Computational Protein Design Force Field Optimization: A Negative Design Approach

Thesis by

Oscar Alvizo

In Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

California Institute of Technology
Pasadena, California

2007

(Defended May 11, 2007)

Acknowledgements

I would like to give the biggest thanks to the two women in my life whose unconditional love and support made this Ph.D. possible, my mother and my wife. My mother installed a love for education from a very early age and nurtured my sense of exploration. My wife has aided my career in countless ways and has always been there to serve as my sounding board.

My path was aided by a handful of wonderful organizations whose objective is to assist individuals with disadvantaged backgrounds reach their potential. The Upward Bound Project at USC opened my eyes to higher education and showed me that a college education was an obtainable goal. The Academic Excellence Honors Program (ACE) at UC Santa Cruz was quintessential in promoting my educational accomplishments in the sciences. I thank Nancy Cox-Konopelski and her team leaders for instilling in me a love for teaching and the sciences. The MARC program at UCSC helped direct my scientific career and opened doors that would ultimately lead to my admission to Caltech.

I would like to thank all the individuals who have helped me become a better scientist. Glenn L. Millhauser was a wonderful undergraduate research advisor and overall a great individual. I could always count on Colin Burns for solid advice and training. My graduate work was only possible with the guidance of Steve Mayo, my

graduate advisor. His agreeable nature and belief in independence gave me the opportunity to come into my own as a scientist. I thank all the following members of the Mayo lab who have all helped me in one way or another: Eric Zollars, Eun Jung Choi, Tom Treynor, Possu Huang, Premal Shah, Geoffrey Hom, Jennifer Keeffe, Kyle Lassila, Christina Vizcarra, Ben Allen, Karin Crowhurst, Alex Perryman, Corey Wilson, Peter Oelschlaeger, Julia Shifman, and JJ Plecs. Special thanks to my bay mates, Jessica Mao and Heidi Privett. The Mayo staff has always been there to make a scientist's life easier and for that I am grateful. I would like to acknowledge Marie Ary for her patience and honest feedback when editing my written work. I thank my thesis committee who could always be counted on for insightful advice.

I thank all my family for their constant encouragement, and give credit to Danielito Alvizo for his unique ability to remove all stress at the sight of him. Finally, I give credit to God for bringing all the mentioned individuals into my life.

Abstract

An accurate force field is essential to computational protein design and protein folding studies. Proper force field tuning is problematic, however, due in part to the incomplete modeling of the unfolded state. The first part of this thesis discusses the optimization of a protein design force field by constraining the amino acid composition of the designed sequences to that of the wild-type protein. According to the random energy model, the unfolded state energies of amino acid sequences with the same composition are identical. Under these constraints, unfolded state energies are inconsequential and any discrepancies between computational predictions and experimental results can be directly attributed to flaws in the force field's ability to properly account for folded state sequence energies. This aspect of fixed composition design allows for force field optimization by focusing solely on the interactions in the folded state. In addition, the fixed composition requirement imposes a large negative design constraint that is used to ensure fold specificity. Several rounds of fixed composition optimization of the β1 domain of protein G yielded force field parameters with significantly greater predictive power: optimized sequences exhibited higher wild-type sequence identity in critical regions of the structure and the wild-type sequence

showed an improved Z-score. Experimental studies revealed a 24-fold mutant to be stably folded with a melting temperature comparable to that of the wild-type protein.

The second part of the thesis discusses the optimization of HIV protease substrate specificity using a combination of positive and negative design. HIV protease is a homodimeric protein with a symmetrical binding region that recognizes and cleaves asymmetrical substrates that exhibit little sequence homology. The designs attempt to increase specificity towards one of HIV protease's wild-type targets by optimizing hydrogen bonds and electrostatic interactions using a positive design approach. Explicit negative design is incorporated by modeling predicted mutations on multiple substrates. A scoring function that selects for mutations that pack favorably with the target substrate but result in large steric clashes in alternate substrates is used. A three point mutant was designed and experimentally shown to have increased specificity towards the target substrate.

Contents

List of Tables

List of Figures

Chapter 1

Introduction to Computational Protein Design

1.1 Protein Design

The fact that many naturally-occurring proteins fold reliably and quickly to their native state despite the astronomical number of possible configurations has come to be known as Levinthal's paradox or the protein folding problem [1]. A protein folding algorithm must exhaustively search conformational space to obtain the one three-dimensional structure with the lowest energy of the sequence in question (Fig. 1-1). In order to accomplish this task, the calculation requires an accurate force field that precisely describes all inter-atomic physical interactions and a search algorithm that eliminates all structures except the native conformation [2].

The same two requirements that apply to protein folding also apply to computational protein design (CPD). Protein design deals with the inverse protein folding problem; instead of predicting a structure from a sequence, the goal is to predict sequences that will fold into a target tertiary structure [3]. However, designing a protein

poses less of a challenge than protein folding, because there are multiple sequences that can fold into the intended structure (Fig. 1-1). As a result, protein design can be carried out successfully with a less stringent force field [4]. The design of difficult structures, however, such as those with a limited pool of compatible sequences, may only become feasible with a highly tuned force field.

1.2 Computational Force Fields

A handful of force fields are commonly used for protein folding and/or protein design: AMBER, CHARMm, GROMOS, OPLS-AA, and DREIDING [5-9]. These force fields are composed of pair-wise decomposable potential functions that are summed to obtain the overall energy of the system (Eq. 1-1). Certain potential functions, such as the one used to calculate van der Waals energies, break down easily into pair-wise interactions; while others require the use of approximations to make them compatible with CPD.

The potential energy functions used in CPD are meant to score inter-residue contacts that have been shown to stabilize the target structure. All CPD force fields optimize for van der Waals (*vdw*), electrostatics (*elec*), and atomic solvation (*as*).

$$E_{nonbonded} = E_{vdw} + E_{elec} + E_{as} \ldots. \qquad \text{(Eq. 1-1)}$$

Force fields differ in their inclusion of other terms such as hydrogen bonding, secondary structure propensities, side-chain conformational probabilities, etc. [10]. Some force fields, (e.g., DREIDING) calculate hydrogen bonds explicitly [9], while others (e.g., AMBER) omit hydrogen bond energies under the assumption that the electrostatics term is inclusive enough to compensate for it. All force fields scale each potential function to account for their relative contributions to the total energy. The main differences between

the force fields are in the values of the scale factors and parameters applied to each of the potential functions. As of yet, no consensus on the optimal set of parameters and potential functions has been reached.

1.3 ORBIT's Force Field

The work presented here was carried out using the Optimization of Rotamers by Iterative Techniques (ORBIT) computational protein design software. Potential functions and scaling factors relevant to what is discussed in upcoming chapters are touched on in the following sections of this chapter. The ORBIT software suite is based on the DREIDING force field and utilizes many of its potential functions to score the fitness of predicted sequences. Details on many of the potential functions have been previously published [11-15]. ORBIT works best on high-resolution crystal structure scaffolds that have been minimized to reduce strain and atomic overlap. The backbone is held fixed while side-chain conformations are optimized to yield the best scoring sequence.

1.3.1 van der Waals Forces

The van der Waals forces are the weak, non-covalent non-ionic forces between atoms that favor close atomic interactions. DREIDING scores these packing interactions with an atomic Lennard-Jones 12-6 potential, which includes a long-range attractive component, and a short-range repulsive component, as described in Eq. 1-2.

$$E_{vdw} = D_0\left[\left(\frac{\alpha R_0}{R}\right)^{12} - 2\left(\frac{\alpha R_0}{R}\right)^{6}\right] \qquad \text{(Eq. 1-2)}$$

The van der Waals energy (E_{vdw}) is a function of the distance between the two atoms (R). R_0 and D_0 are the geometric means of the van der Waals radii and the well depth of the

two atoms at equilibrium, respectively (Fig. 1-2). The van der Waals potential function describes the rapid increase in unfavorable energy as R decreases relative to R_o. As R gets smaller, the electronic clouds of each of the atoms start to overlap and repel each other. As a result, sequences that can pack well in the folded structure without resulting in atomic clashes are strongly favored by the van der Waals potential. The scale factor, α, serves to attenuate overlap between the two atoms. A value of 0.9 for α has been shown to work well when carrying out CPD [12].

1.3.2 Hydrogen Bonds

Like most protein design force fields, ORBIT explicitly accounts for hydrogen bond energies ($E_{H\text{-}bond}$). The hydrogen bond potential is a function of the donor to acceptor distance and includes more restrictive angle-dependent terms to limit unfavorable hydrogen bond geometries (Eq. 1-3). Hydrogen bonds are only calculated for polar heavy atoms that are within 2.6–3.2 Å.

$$E_{H-bond} = D_0\left[5\left(\frac{R_0}{R}\right)^{12} - 6\left(\frac{R_0}{R}\right)^{10}\right]F(\theta,\phi,\varphi) \qquad \text{(Eq. 1-3)}$$

R is the distance between the two heavy atoms involved in the hydrogen bond, and R_0 is set to 2.8 Å, the ideal distance at equilibrium. D_0 is the hydrogen bond well depth, and is historically set to 8.0 kcal/mol, the maximum possible benefit for hydrogen bond formation. The hydrogen bond potential's angle dependence function, F, is elaborated on in Fig. 1-3 *B*.

1.3.3 Electrostatics

The energies due to electrostatic interactions, E_{ele}, are scored using Coulomb's law (Eq. 1-4).

$$E_{ele} = \frac{q_1 q_2}{\varepsilon r} \qquad \text{(Eq. 1-4)}$$

The values q_1 and q_2 pertain to the assigned atomic charges. The dielectric constant, ε, is one of two variables that can be changed to scale the impact of electrostatics on the design. A dielectric constant of 40, similar to what is used for water, is commonly used in ORBIT. The second variable is the distance, r; electrostatic interactions can be set to be inversely proportional to either r or r^2. Squaring the distance emphasizes local interactions over long-range interactions. Every atom in the force field can be assigned a charge and is able to contribute to the electrostatic potential. Naturally, the largest contributors are atoms associated with acidic or basic side chains. In addition to the electrostatic component, salt bridge interactions are reinforced by the hydrogen bond potential.

1.3.4 Atomic Solvation

An atomic solvation potential function is required to recover sequences with a hydrophobic core and a polar surface. It is widely believed that the hydrophobic effect is the main driving force for protein folding [16]. The need to exclude solvent molecules from the core and maximize hydrophobic packing is thought to result in a structural collapse that initiates the folding process [17].

There are two commonly used solvation models in ORBIT, a surface area-based model and a solvent exclusion-based model. The surface area-based solvation model is

conceptually straightforward to understand, but computationally expensive to run. The idea is to quantify atomic burial by calculating how much surface area is buried in the folded state relative to how much is buried in the unfolded state. To obtain the surface area buried in the folded state, a dot surface is calculated for each pair of amino acids using the Connolly algorithm. The algorithm roles a virtual sphere with a set radius over the amino acids; the surface is defined as the area accessed by the sphere. Non-polar surface area is favored by 0.026 kcal/mol/Å^2 when buried and penalized by the same amount when exposed. Polar surface area is penalized by 0.1 kcal/mol/Å ^{2} when buried [18].

The solvent exclusion-based solvation model used in ORBIT is fast since most of the necessary information, such as the volume of surrounding atoms (V_j) and the reference solvation free energy (ΔG_i^{ref}), is predetermined [19]. During the scoring process, the only variable is a function dependent of the atomic distance ($f_i(r_{ij})$) (Eq. 1-5). Solvent exclusion is assumed to be proportional to atom burial. As a result, the distance and volume of surrounding atoms is used to determine the extent of solvent exclusion (Fig. 1-4). Solvent exclusion is favored for non-polar atoms and disfavored for polar atoms.

$$E_{as} = \sum_i \left[\Delta G_i^{ref} - \sum_{j \neq i} f_i(r_{ij}) V_j \right] \qquad \text{(Eq. 1-5)}$$

1.4 Side-Chain Conformational Libraries

In addition to an accurate force field, a complete design procedure requires a search algorithm that will find the global minimum energy conformation (GMEC). This in itself can be a considerable obstacle due to the vast size of conformational space.

Historically, CPD has been carried out on a fixed backbone to simplify the problem. However, efforts to incorporate backbone flexibility are well on their way [20]. The first design of a novel fold was achieved by combining protein design with protein folding technology, which allows backbone flexibility [21].

Amino acid side-chain flexibility is taken into account by using a side-chain conformational library. Instead of dealing with a continuum of side-chain conformations, the library is composed of statistically significant low energy conformations that are obtained from high resolution structures [22]. Two types of amino acid side-chain libraries are available: rotamer libraries, which are composed of side-chain conformations that have been minimized and have idealized bond lengths and angles; and conformer libraries, which are composed of side-chain conformations that take their bond lengths and angles directly from protein crystal structures [23-25].

1.5 Search Algorithms

A search algorithm is used to sort through all the potential conformational sequences and identify those that take on the desired tertiary structure. Designing a small 100-residue protein, for example, would require sorting through 10^{130} unique amino acid sequences if all 20 amino acids are considered at each position. The search problem becomes increasingly complex as more positions and multiple conformations per amino acid are considered.

