.9	T	[151]	[172]	ax
T4 lysozyme	228L	1L63	F104A	-2.7	T	[151]	[172]	ay
T4 lysozyme	235L	1L63	V111A	-1	T	[151]	[173]	az
T4 lysozyme	237L	1L63	V149A	-3.2	T	[151]	[173]	ba
T4 lysozyme	238L	1L63	V103A	-1.6	T	[151]	[173]	bb
T4 lysozyme	239L	1L63	I17A	-2.3	T	[151]	[173]	-
T4 lysozyme	242L	1L63	I50A	-1.6	T	[151]	[173]	-
T4 lysozyme	244L	1L63	I100A	-2.5	T	[151]	[173]	bc
T4 lysozyme	245L	1L63	M6A	-1.6	T	[151]	[173]	bd
T4 lysozyme	247L	1L63	L84A	-3.7	T	[151]	[173]	be
T4 lysozyme	200L	1L63	L121A	-2.3	T	[150]	[150]	bf
T4 lysozyme	252L	1L63	M102A;M106A	-3.7	T	[151]	[172]	bg
T4 lysozyme	253L	1L63	D20A	-0.3	T	[169]	-	-
T4 lysozyme	254L	1L63	D20S	0.7	T	[169]	-	-
T4 lysozyme	255L	1L63	D20N	1.3	T	[169]	-	-
T4 lysozyme	2L78	1L63	V111I	-0.81	T	[117]	-	bh
T4 lysozyme	1L42	2LZM	K16E	0.5	T	[149]	-	-

RNase H	1JL1	1F21	D10A	3.3	U	[174]	-	bi
RNase H	1JL2		Core Switch	2.2	U	[175]	-	-
RNase H	1JXB	1F21	I53A	-2.1	U	[174]	-	-
RNase H	1KVA	2RN2	D134A	1.5	T	[176]	-	-
RNase H	1KVB	1RNH	D134H	1.9	T	[177]	-	-
RNase H	1KVC	1RNH	D134N	0.9	T	[178]	-	-
RNase H	1LAV	2RN2	V74L	0.9	T	[177]	-	-
RNase H	1LAW	2RN2	V74I	0.6	T	[178]	-	-
RNase H	1RBR	2RN2	H62P	1.1	T	[177]	-	-
RNase H	1RBS	2RN2	H62A	0	T	[179]	-	-
RNase H	1RBT	2RN2	K95G	1.9	T	[180]	-	-
RNase H	1RBU	2RN2	K95N	0.9	T	[180]	-	-
RNase H	1RBV	2RN2	K95A	0.1	T	[180]	-	-
cyt c551	1DVV	2PAC	F7A;V13M;F34Y;E43Y;V78I	5.86	G	[181]	-	-
HEWL	1HEM	1HEL	S91T	0.99	T	[182]	-	-
HEWL	1HEN	1HEL	I55V;S91T	0	T	[183]	-	-
HEWL	1HEO	1HEL	I55V	-0.91	T	[183]	-	-
HEWL	1HEP	1HEL	T40S;I55V;S91T	-0.3	T	[183]	-	-

HEWL	1HEQ	1HEL	T40S;S91T	0.61	T	[183]	-	-
HEWL	1HER	1HEL	T40S	-0.27	T	[183]	-	-
HEWL	1IOQ	1RFP	M12F;L56F	-0.35	G	[182]	-	-
HEWL	1IOR	1RFP	M12L;L56F	0.49	G	[182]	-	-
HEWL	1IOS	1RFP	M12F	1.09	G	[182]	-	-
HEWL	1IOT	1RFP	M12L	0.87	G	[182]	-	-
HEWL	1KXW	1RFP	D18N	-0.55	G	[184]	-	-
HEWL	1KXX	1RFP	D18N;N27D	-0.2	G	[184]	-	-
HEWL	1KXY	1RFP	N27D	0.36	G	[184]	-	-
HEWL	1LSM	1HEL	I55L;S91T;D101S	1.25	T	[183]	-	-
HEWL	1LSN	1HEL	S91A	-0.15	T	[183]	-	-
HEWL	1IR8	1RFP	I58M	-1.2	T	[185]	-	bj
HEWL	1IR7	1RFP	I78M	-0.9	T	[185]	-	bj
HEWL	1IR9	1RFP	I98M	-0.9	T	[185]	-	bj
HEWL	1UIC	1UIG	H15A	-0.2	T	[186]	-	-
HEWL	1UID	1UIG	H15F	-0.4	T	[186]	-	-
HEWL	1UIE	1UIG	H15G	-0.8	T	[186]	-	-
HEWL	1UIF	1UIG	H15V	-0.3	T	[186]	-	-

CspB	1HZ9	1C9O	E46A	-0.2		[105]	-	-
CspB	1HZB	1C9O	L66E	-1.2		[105]	-	-
CspB	1HZC	1C9O	R3E;E46A;L66E	-4.1		[105]	-	-
CspB	1I5F	1C9O	R3E	-2.75		[105]	-	-
AmpC	1L0D	1KE4	S64D	1.6		[187]	-	bk
AmpC	1L0E	1KE4	K67Q	1.1		[187]	-	bk
AmpC	1L0F	1KE4	N152H	0.7		[187]	-	bk
AmpC	1L0G	1KE4	S64G	3.6		[187]	-	bk,bl

1 Method of denaturation. T: thermal, G: guanidinium, U: urea.
2 Reference for the thermodynamic data
3 Reference for the structure data (if relevant)
4 Core switch
a buried water in cavity, hbond to F7
b rotation and movement of V51 CB to cavity, mostly (90%) removes cavity
c V76 fills created cavity mostly by linear shift of sidechain
d possible two conformations of V88
e V89 moves to cavity slightly straining chi1
f mutant relative to C30A/C51A
g F14N causes 2.7kcal benefit by itself
h F14N causes 2.7 kcal benefit by itself
i helix 4 moved away from mutation, surprise stabilization
j forms hbond, but completely buried
k could be extraneous effects from diff binding of het
l could be extraneous effects from diff binding of het
m DSC showed irreversibility, used urea, but m value may be issue
n V76 fills cavity by linear shift of sidechain, V38 and I48 also moved

o 172L has disulfide, expt does not
p additive stability, very stable lysozyme
q significant backbone movement near 22
r L121 moves away from mutation, L118,L133,F114 move away, increased cavity volume
s M129 unusual conformation, L121 eclipsed conformer, high thermal factors
t res 152-156 move to 121 cavity
u relaxation, backbone moves towards cavity
v I27 shift, cavity slightly reduced
w M106 Ce moves 2.8 Å
x shift in 107-114 helix, L84 shift
y 107-114 helix increase in mobility, I111 high disorder
z M106 large shift
aa A153 carbonyl moves 1 ang, M102 moves 0.6 Å to cavity
ab M102,M106,L121,L133 large shifts, local bb shift 0.8 Å
ac L99,M102 move
ad M153 strained conformation
ae similar to F153A

af very little bb shift, small shifts in V87 and Y88
ag L84 and N81 move away 0.6A
ah I99 has high B-factor, possibly avg between 2 conformations
ai L84 moves away
aj may be strained
ak M102,M106,L121,L133 large shifts, local bb shift 0.8 Å
al core-10 L102M revertant, other 9 sites show additivity
am core10 I103V revertant, bb returns to WT conf, increase in stability, relieves strain
an core-10, break hbond T109-G113, 2.8 Å bb shift, similar to V111I, new H2O, M102L changes rotamer
ao core-10 A111V revertant, fills cavity caused by core-10, improves stability
ap core-7 M87V revertant
aq strained, 98 bb pushed back, leads to bending of helices
ar T152 pushed away, opposed helices move apart, helix bending
as strained, opposed helices are pushed apart
at strained, bb shift at opposed positions
au strained, torsion strain
av significant strain, bb displacement
aw strained, bb shift and helix bending

ax M106 changed torsion angles, L121 and F153 moved toward cavity
ay E108 rot to cleft 1.9 Å, new H2O hbond to I29 carbonyl
az bb shift to cavity, M102 shift
ba additional solvent bound to cavity
bb V111 move to cavity, M106 changes rotamer
bc res 74-76 rotate
bd bind 2 new H2O
be leads to disorder, bb moves 2 Å
bf F153 shifted 2.4 Å to cavity, L118 changed conformation, increase in thermal factors around cavity
bg res 106-114 2 conformations, alternate is unraveled helix, high glycine content
bh helix 107-114 moved 1.4 Å, breaking H-bond between CO of T109 and NH of G113
bi active site: D10,E48,D70,D134
bj expt done at 35 deg
bk 38% ethylene glycol
bl remove steric strain

Bibliography

[1] Winter, G., Fersht, A. R., Wilkinson, A. J., Zoller, M., and Smith, M. *Nature* **299**(5885), 756–758 (1982).

[2] Matthews, B. W. *Advances In Protein Chemistry* **46**, 249–278 (1995).

[3] Salazar, O., Cirino, P. C., and Arnold, F. H. *Chembiochem* **4**(9), 891–893 (2003).

[4] Zha, D. X., Wilensek, S., Hermes, M., Jaeger, K. E., and Reetz, M. T. *Chemical communications* **7**(24), 2664–2665 (2001).

[5] Bloom, J. D., Meyer, M. M., Meinhold, P., Otey, C. R., MacMillan, D., and Arnold, F. H. *Current Opinion in Structural Biology* **15**(4), 447–452 (2005).

[6] Drexler, K. E. *Proceedings of the National Academy of Sciences of the United States of America* **78**(9), 5275–5278 (1981).

[7] Pabo, C. *Nature* **301**(5897), 200–200 (1983).

[8] Ponder, J. W. and Richards, F. M. *Journal of Molecular Biology* **193**(4), 775–791 (1987).

[9] Lee, C. and Subbiah, S. *Journal of Molecular Biology* **217**(2), 373–388 (1991).

[10] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. *Journal of Chemical Physics* **21**, 1087–1089 (1953).

[11] Kirkpatrick, S., Gelatt, C., and Vecchi, M. *Science* **220**, 671–680 (1983).

[12] Lee, C. and Levitt, M. *Nature* **352**(6334), 448–451 (1991).

[13] Lim, W. and Sauer, R. *Journal of Molecular Biology* **219**(2), 359–376 (1991).

[14] Regan, L. and DeGrado, W. *Science* **241**(4868), 976–978 (1988).

[15] Hecht, M. H., Richardson, J. S., Richardson, D. C., and Ogden, R. C. *Science* **249**(4971), 884–891 (1990).

[16] Hellinga, H. W., Caradonna, J. P., and Richards, F. M. *Journal of Molecular Biology* **222**(3), 787–803 (1991).

[17] Hurley, J. H., Baase, W. A., and Matthews, B. W. *Journal of Molecular Biology* **224**(4), 1143–1159 (1992).

[18] Desjarlais, J. R. and Handel, T. M. *Protein Science* **4**(10), 2006–2018 (1995).

[19] Harbury, P. B., Tidor, B., and Kim, P. S. *Proceedings of the National Academy of Sciences of the United States of America* **92**(18), 8408–8412 (1995).

[20] Nautiyal, S., Woolfson, D. N., King, D. S., and Alber, T. *Biochemistry* **34**(37), 11645–11651 (1995).

[21] Betz, S. F. and DeGrado, W. F. *Biochemistry* **35**(21), 6955–6962 (1996).

[22] Dahiyat, B. I. and Mayo, S. L. *Science* **278**(5335), 82–87 (1997).

[23] Desmet, J., DeMaeyer, M., Hazes, B., and Lasters, I. *Nature* **356**(6369), 539–542 (1992).

[24] Dahiyat, B. I. and Mayo, S. L. *Protein Science* **5**(5), 895–903 (1996).

[25] Dahiyat, B. I., Gordon, D. B., and Mayo, S. L. *Protein Science* **6**(6), 1333–1337 (1997).

[26] Dahiyat, B. I. and Mayo, S. L. *Proceedings of the National Academy of Sciences of the United States of America* **94**(19), 10172–10177 (1997).

[27] Malakauskas, S. M. and Mayo, S. L. *Nature Structural Biology* **5**(6), 470–475 (1998).

[28] Klemba, M., Gardner, K. H., Marino, S., Clarke, N. D., and Regan, L. *Nature Structural Biology* **2**(5), 368–373 (1995).

[29] Hilvert, D. *Annual Review of Biochemistry* **69**, 751–793 (2000).

[30] Bolon, D. N. and Mayo, S. L. *Proceedings of the National Academy of Sciences of the United States of America* **98**(25), 14274–14279 (2001).

[31] Looger, L. L., Dwyer, M. A., Smith, J. J., and Hellinga, H. W. *Nature* **423**(6936), 185–190 (2003).

[32] Dwyer, M. A., Looger, L. L., and Hellinga, H. W. *Science* **304**(5679), 1967–1971 (2004).

[33] Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. *Science* **302**(5649), 1364–1368 (2003).

[34] Dill, K. A. *Biochemistry* **29**(31), 7133–7155 (1990).

[35] Kauzmann, W. *Advances in Protein Chemistry* **14**, 1 (1959).

[36] Street, A. G. and Mayo, S. L. *Folding & Design* **3**(4), 253–258 (1998).

[37] Shimizu, S. and Chan, H. S. *Proteins* **48**(1), 15–30 (2002).

[38] Lee, B., R. F. *Journal of Molecular Biology* **55**, 379–400 (1971).

[39] Eisenberg, D. and McLachlan, A. *Nature* **319**, 199–203 (1986).

[40] Sun, S. J., Brem, R., Chan, H. S., and Dill, K. A. *Protein Engineering* **8**(12), 1205–1213 (1995).

[41] Makhatadze, G. I. and Privalov, P. L. *Advances in Protein Chemistry* **47**, 307–425 (1995).

[42] Karplus, P. A. *Protein Science* **6**, 1302–1307 (1997).

[43] Pauling, L., C. R. B. H. *Proceedings of the National Academy of Sciences of the United States of America* **37**, 205–210 (1951).

[44] Mayo, S., Olafson, B., and Goddard, WA, I. *Journal of Physical Chemistry* **94**, 8897–8909 (1990).

[45] McDonald, I. K. and Thornton, J. M. *Journal Of Molecular Biology* **238**(5), 777–793 (1994).