Two types of search methods are typically used in CPD: stochastic methods and deterministic methods [26]. Monte Carlo simulations are commonly used in combination with simulated annealing for stochastic searches, since a Monte Carlo algorithm can easily be incorporated into the design procedure and the run time is dictated by the user

[27]. However, stochastic methods cannot guarantee that the resulting sequence is the GMEC. On the other hand, a deterministic method such as Dead-End Elimination (DEE) always provides the GMEC if it converges [28]. However, convergence is not assured.

In recent years, the FASTER algorithm has proven to be an effective and useful alternative [29, 30]. FASTER is orders of magnitude more efficient than DEE and can often recover the GMEC. The procedure relies on the idea that optimization of the individual components in the design will result in optimization of the total design. The result is an increase in speed obtained by carrying out multiple rotamer perturbations in each step before evaluating the success of the overall change.

1.6 Energies Independent of the Unfolded State

Since there is no accurate method to calculate the energy of the unfolded state, most force fields, including DREIDING, fail to fully characterize the unfolded state. As a result, the energies predicted are taken to be absolute energies that are, for the most part, independent of the unfolded state energies. It is assumed that the difference in unfolded state energies between distinct sequences is insignificant; hence, unfolded state energies can be omitted without adverse effects on the design [31]. This assumption can pose serious problems if incorrect, since all design sequences are analyzed and compared by determining their thermodynamic properties, and thermodynamic constants such as ΔG (free energy of folding) are relative to the unfolded state energy as shown in Eq. 1-6.

$$\Delta G^{A} = G^{A}_{folded} - G^{A}_{unfolded} \qquad \text{(Eq. 1-6)}$$

The free energy of sequence A in a specific folded conformation is defined as the difference between the energy of sequence A in the folded state and the energy of sequence A in the unfolded state. Hence, sequence energies obtained computationally

can only be accurately compared with those obtained experimentally if the energy of the unfolded state is included in the force field or if it is shown to be insignificant. A situation can be imagined in which a mutation is predicted to be stabilizing in a sequence due to the formation of a favorable inter-residue interaction in the folded state, but is determined to be destabilizing experimentally due to its stabilizing effects on the unfolded state.

Also, the free energy of each of the sequences must be accurately determined in order to rank the compatibility of a group of sequences to a specific scaffold.

$$\Delta\Delta G = \Delta G^B - \Delta G^A = \left(G^B_{folded} - G^A_{folded}\right) - \left(G^B_{unfolded} - G^A_{unfolded}\right) \quad \text{(Eq. 1-7)}$$

$$G^B_{unfolded} \approx G^A_{unfolded} \quad \text{(Eq. 1-8)}$$

$$\Delta\Delta G = \Delta G^B - \Delta G^A = \left(G^B_{folded} - G^A_{folded}\right) \quad \text{(Eq. 1-9)}$$

As Eq. 1-7 states, calculating the difference in ΔG between two sequences (*A* and *B*) requires each of their unfolded state energies as well as their folded state energies. Since force fields do not accurately predict the energy of the unfolded state, the predicted energies will be intrinsically incorrect. Yet, if the energies of the unfolded state can be ignored by keeping them constant (Eq. 1-8), the change in ΔG will only be dependent on the energies of sequences *A* and *B* in the folded conformation (Eq. 1-9). One method that theoretically keeps the unfolded state energies constant is to restrict designs to sequences with a fixed amino acid composition.

1.7 Fixed Amino Acid Composition Protein Design

Fixed composition calculations allow us to disregard energy contributions from the unfolded state, since sequences with fixed composition are assumed to have

equivalent unfolded state energies [32, 33]. This assumption stems from the random energy model (REM) of polymers. REM was first solved for spin glasses by Derrida in 1980 [34]. In the late 1980s, Shakhnovich and Gutin [35] and Byngelson and Wolynes [36] derived models based on REM for proteins. They concluded that the density of energy of all thermodynamically accessible conformations takes on a Gaussian distribution as defined by Eq. 1-9 below.

$$n(E) = \frac{\gamma^N}{\sigma_u^2 \sqrt{2\pi Nz}} \exp\left[-\frac{\left(E - Nz\langle U \rangle\right)^2}{2Nz\sigma_u^2} \right] \qquad \text{(Eq. 1-9)}$$

N = number of residues in the protein

z = average number of contacts per residue

U = average energy per contact

γ^N = number of conformations accessible to the sequence

σ_u = standard deviation (SD) of the contact energies

n = number of conformations

Here, *n(E)* gives the density of energy levels that results from all the conformations accessible to a particular sequence. As a direct result of the Gaussian energy distribution, the energy spectrum for a given sequence is divided into a continuous part and a discrete part (Fig. 1-5). At *n(E)* = 1, the continuous energy levels separate into individual lower energy conformations, forming the discrete part of the spectrum [33]. The energy at which this occurs is called the critical energy (E_c) (Fig. 1-5). The continuous part of the spectrum represents all the conformations accessible to the sequence at higher temperature (unfolded state). At higher energies, the interactions made by the different

conformations average out due to the rapid introversions between the accessible structures, making the energy of the continuous part of the spectrum composition dependent [33]. On the other hand, the discrete part of the spectrum relies on best fit contacts to stabilize the individual conformations, making it sequence dependent. Therefore, by keeping the composition of the predicted sequences constant, we can assume that the energy of the unfolded state stays constant.

Forcing a fixed amino acid composition has the added benefit of imposing specificity to the target structure [37, 38]. The drastic reduction in sequence space will dramatically reduce the number of alternative folds. Fixing the composition essentially serves as a strong negative design constraint on the calculations.

1.8 Negative Design

For a successful computational protein design result, the procedure must consider negative design in addition to positive design. Positive design predicts sequences that will stabilize the target fold. However, positive design does little to rule out sequences that exhibit a lower energy in alternate states or folds. To address this concern, negative design is used to design against alternate conformations [39]. In addition to fixed composition, there are multiple examples in which explicit and implicit negative design have been used in protein engineering.

Explicit negative design is commonly used in the engineering of α-helical structures such as coiled coil domains and α-helical bundles [39-42]. The repetitive nature of α-helices and their well studied inter-helical interactions make them ideal for rational design. Strategically placed charged and polar residues at the protein interfaces have been shown to provide fold specificity by stabilizing hetero-oligomerization or by

destabilizing homo-oligomerization [40, 42]. Negative design of coiled coil domains was taken a step further by using computer automation to design against aggregated and unfolded states [41].

Implicit negative design is used to design against alternate states without explicitly considering any. The design of the first novel fold, Top7, used amino acid reference energies to ensure natural-like amino acid compositions throughout the protein [21]. As a result, the predicted sequences were guaranteed to have hydrophobic cores and polar surfaces. Another approach that has been used is to impose a pseudo-binary pattern by restricting the core to hydrophobic residues and the surface to polar residues [43-45]. A properly folded protein with a polar surface is less likely to aggregate due to exposed hydrophobic surface area.

1.9 Bibliography

1. Levinthal, C. Are there pathways for protein folding? *J Chim Phys* **65**, 44-45 (1968).

2. Baker, D. A surprising simplicity to protein folding. *Nature* **405**, 39-42 (2000).

3. Pokala, N., and Handel, T. M. Review: Protein design--where we were, where we are, where we're going. *J Struct Biol* **134**, 269-281 (2001).

4. Kraemer-Pecore, C. M., Wollacott, A. M., and Desjarlais, J. R. Computational protein design. *Curr Opin Chem Biol* **5**, 690-695 (2001).

5. Jorgensen, W. L., Maxwell, D. S., and TiradoRives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc* **118**, 11225-11236 (1996).

6. MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., and Evanseck, J. D. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem* **102**, 3586-3616 (1998).

7. Ponder, J. W., and Case, D. A. Force fields for protein simulations. *Adv Protein Chem* **66**, 27-85 (2003).

8. Stocker, U., and van Gunsteren, W. F. Molecular dynamics simulation of hen egg white lysozyme: A test of the GROMOS96 force field against nuclear magnetic resonance data. *Proteins* **40**, 145-153 (2000).

9. Mayo, S. L., Olafson, B. D., and Goddard, W. A. DREIDING: A generic force field for molecular simulations. *J Phys Chem* **94**, 8897-8909 (1990).

10. Wang, W., Donini, O., Reyes, C. M., and Kollman, P. A. Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Annu Rev Biophys Biomol Struct* **30**, 211-243 (2001).

11. Dahiyat, B. I., Sarisky, C. A., and Mayo, S. L. De novo protein design: Towards fully automated sequence selection. *J Mol Biol* **273**, 789-796 (1997).

12. Dahiyat, B. I., and Mayo, S. L. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci USA* **94**, 10172-10177 (1997).

13. Dahiyat, B. I., and Mayo, S. L. De novo protein design: Fully automated sequence selection. *Science* **278**, 82-87 (1997).

14. Dahiyat, B. I., and Mayo, S. L. Protein design automation. *Protein Sci* **5**, 895-903 (1996).

15. Dahiyat, B. I., Gordon, D. B., and Mayo, S. L. Automated design of the surface positions of protein helices. *Protein Sci* **6**, 1333-1337 (1997).

16. Chandler, D. Interfaces and the driving force of hydrophobic assembly. *Nature* **437**, 640-647 (2005).

17. Agashe, V. R., Shastry, M. C. R., and Udgaonkar, J. B. Initial hydrophobic collapse in the folding of barstar. *Nature* **377**, 754-757 (1995).

18. Street, A. G., and Mayo, S. L. Pairwise calculation of protein solvent-accessible surface areas. *Fold Des* **3**, 253-258 (1998).

19. Lazaridis, T., and Karplus, M. Effective energy function for proteins in solution. *Proteins* **35**, 133-152 (1999).

20. Desjarlais, J. R., and Handel, T. M. Side-chain and backbone flexibility in protein core design. *J Mol Biol* **290**, 305-318 (1999).

21. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. Design of a novel globular protein fold with atomic-level accuracy. *Science* **302**, 1364-1368 (2003).

22. Ponder, J. W., and Richards, F. M. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol* **193**, 775-791 (1987).

23. Lassila, J. K., Privett, H. K., Allen, B. D., and Mayo, S. L. Combinatorial methods for small-molecule placement in computational enzyme design. *Proc Natl Acad Sci USA* **103**, 16710-16715 (2006).

24. Xiang, Z., and Honig, B. Extending the accuracy limits of prediction for side-chain conformations. *J Mol Biol* **311**, 421-430 (2001).

25. Dunbrack, R. L., Jr., and Karplus, M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol* **230**, 543-574 (1993).

26. Voigt, C. A., Gordon, D. B., and Mayo, S. L. Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. *J Mol Biol* **299**, 789-803 (2000).

27. Metropolis, N., Rosenbluth, A. W., Teller, A. H., and Teller, E. Equation of state calculations by fast computing machines. *J Chem Phys* **21**, 1087-1092 (1953).

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

29. Allen, B. D., and Mayo, S. L. Dramatic performance enhancements for the FASTER optimization algorithm. *J Comput Chem* **27**, 1071-1075 (2006).

30. Desmet, J., Spriet, J., and Lasters, I. Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins* **48**, 31-43 (2002).

31. Gordon, D. B., Marshall, S. A., and Mayo, S. L. Energy functions for protein design. *Curr Opin Struct Biol* **9**, 509-513 (1999).

32. Pande, V. S., Grosberg, A. Y., and Tanaka, T. Statistical mechanics of simple models of protein folding and design. *Biophys J* **73**, 3192-3210 (1997).

33. Shakhnovich, E. I., and Gutin, A. M. A new approach to the design of stable proteins. *Protein Eng* **6**, 793-800 (1993).

34. Derrida, B. Random-energy model - limit of a family of disordered models. *Phy Rev Let* **45**, 79-82 (1980).

35. Shakhnovich, E. I., and Gutin, A. M. Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of a replica approach. *Biophys Chem* **34**, 187-199 (1989).

36. Bryngelson, J. D., and Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci USA* **84**, 7524-7528 (1987).

37. Koehl, P., and Levitt, M. De novo protein design. II. Plasticity in sequence space. *J Mol Biol* **293**, 1183-1193 (1999).

38. Koehl, P., and Levitt, M. De novo protein design. I. In search of stability and specificity. *J Mol Biol* **293**, 1161-1181 (1999).

39. Hecht, M. H., Richardson, J. S., Richardson, D. C., and Ogden, R. C. De novo design, expression, and characterization of Felix: a four-helix bundle protein of native-like sequence. *Science* **249**, 884-891 (1990).

40. Deng, Y., Liu, J., Zheng, Q., Eliezer, D., Kallenbach, N. R., and Lu, M. Antiparallel four-stranded coiled coil specified by a 3-3-1 hydrophobic heptad repeat. *Structure* **14**, 247-255 (2006).

41. Havranek, J. J., and Harbury, P. B. Automated design of specificity in molecular recognition. *Nat Struct Biol* **10**, 45-52 (2003).

42. Nautiyal, S., Woolfson, D. N., King, D. S., and Alber, T. A designed heterotrimeric coiled coil. *Biochemistry* **34**, 11645-11651 (1995).

43. Yue, K., Fiebig, K. M., Thomas, P. D., Chan, H. S., Shakhnovich, E. I., and Dill, K. A. A test of lattice protein-folding algorithms. *Proc Natl Acad Sci USA* **92**, 325-329 (1995).

44. Yue, K., and Dill, K. A. Inverse protein folding problem - designing polymer sequences. *Proc Natl Acad Sci USA* **89**, 4163-4167 (1992).

45. Marshall, S. A., and Mayo, S. L. Achieving stability and conformational specificity in designed proteins via binary patterning. *J Mol Biol* **305**, 619-631 (2001).

Fig. 1-1: Inverse Protein Folding

The concept behind protein folding is shown in orange. The purpose is to take one sequence (depicted as an orange dot in sequence space) and find the lowest energy conformation. Protein design strives to take a low energy conformation and identify any sequence that will fold to it (compatible sequences are represented by the purple region within sequence space).

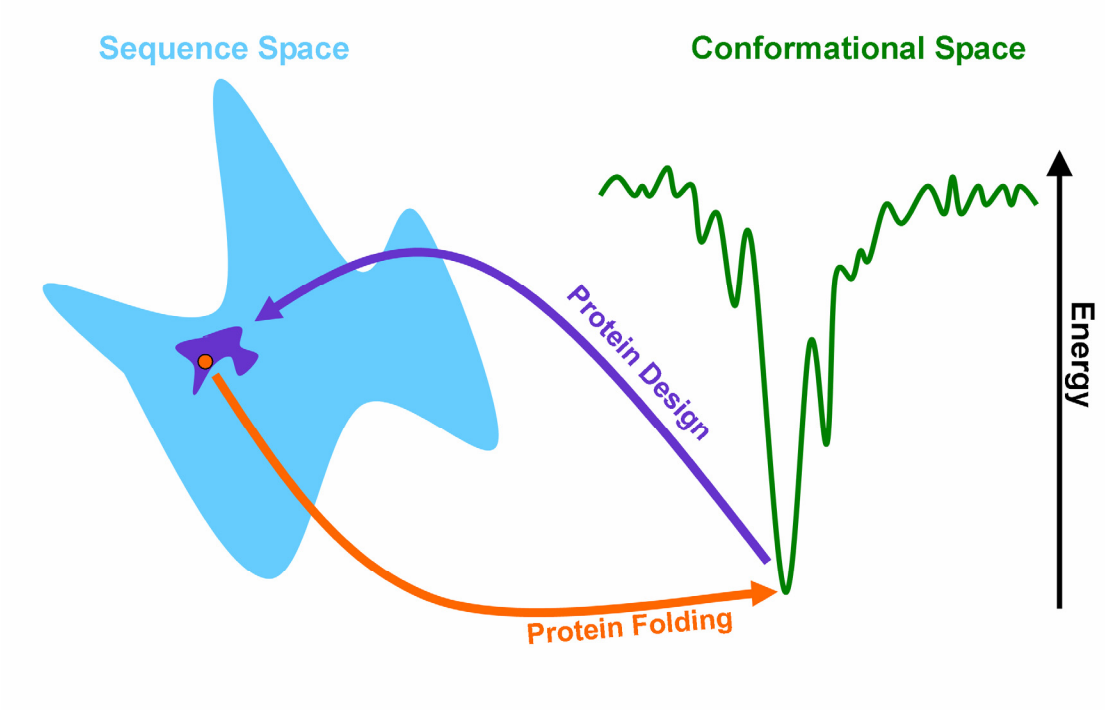


Fig. 1-2: van der Waals Potential Function

The form of the Lennard-Jones 12-6 potential used to calculate the van der Waals interaction energy, E_{vdw}, is shown below. The two spheres represent atoms for which E_{vdw} is being calculated. R_1 and R_2 are the atomic radii for each of the respective atoms. R_o is the average of R_1 and R_2. D_o is the average of the well depth of the two atoms. The distance between the two atoms (R) dictates the atomic van der Waals energy.

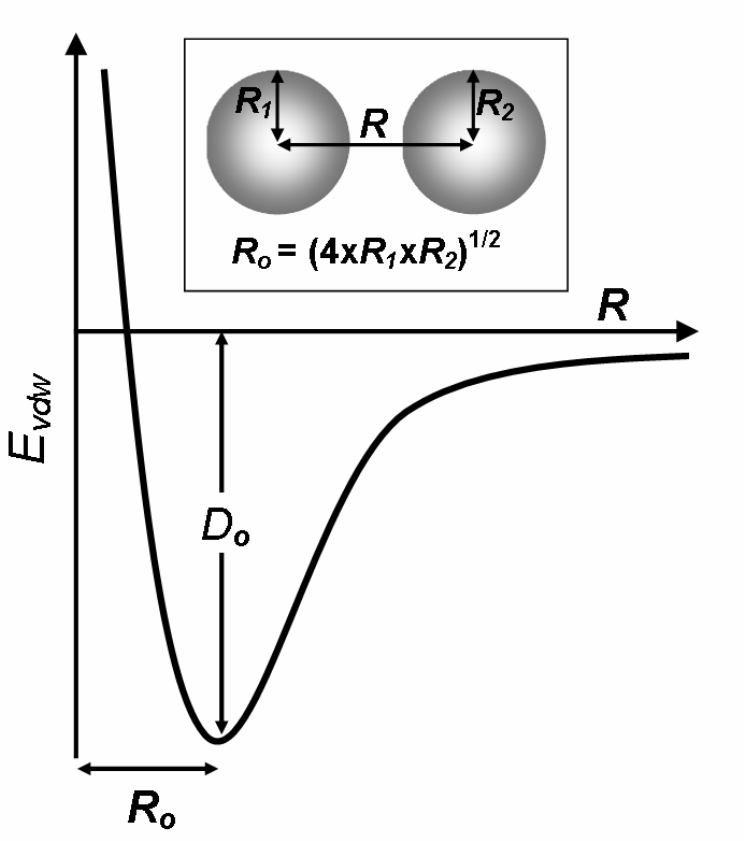


Fig. 1-3: Hydrogen Bond Potential

The form of the hydrogen bond potential (panel *A*) is similar to that of the van der Waals potential. R_0, however, is not a variable, but is set to 2.8 Å, the optimal donor to acceptor distance for a hydrogen bond. The three angles used to calculate the angle dependence function, F, are illustrated in the diagram in panel *B*.

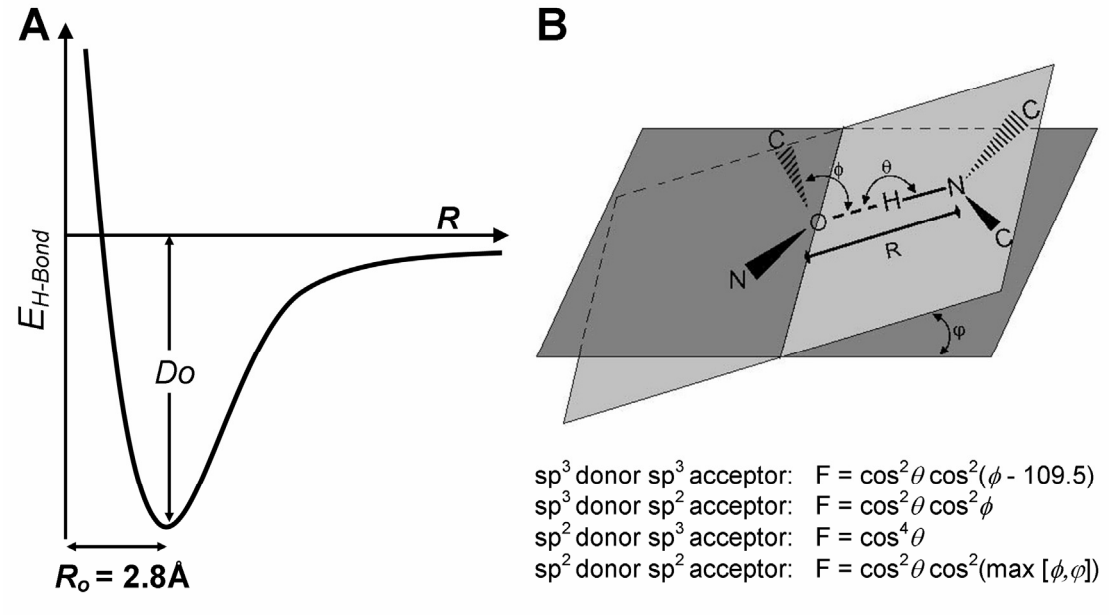


Fig. 1-4: Solvent Exclusion-Based Solvation

Solvent exclusion is assumed to be proportional to atom burial. In this example, the yellow atom's solvation is being calculated. Panel *A* shows the yellow atom in its fully solvated reference state that is used to calculate the reference solvation energy (ΔG_i^{ref}). The atoms that make up the rest of the side-chain are shown in gray. Panel *B* shows the yellow atom fully buried. Surrounding atoms in the folded state are shown in white and are depicted taking up the space once held by the solvent. The distance (r_{ij}) and volume (V_j) of the surrounding atoms are used to determine the extent of solvent exclusion.

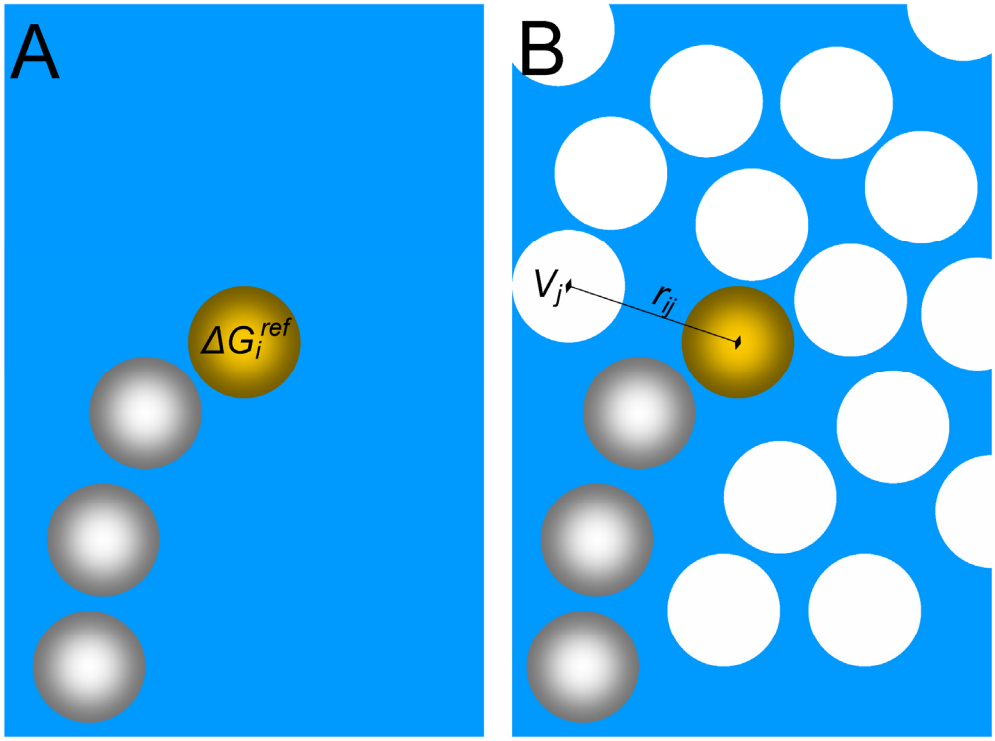
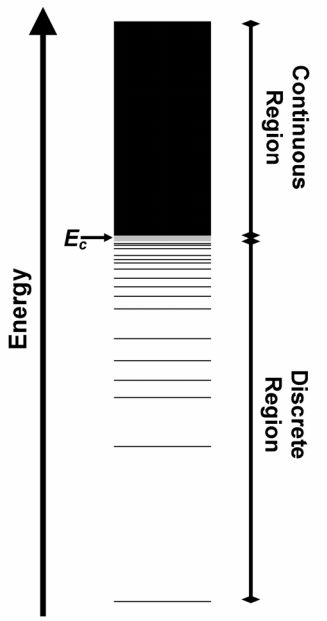


Fig. 1-5: Conformational Energy Spectrum

The energy spectrum is divided into a continuous region and a discrete region. Conformations above the critical energy, E_c, are only accessible at high temperature and are in the continuous region. Conformations with energy below E_c are in the discrete region. Sequences with identical amino acid composition have identical continuous regions and sequence specific discrete regions.



Chapter 2

*Application of Fixed Composition Protein Design**

2.1 Introduction

A major aim of CPD is to reproducibly design sequences that adopt a desired tertiary structure. This requires a CPD procedure yielding sequence scores that accurately reflect experimentally determined stabilities. Since experimental energies are determined with respect to an unfolded state, a CPD force field should accurately model interactions in both the folded and unfolded states. However, modeling the unfolded state in a useful way has proven difficult. As a result, most CPD force fields omit the specific effects of sequence changes on the unfolded state and only optimize interactions in the folded state [1]. This disregard of the unfolded state is partly to blame for discrepancies between computationally derived and experimentally determined protein stabilities and for the difficulty of developing a properly tuned CPD force field [2, 3].