[46] Cohen, B. E., McAnaney, T. B., Park, E. S., Jan, Y. N., Boxer, S. G., and Jan, L. Y. *Science* **296**(5573), 1700–1703 (2002).

[47] Richards, F. *Annual Review of Biophysics and Bioengineering* **6**, 151–176 (1977).

[48] Colonna-Cesari, F. and Sander, C. *Biophysical Journal* **57**, 1103–1107 (1990).

[49] Holm, L. and Sander, C. *Journal of Molecular Biology* **225**(1), 93–105 (1992).

[50] Lazaridis, T. and Karplus, M. *Proteins* **35**(2), 133–152 (1999).

[51] Ramachandran GN, Ramakrishnan C, S. V. *Journal of Molecular Biology* **7**, 95–99 (1963).

[52] Avbelj, F. and Baldwin, R. L. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **100**(10), 5742–5747 (2003).

[53] Avbelj, F., Grdadolnik, S. G., Grdadolnik, J., and Baldwin, R. L. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **103**(5), 1272–1277 (2006).

[54] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. *Nucleic Acids Research* **28**(1), 235–242 (2000).

[55] Shortle, D. *Protein Science* **12**(6), 1298–1302 (2003).

[56] Shortle, D. *Protein Science* **11**(1), 18–26 (2002).

[57] Janin, J., Wodak, S., Levitt, M., and Maigret, B. *Journal of Molecular Biology* **125**(3), 357–386 (1978).

[58] Dunbrack, R. L. and Karplus, M. *Journal of Molecular Biology* **230**(2), 543–574 (1993).

[59] Dunbrack, R. L. and Karplus, M. *Nature Structural Biology* **1**(5), 334–340 (1994).

[60] Dunbrack, R. L. and Cohen, F. E. *Protein Science* **6**(8), 1661–1681 (1997).

[61] Lovell, S. C., Word, J. M., Richardson, J. S., and Richardson, D. C. *Proteins: Structure, Function, and Genetics* **40**(3), 389–408 (2000).

[62] Gordon, D. B., Hom, G. K., Mayo, S. L., and Pierce, N. A. *Journal of Computational Chemistry* **24**(2), 232–243 (2003).

[63] Sarisky, C. *Exploration of the determinants of protein structure and stability by protein design.* PhD thesis, California Institute of Technology, (2005).

[64] Dunbrack, R. L. *Current Opinions in Structural Biology* **12**(4), 431–440 (2002).

[65] Johnson, B. H. and Hecht, M. H. *Bio-technology* **12**(13), 1357–1360 (1994).

[66] Minor, D. and Kim, P. *Nature* **367**, 660–663 (1994).

[67] Honig, B. and Nicholls, A. *Science* **268**(5214), 1144–1149 (1995).

[68] Havranek, J. J. and Harbury, P. B. *Proceedings of the National Academy of Sciences of the United States of America* **96**(20), 11145–11150 (1999).

[69] Havranek, J. J. and Harbury, P. B. *Nature Structural Biology* **10**(1), 45–52 (2003).

[70] Pokala, N. and Handel, T. M. *Protein Science* **13**(4), 925–936 (2004).

[71] Wisz, M. S. and Hellinga, H. W. *Proteins: Structure, Function, and Genetics* **51**(3), 360–377 (2003).

[72] Marshall, S. A., Vizcarra, C. L., and Mayo, S. L. *Protein Science* **14**(5), 1293–1304 (2005).

[73] Marshall, S. A., Morgan, C. S., and Mayo, S. L. *Journal of Molecular Biology* **316**(1), 189–199 (2002).

[74] Rocchia, W., Alexov, E., and Honig, B. *Journal of Physical Chemistry B* **105**(28), 6507–6514 (2001).

[75] Gilson, M. and Honig, B. *Proceedings of the National Acaademy of Sciences USA* **86**, 1524–1528 (1989).

[76] Sengupta, D., G. R. S. J. and Ullman, G. *Structure* **13**, 849–855 (2005).

[77] Alexov, E. *Proteins-Structure Function And Bioinformatics* **50**(1), 94–103 (2003).

[78] Sitkoff, D., Sharp, K. A., and Honig, B. *Journal of Physical Chemistry* **98**(7), 1978–1988 (1994).

[79] Gordon, D. B., Marshall, S. A., and Mayo, S. L. *Current Opinions in Structural Biology* **9**(4), 509–513 (1999).

[80] Goldstein, R. F. *Biophysical Journal* **66**(5), 1335–1340 (1994).

[81] Gordon, D. B. and Mayo, S. L. *Journal Of Computational Chemistry* **19**(13), 1505–1514 (1998).

[82] Prodromou, C. and Pearl, L. *Protein Engineering* **5**, 827–829 (1992).

[83] Pilon, A., Yost, P., Chase, T., Lohnas, G., Burkett, T., Roberts, S., and Bentley, W. *Biotechnology Progress* **13**, 374–379 (1997).

[84] Hendsch, Z. S. and Tidor, B. *Protein Science* **3**(2), 211–226 (1994).

[85] Hendsch, Z. S., Jonsson, T., Sauer, R. T., and Tidor, B. *Biochemistry* **35**(24), 7621–7625 (1996).

[86] Giletto, A. and Pace, C. N. *Biochemistry* **38**, 13379–13384 (1999).

[87] Takano, K., Scholtz, J. M., Sacchettini, J. C., and Pace, C. N. *Journal of Biological Chemistry* **278**(34), 31790–31795 (2003).

[88] Alber, T., Sun, D. P., Wilson, K., Wozniak, J. A., Cook, S. P., and Matthews, B. W. *Nature* **330**(6143), 41–46 (1987).

[89] Guerois, R., Nielsen, J. E., and Serrano, L. *Journal of Molecular Biology* **320**(2), 369–387 (2002).

[90] Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K., and Sarai, A. *Nucleic Acids Research* **32**, D120–D121 (2004).

[91] Vriend, G. *Journal of Molecular Graphics* **8**, 52–56 (1990).

[92] Stark, P. B. and Parker, R. *Computational Statistics* **10**(2), 129–141 (1995).

[93] Munoz, V. and Serrano, L. *Proteins* **20**(4), 301–311 (1994).

[94] Street, A. G. and Mayo, S. L. *Proceedings of the National Academy of Sciences of the United States of America* **96**(16), 9074–9076 (1999).

[95] Abagyan, R. and Totrov, M. *Journal of Molecular Biology* **235**(3), 983–1002 (1994).

[96] Koehl, P. and Delarue, M. *Journal Of Molecular Biology* **239**(2), 249–275 (1994).

[97] Chen, J. M. and Stites, W. E. *Journal Of Molecular Biology* **344**(1), 271–280 (2004).

[98] Fauchere, J. L. and Pliska, V. *European Journal Of Medicinal Chemistry* **18**(4), 369–375 (1983).

[99] Mooers, B. H. M., Datta, D., Baase, W. A., Zollars, E. S., Mayo, S. L., and Matthews, B. W. *Journal of Molecular Biology* **332**(3), 741–756 (2003).

[100] Mooers, B. H. M., Datta, D., Baase, W. A., Zollars, E. S., Mayo, S. L., and Matthews, B. W. *Journal Of Molecular Biology* **332**(3), 741–756 (2003).

[101] Chiu, T. L. and Goldstein, R. A. *Protein Engineering* **11**(9), 749–752 (1998).

[102] Goldstein, R. A., Luthey-Schulten, Z. A., and Wolynes, P. G. *Proceedings of the National Academy of Sciences of the United States of America* **89**(11), 4918–22 (1992).

[103] Yue, K. and Dill, K. A. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **89**(9), 4163–4167 (1992).

[104] Kuhlman, B. and Baker, D. *Proceedings of the National Academy of Sciences of the United States of America* **9**(19), 10383–10388 (2000).

[105] Perl, D. and Schmid, F. X. *Journal of Molecular Biology* **313**(2), 343–357 (2001).

[106] Rees, D. C. and Robertson, A. D. *Protein Science* **10**, 1187–1194 (2001).

[107] Kumar, S., Tsai, C. J., and Nussinov, R. *Biochemistry* **41**(17), 5359–5374 (2002).

[108] Kumar, S. and Nussinov, R. *Biophysical Chemistry* **111**, 235–246 (2004).

[109] Street, A. G., Datta, D., Gordon, D. B., and Mayo, S. L. *Physical Review Letters* **84**(21), 5010–5013 (2000).

[110] Mombelli, E., Afshar, M., Fusi, P., Mariani, M., Tortora, P., Connelly, J. P., and Lange, R. *Biochemistry* **36**(29), 8733–8742 (1997).

[111] Consonni, R., Santomo, L., Fusi, P., Tortora, P., and Zetta, L. *Biochemistry* **38**(39), 12709–12717 (1999).

[112] Serrano, L., Kellis, J. T., Cann, P., Matouschek, A., and Fersht, A. R. *Journal of Molecular Biology* **224**(3), 783–804 (1992).

[113] Kellis, J. T., Nyberg, K., Sali, D., and Fersht, A. R. *Nature* **333**(6175), 784–786 (1988).

[114] Buckle, A. M., Cramer, P., and Fersht, A. R. *Biochemistry* **35**(14), 4298–4305 (1996).

[115] Kellis, J. T., Nyberg, K., and Fersht, A. R. *Biochemistry* **28**(11), 4914–4922 (1989).

[116] Buckle, A. M., Henrick, K., and Fersht, A. R. *Journal of Molecular Biology* **234**(3), 847–860 (1993).

[117] Hurley, J. H., Baase, W. A., and Matthews, B. W. *Journal of Molecular Biology* **224**(4), 1143–1159 (1992).

[118] Tissot, A. C., Vuilleumier, S., and Fersht, A. R. *Biochemistry* **35**(21), 6786–6794 (1996).

[119] Vaughan, C. K., Harryson, P., Buckle, A. M., and Fersht, A. R. *Acta Crystallography D Biological Crystallography* **58**(Pt 4), 591–600 (2002).

[120] Liu, Y., Breslauer, K., and Anderson, S. *Biochemistry* **36**(18), 5323–5335 (1997).

[121] Solà, M., López-Hernández, E., Cronet, P., Lacroix, E., Serrano, L., Coll, M., and Párraga, A. *Journal of Molecular Biology* **303**(2), 213–225 (2000).

[122] Wilcock, D., Pisabarro, M. T., López-Hernandez, E., Serrano, L., and Coll, M. *Acta Crystallography D Biological Crystallography* **54**(Pt 3), 378–385 (1998).

[123] Chrunyk, B. A., Evans, J., Lillquist, J., Young, P., and Wetzel, R. *Journal of Biological Chemistry* **268**(24), 18053–18061 (1993).

[124] Lim, W. A., Farruggio, D. C., and Sauer, R. T. *Biochemistry* **31**(17), 4324–4333 (1992).

[125] Lim, W. A., Hodel, A., Sauer, R. T., and Richards, F. M. *Proceedings of the National Academy of Sciences of the United States of America* **91**(1), 423–427 (1994).

[126] Thapar, R., Nicholson, E. M., Rajagopal, P., Waygood, E. B., Scholtz, J. M., and Klevit, R. E. *Biochemistry* **35**(35), 11268–11277 (1996).

[127] Hebert, E. J., Giletto, A., Sevcik, J., Urbanikova, L., Wilson, K. S., Dauter, Z., and Pace, C. N. *Biochemistry* **37**(46), 16192–16200 (1998).

[128] Takano, K., Scholtz, J. M., Sacchettini, J. C., and Pace, C. N. *Journal of Biological Chemistry* **278**(34), 31790–31795 (2003).

[129] Pace, C. N., Horn, G., Hebert, E. J., Bechert, J., Shaw, K., Urbanikova, L., Scholtz, J. M., and Sevcik, J. *Journal of Molecular Biology* **312**(2), 393–404 (2001).

[130] Shirley, B. A., Stanssens, P., Steyaert, J., and Pace, C. N. *Journal of Biological Chemistry* **264**(20), 11621–11625 (1989).

[131] Schubert, W. D., Schluckebier, G., Backmann, J., Granzin, J., Kisker, C., Choe, H. W., Hahn, U., Pfeil, W., and Saenger, W. *European Journal of Biochemistry* **220**(2), 527–534 (1994).

[132] Huyghues-Despointes, B. M., Langhorst, U., Steyaert, J., Pace, C. N., and Scholtz, J. M. *Biochemistry* **38**(50), 16481–16490 (1999).

[133] Vos, S. D., Backmann, J., Prévost, M., Steyaert, J., and Loris, R. *Biochemistry* **40**(34), 10140–10149 (2001).

[134] Vos, S. D., Doumen, J., Langhorst, U., and Steyaert, J. *Journal of Molecular Biology* **275**(4), 651–661 (1998).

[135] Langhorst, U., Backmann, J., Loris, R., and Steyaert, J. *Biochemistry* **39**(22), 6586–6593 (2000).

[136] Chen, J., Lu, Z., Sakon, J., and Stites, W. E. *Journal of Molecular Biology* **303**(2), 125–130 (2000).

[137] Eftink, M. R., Ghiron, C. A., Kautz, R. A., and Fox, R. O. *Biochemistry* **30**(5), 1193–1199 (1991).

[138] Hodel, A., Kautz, R. A., Jacobs, M. D., and Fox, R. O. *Protein Science* **2**(5), 838–850 (1993).

[139] Hodel, A., Kautz, R. A., and Fox, R. O. *Protein Science* **4**(3), 484–495 (1995).

[140] Truckses, D. M., Somoza, J. R., Prehoda, K. E., Miller, S. C., and Markley, J. L. *Protein Science* **5**(9), 1907–1916 (1996).

[141] Hynes, T. R., Hodel, A., and Fox, R. O. *Biochemistry* **33**(17), 5021–5030 (1994).

[142] Sandberg, W. S. and Terwilliger, T. C. *Proceedings of the National Academy of Sciences of the United States of America* **88**(5), 1706–1710 (1991).

[143] Sandberg, W. S. and Terwilliger, T. C. *Proceedings of the National Academy of Sciences of the United States of America* **90**(18), 8367–8371 (1993).