Separating the tuning of a CPD force field into two logical components, the unfolded and folded states, could ultimately lead to force fields with significantly

* This chapter was adapted from a manuscript in preparation for PNAS coauthored with Stephen L. Mayo.

improved predictive power. The work presented here demonstrates a procedure for achieving this separation by invoking the random energy model (REM) [4] in order to minimize the influence of the unfolded state in determining sequence designs. In this way, force field evaluation and tuning can be focused on the more tractable folded state. REM was initially developed for spin glass models and later adapted for proteins [5-7]. REM asserts that the energy spectrum for any specific amino acid sequence is divided into continuous and discrete regions (Fig. 2-1). The conformational energies in the discrete region rely on best-fit contacts, making them sequence-specific. The continuous region, however, represents conformations that are only accessible at higher temperatures where the rapid inter-conversion between conformations leads to a distribution of conformational energies that depends solely on the amino acid composition. Consequently, all sequences with identical amino acid composition are expected to have identical continuous region distributions and, thus, unfolded state energies (Fig. 2-1 *D*, *E,* and *F*) [8]. As a result, the free energy of folding of fixed composition sequences are directly correlated to their folded state energies. The same cannot be said when comparing sequences with varied composition (Fig. 2-1 *A*, *B*, and *C*). In this case, the continuous region varies between sequences and the free energy of folding cannot be directly compared without explicit consideration of the unfolded state. A sequence can potentially have the best energy in the folded state and fail to have the largest change in energy (Fig. 2-1 *C*).

Here, we exploit the fixed composition concept by limiting designs to sequences with fixed amino acid composition [9, 10]. By doing so, we can eliminate unfolded state contributions and focus on evaluating and optimizing the force field for the folded state.

If the unfolded states for fixed sequence designs are inconsequential, any discrepancies between experimental and computational stabilities can be attributed to the force field's inability to predict the impact of sequence variation on the folded state.

Application of a fixed composition method imposes a large negative design constraint on the system [9, 10]. The importance of negative design for protein sequence selection was revealed with hydrophobic/polar lattice model simulations [11, 12]. Early studies on lattice models demonstrated that in order to recover sequences that specifically folded to the target structure, polar monomers had to be explicitly considered at surface positions even though they did not impart favorable energy to the system [11]. The alternative led to sequences dominated by solvent exposed hydrophobic monomers. Incorporation of an explicit negative design constraint on amino acid sequence selection was demonstrated by Dahiyat and Mayo [13] who went on to show that CPD could be successfully applied to complete protein domains [14]. In that and related work, either a pseudo-binary pattern or an explicit binary pattern of polar and non-polar amino acids was used to impose fold specificity [14, 15]. Less restrictive negative design alternatives include the use of amino acid reference energies to control amino acid composition [16-18].

In addition to normalizing folded state energies, fixed composition design directly considers the fold specificity of sequences (i.e., the ability of an amino acid sequence to adopt a single or limited number of structures). REM theory indicates that as ΔE increases, the accessible conformations for a sequence exponentially decrease [19]. Since the unfolded state energies are identical for amino acid sequences with the same composition, finding sequence arrangements that optimize the energy of the folded state

is equivalent to maximizing ΔE (Fig. 2-1 *D*, *E,* and *F*). Optimized sequences with large favorable scores on the target fold are thus expected to exhibit an energy spectrum in which it is improbable to achieve an alternative conformation with lower energy.

Explicitly fixing the amino acid composition for a design has the inherent problem of requiring knowledge of the composition before the design calculation is started. For the work presented here, the wild-type sequence of the β1 domain of streptococcal protein G (Gβ1) is used. Because Gβ1 has a high thermal stability (T_m of 88°C) its wild-type amino acid sequence is expected to be near optimal (given the constraint of maintaining the wild-type amino acid composition). Consequently, the CPD force field can be evaluated and optimized based on its ability to recover the wild-type sequence prior to laborious experimental testing of designed sequences. More specifically, the use of a wild-type sequence bias can be used in a stepwise fashion to force recovery of the wild-type sequence and to identify problematic force field components. The computed Z-score of the wild-type sequence and the experimental testing of unbiased designs can then be used to assess the overall quality of the CPD force fields.

2.2 Results and Discussion

2.2.1 The Standard Force Field: Identifying the Inaccuracies

Standard force field parameters and potential functions [13, 14, 20-22] were used for our initial fixed composition designs, since these have been previously tested and successfully applied to a wide range of protein design problems [14, 15, 23]. The initial force field included terms for van der Waals interactions [13], hydrogen bond formation,

and electrostatic interactions. Solvation was modeled using a solvent-accessible surface area-based term that encourages hydrophobic burial and polar exposure. Side-chain flexibility was taken into account using expanded versions of the backbone-dependent rotamer library of Dunbrack and Karplus [24].

The success of the standard force field has required imposing some type of binary pattern, either explicitly or by restricting buried positions to non-polar amino acids and exposed positions to polar residues [14, 15, 23]. In our fixed composition designs, however, we removed these restrictions, and within the fixed composition limits, allowed all amino acids at all positions. Without any binary pattern or regional restrictions, we expected the resulting fixed composition sequences to reveal previously hidden inaccuracies in the standard force field, and allow us to identify aspects that could be improved.

Fixed composition designs were first performed on Gβ1 using the initial (standard) force field. All non-Gly and non-Met positions were included in the design and the amino acid composition was fixed to that of the wild-type protein. A wild-type sequence bias was imposed and incrementally increased until the wild-type sequence was recovered. Fig. 2-2 *A* shows the top-ranked sequences obtained from each calculation. At lower sequence biases, the predicted sequences exhibited poor recovery of the wild-type amino acids, revealing substantial inaccuracies in the initial force field. The unbiased design (sbias0.0) had 16% sequence identity with the wild type, an increase of only 5 percentage points over random fixed composition sequences (Table 1). Only two out of ten designed core positions were predicted to take on wild-type amino acids, and

even fewer boundary and surface positions recovered the wild-type amino acid identities (8% and 17%, respectively).

The inaccuracies in the initial force field were further highlighted by the poor quality of the sequences predicted using lower sequence biases. All the predicted sequences contained charged and polar amino acids at core positions (Fig. 2-2 *A*). The sequence recovered at a sequence bias of 5.0 replaced a core Leu with a Glu. In a small protein with a well-packed hydrophobic core, it is unlikely that substituting non-polar amino acids with charged residues would result in a more stable variant [25]. Exploratory modifications to the force field suggested that changing to a solvent exclusion-based solvation model would result in improved prediction of core residues. This model emphasizes polar interactions, which results in a larger penalty for burial of polar atoms; consequently, charged or polar amino acids in the core are strongly disfavored.

Further inspection of the predicted sequences revealed a bias towards sequence arrangements that benefit from the strong hydrogen bond potential contained in the initial force field. Certain core positions were predicted to take on polar side chains, partly because they were able to form strong inter-residue hydrogen bonds. For example, core position 20 mutated from Ala to Gln in order to form two hydrogen bonds with surface residues. The predicted Gln side chain assumes a strained conformation in order to satisfy the interactions. Similarly, Thr at core position 30 is predicted to form a hydrogen bond with the α-helical backbone. Due to Thr's low α-helical propensity, this mutation is likely to be destabilizing. We anticipated that lowering the benefit for hydrogen bond formation would reduce this unwanted preference for polar side chains in the core.

The discrete nature of rotamer libraries also appears to be problematic in that a suitable conformation for the wild-type amino acid may not be available for certain positions. For example, the absence of a rotamer with chi angles similar to those seen in the crystal structure for core Leu7 resulted in spurious predictions: in all of the predicted sequences, Trp 43 swings out into the solvent, even though the crystal structure clearly depicts the Trp to be buried. The poor choice of conformation at position 7 propagates throughout the core and results in the expulsion of Trp 43 (Fig. 2-3 *A*). The use of a larger, more representative rotamer library should mitigate this type of problem, since it is more likely to contain conformations comparable to those observed in the crystal structures.

Three of the predicted sequences obtained with the initial force field were selected for further study. Sequences obtained at a sequence bias of 0.0, 2.0, and 5.0 (sbias0.0, sbias2.0, and sbias5.0) were chosen for physical characterization. Not surprisingly, circular dichroism (CD) experiments showed that proteins with the largest difference from wild type (sbias0.0 and sbias2.0) were unfolded (data not shown). Sbias5.0, on the other hand, was folded but significantly destabilized compared to wild type (Fig. 2-4).

2.2.2 Optimizing the Force Field

Multiple rounds of optimization were required to obtain a set of parameters that yielded viable sequences. In an effort to hinder the selection of charged or polar residues in the core, we first changed the model used to calculate atomic solvation: the surface area-based model was replaced by a solvent exclusion-based model [26]. In addition, the benefit for hydrogen bond formation was decreased, especially at the surface. These changes, particularly switching to a different solvation model, yielded the most dramatic

improvements. To increase the chances of recovering native conformations, we also replaced the rotamer library with a larger conformer library [27, 28] consisting of side-chain conformations closely resembling those observed in wild-type crystal structures. Since the reduction in the benefit for hydrogen bond formation also affects the benefit for salt bridges, we compensated by adjusting the dielectric constant (from 40 to 20) to improve electrostatic interactions throughout the protein. Further details on the impact of the modifications are discussed in Chapter 3.

The sequences predicted with the improved force field are shown in Fig. 2-2 *B*. Unlike the sequences obtained with the initial parameters, the improved parameter sequences exhibited no obvious irregularities that discouraged further evaluation. All core positions in every sequence adopted wild-type amino acid identities, ensuring that the core was well-packed and hydrophobic (Table 2-1 and Fig. 2-2 *B*). In addition, the larger conformer library selected core conformations that overlaid nicely with those seen in the wild-type crystal structure, even though the Gβ1 structure was not included in the set of structures used to generate the conformer library (Fig. 2-3 *B*). Trp43 swung out into the solvent with the smaller rotamer library, but with the conformer library, Trp43 was packed into the core.

Further improvements in our CPD procedure are illustrated in Table 1. The total wild-type recovery of the unbiased sequence (sbias0.0) increased from 16% to 53%. Wild-type recovery was 100% for core, 50% for boundary, and 38% for surface positions, representing 5-fold, 6-fold, and 2-fold improvements, respectively. The sequence bias required to recover the wild-type sequence decreased from 6.0 to 2.0

kcal/mol. The improvement in the predictive power of the force field was further confirmed by improved Z-scores.

The Z-score, a value previously used for force field optimization [29], was calculated for the wild-type sequence. As the force field improves, the Z-score for the wild-type sequence should increase [30, 31]. Initial parameters gave a Z-score of 2.7 for the wild-type sequence, while the improved parameters yielded a Z-score of 3.5 (Fig. 2-5). A clearer picture of the improvement is seen by comparing the wild-type sequence with sequences obtained using no sequence bias. Initial parameters yielded a Z-score of 8.6 for the predicted unbiased sequence (sbias0.0), a difference of 5.9 compared to the 2.7 value obtained for the wild-type sequence (Fig. 2-5 *A*). The fact that a sequence resulting in an unfolded protein had such a large Z-score relative to that calculated for the wild-type sequence is further evidence of the poor predictive power of the initial force field. In contrast, the improved parameters produced a Z-score for the unbiased sequence (sbias0.0) of 4.7, a difference of only 1.2 relative to the wild-type sequence (Fig. 2-5 *B*).

Definitive validation of the improved parameters was provided by experimental analysis of the predicted sequences. Proteins corresponding to sbias0.0, sbias0.5, sbias1.0, and sbias1.5 were all shown to be folded by CD. The unbiased sequence, which had the lowest wild-type recovery (53%), was also shown to be folded by NMR (data not shown). Temperature denaturation experiments revealed all the predicted proteins to be highly thermostable (Fig. 2-6). Sbias0.0, sbias0.5, sbias1.0, and sbias1.5 exhibited T_ms of 73.6, 83.1, 85.0, and 83.7ºC, respectively. In contrast to the unbiased sequence obtained with the initial parameters, the unbiased sequence obtained with the improved parameters resulted in a protein that was folded and well-behaved.

2.2.3 Evaluating Improved Parameters on Engrailed Homeodomain

To determine whether the improved force field parameters showed a preference for the Gβ1 scaffold, fixed composition designs were carried out on Engrailed homeodomain from *Drosophila melanogaste*r (ENH). ENH is a small globular protein with no sequence or structure homology to Gβ1.

Table 2-1 shows the resulting sequence statistics from the fixed composition designs on ENH. The wild-type sequence was recovered at a sequence bias of 2.5 kcal/mol/pos. The unbiased design (sbias0.0) predicted a sequence with 42% wild-type sequence identity, with the core recovering 80% of the wild-type amino acids. Core position 34 mutated from Leu to Phe to achieve improved packing interactions (Fig. 2-7). In the absence of a Gln rotamer that can hydrogen bond with the N-terminus at position 44, the force field selected a polar residue that can hydrogen bond with the backbone at residue 41. Using the improved force field parameters, the Z-score difference for the unbiased design vs. wild type was only 1.65 (4.59 – 2.94). This is in contrast to a Z-score difference of 5.51 (8.0 – 2.49) when using the initial parameters.

The improved force field predicted reasonable sequences for ENH with improved Z-scores when compared to the initial force field. These results support an assertion that the idea that the modifications to the force field are not specific for the Gβ1 scaffold. However, in order to best eliminate bias towards a particular scaffold, the optimization procedure would need to be carried out simultaneously on multiple scaffolds.