[144] Jackson, S. E., Moracci, M., elMasry, N., Johnson, C. M., and Fersht, A. R. *Biochemistry* **32**(42), 11259–11269 (1993).

[145] Itzhaki, L. S., Otzen, D. E., and Fersht, A. R. *Journal of Molecular Biology* **254**(2), 260–288 (1995).

[146] Pjura, P., Matsumura, M., Baase, W. A., and Matthews, B. W. *Protein Science* **2**(12), 2217–2225 (1993).

[147] Zhang, X. J., Baase, W. A., Shoichet, B. K., Wilson, K. P., and Matthews, B. W. *Protein Engineering* **8**(10), 1017–1022 (1995).

[148] Matsumura, M., Becktel, W. J., and Matthews, B. W. *Nature* **334**(6181), 406–410 (1988).

[149] Sun, D. P., Söderlind, E., Baase, W. A., Wozniak, J. A., Sauer, U., and Matthews, B. W. *Journal of Molecular Biology* **221**(3), 873–887 (1991).

[150] Baldwin E., Xu J., H. O. B. W. M. B. *Journal of Molecular Biology* **259**, 542–59 (1996).

[151] Gassner, N. C., Baase, W. A., Lindstrom, J. D., Lu, J., Dahlquist, F. W., and Matthews, B. W. *Biochemistry* **38**(44), 14451–14460 (1999).

[152] Gassner, N. C., Baase, W. A., and Matthews, B. W. *Proceedings of the National Academy of Sciences of the United States of America* **93**(22), 12155–12158 (1996).

[153] Gassner, N. C., Baase, W. A., Mooers, B. H. M., Busam, R. D., Weaver, L. H., Lindstrom, J. D., Quillin, M. L., and Matthews, B. W. *Biophysical Chemistry* **100**(1-3), 325–340 (2003).

[154] Gassner, N. C., Baase, W. A., Hausrath, A. C., and Matthews, B. W. *Journal of Molecular Biology* **294**(1), 17–20 (1999).

[155] Pjura, P., McIntosh, L. P., Wozniak, J. A., and Matthews, B. W. *Proteins* **15**(4), 401–412 (1993).

[156] Grütter, M. G., Gray, T. M., Weaver, L. H., Wilson, T. A., and Matthews, B. W. *Journal of Molecular Biology* **197**(2), 315–329 (1987).

[157] Gray, T. M. and Matthews, B. W. *Journal of Biological Chemistry* **262**(35), 16858–16864 (1987).

[158] Nicholson, H., Anderson, D. E., Dao-pin, S., and Matthews, B. W. *Biochemistry* **30**(41), 9816–9828 (1991).

[159] Nicholson, H., Söderlind, E., Tronrud, D. E., and Matthews, B. W. *Journal of Molecular Biology* **210**(1), 181–193 (1989).

[160] Nicholson, H., Tronrud, D. E., Becktel, W. J., and Matthews, B. W. *Biopolymers* **32**(11), 1431–1441 (1992).

[161] Alber, T., Bell, J. A., Sun, D. P., Nicholson, H., Wozniak, J. A., Cook, S., and Matthews, B. W. *Science* **239**(4840), 631–635 (1988).

[162] Wetzel, R., Perry, L. J., Baase, W. A., and Becktel, W. J. *Proceedings of the National Academy of Sciences of the United States of America* **85**(2), 401–405 (1988).

[163] Sun, D. P., Sauer, U., Nicholson, H., and Matthews, B. W. *Biochemistry* **30**(29), 7142–7153 (1991).

[164] Dao-pin, S., Anderson, D. E., Baase, W. A., Dahlquist, F. W., and Matthews, B. W. *Biochemistry* **30**(49), 11521–11529 (1991).

[165] Heinz, D. W., Baase, W. A., and Matthews, B. W. *Proceedings of the National Academy of Sciences of the United States of America* **89**(9), 3751–3755 (1992).

[166] Sauer, U. H., San, D. P., and Matthews, B. W. *Journal of Biological Chemistry* **267**(4), 2393–2399 (1992).

[167] Eriksson, A. E., Baase, W. A., and Matthews, B. W. *Journal of Molecular Biology* **229**(3), 747–769 (1993).

[168] Bell, J. A., Becktel, W. J., Sauer, U., Baase, W. A., and Matthews, B. W. *Biochemistry* **31**(14), 3590–3596 (1992).

[169] Shoichet, B. K., Baase, W. A., Kuroki, R., and Matthews, B. W. *Proceedings of the National Academy of Sciences of the United States of America* **92**(2), 452–456 (1995).

[170] Lu, J., Baase, W. A., Muchmore, D. C., and Dahlquist, F. W. *Biochemistry* **31**(34), 7765–7772 (1992).

[171] Liu, R., Baase, W. A., and Matthews, B. W. *Journal of Molecular Biology* **295**(1), 127–145 (2000).

[172] Baldwin, E., Baase, W. A., Zhang, X., Feher, V., and Matthews, B. W. *Journal of Molecular Biology* **277**(2), 467–485 (1998).

[173] Xu, J., Baase, W. A., Baldwin, E., and Matthews, B. W. *Protein Science* **7**(1), 158–177 (1998).

[174] Raschke, T. M., Kho, J., and Marqusee, S. *Nature Structural Biology* **6**(9), 825–831 (1999).

[175] Robic, S., Berger, J. M., and Marqusee, S. *Protein Science* **11**(2), 381–389 (2002).

[176] Haruki, M., Noguchi, E., Nakai, C., Liu, Y. Y., Oobatake, M., Itaya, M., and Kanaya, S. *European Journal of Biochemistry* **220**(2), 623–631 (1994).

[177] Akasako, A., Haruki, M., Oobatake, M., and Kanaya, S. *Biochemistry* **34**(25), 8115–8122 (1995).

[178] Ishikawa, K., Nakamura, H., Morikawa, K., and Kanaya, S. *Biochemistry* **32**(24), 6171–6178 (1993).

[179] Kanaya, S., Oobatake, M., Nakamura, H., and Ikehara, M. *Journal of Biotechnology* **28**(1), 117–136 (1993).

[180] Kimura, S., Kanaya, S., and Nakamura, H. *Journal of Biological Chemistry* **267**(31), 22014–22017 (1992).

[181] Hasegawa, J., Uchiyama, S., Tanimoto, Y., Mizutani, M., Kobayashi, Y., Sambongi, Y., and Igarashi, Y. *Journal of Biological Chemistry* **275**(48), 37824–37828 (2000).

[182] Ohmura, T., Ueda, T., Ootsuka, K., Saito, M., and Imoto, T. *Protein Science* **10**(2), 313–320 (2001).

[183] Shih, P., Holland, D. R., and Kirsch, J. F. *Protein Science* **4**(10), 2050–2062 (1995).

[184] Motoshima, H., Mine, S., Masumoto, K., Abe, Y., Iwashita, H., Hashimoto, Y., Chijiiwa, Y., Ueda, T., and Imoto, T. *Journal of Biochemistry (Tokyo)* **121**(6), 1076–1081 (1997).

[185] Ohmura, T., Ueda, T., Hashimoto, Y., and Imoto, T. *Protein Engineering* **14**(6), 421–425 (2001).

[186] Ohmura, T., Ueda, T., Motoshima, H., Tamura, T., and Imoto, T. *Journal of Biochemistry (Tokyo)* **122**(3), 512–517 (1997).

[187] Beadle, B. M. and Shoichet, B. K. *Journal of Molecular Biology* **321**(2), 285–296 (2002).

A. Lysozyme Core Design

The calculations described here were an attempt to repack the core of lysozyme. No design was able to repack to a more stable structure than wildtype. A methionine penalty is used to remove some (but not all) methionines from the design. This led to a slight more stable design. ORBIT calculated energies correlate poorly with experimentally determined energies.

doi:10.1016/S0022-2836(03)00856-8 *J. Mol. Biol.* (2003) **332**, 741–756

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Repacking the Core of T4 Lysozyme by Automated Design

Blaine H. M. Mooers[1], Deepshikha Datta[2], Walter A. Baase[1] Eric S. Zollars[2], Stephen L. Mayo[3]* and Brian W. Matthews[1]*

[1]*Department of Physics Institute of Molecular Biology Howard Hughes Medical Institute, 1229 University of Oregon, Eugene, OR 97403-1229, USA*

[2]*Biochemistry and Molecular Biophysics Option, California Institute of Technology, Mail Code 114-76, Pasadena, CA 91125, USA*

[3]*Howard Hughes Medical Institute and Divisions of Biology and Chemistry and Chemical Engineering California Institute of Technology, Mail Code 114-76 Pasadena, CA 91125, USA*

**Corresponding authors*

Automated protein redesign, as implemented in the program ORBIT, was used to redesign the core of phage T4 lysozyme. A total of 26 buried or partially buried sites in the C-terminal domain were allowed to vary both their sequence and side-chain conformation while the backbone and non-selected side-chains remained fixed. A variant with seven substitutions ("Core-7") was identified as having the most favorable energy. The redesign experiment was repeated with a penalty for the presence of methionine residues. In this case the redesigned protein ("Core-10") had ten amino acid changes. The two designed proteins, as well as the constituent single mutants, and several single-site revertants were over-expressed in *Escherichia coli*, purified, and subjected to crystallographic and thermal analyses. The thermodynamic and structural data show that some repacking was achieved although neither redesigned protein was more stable than the wild-type protein. The use of the methionine penalty was shown to be effective. Several of the side-chain rotamers in the predicted structure of Core-10 differ from those observed. Rather than changing to new rotamers predicted by the design process, side-chains tend to maintain conformations similar to those seen in the native molecule. In contrast, parts of the backbone change by up to 2.8 Å relative to both the designed structure and wild-type.

Water molecules that are present within the lysozyme molecule were removed during the design process. In the redesigned protein the resultant cavities were, to some degree, re-occupied by side-chain atoms. In the observed structure, however, water molecules were still bound at or near their original sites. This suggests that it may be preferable to leave such water molecules in place during the design procedure. The results emphasize the specificity of the packing that occurs within the core of a typical protein. While point substitutions within the core are tolerated they almost always result in a loss of stability. Likewise, combinations of substitutions may also be tolerated but usually destabilize the protein. Experience with T4 lysozyme suggests that a general core repacking methodology with retention or enhancement of stability may be difficult to achieve without provision for shifts in the backbone.

Keywords: ORBIT; single-site revertants; T4 lysozyme; repacking

Introduction

The cores of proteins are generally well packed.[1,2] They have shown a remarkable ability to accommodate changes in buried hydrophobic residues although generally with some loss of stability.[3–5] It has been suggested that protein core packing is not like a jigsaw puzzle. Rather, it is more like nuts and bolts in a jar.[6] If this is the case there may be opportunities to improve the stability of native proteins by optimizing the packing of buried amino acids. An early test with phage T4 lysozyme showed that the effectiveness of doing

Abbreviations used: WT*, cysteine-free pseudo-wild-type T4 lysozyme.

E-mail addresses of the corresponding authors: steve@mayo.caltech.edu; brian@uoxray.uoregon.edu

so by single amino acid substitutions seemed limited.[7] A more general and possibly more powerful approach is by using automated design procedures that permit the consideration of multiple substitutions with alternative side-chain packing arrangements.

Several side-chain packing algorithms have been developed in which core redesign has been simplified by placing the side-chains on a rigid template. The side-chain conformations are usually varied by selecting from a library of rotamers, which are defined as statistically significant combinations of dihedral angles of a side-chain.[8] One of the earliest attempts at automated side-chain repacking was implemented in the program, known as propack, developed by Ponder & Richards.[9] Hurley *et al.*[10] used a modification of this program to redesign the C-terminal domain of T4 lysozyme. They considered several hundred promising sequences and energy minimized the best candidates. When constructed, these redesigned proteins folded into native-like structures, but their stabilities were less than that of the wild-type protein.

Programs such as propack make a direct attack on the combinatorial problem of finding the globally optimal arrangement of side-chains on a fixed template. The astronomical number of possible rotamer combinations limits the size of the rotamer library and the number of positions that are allowed to vary in sequence. In addition, these algorithms have no guarantee of finding the structure with the lowest energy.

A different approach that has been developed recently is to iteratively eliminate the so-called dead-ending rotamers, i.e. those rotamers that cannot be part of the lowest-energy structure.[11–13] This improvement allows the extremely rapid testing of the 10^{40} to 10^{60} possible rotamer sequences in a reasonable amount of time, thereby permitting the use of more detailed rotamer libraries and the consideration of larger numbers of sites for repacking.

The optimization of rotamers by iterative techniques (ORBIT) protein redesign program allows use of several alternative versions of the dead-end elimination theorem.[13–16] Several optional terms in the force field and alternative design strategies were developed using feedback from the redesign of two small proteins: the 56 residue β1 domain of streptococcal protein G[14,15,17] and 33 residue peptides that form homodimeric coiled-coils based on GCN4-p1.[13,16] By implementing these strategies, the β1 domain of streptococcal protein G was successfully redesigned with substantially enhanced thermal stability. One variant had a melting temperature in excess of 100 °C and an increase in thermal stability of 4.3 kcal/mol at 50 °C.[18]

For several reasons, it was unclear whether the success of ORBIT with small proteins would be directly transferable to larger ones. For example, the change in exposed surface area on unfolding, as well as the change in heat capacity on unfolding both increase essentially linearly with protein size.[19] Thus a given number of substitutions is likely to have a larger effect on stability when the total number of residues is small. Also, a larger proportion of residues is buried in larger proteins compared to smaller ones.[20,21] This may require the design process to be more stringent. Despite these concerns, successful core designs using ORBIT with an approximately 200 amino acid residue protein have been reported.[22]

To further test the applicability of ORBIT to designs in larger proteins, we used it to redesign the C-terminal domain of T4 lysozyme. Two designs were developed: one without and one with a penalty for the incorporation of methionine. The proteins were constructed, their thermal stabilities measured and their crystal structures determined. To determine the contributions made by individual substitutions, we studied proteins with constituent single mutations as well as proteins with the designed sequences but with a single site changed back to the wild-type sequence.

Results

Redesigned T4 lysozyme

The coordinates of the starting model were from the atomic-resolution crystal structure of the cysteine-free pseudo-wild-type T4 lysozyme, referred to as WT*.[23,24] To obtain the greatest possible accuracy the X-ray diffraction data were collected to 1.05 Å resolution at 100 K (B.H.M.M. & B.W.M., unpublished results). After removal from the coordinate file of the solvent molecules and the alternative side-chain conformations, the crystal structure was partially energy minimized to relieve possible van der Waals clashes and internal-coordinate strain before its use as the starting model in the redesign exercise. The discrepancy between the backbone atom positions in the crystal structure and in the energy minimized structure was 0.21 Å, which is less than that between the 100 K and 293 K crystal structures (0.30 Å) (data not shown). Thus the energy minimization resulted in only small changes in the crystal structure.

T4 lysozyme has an N-terminal and a C-terminal domain. The latter is composed of a tightly packed α-helical bundle and includes residues 1–11 plus 70–164. It includes the most extensive and well-defined hydrophobic core and the redesign was in this part of the molecule. A total of 26 buried or largely buried residues were selected as contributing to the core (see Materials and Methods and Figure 1(a)). The amino acids at these positions were allowed to vary with regard to both their amino acid identity and their side-chain conformation while the remaining residues were held fixed (see Materials and Methods). This resulted in about 3×10^{23} amino acid sequences. By also allowing differing side-chain conformations (see Materials and Methods) the overall number of possible combinations increased to about 4×10^{59}.

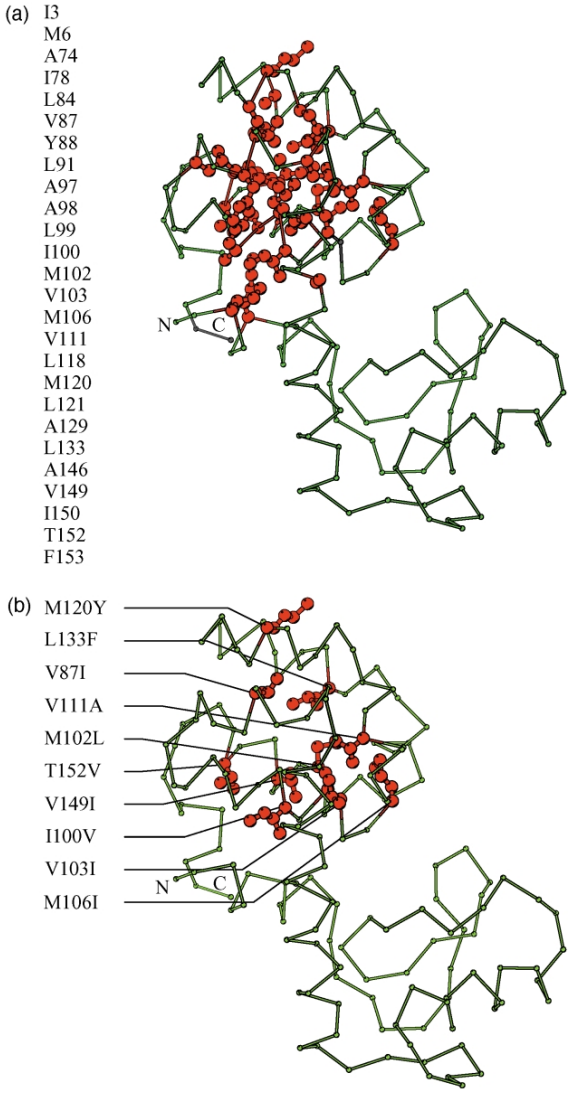


Figure 1. (a) C^{α} trace of the WT* T4 lysozyme backbone showing, in red, the 26 sites that were allowed to vary during the design process. The sites are identified at the left. (b) Structure of T4 lysozyme showing the ten sites that were substituted in Core-10.

Based on the most favorable calculated energy the optimal design selected by ORBIT had seven substitutions (I78V, V87M, L118I, M120Y, L133F, V149I and T152V) and is referred to as Core-7.

This design protocol has been found to lead to an over-representation of methionine residues compared to the occurrence of methionine in natural protein cores (C. A. Sarisky & S.L.M., unpublished results). The larger number of possible rotameric states for methionine (and the lack of an entropy penalty in the force field) leads to a proportionately over-representation of methionine in the rotamer library in comparison to other amino acids. It is also known that methionine-to-leucine substitutions at geometrically appropriate sites can enhance stability.[25] To take these factors into account, the design procedure was repeated with a penalty of 8 kcal/mol for each methionine residue included. With this crude entropy penalty in place, ORBIT selected the tenfold mutant (Core-10), which has the mutations shown in Figure 1(b). In the present instance the effect of the penalty was to both prevent the selection of new methionine residues and eliminate three of the four methionine residues present in the wild-type protein.

In order to obtain calculated energies for the various single, double and other mutants that had been constructed, the same procedure was applied without allowing amino acid sequence variation at the 26 sites. Energies were determined in the presence and absence of the methionine penalty (Table 1).

Thermal stability

Table 1 includes the thermodynamic data for Core-7, Core-10, and the other variants. Neither Core-7 or Core-10, nor any of the revertants is as stable as WT*. The pH of maximum stability for both Core-7 and Core-10 is between pH 5 and pH 5.5 (data not shown). This is similar to WT*[26] and suggests that the strong salt-bridges, especially that between His31 and Asp70, are not significantly perturbed by either set of mutations.

Crystal structures

Structures were determined for almost all of the proteins that had not been analyzed previously (Table 2). Most crystallized isomorphously with WT* in space group $P3_221$. Diffraction data were generally to high resolution with an estimated uncertainty in the main-chain atom positions of 0.1 Å. Although the diffraction data were collected at 100 K, the crystal structures are assumed to be accurate representations of the structure at room temperature. This is supported by comparisons of pairs of 100 K and 293 K crystal structures for the wild-type protein, WT*, and several mutants not included in this study (B.H.M.M. & B.W.M., unpublished results).

Core-7 crystallized in space group *F*222 with two or three molecules in the asymmetric unit and diffracted to 2.4 Å resolution, but it has not been possible to use molecular replacement to solve the structure. Crystals of M87V/Core-7 were also non-isomorphous with WT*. In this case there were three molecules per asymmetric unit and it was possible to determine the structure to 1.56 Å resolution. Crystals of the single-site revertant I118L/Core-7 were isomorphous with WT* and the structure was determined to high resolution (Table 2). As will be apparent from the behavior of Core-10 revertants, however, the structures of M87V/Core-7 and I118L/Core-7 cannot be reliably used to infer the structure of Core-7 itself.

The redesigned protein Core-10 crystallized isomorphously with WT* and its structure was

Table 1. Stabilities of mutant lysozymes

Mutant	ORBIT score (kcal/mol)	ORBIT score with methionine penalty (kcal/mol)	Δt_m (deg. C)	ΔH (kcal/mol)	$\Delta\Delta G$ (kcal/mol)	Non-additivity of $\Delta\Delta G$ (kcal/mol)
I78V	−364	−332	−2.1	127	−0.8	–
V87M	−349	−309	−6.3[a]	113[a]	−2.3[a]	–
L118I	−352	−320	−3.1	123	−1.2	–
V87I	−362	−330	−0.8	127	−0.3	–
I100V	−365	−333	−1.1	129	−0.4	–
M102L	−316	−292	−2.3	118	−1.0	–
V103I	−295	−263	−1.5	130	−0.5	–
M106I	−362	−338	0.6	132	0.2	–
V111A	−354	−322	−2.9	121	−1.1	–
M120Y	−365	−341	−0.1	126	−0.1	–
L133F	−368	−336	−0.7	130	−0.3	–
V149I	−362	−330	−0.3	128	0.0	–
T152V	−365	−333	0.8[b]	127[b]	0.2[b]	–
Core-7	−382	−350	−9.8	103	−3.5	1.0
M87V/Core-7	−368	−344	−5.0	117	−3.0	−0.8
I118L/Core-7	−380	−348	−9.5	103	−3.3	0.0
Core-10	−371	−363	−6.4	97	−2.4	1.1
L102M/Core-10	−372	−356	−7.2	101	−2.6	−0.1
I103V/Core-10	−370	−362	−4.0	110	−1.6	1.3
A111V/Core-10	−269	−261	−4.8	106	−1.8	0.5
WT*	−362	−330	0.0	132	0.0	–

The first two columns give the score calculated by ORBIT, respectively, without and with a penalty for incorporation of methionine residues (see the text). Δt_m is the change in melting temperature relative to WT* which is 65.5 °C under these conditions. ΔH is the enthalpy of unfolding at t_m. $\Delta\Delta G$ is the change in the free energy of unfolding relative to WT*. Non-additivity of $\Delta\Delta G$ is the difference between $\Delta\Delta G$ measured for the multiple construct and the sum of the $\Delta\Delta G$ values for the constituent single mutants. Uncertainties in Δt_m are about ±0.2 deg. C, in ΔH about ±5% and in $\Delta\Delta G$ about 0.15–0.4 kcal/mol (increasing from the most stable to the least stable mutants). As is also explained in the text, more negative ORBIT scores correspond to proteins that are predicted to be more stable, whereas more negative $\Delta\Delta G$ values correspond to proteins that are of lesser stability.

[a] From Gassner *et al.*[51]

[b] From Xu *et al.*[31] Note that ΔH is a corrected value.

determined to 1.65 Å resolution. The structure is generally similar to WT* but also has some distinct differences in both the backbone structure and the side-chain conformations (Figure 2(a)). The average discrepancy between the main-chain atoms of residues 81–161 in Core-10 and WT* is 0.49 Å (Table 4), which is about three times the combined uncertainty in the positions of the backbone atoms in each structure. When sites 106–123 are excluded from the least-squares superimposition to avoid incorporating the effect of the shifts in helices F and G, the discrepancy is 0.21 Å

Table 2. Crystal and refinement statistics

Protein	Cell dimensions a, b (Å)	Cell dimensions c (Å)	Resolution (Å)	R_{merge} (%)	Completeness (%)	R-factor (%)	Δ_{bonds} (Å)	Δ_{angles} (deg.)	PDB code
I78V	60.0	95.23	1.58	4.6	94.6 (88)	18.8	0.011	2.3	1P2R
V87I	59.6	95.3	1.58	5.5	94.1 (77)	16.8	0.015	2.3	1P2L
I100V	59.8	95.6	1.45	6.0	97.9 (86)	18.7	0.015	2.4	1P36
V103I	[a]	[a]	1.5	5.6	95.3 (78)	19.2	0.012	2.1	1P7S
M106I	60.1	95.6	1.67	4.6	96.9 (84)	17.9	0.014	2.3	1P46
L118I	60.2	95.9	1.65	4.9	94.0 (91)	20.1	0.013	2.3	1PQO
M120Y	60.3	95.3	1.54	5.1	97.7 (96)	18.6	0.013	2.4	1P6Y
L133F	60.1	96.2	1.62	4.4	96.5 (79)	18.6	0.012	2.3	1P64
V149I/T152V	59.8	95.4	1.52	5.8	93.5 (71)	17.6	0.016	2.5	1PQM
M87V/Core-7	[b]	[b]	2.0	5.2	96.0 (91)	18.6	0.020	2.9	1PQK
I118L/Core-7	60.0	95.6	1.56	4.9	91.1 (91)	19.8	0.016	2.7	1PQI
Core-10	60.0	96.6	1.65	7.6	96.9 (85)	17.8	0.016	2.7	1PQD
L102M/Core-10	59.5	96.2	1.57	6.1	90.6 (75)	17.7	0.018	2.5	1P37
I103V/Core-10	60.0	95.9	1.55	6.0	96.7 (97)	18.7	0.015	2.3	1P3N
A111V/Core-10	59.5	95.5	1.90	6.1	99.0 (100)	18.8	0.016	2.6	1PQJ

[a] WT crystallized in space group $P3_221$. V103I crystallized in space group $P2_12_12_1$ with cell dimensions $a = 30.8$ Å, $b = 54.9$ Å and $c = 88.4$ Å.

[b] M87V/Core-7 crystallized in space group $C2$ with cell dimensions $a = 156.5$ Å, $b = 61.9$ Å, $c = 67.4$ Å, $\beta = 112.3°$.

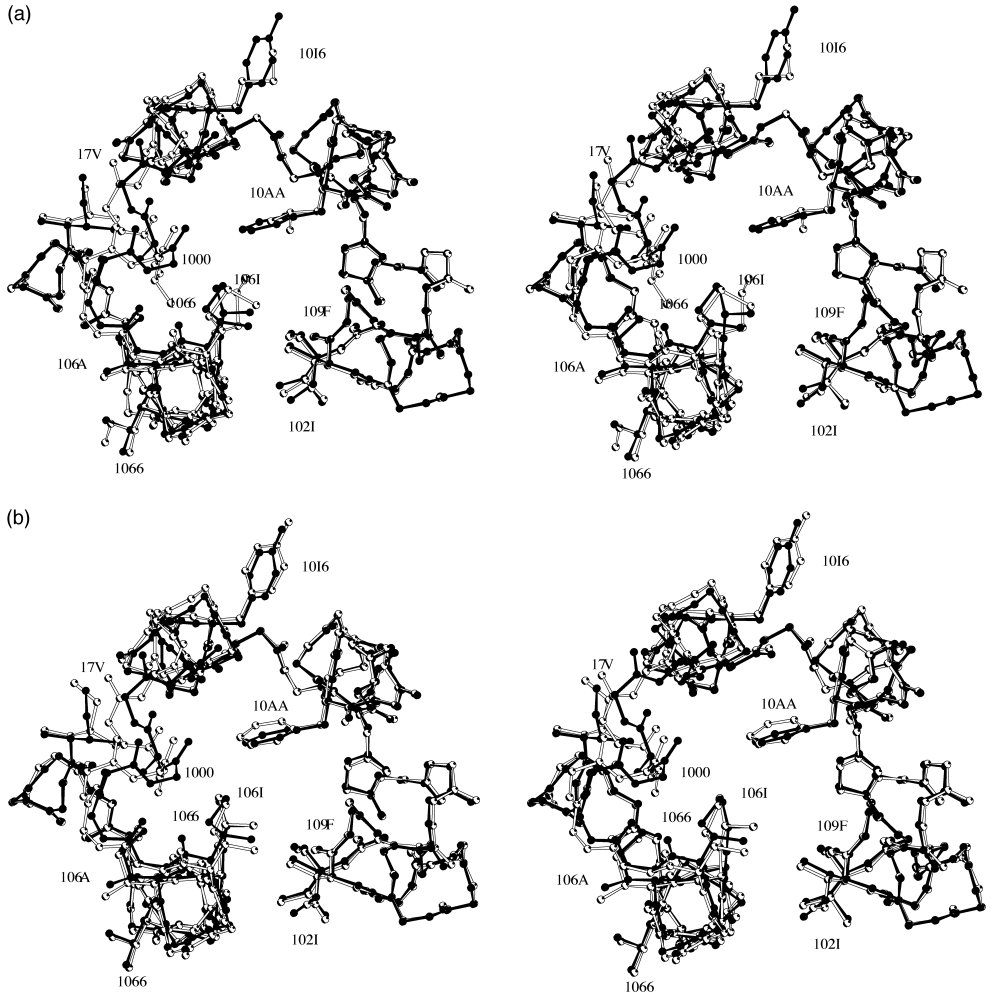


Figure 2. (a) Stereo view showing the superposition of the crystal structure of Core-10 (filled bonds) on the crystal structure of WT* (open bonds). For clarity only the side-chains of the ten substituted residues are shown. (b) Superposition of designed Core-10 (open bonds) onto the observed crystal structure (filled bonds).