2.2.4 Removal of Fixed Composition Constraint

A simple and straightforward way to show the importance of negative design in CPD is to compare the results obtained under a fixed composition restraint with the

results obtained without it. We therefore used the improved parameters to design Gβ1, except this time we removed the requirement for a fixed composition. This resulted in a sequence dominated by large hydrophobic residues (Fig. 2-8). Similar to what is observed in lattice models when the composition at surface residues is not restricted, the CPD procedure predicted a hydrophobic surface to be stabilizing.

This result is further evidence that negative design must be explicitly considered in CPD to promote specificity of the target fold [10, 15, 16, 32]. The hydrophilic composition at the surface in native proteins is crucial for dissuading aggregation, increasing solubility and ensuring proper folding [33]. Although not always energetically favorable, polar and charged residues have naturally evolved at the surface to ensure highly specific folding [15]. In the absence of any negative design terms, physically based force fields will select hydrophobic residues at the surface since these residues can form strong packing interactions. Consequently, a strong negative design term is required to provide folding specificity in designed sequences.

Negative design can be incorporated in a number of ways. Surface positions can be limited to polar side chains, or reference energies that ensure a native-like composition at surface positions can be used. The solvation term can also include negative design terms that serve as an incentive to select polar residues at the surface (e.g., a benefit for polar solvent exposure can be included).

2.3 Conclusion

Fixed composition design proved to be an effective way to optimize the positive design parameters in our CPD force field. By limiting the designs to sequences with

identical amino acid composition, inconsistencies between experimental and computational results could be attributed to the force field's inability to accurately model the folded state. Direct comparison between experimental and computational results was possible since the unfolded state energy is presumed to be equal for all sequences.

Iterative use of fixed composition design allowed for a set of force field parameters to be identified that consistently predicted folded and well-behaved sequences. The implementation of this procedure is straightforward and generalizable to any protein design force field, provided the sequence for the selected scaffold is shown to be near optimal for its folded structure. To dissuade unintentional bias towards the target scaffold, the fixed composition optimization procedure should be carried out on multiple scaffolds.

Limiting the force field optimization procedure to fixed composition designs is a conceptually straightforward way to optimize positive design parameters; however, it does little for negative design parameters. To address this issue, it would be useful to include a step in the procedure that optimizes the negative design parameters in the absence of a fixed composition restraint.

The work presented here describes a procedure for systematically optimizing a CPD force field. The experimental results clearly show a dramatic improvement in sequence prediction after optimization. Whereas the starting force field failed to predict a stable sequence in the absence of a sequence bias, the improved force field predicted a 23-fold mutant that was stably folded, and displayed 6-fold, 3-fold, and 2-fold improvements in predictive power at core, boundary, and surface, respectively.

2.4 Materials and Methods

2.4.1 Fixed Composition Scaffolds

Coordinates for the backbone structure of Gβ1 and ENH were obtained from the Protein Data Bank entry 1PGA and 1ENH, respectively. Any strain or steric clashes in the structure were removed by performing 50 steps of energy minimization. Residue classification into core, boundary, and surface groups was performed as described previously [14]. All 51 non-Gly and non-Met positions were included in the design, and within fixed composition restraints, all amino acids found in the wild-type Gβ1 sequence list were allowed at all designed positions.

2.4.2 Fixed Composition Force Fields

The initial force field used standard potential functions and parameters including scaled van der Waals, hydrogen bonding, electrostatic, and surface area-based solvation terms, as described previously [13, 14, 20-22]. Expanded versions of Dunbrack and Karplus' 1995 backbone-dependent rotamer library were used [24]. Aromatic residues were expanded 1 SD about their χ_1 and χ_2 values, and hydrophobic residues were expanded 1 SD about their χ_1 values; polar residues were not expanded.

The improved force field used a solvent exclusion-based solvation potential [26]. All published solvation parameters were used with the exception of polar burial, which was decreased by 40% [26]. The benefit for side chain-side chain hydrogen bond formation was decreased by 50% for core and boundary residues. Hydrogen bond energies were decreased by an additional 75% if predicted between immediate neighbors (n+1 and n–1 positions). Hydrogen bonds at surface positions received a benefit from the

electrostatic potential, but not from the hydrogen bond potential. The dielectric constant was reduced from 40 to 20.

The improved force field used a larger backbone-dependent conformer library [28] instead of the rotamer library. The conformer library was constructed using Cartesian coordinates taken directly from high-resolution crystal structures as described by Lassila et al. [27]. Conformer probabilities were taken into account: the *p* value for non-polar amino acids was set to 0.3; a *p* value of 0.6 was used for Asp, Glu, Asn, and Gln; and representative conformers for Arg and Lys were obtained with a *p* of 0.8.

2.4.3 Fixed Composition Sequence Optimization

Prior to sequence optimization, an energy matrix containing all one-body and two-body interactions was created. The one-body term for each rotamer was modified to reflect a specified sequence bias energy. Each rotamer that differed in identity from the wild-type amino acid at a particular position received a penalty. The resulting sequence was thus penalized for each residue that differed from the wild-type sequence. All calculations were first carried out in the absence of a sequence bias. The bias energy was then incrementally increased by 1.0 or 0.5 kcal/mol/position, while keeping all other parameters fixed, until the wild-type sequence was recovered.

A stochastic algorithm, Monte Carlo simulated annealing, was used for the fixed composition Gβ1 designs. A fixed composition restraint was imposed by creating a new version of Monte Carlo, FMONTE. The FMONTE algorithm randomly picks four positions and arbitrarily switches the amino acids at two, three, or all four of the positions. A random rotamer is chosen at each of the switched positions, and the sequence energies are compared. All calculations were carried out for 1,000 annealing

cycles at 1,000,000 steps per cycle and the temperature was cycled from 4,000 K to 150 K. Fixed composition designs on ENH were performed using a fixed composition version of the FASTER algorithm [34].

2.4.4 Sequence Optimization in the Absence of a Fixed Composition Constraint

Designs were carried out on Gβ1 using the improved parameters. All amino acids were allowed at all positions with the exception of Met, Gly, Cys, and Pro. The sequence search was carried out using the FASTER algorithm [35].

2.4.5 Protein Expresion and Purification

Mutant plasmids were created by site-directed mutagenesis of the wild-type gene in pET-11a or ordered from Blue Heron Biotechnology. Electroporation was used to transform the completed plasmids into BL21 (DE3) cells. Cells were allowed to express protein for 3 hr after induction with IPTG, then harvested and lysed by sonication. Cell extracts were spun down and precipitated by addition of 50% acetonitrile. The soluble protein was separated from the precipitate by centrifugation and HPLC purified. Pure proteins were analyzed by either trypsin digest or by collision-induced dissociation mass spectrometry.

2.4.6 Experimental Studies

CD studies were done in an Aviv 62A DS spectropolarimeter with a thermoelectric cell holder. Samples were prepared in 50 mM sodium phosphate buffer at pH 5.5. Guanidinium denaturations were carried out in a 1 cm path length cuvette at a protein concentration of 5 μM (1,800 μl). An autotitrator was used for the chemical denaturations and data was collected at 218 nm and 25ºC. After each injection of

denaturant, samples were stirred for 10 min before data was collected (100 sec averaging time). Wavelength scans and temperature denaturations were carried out in cuvettes with a 0.1 cm path length at a concentration of 50 μM (300 μl). Three wavelength scans were done at 25ºC. Data was collected from 200 nm to 250 nm at 1 nm intervals and averaged for 1 sec. Temperature denaturations were carried out from 0ºC to 99ºC, sampling every 1ºC. Samples were equilibrated for 90 sec before data was collected (averaging time 30 sec). 1D NMR experiments were done on a Varian Unityplus 600-MHz spectrometer at 25ºC. Samples were prepared in 50 mM sodium phosphate buffer pH 5.5 using 9:1 $H_2O/^2H_2O$.

2.5 Bibliography

1. Gordon, D. B., Marshall, S. A., and Mayo, S. L. Energy functions for protein design. *Curr Opin Struct Biol* **9**, 509-513 (1999).

2. Dill, K. A., and Shortle, D. Denatured states of proteins. *Annu Rev Biochem* **60**, 795-825 (1991).

3. Lazar, G. A., Desjarlais, J. R., and Handel, T. M. De novo design of the hydrophobic core of ubiquitin. *Protein Sci* **6**, 1167-1178 (1997).

4. Derrida, B. Random-energy model: Limit of a family of dosordered models. *Phys Rev Lett* **45**, 79-82 (1980).

5. Bryngelson, J. D., and Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci USA* **84**, 7524-7528 (1987).

6. Pande, V. S., Grosberg, A. Y., and Tanaka, T. Statistical mechanics of simple models of protein folding and design. *Biophys J* **73**, 3192-3210 (1997).

7. Shakhnovich, E. I., and Gutin, A. M. Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of a replica approach. *Biophys Chem* **34**, 187-199 (1989).

8. Shakhnovich, E. I., and Gutin, A. M. Engineering of stable and fast-folding sequences of model proteins. *Proc Natl Acad Sci USA* **90**, 7195-7199 (1993).

9. Koehl, P., and Levitt, M. De novo protein design. II. Plasticity in sequence space. *J Mol Biol* **293**, 1183-1193 (1999).

10. Koehl, P., and Levitt, M. De novo protein design. I. In search of stability and specificity. *J Mol Biol* **293**, 1161-1181 (1999).

11. Yue, K., and Dill, K. A. Inverse protein folding problem: Designing polymer sequences. *Proc Natl Acad Sci USA* **89**, 4163-4167 (1992).

12. Yue, K., Fiebig, K. M., Thomas, P. D., Chan, H. S., Shakhnovich, E. I., and Dill, K. A. A test of lattice protein folding algorithms. *Proc Natl Acad Sci USA* **92**, 325-329 (1995).

13. Dahiyat, B. I., and Mayo, S. L. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci USA* **94**, 10172-10177 (1997).

14. Dahiyat, B. I., and Mayo, S. L. De novo protein design: Fully automated sequence selection. *Science* **278**, 82-87 (1997).

15. Marshall, S. A., and Mayo, S. L. Achieving stability and conformational specificity in designed proteins via binary patterning. *J Mol Biol* **305**, 619-631 (2001).

16. Kuhlman, B., and Baker, D. Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci USA* **97**, 10383-10388 (2000).

17. Raha, K., Wollacott, A. M., Italia, M. J., and Desjarlais, J. R. Prediction of amino acid sequence from structure. *Protein Sci* **9**, 1106-1109 (2000).

18. Wernisch, L., Hery, S., and Wodak, S. J. Automatic protein design with all atom force-fields by exact and heuristic optimization. *J Mol Biol* **301**, 713-736 (2000).

19. Shakhnovich, E. I., and Gutin, A. M. A new approach to the design of stable proteins. *Protein Eng* **6**, 793-800 (1993).

20. Dahiyat, B. I., Gordon, D. B., and Mayo, S. L. Automated design of the surface positions of protein helices. *Protein Sci* **6**, 1333-1337 (1997).

21. Dahiyat, B. I., and Mayo, S. L. Protein design automation. *Protein Sci* **5**, 895-903 (1996).

22. Dahiyat, B. I., Sarisky, C. A., and Mayo, S. L. De novo protein design: towards fully automated sequence selection. *J Mol Biol* **273**, 789-796 (1997).

23. Malakauskas, S. M., and Mayo, S. L. Design, structure and stability of a hyperthermophilic protein variant. *Nat Struct Biol* **5**, 470-475 (1998).

24. Dunbrack, R. L., Jr., and Karplus, M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol* **230**, 543-574 (1993).

25. Meyer, S. C., Huerta, C., and Ghosh, I. Single-site mutations in a hyperthermophilic variant of the β1 domain of protein G result in self-assembled oligomers. *Biochemistry* **44**, 2360-2368 (2005).

26. Lazaridis, T., and Karplus, M. Effective energy function for proteins in solution. *Proteins* **35**, 133-152 (1999).

27. Lassila, J. K., Privett, H. K., Allen, B. D., and Mayo, S. L. Combinatorial methods for small-molecule placement in computational enzyme design. *Proc Natl Acad Sci USA* **103**, 16710-16715 (2006).

28. Shetty, R. P., De Bakker, P. I., DePristo, M. A., and Blundell, T. L. Advantages of fine-grained side chain conformer libraries. *Protein Eng* **16**, 963-969 (2003).

29. Chiu, T. L., and Goldstein, R. A. Optimizing potentials for the inverse protein folding problem. *Protein Eng* **11**, 749-752 (1998).

30. Bowie, J. U., Luthy, R., and Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164-170 (1991).

31. Street, A. G., Datta, D., Gordon, D. B., and Mayo, S. L. Designing protein β-sheet surfaces by Z-score optimization. *Phys Rev Lett* **84**, 5010-5013 (2000).

32. Hellinga, H. W. Rational protein design: combining theory and experiment. *Proc Natl Acad Sci USA* **94**, 10015-10017 (1997).

33. Hecht, M. H., Richardson, J. S., Richardson, D. C., and Ogden, R. C. De novo design, expression, and characterization of Felix: A four-helix bundle protein of native-like sequence. *Science* **249**, 884-891 (1990).