(Table 3). This shows that the backbone structure of most of the C-terminal domain is well conserved, but within helices F and G some atoms move substantially (up to about 2.8 Å) (Figure 3(a)). The shift in helix F is associated with the breaking of the hydrogen bond between Thr109 O and Gly113 N. This distance increases from 3.0 Å to 4.2 Å. The breaking of this hydrogen bond was also observed in the crystal structure of the single mutant Val111 → Ile.[10] The outward shift of helix F creates a cavity to which a water molecule, HOH508, binds and is within hydrogen bonding distance of Ala111 O (2.9 Å with a C–O···HOH angle of 100°).

The temperature factors for the side-chain atoms at the ten sites of mutation in the crystal structures of Core-10 and of WT* are quite similar, indicating that these side-chains are not disordered; nor is there any indication of a molten globular state (Table 4).

Comparison of the crystal structures of WT* and Core-10 reveals that the side-chain rotameric states are completely conserved at all of the non-substituted sites (Table 5). Conservation also occurred at all but one of the substitution sites, the single exception being M102L, where both χ_1 and χ_2 changed (Table 5; Figure 2(a)).

Discussion

The overall objective of the present experiments was to use ORBIT to identify variants of T4 lysozyme that had repacked cores and were more stable than wild-type. The most promising variant identified by the design process, Core-7, was found to be a functional lysozyme but with melting temperature reduced by 9.8 deg. C, which corresponds to a destabilization of 3.5 kcal/mol relative to WT*. Change in the design procedure to include a penalty for methionine residues led to a modified design, Core-10, which was 1.1 kcal/mol more stable than Core-10 but still not equal to WT*. In

Table 3. Backbone shifts in designed and mutant T4 lysozymes

Protein	Shift, C-terminal domain (Å)	Shift, C-terminal domain without helices F and G (Å)
Core-10 design	0.19	0.23
Core-10 crystal	0.49	0.21
L102M/Core-10	0.49	0.22
I103V/Core-10	0.22	0.18
A111V/Core-10	0.55	0.26
M87V/Core-7		
Molecule A	0.40	0.36
Molecule B	0.49	0.50
Molecule C	0.44	0.46
I118L/Core-7	0.28	0.25

Each entry in the Table gives the root-mean-square (rms) difference between the main-chain atoms of the specified structure and WT*. The column labeled C-terminal domain gives the rms shift for essentially the whole C-terminal domain (i.e. for residues 81–161). The column labeled C-terminal domain without helices F and G gives the rms shifts for residues 81–105 plus 124–161. Superpositions were carried out using EDPDB.[52]

the following sections we discuss these findings in more detail with their implications for future design initiatives.

Energetics of the designed variants

As noted above, neither of the designed variants was as stable as wild-type lysozyme. This is at variance with the success of ORBIT in predicting stabilized variants of the β1 domain of protein G and coiled-coils based on GCN4.[13–15,17,18] It is, however, in agreement with earlier experiments on T4 lysozyme. Hurley *et al.*[10] used a computational procedure to identify combinations of amino acids that would repack the core. Some possible combinations were suggested but their stability was, at best, slightly less than the native molecule. Also Baldwin *et al.*[3] used a genetic approach to select variants that had repacked cores. Again, a large number of variants were identified, but none had stability greater than that of WT*.

One possible inference of these results is that it may be energetically more costly to repack larger proteins than smaller ones. In a very small protein most side-chains may be at least partly in contact with solvent. This may allow them freedom to be substituted, or to adjust their positions in response to substitutions at nearby sites. Within the core of a larger protein the side-chains tend to be tightly packed by their neighbors and it is more difficult for the structure to relax in response to introduced changes.

Table 4. Comparison of temperature factors at mutated sites in WT* and Core-10

	Main-chain B (Å^2)		Side-chain B (Å^2)	
Residue	WT*	Core-10	WT*	Core-10
87	13.1	20.0	17.7	26.0
100	11.6	17.2	13.0	15.4
102	12.3	17.0	13.5	17.3
103	13.5	19.9	15.7	23.3
106	17.6	21.4	16.6	21.5
111	18.1	33.5	16.7	27.8
120	12.1	15.9	16.4	19.4
133	11.1	13.9	12.3	14.4
149	10.2	16.3	11.4	15.1
152	11.3	19.2	11.4	14.1

The Wilson B-value is 14.4 Å^2 for WT* and 17.9 Å^2 for Core-10.

Calculated and observed stabilities

The stabilities of the various T4 lysozymes predicted by ORBIT are compared with those determined experimentally in Table 1 and Figure 4. These two energy terms do not have the same definition, but they are expected to correlate. For T4 lysozyme the experimental $\Delta\Delta G$ is traditionally defined to be the free energy of unfolding relative to the WT* protein.[27,28] $\Delta\Delta G$ refers to the free energy of unfolding and a positive value indicates that the protein is more stable than WT*. The ORBIT score is the sum of the calculated energies of interactions for the side-chains that are allowed to vary. The energetically more favorable ORBIT scores are in the negative direction. For the individual mutations, excluding sites 102 and 103, there is a possible correlation between the calculated and observed energies but when all sites and all constructs are considered no clear-cut relation emerges (Figure 4(a) and (b)). This lack of agreement between the experimental and the predicted energies could be due to a number of factors, including the following. (1) Some of the mutant proteins experience significant changes in the main chain (see below). These may invalidate the rigid template assumption. (2) The rotameric states of some of the side-chains in the calculated structures (in particular, in Core-10) do not agree with those in the actual proteins, leading to inaccurate energies. (3) The rotamer library used in the design process may not be detailed enough to capture the tight packing seen in the core of T4 lysozyme. (4) The force field, which was based in part on experience with smaller proteins, may not be optimal for all proteins.

Predicted and observed structure of Core-10

Figure 2(b) compares the backbone of the predicted and observed structure of Core-10. For residues 81–105 plus 124–161 the backbone agreement is generally good but in the remaining region there are shifts up to 2.8 Å. Likewise, most but not all of the side-chain conformations are correctly predicted. Eight of the ten modified side-chains adopt the rotameric state that was predicted (Table 5). The two exceptions are Ile87 and Ile149, in which cases the differences are restricted to the χ_2 torsion angle. At another site (V103I) the χ_2 angle differs from the predicted value by more than 30°. Of the ten non-alanine residues that were included in the design process but did not change

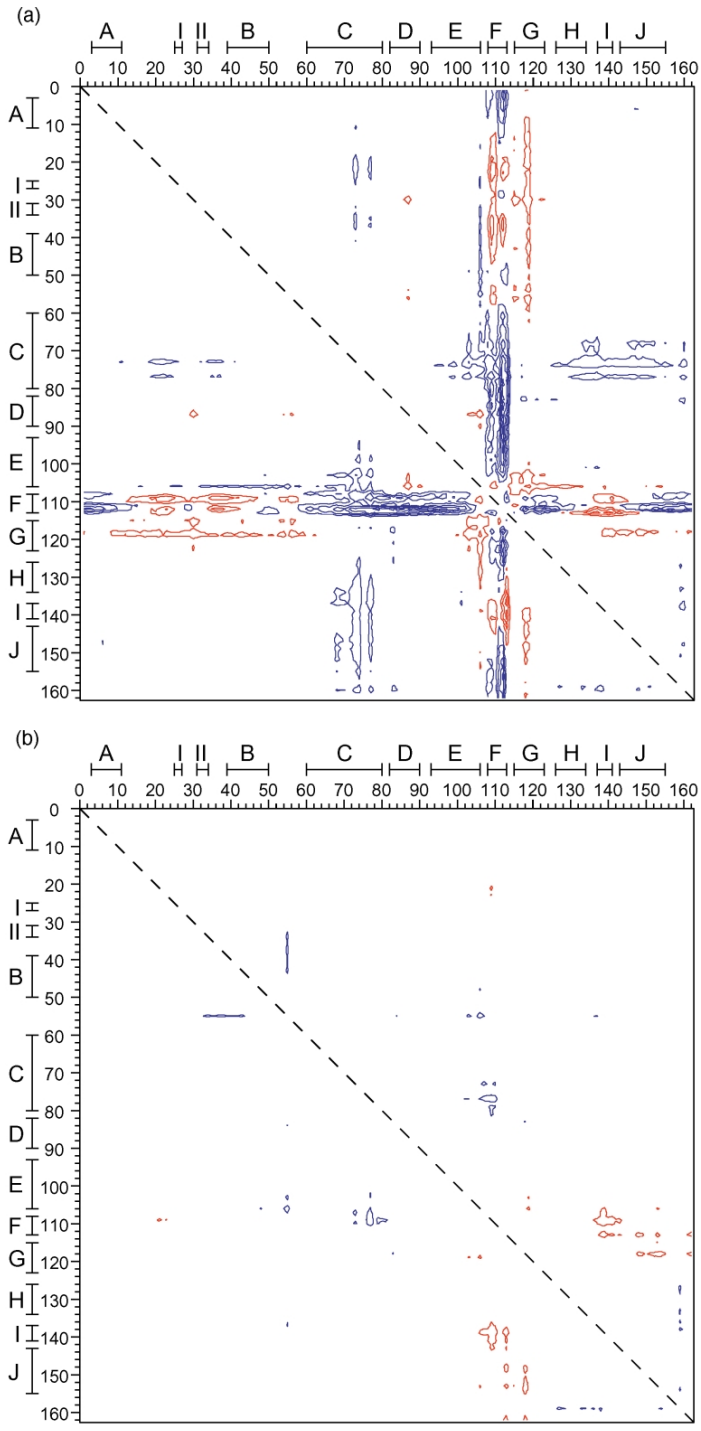


Figure 3. Plots showing differences in C^{α}–C^{α} separation in different crystal structures. The contours start at ± 0.5 Å and have 0.5 Å intervals. The red contours correspond to decreased separation and the blue contours correspond to an increase in distance. (a) Core-10 *versus* WT*. (b) I103V/Core-10 *versus* WT*.

identity, nine had correctly predicted rotamers. The three incorrect predictions plus the prediction that is somewhat in error are discussed briefly below.

$$\text{Val87} \rightarrow \text{Ile}(\Delta\chi_1 = 25°, \Delta\chi_2 = -96°)$$

With reference to the crystal structure of WT*, the introduction of the CD1 carbon atom of Ile87 is associated with the outward movement of the side-chains of Leu118 and Glu122 as well as other shifts (Figure 5(a)). Notwithstanding these shifts, potential steric clashes appear to cause the CD1 methyl group of Ile87 to adopt a rotameric state that is fairly uncommon (frequency of 14%).[29] Before the design process, the starting model was partially energy minimized. Comparison of the design to the crystal structure of WT* shows that

Table 5. Comparison of the side-chain torsion angles at the 26 sites open to modification in T4 lysozyme

Site	WT* (χ_1,χ_2)	Energy-minimized WT* (χ_1,χ_2)	Core-10 design (χ_1,χ_2)	Core-10 crystal structure (χ_1,χ_2)
3	Ile (183, 57)	Ile (186, 52)	Ile (185, 62)	Ile (186, 60)
6	Met (184, 201)	Met (188, 197)	Met (185, 191)	Met (191, 213)
78	Ile (285, 163)	Ile (290, 161)	Ile (285, 168)	Ile (277, 163)
84	Leu (302, 175)	Leu (305, 175)	Leu (300, 174)	Leu (304, 173)
87	Val (307, –)	Val (309, –)	Ile (282, 49)	Ile (307, 313)
88	Tyr (184, 82)	Tyr (185, 82)	Tyr (178, 93)	Tyr (186, 85)
91	Leu (297, 168.5)	Leu (291, 173.1)	Leu (268, 32.1)	Leu (292, 170.8)
100	Ile (296, 163)	Ile (305, 158)	Val (303, –)	Val (292, –)
102	Met (293, 186)	Met (297, 186)	Leu (193, 63)	Leu (178, 68)
103	Val (293, –)	Val (293, –)	Ile (282, 49)	Ile (283, 17)
106	Met (76, 182)	Met (72, 179)	Ile (65, 172)	Ile (78, 183)
111	Val (303, –)	Val (305, –)	Ala (–, –)	Ala (–, –)
118	Leu (293, 169)	Leu (288, 167)	Leu (289, 176)	Leu (291, 167)
120	Met (300, 175)	Met (294, 171)	Tyr (276, 153*)	Tyr (283, 124*)
121	Leu (290, 172)	Leu (295, 170)	Leu (282, 174)	Leu (289, 177)
133	Leu (282, 164)	Leu (286, 161)	Phe (271, 104*)	Phe (267, 114*)
149	Val (296, –)	Val (299, –)	Ile (298, 169)	Ile (296, 280)
150	Ile (292, 171)	Ile (289, 168)	Ile (286, 169)	Ile (288, 176)
152	Thr (307, –)	Thr (311, –)	Val (296, –)	Val (297, –)
153	Phe (280, 303)	Phe (280, 299)	Phe (275, 324)	Phe (282, 308)