34. Hom, G. K., and Mayo, S. L. A search algorithm for fixed-composition protein design. *J Comput Chem* **27**, 375-378 (2006).

35. Desmet, J., Spriet, J., and Lasters, I. Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins* **48**, 31-43 (2002).

Table 2-1: Percent Wild-Type Sequence Identity Before and After Force Field Optimization

	Percent Sequence Identity*			
Predicted Sequences	Total	Core	Boundary	Surface
Gβ1 - Initial Parameters				
sbias0.0	16	20	8	17
sbias2.0	55	70	33	59
sbias4.0	84	80	67	93
sbias5.0	92	90	92	93
sbias6.0	100	100	100	100
Gβ1 - Improved Parameters				
sbias0.0	53	100	50	38
sbias0.5	76	100	50	86
sbias1.0	92	100	83	97
sbias1.5	96	100	92	97
sbias2.0	100	100	100	100
ENH – Initial Parameters				
sbias0.0	22	20	9	28
sbias2.0	46	30	27	59
sbias4.0	66	60	55	72
sbias6.0	96	90	91	100
sbias8.0	96	90	91	100
sbias9.0	100	100	100	100
ENH – Improved Parameters				
sbias0.0	42	80	45	28
sbias0.5	64	80	55	62
sbias1.0	86	90	82	86
sbias1.5	90	90	91	90
sbias2.0	94	90	100	93
sbias2.5	100	100	100	100

*Wild-type sequence identity was determined using only the positions in the design. Values are rounded to the nearest integer. Wild-type sequence identities for random fixed composition sequences for Gβ1 were calculated to be 11, 8, 11 and 12 percent for total, core, boundary, and surface positions, respectively.

Fig. 2-1: Conformational Energy Spectra for Six Sequences

Each spectrum is divided into a continuous and a discrete region. The continuous region is depicted as a solid black bar above the red marker, while the discrete region is shown below the red marker. The lowest energy conformation for each of the sequences is shown in green. ΔE is defined as the energy difference between the lowest energy conformation and the energy at the transition between the continuous and discrete regions. Energy spectra *A*, *B,* and *C* represent sequences with different amino acid compositions obtained from standard, non-fixed composition designs. Energy spectra *D*, *E,* and *F* represent sequences with identical amino acid composition.

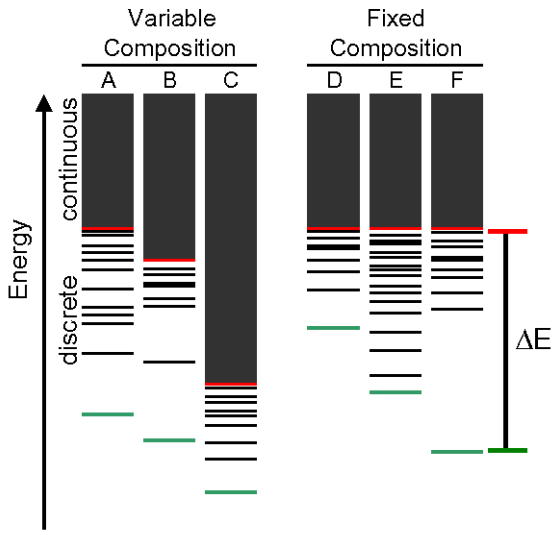


Fig. 2-2: Predicted Sequences for Gβ1 Fixed Composition Designs

Designed core, boundary and surface positions are shown in red, green, and blue, respectively. The wild-type sequence is shown followed by the sequences predicted at increasing sequence bias values (kcal/mol/position). Dots represent wild-type amino acids. (*A*) Sequences obtained using the initial parameters. (*B*) Sequences obtained using the improved parameters.

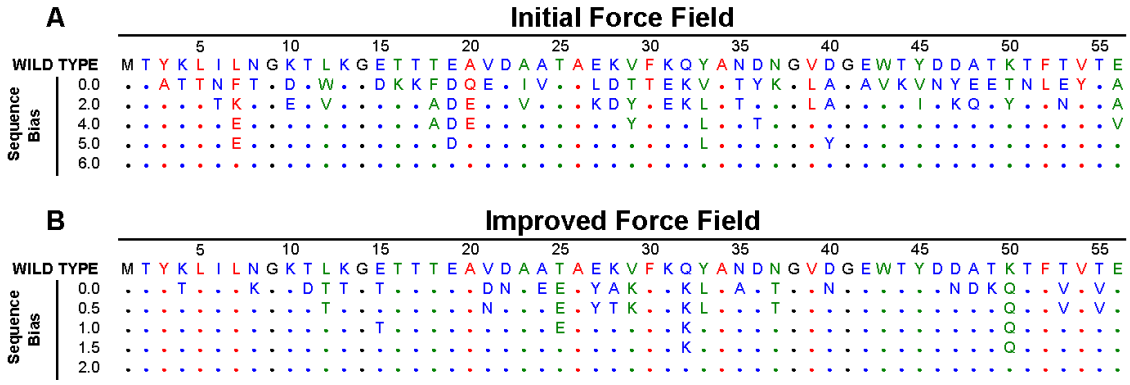


Fig. 2-3: Predicted and Wild-Type Crystal Structure Conformations for Four Gβ1 Designed Core Residues

(*A*) Conformations obtained using a standard rotamer library. (*B*) Conformations obtained using the conformer library. Predicted conformations are shown in orange; wild-type crystal structure conformations are depicted in gray. The conformer library does a much better job at recapitulating the conformations seen in the wild-type crystal structure. With the standard rotamer library, the absence of a rotamer with a conformation similar to that in the crystal structure at position 7 results in Trp 43 swinging out into the solvent.

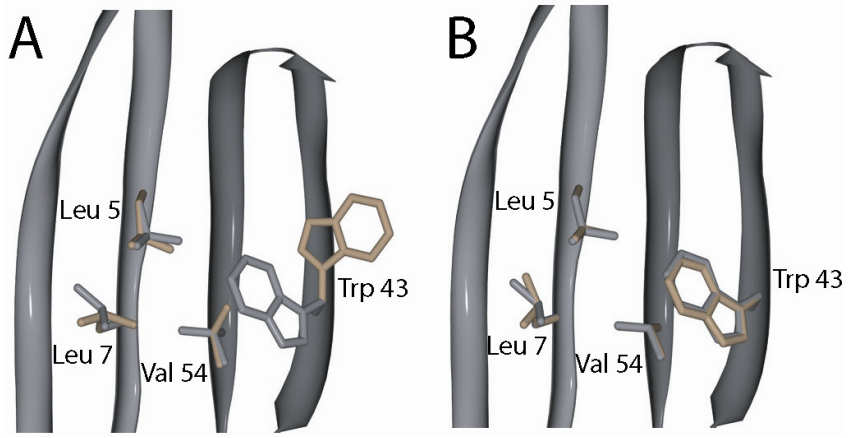


Fig. 2-4: Denaturation Curves of the Wild Type and a Mutant Obtained With the Standard Parameters

The sequence predicted at a sequence bias of 5.0 kcal/mol/pos using the standard parameters is shown. Thermal denaturation (*A*) and chemical detautration (*B*) experiments clearly depict the decrease in stability of sbias5.0 relative to the wild-type protein.

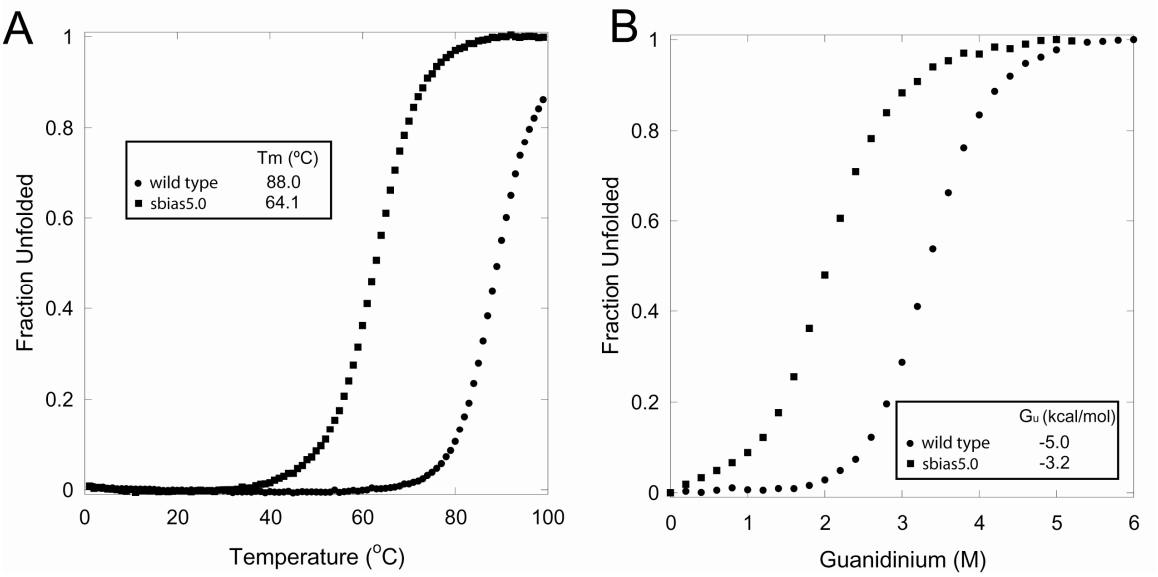


Fig. 2-5: Energies and Z-Scores for Wild Type and Unbiased Sequences Predicted in Gβ1 Fixed Composition Designs

(*A*) Designs used initial force field parameters. (*B*) Designs used improved force field parameters. The energy distributions were obtained by evaluating the energy of 1000 random sequences with the wild type's amino acid composition. The energy scores omit contributions from the van der Waals potential function.

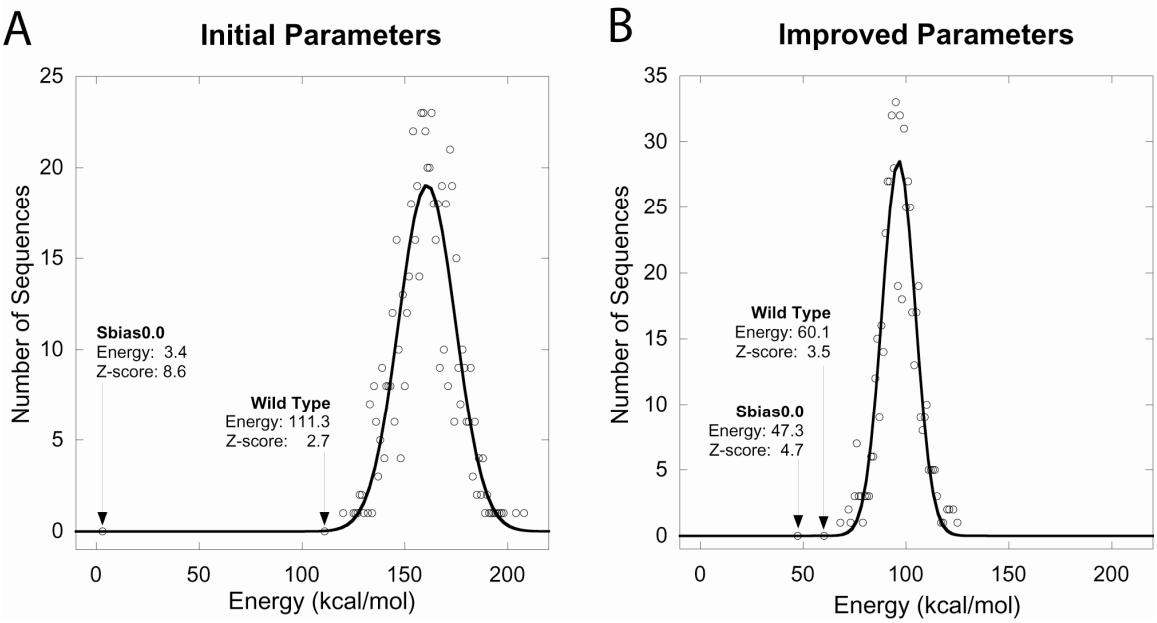


Fig. 2-6: Temperature Denaturation of Gβ1 Mutants Obtained With the Improved Parameters

The data clearly show that mutants predicted at sequence biases of 0.5 or higher have thermal stabilities comparable to wild type. Sbias0.0 has 53% sequence identity with the wild-type sequence, is folded and exhibits a T_m that is only 10ºC lower than wild type.