The torsion angles are listed for the crystal structure of WT*, the energy minimized model of WT* used in the design process, the predicted structure of Core-10, and the observed crystal structure. The five sites that started as alanine and remained alanine (sites 74, 97, 98, 129, 146) are not shown. The IUPAC conventions for determining χ_1 and χ_2 were followed except for the following two changes, which were made to simplify comparison of unlike side-chains. (1) Following Blaber *et al.*[29] the χ_1 torsion angle of valine was measured using the CG2 carbon atom rather than CG1 as in the standard IUPAC nomenclature. This is about the same as increasing χ_1 by 120° and makes the *gauche* − , *trans* and *gauche* + conformations for valine the same as for the other amino acids. (2) For the phenylalanine and tyrosine side-chains marked with an asterisk, the χ_2 value was decreased by 180°. This change essentially corresponds to a renaming of the ring atoms. χ_3 values are not shown but in general agree fairly well at any given site. At sites 6, 102, 106 and 120 the maximum discrepancy in χ_3 among the structures being compared is, respectively, 21°, 2°, 19°, and 3°.

the distal part of the side-chain of Gln122 has moved outward in the designed structure. At least in part, this suggested that an isoleucine residue in a common rotameric state could be accommodated at this site. (Gln122 is a surface residue that was held fixed during the design process. Thus, its outward movement is the result of the energy minimization step and not the rotamer selection step.)

$$\text{Leu91}(\Delta\chi_1 = 24°, \Delta\chi_2 = 139°)$$

Leu91 was included in the design process but its identity remained unchanged. It was predicted, however, that the two methyl groups at the end of the leucine side-chain would flip by about 180°. This change is not observed. Rather, the conformation of Leu91 in Core-10 is essentially identical with WT*. The change in conformation in the designed structure presumably occurs in concert with the introduction of isoleucine at site 87. As mentioned above, Ile87 is predicted to have an altered conformation in Core-10. To avoid close contact with this residue Leu91, in the designed structure Ile87 adopts a different rotamer. Thus the error in prediction at the two sites seems to be coupled

$$\text{Val103} \rightarrow \text{Ile}(\Delta\chi_1 = 1°, \Delta\chi_2 = -32°)$$

The side-chain of Ile103 adopts a rotameric state, which has a frequency of only 3% among proteins in general. This is essentially as predicted although the observed χ_2 is 32° from that anticipated. The distal methyl groups of the side-chain adopt positions that are close to those predicted (Figure 5(b)). This coincidence occurs in spite of the change in the side-chain torsion angle and extensive shifts in several of the surrounding residues (especially 106–111). The superimposition of the crystal structure of WT* on the crystal structure of Core-10 suggests that these shifts may be caused in part by the need for Val111 to avoid a close contact with the CD atom of Ile103. Since the design process assumes a rigid framework, such backbone shifts are not anticipated.

$$\text{Val149} \rightarrow \text{Ile}(\Delta\chi_1 = -2°, \Delta\chi_2 = 111°)$$

Ile149 was predicted to adopt the most common rotameric state for isoleucine, which has a frequency of 57%. Instead, it adopts a rotameric state that has a frequency of 14%. The design procedure deleted the four water molecules that are bound within the T4 lysozyme molecule.[30] The removal of one of these resulted in a cavity that the CD methyl group of Ile149 was predicted to occupy. In actuality, the water molecule remains bound to Core-10 and forces the isoleucine to adopt an alternative rotamer. (The water HOH197 shifts by 0.7 Å but retains its hydrogen bonding partners (Figure 5(c)).)

Internal water molecules

Four buried water moleculess occupy three

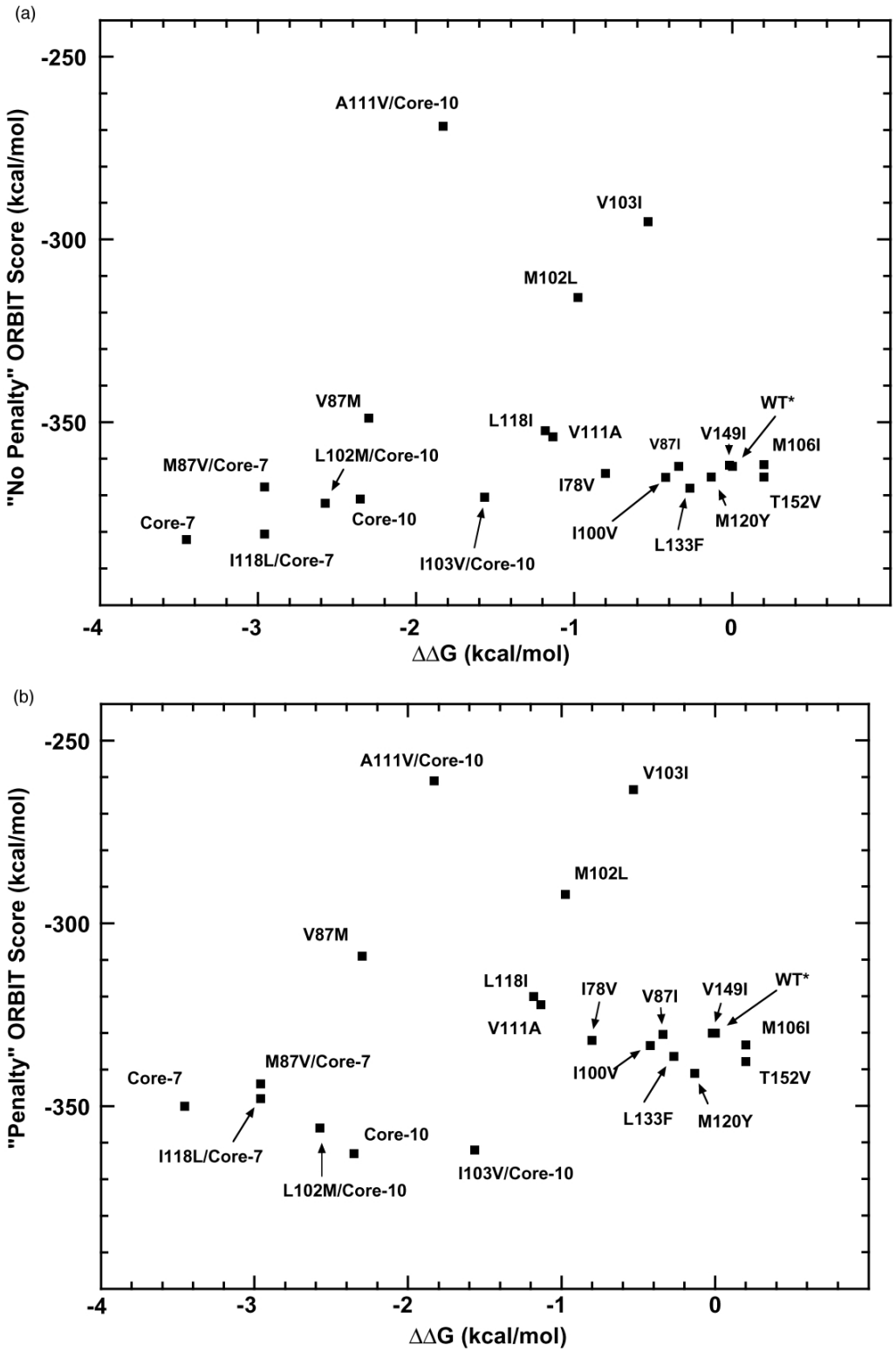


Figure 4. Comparison of the energies calculated using ORBIT with the observed protein stability (Table 1). (a) Comparison of single and multiple mutants with the ORBIT score determined without a penalty for incorporation of methionine. (b) Comparison of single and multiple mutants with the ORBIT score determined with a penalty for incorporation of methionine.

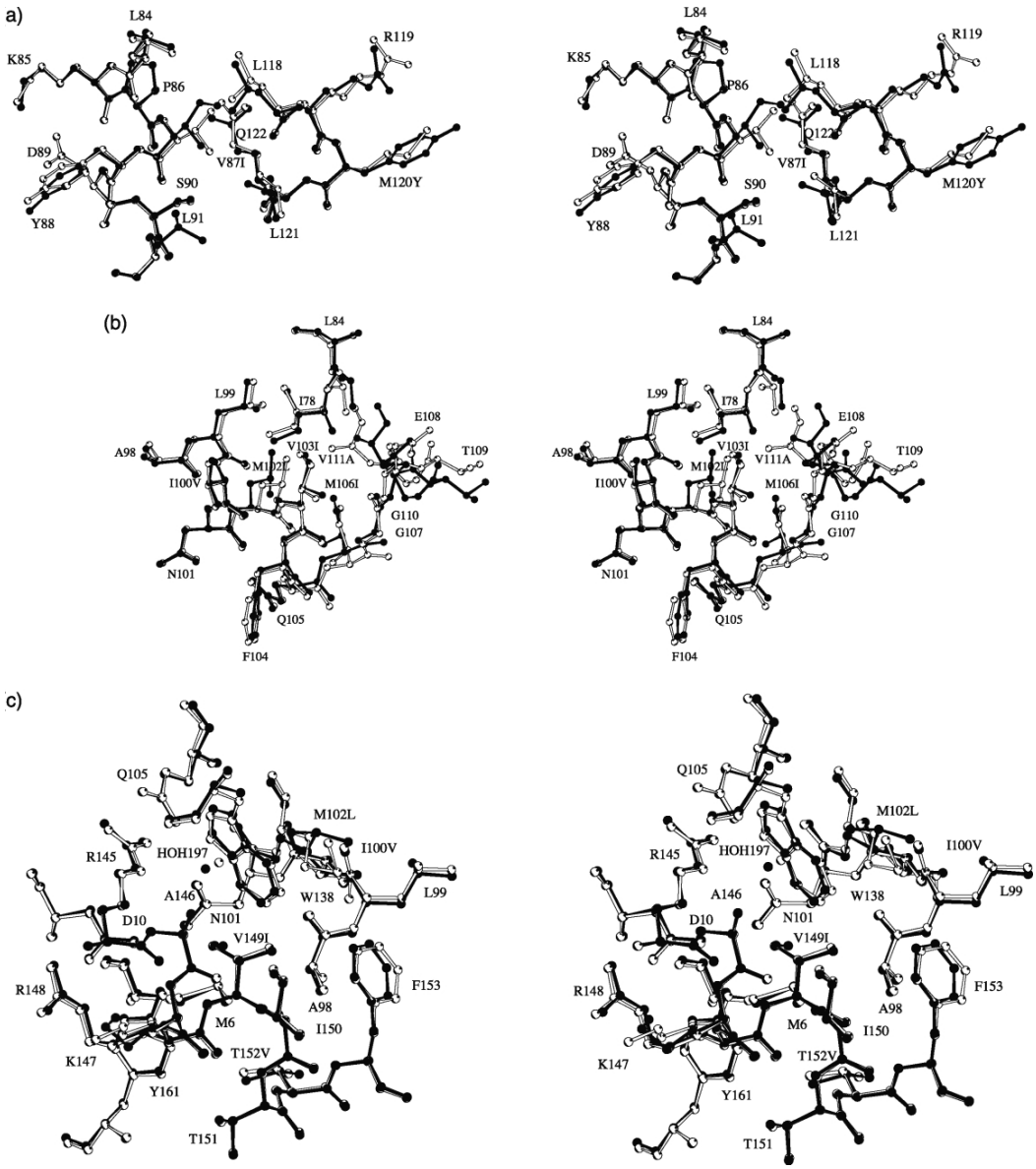


Figure 5. Stereo views of the sites of discrepancies between the predicted and observed structure of Core-10. (a) Crystal structure of WT* (open bonds) superimposed on the crystal structure of Core-10 (filled bonds) in the vicinity of site 87. (b) Design of Core-10 (open bonds) superimposed on the crystal structure of Core-10 (filled bonds) in the vicinity of site 102. (c) Crystal structure of WT* (open bonds) superimposed on the crystal structure of Core-10 (filled bonds) in the vicinity of site 149.

cavities in WT*.[30,31] These four water molecules were removed from the coordinate file during the design process. Two of the cavities are in the C-terminal domain and were therefore available for repacking by side-chain atoms. The first of these two cavities is next to site 149 and has already been discussed. The second cavity decreases slightly in the designed structure following the replacement of Thr152 with valine. In the crystal structure, however, the water molecule (HOH173) still appears in the cavity although it is displaced towards the surface of the protein by about 1 Å.

Core-10 revertants

Selected single-site revertants were constructed to address, both energetically and structurally, how the different sites interact with each other. The sites chosen for reversion were those where the point mutant had the largest effect on the stability of WT* (Table 1). The single-site revertants of Core-10 are discussed briefly below.

Leu102Met/Core-10

In the revertant L102M/Core-10, the leucine

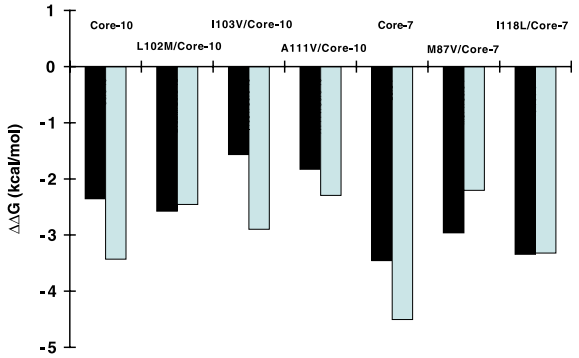


Figure 6. Comparison of the observed stabilities of the multiple mutants (ΔΔ*G*; Table 1; filled bars) with the sums of the ΔΔ*G* values of the constituent single mutants (grey bars).

residue at site 102 in Core-10 was changed back to methionine as in the wild-type sequence. The structure, however, remains very similar to that of Core-10 (rmsd of 0.14 Å). Thus, the amino acid change at site 102 back to that of the wild-type sequence does not recover the backbone atoms positions of the WT* structure. Met102 in the revertant adopts a side-chain conformation that is very similar to that of Met102 in WT* but that differs from that of Leu102 in Core-10 by a rotation of 85° about the χ_1 torsion angle.