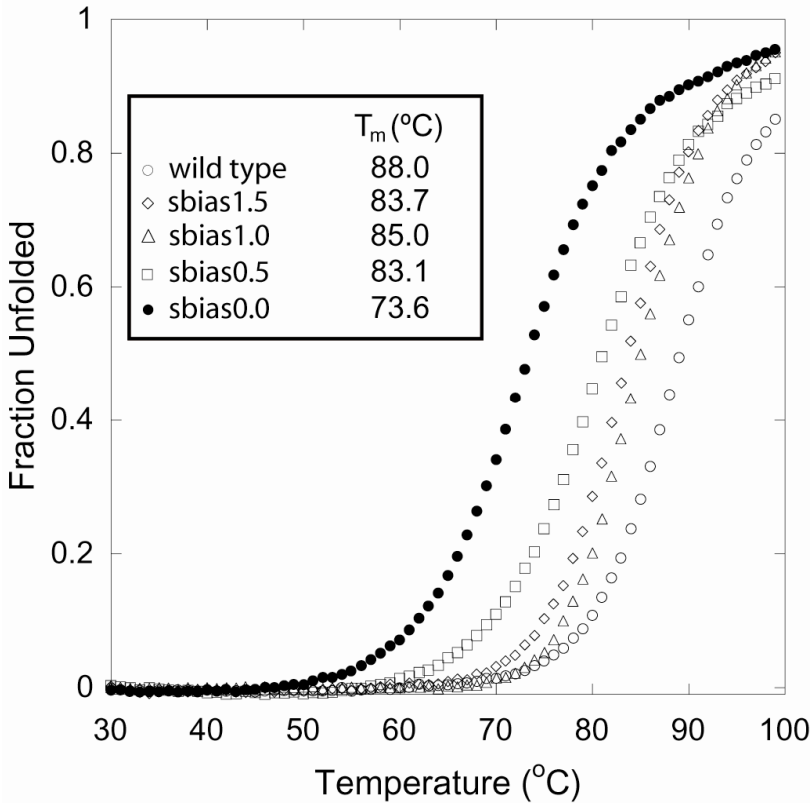


Fig. 2-7: Predicted Sequences for ENH Fixed Composition Designs

Designed core, boundary, and surface positions are shown in red, green and blue, respectively. The wild-type sequence is shown followed by the sequences predicted at increasing sequence bias values (kcal/mol/position). Dots represent wild-type amino acids. (*A*) Sequences obtained using the initial parameters. (*B*) Sequences obtained using the improved parameters.

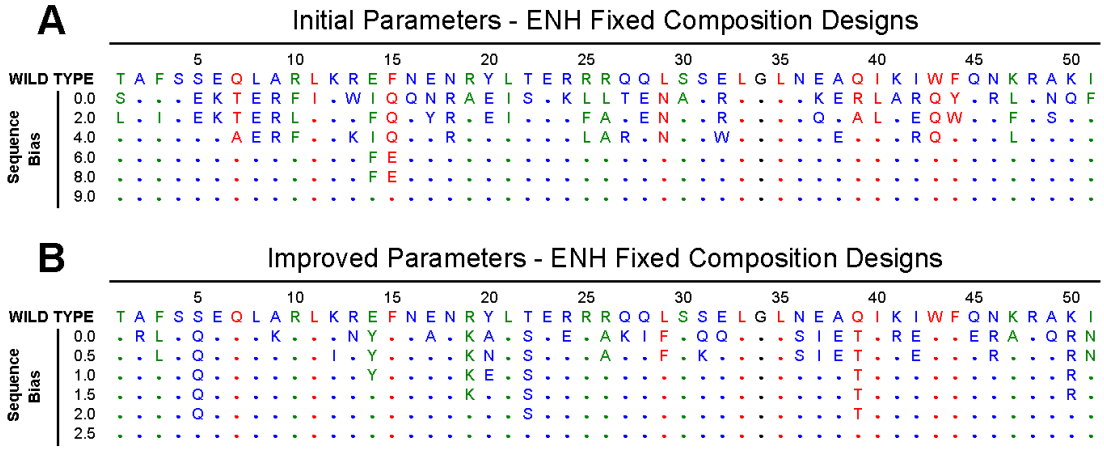
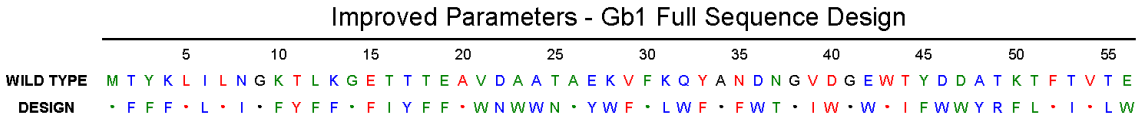


Fig. 2-8: Full Sequence Design of Gβ1 With the Improved Parameters

The figure shows the resulting sequence from a design that omits the fixed composition constraint. Designed core, boundary, and surface positions are shown in red, green, and blue, respectively. Dashes in the predicted sequences represent wild-type amino acids. The majority of positions are selected to be Phe or Trp.



Chapter 3

Step by Step Force Field Optimization

3.1 Introduction

This chapter is in many ways a continuation of Chapter 2, since it deals with the same topic of force field optimization using fixed composition protein design. However, while Chapter 2 addresses the general applicability of fixed composition protein design, this chapter takes a close look at each modification that led to the final improved force field. It includes a detailed discussion of the changes in the force field, the reasoning behind each modification, and the resulting sequences. It is important to show the logical progression of the process that resulted in successful optimization of ORBIT's force field.

The computational protein design cycle is an iterative process of predicting sequences, analyzing the sequences experimentally, and using this information to improve the design [1, 2]. Fixed composition force field optimization is implemented in a similar fashion (Fig. 3-1). Fixed composition sequences are obtained at increasing

wild-type sequence biases. The predicted sequences are analyzed, and if no obvious discrepancies are observed, experimental testing is performed. The information obtained from sequence analysis and experiments is then used to decide on modifications that are most likely to improve the predictive power of the force field. This stepwise approach to force field optimization is expected to yield improvements with each round.

3.2 Results and Discussion

3.2.1 Solvent Exclusion-Based Solvation Model

The standard force field (also referred to as the initial force field) was the first force field tested using fixed composition protein design. The predicted sequences and the experimental results are thoroughly discussed in the previous chapter. In an effort to prevent the selection of polar amino acids at core positions, the surface area-based solvation model was replaced with a solvent exclusion-based one [3]. The change in solvation model had a dramatic effect on the quality of the predicted sequences (Fig. 3-2 *B*). With the exception of one position in one of the sequences, all core positions took on a hydrophobic residue. The unbiased design exhibited 2.5-, 4.0-, and 1.4-fold improvements in wild-type sequence recovery at core, boundary, and surface positions, respectively.

While large improvements were observed at the protein core, little change was seen at the surface. The initial parameters, as well as those using the new solvation model, predicted surface mutations that form hydrogen bonds. Both sets of parameters also predicted the unbiased design to have side-chain-side-chain hydrogen bond energies four times greater than the wild-type sequence. However, the large entropic cost

associated with hydrogen bond formation at the surface makes these bonds energetically unfavorable. As a result, the predicted surface mutations were not expected to increase surface stability.

An example of this is seen with E19D/D36E, which both sets of parameters produced as the highest scoring two-point mutant. This mutant was predicted to form two new hydrogen bonds at the protein surface (Fig. 3-3 *A*). Experimental analysis, however, revealed no increase in stability relative to the wild-type protein (Fig. 3-3 *B, C*), suggesting that the benefit for hydrogen bond formation between side chains at the surface was too high. Consequently, the hydrogen bond potential was modified to eliminate the benefit for inter-rotamer hydrogen bonds at surface positions.

3.2.2 Rotamer Library and Side-Chain-Side-Chain Hydrogen Bonds

The van der Waals and hydrogen bond potentials are both extremely sensitive to small changes in rotamer conformation [4]. The presence or absence of a specific rotamer can mean the difference between a design with a well-packed core and one with cavities in the core. It is beneficial to use the largest possible library to maximize the possibility of identifying the lowest energy conformation for each predicted sequence. However, as the size of the rotamer library increases, the sequence search becomes exponentially more difficult. One must always weigh the advantages of using the largest possible library with the tractability of the optimization. For our next set of designs, we employed a larger rotamer library that included rotamer probabilities [5].

The larger rotamer library and modification to the hydrogen bond potential resulted in an unbiased design sequence with lower wild-type sequence recovery (Table 3-1 *C*). This result was expected, since the design was more stringent, and fixed

composition Monte Carlo (FMONTE) was run for the same number of cycles and steps. As anticipated, surface mutations were no longer driven by the hydrogen bond potential; instead, they were determined by the van der Waals potential and the rotamer polar burial solvation term.

3.2.3 Rotamer Probability Scale Factor

A rotamer library is composed of side-chain conformations in local minima [6]. Each rotamer is a side-chain representation of all the conformations in that specific minimum. During optimization, the current implementation of fixed composition design considers each rotamer equally, and thus does not take the density of conformations in each of the minima into account. Rotamer probabilities can be used, however, to incorporate the likelihood of finding a specific amino acid side chain in a particular conformation into the side-chain selection procedure. A penalty is assigned to each rotamer based on its probability of occurrence; the lower the rotamer probability, the higher the penalty, and the less likely it will be selected. A scaling factor is used to control the magnitude of the penalty.

We investigated the ideal scale factor for rotamer probability penalties by performing side-chain placement designs using Gβ1 as a scaffold. The force field was allowed to select the lowest energy conformation for the wild-type amino acid at each position. The ideal scale factor was expected to yield the conformation with the lowest RMSD compared to the crystal structure. Scale factors of 0.8 and 1.0 provided the lowest overall RMSDs (Table 3-2). Interestingly, the optimal scale factor varied for different regions in the protein. The core exhibited the lowest RMSD in the absence of a rotamer probability term (scale factor of 0.0), while surface and boundary residues required a

scale factor of 0.8 or higher to achieve their lowest RMSDs. It became apparent that, unlike core designs, fixed composition design at the surface and boundary would benefit from the use of rotamer probabilities. Due to the results observed for core positions, however, the optimal scale factor would most likely be below 0.8.

The fact that core residues exhibited the lowest RMSD in the absence of rotamer probability penalties is not surprising. It is not uncommon to have core residues assume less than optimal conformations in order to tightly pack into the core. In such instances, the closest rotamer to the crystal structure would receive a large penalty. Gβ1's core Leu7 for example, takes on a conformation so rare that it is not included in any of the rotamer libraries tested. However, the only way to identify the best scaling factor for the entire protein is to systematically vary the scale factor in fixed composition designs.

Six sets of designs were carried out, each with a different rotamer probability scale factor (RPSF) that varied from 0.0 to 0.5. Surprisingly, the best RPSF for core residues proved to be larger than the best for boundary or surface residues (Table 3-3). Wild-type recovery at boundary and surface positions was highest at an RPSF of 0.1, while an RPSF of 0.2 or 0.3 was required to recover the largest number of core residues. At an RPSF of 0.3, 70% of core positions were recovered, compared to 50% using no RPSF; similarly, recovery for surface residues jumped from 17% to 28%. Unfortunately, recovery at boundary positions decreased from 25% to 17%. Nevertheless, the high recovery at core positions made 0.3 extremely attractive for the RPSF, so this value was chosen for subsequent designs.

One weakness in the sequences obtained with an RPSF of 0.3 was the persistent prediction of a polar residue at core position 52 (Fig. 3-2 *D*). Thr at position 52 was

considered to be more stabilizing than Phe, since it was predicted to form a hydrogen bond with Asn45. However, the mutation from a large hydrophobic amino acid to a much smaller polar residue is drastic and results in the formation of a cavity. Force field energy analysis revealed a score of –6.6 kcal/mol for this hydrogen bond, only 1.4 kcal/mol shy of the maximum allowed. Despite the decrease in van der Waals energy, the mutation was predicted to be stabilizing due to the high score from the hydrogen bond potential. To mitigate this problem, we decreased the hydrogen bond well depth from 8.0 kcal/mol to 4.0 kcal/mol, expecting that this would sway the force field towards predicting mutations with improved van der Waals and solvation scores.

3.2.4 Decrease of Hydrogen Bond Well Depth

The predicted sequences obtained after lowering the hydrogen bond well depth contained no polar residues at the core (Fig. 3-2 *E*). The wild-type amino acid was recovered for 70% of core positions in the unbiased design, and total core recovery was achieved at a sequence bias of 1.0 kcal/mol/pos. The gap in energy between the unbiased design and the wild-type sequence was calculated to be only -20.3 kcal/mol, the smallest thus far. Energy analysis revealed the unbiased design had a van der Waals energy 15.6 kcal/mol less than the wild-type sequence, suggesting that the unbiased sequence was not as well packed. The poor van der Waals energy for the unbiased design was largely compensated for by the rotamer polar burial score, which was 20.1 kcal/mol more favorable than the polar burial score for the wild-type sequence. This observation indicated that the force field would benefit from scaling down the polar burial term in the solvation model.

3.2.5 Polar Burial Scale Factor

Reduction of the polar burial term resulted in a significant improvement in the quality of the predicted sequences (Fig. 3-2 *F*). For the first time, the unbiased design recovered 90% of the wild-type core residues (Table 3-1 *F*). Wild-type sequence recovery at the surface jumped 10%. In addition, there was a significant improvement in the calculated Z-score, which increased from 2.1 to 3.2 [7, 8]. The same two-point mutant was predicted at a sequence bias of 1.0 and 1.5 (Fig. 3-2 *F*). We expected these two mutations would stabilize the protein by improving van der Waals interactions and by forming a new hydrogen bond with the backbone. Experimental analysis of the two-point mutant showed a slight increase in stability relative to the wild-type (Fig. 3-4 *A* and *B*). The sequence predicted at a sequence bias of 0.5 was a 15-fold mutant; we found it to be folded and slightly destabilized, with a T_m only 6ºC lower than wild type (Fig. 3-4 *A* and *B*). Unfortunately, thermodynamic analysis of the unbiased design (sbias0.0) was not possible, since the protein aggregated in solution. Further refinement of the force field was clearly required before a full set of predicted sequences could be shown to be properly folded.