In the L102M/Core-10 revertant, the sum of the ΔΔ*G* values of the remaining nine constituent mutants is essentially the same as the measured ΔΔ*G* for the revertant (Table 1, Figure 6). This suggests that each of these nine sites is acting independently and that there is no interaction between them.

Ile103Val/Core-10

The discrepancy between revertant I103V/Core-10 and Core-10 for the main-chain atoms from sites 81–161 is 0.47 Å, while it is only 0.22 Å relative to WT*. Thus the change in this single site back to the wild-type sequence is sufficient to revert the C^{α} positions in Core-10 essentially back to those of WT* (Figure 3(b)). (It should be noted that the discrepancy between Core-10 and V103I is 0.68 Å, showing that the introduction of this single mutation is not sufficient to cause all the structural changes seen in Core-10. At the same time, the single mutant V103I crystallized in a different space group and has a hinge-bending motion relative to WT*. This results in shifts in the C terminus of helix C, which makes detailed structure comparison more difficult.)

The change back to a valine from an isoleucine residue at site 103 removes a buried methyl group. This is correlated with Ala111 moving into a position similar to that occupied by Val111 in WT* and with helix F reverting to its wild-type conformation. It appears that the potential clash between the Ile103 CD1 methyl group and the CB methyl group of Ala111 causes helix F to move outwards. The I103V revertant resulted in an 0.8 kcal/mol increase in stability relative to Core-10. This is notwithstanding the decrease in hydrophobicity resulting from the Ile to Val substitution and clearly suggests that the original V103I replacement introduces strain in the Core-10 structure.

The I103V/Core-10 revertant shows the largest non-additivity in ΔΔ*G* of all the variants studied (Figure 6). This also suggests that the remaining nine sites have the greatest degree of repacking and synergistic interaction.

Ala111Val/Core-10

In the Core-10 revertant A111V/Core-10, the alanine at position 111 in the Core-10 background is changed back to valine as in the wild-type sequence. If residues in the vicinity of site 111 in Core-10 were tightly packed, it would be expected that the introduction of two methyl groups would result in large structural changes. This, however, is not the case. The observed changes are actually modest. Val111 moves closer to the core by about 0.3 Å compared to Ala111 in Core-10, and atoms surrounding the reintroduced valine side-chain move by at most a few tenths of an ångström unit (Figure 7(a)). The two methyl groups of the valine essentially refill the cavity that was created by the V111A substitution in Core-10. The most dramatic change in atomic position in the revertant is a 2 Å movement of the CD1 atom in the side-chain of Ile103. This movement occurs largely by a rotation about the χ_2 angle to an energetically unfavorable rotameric state which places the CD1 atom at a distance of 2.7 Å from Ile103 CG2 atom (as opposed to 3.8 Å in Core-10).

The A111V reversion increases the stability of Core-10 by 0.6 kcal/mol (Table 1). The fact that this is an increase rather than a decrease also suggests that the valine side-chain occupies a preformed cavity and does not introduce any serious steric clashes.

Evidence for synergy between the mutation sites

One can ask whether the ORBIT procedure results in genuine repacking of the core or, conversely, the individual substitutions act independently. In the case of Core-10 none of the constituent point mutations causes a large change in stability. Six of the ten mutations change the melting temperature by less than 1.0 deg. C and the largest effect is for V111A, for which the change is 2.9 deg. C (Table 1). If each of the substitutions acts independently of the others the change in stability of the multiple mutant should equal the sum of the ΔΔ*G* values of its single-site constituents. As can be seen in Table 1 and Figure 6, the sum of the ΔΔ*G* values for Core-10 is numerically 1.1 kcal/mol greater than the observed ΔΔ*G*.

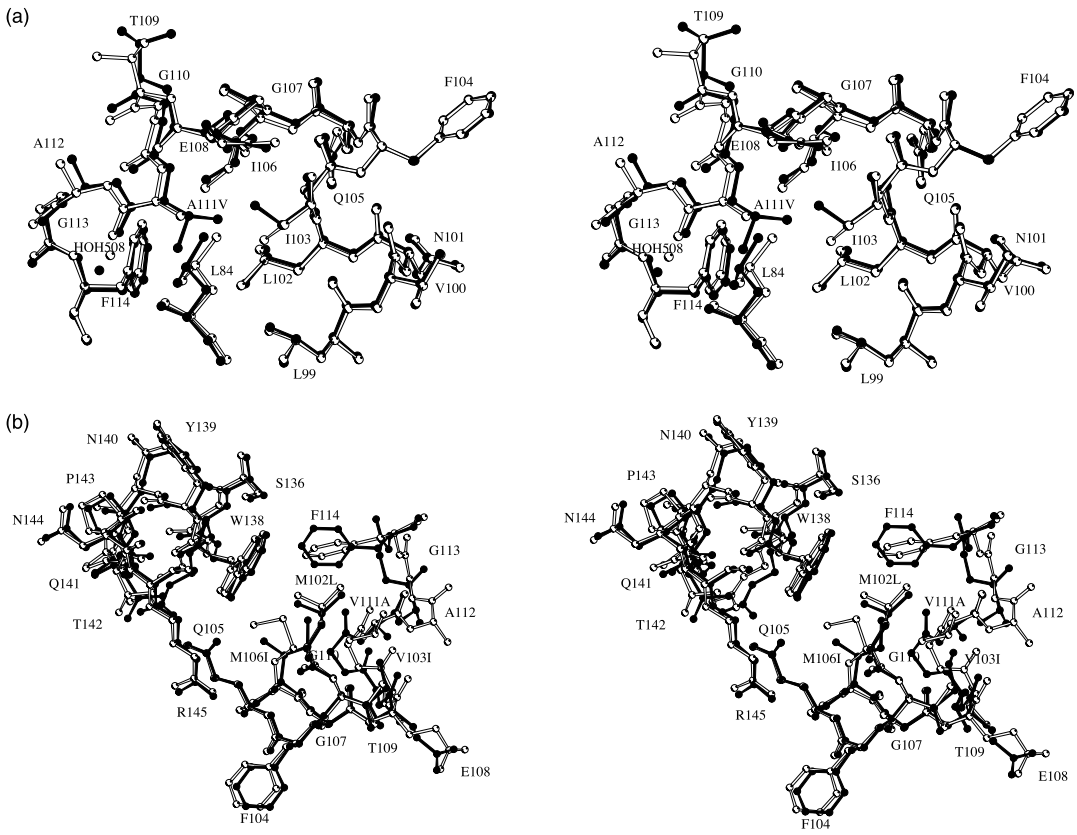


Figure 7. (a) Stereo diagram showing the superposition of the single-site revertant A111V/Core-10 (filled bonds) on Core-10 (open bonds). (b) Superposition of the structure of the single mutant, M102L[10] (open bonds) on Core-10 (filled bonds).

This shows that there is some favorable interaction among the redesigned sites, although the effect is modest. By way of comparison, in the "size switch" mutant in which the sizes of adjacent residues were switched by the substitutions L21A and A129L, the thermodynamic compensation was substantially larger (2.5 kcal/mol).[4]

Cooperativity between substitutions at different sites can also be evaluated structurally. Using a cut-off distance of 4.0 Å the average number of residues among the 26-residue set that are in contact or almost in contact with any given residue is 2.4 (or 1.4 residue–residue contacts if the threshold is reduced to 3.5 Å). Thus, even though the 26 residues are all within the most pronounced hydrophobic core of T4 lysozyme there do not tend to be multiple close contacts between each residue and a multitude of neighbors. This separation of the sites may make cooperativity difficult to achieve. In the present case the design algorithm assumes that selected variants will retain the same backbone structure as the parent molecule. As noted above, this is true for much of the C-terminal domain of Core-10, but not in the vicinity of the F and G helices.

In this context it is instructive to contrast the behavior of the Core-10 revertant L102M/Core-10 with that of I103V/Core-10. When the single-site reversion I103V is made in Core-10 the structure reverts much closer to that of WT* (Figure 3(b)). Also the stability of the protein is increased by 0.8 kcal/mol and, in addition, the non-additivity of the $\Delta\Delta G$ values increases by 0.2 kcal/mol (Table 1, Figure 6). When the V103I mutation is included in the full Core-10 construct, the addition of the CD1 methyl group introduces a steric clash that is not compensated by the other replacements and, therefore, leads to a relatively large change in the structure.

In contrast, the behavior of the L102M/Core-10 revertant is quite different. Here the reversion of Leu102 to Met causes almost no change in the Core-10 structure. At the same time (as judged by the equivalence of the $\Delta\Delta G$ values; Figure 6), it eliminates any synergistic interaction between the remaining nine sites. The L102M/Core-10 structure seems "poised" to accept the M102L substitution without structural perturbation, and, in so doing, the Leu102 side-chain contributes to the synergistic interaction that is observed in Core-10.

Since the L102M revertant in Core-10 eliminates synergistic interaction between the remaining nine sites, it implies that the M102L substitution does contribute to cooperativity in Core-10. There is

some structural evidence for this. When the M102L mutation is made in WT* it results in a rotation of the side-chain of Phe114 by almost 70° into a strained conformation. (This rotation appears to be mediated indirectly *via* Trp138 and possibly other residues as well.) In Core-10 (and in M102L/Core-10), however, the combination of substitutions allows the side-chain of Phe114 to revert to the angle seen in WT* (Figure 7(b)), relaxing the strain that had been introduced.

Success of the methionine penalty

Because of their conformational adaptability, methionine side-chains tend to be more readily accommodated within a designed protein. At the same time incorporation of multiple methionine residues can result in a loss of stability.[5,32] Conversely, under favorable circumstances substitutions from methionine to leucine can increase stability.[25] For these reasons it would seem desirable to avoid the introduction of methionine residues into the designed protein.

In the present case the imposition of a methionine penalty resulted in four positions in Core-7 being retained in Core-10 while I78V and I118L were lost and V87M was replaced with V87I. Meanwhile, five new positions were added, resulting in the loss of two methionine residues: I100V, M102L, V103I, M106I, and V111A. In total, Core-10 has three fewer methionine residues than Core-7. The M102L substitution is known to introduce steric clashes[10] and it could be that the additional sites of substitution in Core-10 arise from the need to minimize this steric interference. In any event, the incorporation of the methionine penalty did increase the stability of the protein by 1.1 kcal/mol (relative to Core-7).

Conclusions

One of the main findings of this work is that the introduction of the designed core-repacking mutations resulted in changes of the backbone up to 2.8 Å. Also both of the designed variants were less stable than the wild-type protein. Taken together, these results suggest that genuine core repacking with retention or enhancement of stability may be difficult if not impossible to achieve without provision for shifts in the backbone.

A second finding is that the rotamer angles that occur in WT* are strongly conserved in the mutant. For the substituted and non-substituted sites in Core-10 there is only one case (Met102Leu) where there is a change of rotamer (Table 5, Figures 2(a) and 5(c)). Conservation of rotamers was also observed in genetically selected core-repacking variants of T4 lysozyme.[3] This suggests that core redesign might be improved by favoring models that maintain the side-chain rotamers present in the reference structure.

If, as was the case with the Core-10 design, a total of 26 sites were allowed to vary, the overall number of possible sequence combinations is astronomical. At a given site, however, the packing is typically determined by the side-chain itself plus two or three neighbors. Here, the number of choices is more limited. Also since the number of hydrophobic amino acids is fairly small, and each amino acid is restricted to distinct rotamers, the choice of substitutions is "quantized".[7] On the other hand, if the backbone were allowed to move it would allow a wider range of substitutions to be considered.

Materials and Methods

Redesign by ORBIT

All residues of cysteine-free pseudo-wild-type T4 lysozyme, referred to as WT*, were classified as surface, core, or boundary, using a residue classification program, RESCLASS.[14,15] RESCLASS classifies the residues based on their C^{α} and C^{β} distances from a solvent-accessible surface, which is calculated using the Connolly algorithm.[33]

We selected 26 core positions located in the C-terminal domain of WT* for design. The selected positions were I3, M6, A74, I78, L84, V87, Y88, L91, A97, A98, L99, I100, M102, V103, M106, V111, L118, M120, L121, A129, L133, A146, V149, I150, T152 and F153. Positions 3 and 91 were classified as boundary residues but were nevertheless included in the core calculations as visual inspection showed them to be significantly buried. Positions 3 and 6 are close to the N terminus but were considered for design because they contribute to the core of the C-terminal domain (Figure 1(a)). The hydrophobic amino acids allowed at all 26 positions were Ala, Val, Leu, Ile, Phe, Tyr, Trp and Met. Proline, glycine and cysteine were omitted from consideration to avoid possible disruption of secondary structure and the formation of disulfide bonds. An expanded version of the backbone-dependent rotamer library of Dunbrack and Karplus was used for the calculations.[34] For aromatic residues, the expansions included the mean χ values ± 1 standard deviation about χ_1 and χ_2 torsional angles. For other hydrophobic groups, a similar expansion was performed, but was limited only to the χ_1 torsional angle. Energies for the point mutants were calculated by fixing the identities of amino acids at all 26 positions while allowing their rotameric conformations to vary based on the rotamer library. The design calculations were run using an optimization procedure based on the Dead-End Elimination algorithm.[11,35]

The energy terms included in the calculations were van der Waals interactions, hydrogen bond, electrostatic interactions and solvation. The reported energies include the interaction energies of the 26 positions considered in the calculations with each other and with the remaining portion of the protein not directly considered in the sequence optimization. The van der Waals radii of all atoms were scaled by 0.9.[14] Hydrogen bonds were represented by a distance, angle, and hybridization-dependent, 12–10 potential, and electrostatic interactions were treated using Coulomb's law with a distance-dependent dielectric constant.[16] Hydrophobic solvation energies were calculated by a surface area burial method.[36]

Mutagenesis, protein expression, and purification

The two redesigns of the C-terminal core of bacteriophage T4 lysozyme, Core-7 and Core-10, were made by iterative two-stage PCR[37] using the gene for the cysteine-free (C54T/C97A) pseudo-wild-type (WT*) T4 lysozyme as the template.[23] The *Bam*HI/*Hin*dIII-digested PCR products were ligated into the vector PH1403. The single (where they did not previously exist), double, and revertant mutants were made by the inverse PCR.[38] The gene for WT*, Core-10, or Core-7 in the vector PH1403 was used as the template. The individual single-site mutants (relative to WT*) were drawn from existing stocks except for I78V, V87I, I100V, V103I, M106I, L118I, M120Y, and L133F. The double mutant V149/T152V was made in the WT* background. The DNA sequences of the new constructs were confirmed by automated methods incorporating the polymerase chain reaction (Perkin–Elmer ABI PRISM 377 DNA sequencer). The vectors were transformed into *Escherichia coli* RR1 cells for over-expression. The mutant proteins were over-expressed and purified by standard methods.[39–41] The molecular mass of the mutant proteins were checked with a Perspective Biosystems Voyager-DE MALDI/TOF mass spectrometer. The buffer used for protein storage was 0.1 M sodium phosphate (pH 6.5), 0.55 M NaCl, 0.02% (w/v) NaN_3. As judged by the fact that each lysozyme caused cell lysis and behaved similarly during purification, we assume that all have activity similar to that of WT*.

Thermal unfolding

Circular dichroism-monitored thermal stability data were collected at 223 nm using a JASCO model J-600 spectropolarimeter and the Hewlett–Packard model HP89100 thermal control system.[42] The buffer was 0.10 M sodium chloride, 1.4 mM acetic acid, 8.6 mM sodium acetate (pH 5.35), with protein concentrations of 0.01–0.03 mg/ml as determined from absorbance at 280 nm.[27] Unfolding profiles were analyzed by means of the two-state model to determine the temperature of melting (t_m) and the van't Hoff enthalpy at the melting temperature (ΔH).[43] At least three independent trials were done for each mutant. Averaged values of t_m and ΔH were used to calculate $\Delta G°$ at 61 °C by means of an integrated form of the Gibbs–Helmholtz equation[44] assuming a ΔC_p of 2.5 kcal mol^{-1} K^{-1}. $\Delta\Delta G$ values were computed as $\Delta G°$(mutant) − $\Delta G°$(WT*).

Crystallization

It was possible to crystallize the two designed proteins, selected single mutant back-revertant proteins, and the previously unpublished single mutants. In all, 13 of the 16 new proteins were crystallized in space group $P3_221$ isomorphously with the wild-type protein in 2 M K/Na phosphate buffers as described.[42] Core-7 crystallized in space group $F222$ in 100 mM Na/K phosphate buffer (pH 6.7) and 20% (v/v) MPD. V103I crystallized in space group $P2_12_12_1$ in solutions of 0.1 M Hepes (pH 7.5), 20% (w/v) PEG3400, 5% (v/v) isopropanol. M87V/Core-7 crystallized in space group $C2$ in 25% PEG3400, 5% PEG600, 200 mM NaCl, 100 mM Na/K phosphate (pH 6.7) and Fos-choline 12 at its critical micelle concentration.

X-ray data collection

Since the 100 K structure of the pseudo wild-type had been used as the template in the design process, X-ray data of the new proteins were collected at 100 K. Crystals of proteins grown from the high-salt solutions were mounted in paratone and flash-cooled. Crystals of Core-7 and of V103I were flash-cooled in rayon loops containing cryogenic reservoir solutions. X-ray data for Core-7 and I103V/Core-10 were collected at beamline 7-1 at SSRL with monochromatic radiation having a wavelength of 1.06 Å and a MAR image plate. X-ray data for the remaining structures were collected in-house with 1.54 Å radiation and a Rigaku RAXIS4 image plate. The data were integrated with Mosflm and scaled with Scala.[45,46]

Structure determination

The structures of V103I and M87V/Core-7 were solved by molecular replacement using the program EPMR[47] while the remaining structures were determined by molecular substitution using the coordinates of WT* (Table 2) as the starting model.

Structure refinement

The crystal structures were refined using the refinement package TNT[48,49] following the procedures described previously.[42] The Xfit molecular graphics module of XtalView was used for model rebuilding.[50] The PDB codes are given in Table 2.

Acknowledgements

We thank Hong Xiao, Leslie Gay, and Andy Fields for making the mutant proteins and for crystallizing them, Cathy Sarisky for help with the calculations, Doug Juers for collecting preliminary X-ray data for the mutant proteins V103I and Core-7, and the user support staff at SSRL and ALS for their assistance. This work was supported in part by grants from the NIH (GM21967 to B.W.M.), the Howard Hughes Medical Institute (to S.L.M. & B.W.M.), the Ralph M. Parsons Foundation (to S.L.M.) and an IBM Shared University Research grant (to S.L.M.).

References

1. Richards, F. M. (1986). Protein design: are we ready? *Protein Struct. Fold. Des.*, 171–196.
2. Richards, F. M. & Lim, W. A. (1994). An analysis of packing in the protein folding problem. *Quart. Rev. Biophys.* **26**, 423–498.
3. Baldwin, E. P., Hajiseyedjavadi, O., Baase, W. A. & Matthews, B. W. (1993). The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme. *Science*, **262**, 1715–1718.
4. Baldwin, E. P., Xu, J., Hajiseyedjavadi, O., Baase, W. A. & Matthews, B. W. (1996). Thermodynamic and structural compensation in "size-switch" core

repacking variants of bacteriophage T4 lysozyme. *J. Mol. Biol.* **259**, 542–559.
5. Gassner, N. C., Baase, W. A. & Matthews, B. W. (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.
6. Liang, J. & Dill, K. A. (2001). Are proteins well-packed? *Biophys. J.* **81**, 751–766.
7. Karpusas, M., Baase, W. A., Matsumura, M. & Matthews, B. W. (1989). Hydrophobic packing in T4 lysozyme probed by cavity-filling mutants. *Proc. Natl Acad. Sci. USA*, **86**, 8237–8241.
8. Janin, J., Wodak, S., Levitt, M. & Maigret, B. (1978). Conformation of amino acid side-chains in proteins. *J. Mol. Biol.* **125**, 357–386.
9. Ponder, J. W. & Richards, F. M. (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.
10. Hurley, J. H., Baase, W. A. & Matthews, B. W. (1992). Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme. *J. Mol. Biol.* **224**, 1143–1159.
11. 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.
12. Goldstein, R. F. (1994). Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys. J.* **66**, 1335–1340.
13. Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci.* **5**, 895–903.
14. Dahiyat, B. I. & Mayo, S. L. (1997). Probing the role of packing specificity in protein design. *Proc. Natl Acad. Sci. USA*, **94**, 10172–10177.
15. Dahiyat, B. I. & Mayo, S. L. (1997). De novo protein design: fully automated sequence selection. *Science*, **278**, 82–87.
16. Dahiyat, B. I., Gordon, B. D. & Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Sci.* **6**, 1333–1337.
17. Su, A. & Mayo, S. L. (1997). Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci.* **6**, 1701–1707.
18. Malakauskas, S. M. & Mayo, S. L. (1998). Design, structure and stability of a hyperthermophilic protein variant. *Nature Struct. Biol.* **5**, 470–475.
19. Myers, J. K., Pace, C. N. & Scholtz, J. M. (1995). Denaturant *m* values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci.* **4**, 2138–2148.
20. Janin, J. (1979). Surface and inside volumes in globular proteins. *Nature*, **277**, 491–492.
21. Miller, S., Janin, J., Lesk, A. M. & Chothia, C. (1987). Interior and surface of monomeric proteins. *J. Mol. Biol.* **196**, 641–656.
22. Shimaoka, M., Shifman, J. M., Jing, H., Takagi, J., Mayo, S. L. & Springer, T. A. (2000). Computational design of an integrin I domain stabilized in the open high affinity conformation. *Nature Struct. Biol.* **7**, 674–678.
23. Matsumura, M. & Matthews, B. W. (1989). Control of enzyme activity by an engineered disulfide bond. *Science*, **243**, 792–794.
24. Eriksson, A. E., Baase, W. A., Zhang, X.-J., Heinz, D. W., Blaber, M., Baldwin, E. P. & Matthews, B. W. (1991). A cavity-containing mutant of T4 lysozyme is stabilized by buried benzene. *Science*, **255**, 178–183.
25. Lipscomb, L. A., Gassner, N. C., Snow, S. D., Eldridge, A. M., Basse, W. A., Drew, D. I. & Matthews, B. W. (1998). Context-dependent protein stabilization by methionine-to-leucine substitution shown in T4 lysozyme. *Protein Sci.* **7**, 765–773.
26. Anderson, D. E., Becktel, W. J. & Dahlquist, F. W. (1990). pH-induced denaturation of proteins: a single salt bridge contributes 3–5 kcal/mol to the free energy of folding of T4 lysozyme. *Biochemistry*, **29**, 2403–2408.
27. Elwell, M. & Schellman, J. A. (1975). Phage T4 lysozyme. Physical properties and reversible unfolding. *Biochim. Biophys. Acta*, **386**, 309–323.
28. Grütter, M. G., Gray, T. M., Weaver, L. H., Alber, T., Wilson, K. & Matthews, B. W. (1987). Structural studies of mutants of the lysozyme of bacteriophage T4. The temperature sensitive mutant protein Thr157 → Ile. *J. Mol. Biol.* **197**, 315–329.
29. Blaber, M., Zhang, X.-J., Lindstrom, J. D., Pepiot, S. D., Baase, W. A. & Matthews, B. W. (1994). Determination of α-helix propensity within the context of a folded protein: sites 44 and 131 in bacteriophage T4 lysozyme. *J. Mol. Biol.* **235**, 600–624.
30. Weaver, L. H. & Matthews, B. W. (1987). Structure of bacteriophage T4 lysozyme refined at 1.7 Å resolution. *J. Mol. Biol.* **193**, 189–199.
31. Xu, J., Baase, W. A., Quillin, M. L., Baldwin, E. P. & Matthews, B. W. (2001). Structural and thermodynamic analysis of the binding of solvent at internal sites in T4 lysozyme. *Protein Sci.* **10**, 1067–1078.
32. Gassner, N. C., Baase, W. A., Mooers, B. H. M., Busam, R. D., Weaver, L. H., Lindstrom, J. D. *et al.* (2003). Multiple methionine substitutions are tolerated in T4 lysozyme and have coupled effects on folding and stability. *Biophys. Chem.* **100**, 325–340.
33. Connolly, M. L. (1983). Analytical molecular surface calculation. *J. Appl. Crystallog.* **16**, 548–558.
34. Dunbrack, R. L., Jr & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J. Mol. Biol.* **230**, 543–574.
35. Pierce, N. A., Spriet, J. A., Desmet, J. & Mayo, S. L. (2000). Conformational splitting: a more powerful criterion for dead-end elimination. *J. Comput. Chem.* **21**, 999–1009.
36. Street, A. G. & Mayo, S. L. (1998). Pairwise calculation of protein solvent-accessible surface areas. *Fold Des.* **3**, 253–258.
37. Landt, O., Grunert, H. & Hahn, U. (1990). A general method for rapid site-directed mutagenesis using the polymerase chain reaction. *Gene*, **96**, 125–128.
38. Hemsley, A., Arnheim, N., Toney, M. D., Cortopassi, G. & Galas, D. J. (1989). A simple method for site-directed mutagenesis using the polymerase chain reaction. *Nucl. Acids Res.* **17**, 6545–6551.
39. Alber, T. & Matthews, B. W. (1987). Temperature-sensitive mutation of bacteriophage T4 lysozyme occur at sites with low mobility and low solvent accessibility in the folded protein. *Biochemistry*, **26**, 3754–3758.
40. Muchmore, D. C., McIntosh, L. P., Russell, C. B., Anderson, D. E. & Dahlquist, F. W. (1989). Expression and 15N labelling of proteins for proton and nitrogen-15 NMR. *Methods Enzymol.* **177**, 44–73.
41. Poteete, A. R., Dao-pin, S., Nicholson, H. & Matthews, B. W. (1991). Second-site revertants of an inactive T4 lysozyme mutant restore activity

structuring the active site cleft. *Biochemistry*, **30**, 1425–1432.
42. Eriksson, A. E., Baase, W. A. & Matthews, B. W. (1993). Similar hydrophobic replacements of Leu99 and Phe153 within the core of T4 lysozyme have different structural and thermodynamic consequences. *J. Mol. Biol.* **229**, 747–769.
43. Zhang, X.-J., Baase, W. A., Shoichet, B. K., Wilson, K. P. & Matthews, B. W. (1995). Enhancement of protein stability by the combination of point mutations in T4 lysozyme is additive. *Protein Eng.* **8**, 1017–1022.
44. Hawkes, R., Grutter, M. G. & Schellman, J. (1984). Thermodynamic stability and point mutations of bacteriophage T4 lysozyme. *J. Mol. Biol.* **175**, 195–212.
45. Leslie, A. G. W. (1992). Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4 and ESF-EAMCB Newsletter on Protein Crystallography No. 26*
46. Evans, P. R. (1994). Scala. *Joint CCP4 and ESF-EACBM Newsletter*, **33**, 22–24.
47. Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). Rapid automated molecular replacement by evolutionary search. *Acta Crystallog. sect. D*, **55**, 484–491.
48. Tronrud, D. E., Ten Eyck, L. F. & Matthews, B. W. (1987). An efficient general-purpose least-squares refinement program for macromolecular structures. *Acta Crystallog. sect. A*, **43**, 489–501.
49. Tronrud, D. E. (1997). TNT refinement package. *Methods Enzymol.* **277**, 306–319.
50. MacRee, D. E. (1992). A visual protein crystallographic software system for X11/Xview. *J. Mol. Graph.* **10**, 44–46.
51. Gassner, N. C., Baase, W. A., Lindstrom, J. D., Lu, J., Dalqusit, F. W. & Matthews, B. W. (1999). Methionine and alanine substitutions show that the formation of wild-type-like structure in the carboxy-terminal domain of T4 lysozyme is a rate-limiting step in folding. *Biochemistry*, **38**, 14451–14460.
52. Zhang, X.-J. & Matthews, B. W. (1995). EDPDB: a multi-functional tool for protein structure analysis. *J. Appl. Crystallog.* **28**, 624–630.

Edited by F. E. Cohen

(Received 25 March 2003; received in revised form 1 July 2003; accepted 1 July 2003)