3.2.6 Hydrogen Bonds Between Immediate Neighbors

Structural analysis of the sequences obtained after reducing polar burial revealed minor problems with the predicted hydrogen bonds. Specific positions were consistently mutated to residues predicted to form local side chain-backbone hydrogen bonds. Take for example the T55Q mutation predicted for sbias0.0 and sbias0.5 or the T18D mutation predicted for sbias0.5 (Fig. 3-4 *C*); both mutations form hydrogen bonds with their respective +1 position. Neither mutation is expected to increase specificity for the target

fold due to the local nature of the hydrogen bond. It is more likely that formation of such a hydrogen bond would be destabilizing due to its entropic cost. We anticipated that lowering the benefit for hydrogen bonds between immediate neighbors (+1 and -1 positions) would reduce the frequency of mutations similar to those observed at positions 55 and 18.

This latest tweak to the force field was expected to have a minor effect on the overall predictions. The resulting sequences, however, had significantly worse wild-type sequence recovery (Table 3-1 *G*). In the unbiased design (sbias0.0), wild-type recovery at core positions dropped from 90% to 80%. Most concerning was the fact that sbias0.5 replaced a core Leu at position 7 with an Asn (Fig. 3-2 *G*). The mutation was predicted to alleviate steric strain at the core, since the proper rotamer for Leu7 was absent in the rotamer library used. We therefore repeated the design using a rotamer library that included the crystallographic conformation for Leu7.

3.2.7 Rare Crystallographic Conformations

The sequences predicted with the new rotamer library are shown in Fig. 3-2 *H*. Unlike previous unbiased sequences, all wild-type core positions were recovered, ensuring that the core was well-packed and hydrophobic. The total wild-type sequence recovery increased from 37% to 53% (Table 3-1 *H*). There was a two-fold improvement at boundary positions and a 10% improvement at the surface. The two-point mutant predicted at a sequence bias of 1.0 and 1.5 is identical to that previously tested, which was shown to be as stable as the wild-type protein (Fig. 3-5). CD wavelength scans of the other two sequences in the set (sbias0.0 and sbias0.5) indicated they were well folded and well-behaved. The unbiased protein, which had the lowest wild-type recovery

(53%), was also shown to be folded by NMR. In addition, thermodynamic analysis of the protein revealed a T_m of 79.5ºC, only 10º lower than the wild-type sequence (Fig. 3-5).

3.2.8 Fixed Composition Design on Alternative Scaffolds

The overwhelming success obtained with this latest set of parameters required validation on alternate scaffolds. We chose engrailed homeodomain from *Drosophila melanogaste*r (ENH) and the β1 domain of *Pepetotreptococcus magnus* protein L (Lβ1), both small soluble proteins with distinct characteristics. Like Gβ1, Lβ1 is very stable with both β-sheet and α-helical secondary structure. Its tertiary structure and hydrophobic core are also similar to those of Gβ1. ENH, on the other hand, is mostly α–helical and significantly less stable than either Lβ1 or Gβ1.

Table 3-4 shows the resulting sequence statistics from fixed composition designs on these two scaffolds. The wild-type sequence was recovered at a sequence bias of 1.0 for Lβ1 and 1.5 for ENH, both less than the 2.0 value required for Gβ1. Lβ1's unbiased design predicted a well-packed hydrophobic core with 60% sequence identity with the wild-type core, and ENH's unbiased design predicted a core with 70% sequence identity with the wild-type sequence.

ENH's unbiased design predicted two polar residues in the core. Position 12 took on the wild-type Asn, which serves as a bridge between two helices by hydrogen bonding the backbone. Position 44 replaced a Glu with a Thr. Similar to the wild-type Gln, Thr forms a hydrogen bond with position 1, stabilizing the N-terminus (Fig. 3-6 *B*). Unlike with the designs on Gβ1, the prediction of polar side chains in ENH's core is a positive result, since it recapitulates what is seen in the wild type. It shows that the optimized force field was able to predict stabilizing polar interactions in the core when required.

The unbiased fixed composition designs on Lβ1 and ENH predicted reasonable sequences. In both cases, the interactions in the core were consistent with what is observed in the wild-type structure. These results, in conjunction with recovering the wild-type sequence at a low sbias (1.5 or less), suggested that the optimized force field parameters were not strongly biased towards the Gβ1 scaffold.

3.2.9 Conformer Library

One drawback with the fixed composition designs at this point was the required use of the crystallographic rotamer at position 7 for Gβ1. Ideally, one would like to provide as little information on the wild-type conformation as possible. However, the lack of a rotamer that could satisfy the constraints required for proper packing at core position 7 left few choices. Fortunately, the recent availability of conformer libraries provided a good alternative to rotamer libraries [6, 9]. The different approach used in generating conformer libraries allows for the incorporation of new side-chain conformations.

Fixed composition designs were performed on Gβ1 using the optimized force field parameters, but with a conformer library. The predicted sequences showed slightly less wild-type sequence recovery than those obtained with the rotamer library (Table 3-1 *I*). All the core residues were still recovered in the absence of any sequence bias (Fig. 3-2 *I*). Despite the slightly worse performance of the conformer library, it is preferred since it doesn't rely on information on the Gβ1 crystal structure. We expected additional force field modifications could easily improve sequence recovery and result in sequences comparable to those predicted with the "informed" rotamer library.

3.2.10 Dielectric Constant

The modifications to the hydrogen bond potential had rather far-reaching effects, causing salt bridges to be penalized to the same extent as traditional hydrogen bonds. In order to compensate, we next decreased the electrostatic potential's dielectric constant by one-half, thus increasing the interaction energy of salt bridges by a factor of two. Fixed composition designs using the lower dielectric showed an increase in wild-type sequence recovery at surface and boundary positions (Table 3-1 *J*). The unbiased design had identical statistics to those of the best unbiased design obtained using the rotamer library (Table 3-1 *H*). The experimental results were also comparable to those obtained for the proteins predicted with the rotamer library. A detailed discussion of the sequences and experimental results obtained with these final force field parameters is given in Chapter 2 in the section on improved force field parameters.

3.3 Conclusion

Fixed composition protein design was successful in improving the ORBIT force field. The optimization procedure exploited the iterative process of the protein design cycle, producing two final sets of parameters. Both include the same modifications to the solvation, hydrogen bonding, rotamer probabilities and polar burial terms, but differ in their dielectric constant and side-chain library. Each set of final parameters predicted sequences that are folded and well behaved. The stepwise approach to force field optimization provided insightful information that can be incorporated into future applications of the process.

The protein design cycle typically requires experimental verification at each step. However, experimental validation is time consuming and costly. The stepwise

improvement of the force field revealed that optimization can be streamlined by bypassing the experimental steps and focusing solely on sequence and energy analysis. Success at each step can be measured by calculating the Z-score for the wild-type sequence [8]. Elimination of the experimental step results in a computational optimization procedure that can be easily automated.

This fixed composition optimization procedure should be carried out using multiple scaffolds to avoid bias to a particular structure or fold. The automated procedure starts by selecting a random force field parameter to adjust. Simultaneous fixed composition designs can then be carried out on all the protein scaffolds. The Z-score for the wild-type sequences can be calculated and used to determine if the modification to the force field will be kept. The iterative process can be repeated for multiple rounds until no improvement in Z-score is obtained. The final force field can then be experimentally tested on a scaffold not used in the optimization procedure to verify the accuracy of the predictions. The major drawback with the automated procedure is the computational cost of the designs. The large scale of the optimization procedure will require the use of FC-FASTER on multiple processors to limit the run time [10].

3.4 Materials and Methods

3.4.1 Fixed Composition Scaffolds

Coordinates for the backbone structure of Gβ1, Lβ1, and ENH were obtained from the Protein Data Bank entry 1PGA, 1HZ5, and 1ENH, respectively. Any strain or steric clashes in the structure were removed by performing 50 steps of energy

minimization. Residue classification into core, boundary, and surface groups was performed as described previously [11]. All non-Gly and non-Met positions were included in the design, and within fixed composition restraints, all amino acids found in the wild-type sequence were allowed at all designed positions.

3.4.2 Fixed Composition Force Fields

The initial force field used standard potential functions and parameters including scaled van der Waals, hydrogen bonding, electrostatic, and surface area-based solvation terms, as described previously [1, 11-14]. Expanded versions of Dunbrack and Karplus' 1995 backbone-dependent rotamer library were used [15]. Aromatic residues were expanded 1 SD about their χ_1 and χ_2 values, and hydrophobic residues were expanded 1 SD about their χ_1 values; polar residues were not expanded.

Sequences shown in Fig. 3-2 *B-J* were obtained using a solvent exclusion-based solvation potential [3]; for *B-E*, all published solvation parameters were used; for *F-J*, the polar burial scale factor was decreased to 0.6 [3]. For sequences listed in *C-J*, hydrogen bonds at surface positions received a benefit from the electrostatic potential, but not from the hydrogen bond potential; for *C-H*, a larger rotamer library from Dunbrack and Cohen was also used [5]. The rotamers in the library were expanded in a similar fashion as with the smaller Dunbrack and Karplus library. The force field used to obtain sequences under *D-H* included the following rotamer probability term:

$$E_{rotamer\ probability} = - W \log (\rho)$$

Rotamer probabilities (ρ) were taken directly from the published rotamer library and the scale factor, W, was set to 0.3. The benefit for hydrogen bond formation between side chains was decreased by 50% for core and boundary residues to predict sequences

shown in *E-J*. Sequences under *G-J* required the hydrogen bond energies to be decreased by an additional 75% if predicted between immediate neighbors (n+1 and n–1 positions). The force field that predicted sequences under *I* and *J* used a larger backbone-dependent conformer library [6] instead of the rotamer library. The conformer library was constructed using Cartesian coordinates taken directly from high-resolution crystal structures, as described by Lassila et al. [9]. A *p* value for non-polar amino acids was set to 0.3; a *p* value of 0.6 was used for Asp, Glu, Asn, and Gln; and representative conformers for Arg and Lys were obtained with a *p* of 0.8. For the final set of sequences, listed under *J*, the dielectric constant was decreased from 40 to 20.

3.4.3 Fixed Composition Sequence Optimization

Prior to sequence optimization, an energy matrix containing all one-body and two-body interactions was created. The one-body term for each rotamer was modified to reflect a specified sequence bias energy. Each rotamer that differed in identity from the wild-type amino acid at a particular position received a penalty. The resulting sequence was thus penalized for each residue that differed from the wild-type sequence. All calculations were first carried out in the absence of a sequence bias. The bias energy was then incrementally increased by 1.0 or 0.5 kcal/mol/position, while keeping all other parameters fixed, until the wild-type sequence was recovered.

A stochastic algorithm, Monte Carlo simulated annealing, was used for the fixed composition Gβ1 designs. A fixed composition restraint was imposed by creating a new version of Monte Carlo, FMONTE. The FMONTE algorithm randomly picks four positions and arbitrarily switches the amino acids at two, three, or all four of the positions. A random rotamer is chosen at each of the switched positions, and the

sequence energies are compared. All calculations were carried out for 1,000 annealing cycles at 1,000,000 steps per cycle and the temperature was cycled from 4,000 K to 150 K. Fixed composition designs on ENH were performed using a fixed composition version of the FASTER algorithm [10].

3.4.4 Protein Expression and Purification

Mutant plasmids were created by site-directed mutagenesis of the wild-type gene in pET-11a or ordered from Blue Heron Biotechnology. Electroporation was used to transform the completed plasmids into BL21 (DE3) cells. Cells were allowed to express protein for 3 hr after induction with IPTG, then harvested and lysed by sonication. Cell extracts were spun down and precipitated by addition of 50% acetonitrile. The soluble protein was separated from the precipitate by centrifugation and HPLC purified. Pure proteins were analyzed by either trypsin digest or by collision-induced dissociation mass spectrometry.

3.4.5 Experimental Studies

CD studies were done on an Aviv 62A DS spectropolarimeter with a thermoelectric cell holder. Samples were prepared in 50 mM sodium phosphate buffer at pH 5.5. Guanidinium denaturations were carried out in a 1 cm path length cuvette at a protein concentration of 5 μM (1,800 μl). An autotitrator was used for the chemical denaturations and data was collected at 218 nm and 25ºC. After each injection of denaturant, samples were stirred for 10 min before data was collected (100 sec averaging time). Wavelength scans and temperature denaturations were carried out in cuvettes with a 0.1 cm path length at a concentration of 50 μM (300 μl). Three wavelength scans were

done at 25ºC. Data was collected from 200 nm to 250 nm at 1 nm intervals and averaged for 1 sec. Temperature denaturations were carried out from 0ºC to 99ºC, sampling every 1ºC. Samples were equilibrated for 90 sec before data was collected (averaging time 30 sec). 1D NMR experiments were done on a Varian Unityplus 600-MHz spectrometer at 25ºC. Samples were prepared in 50 mM sodium phosphate buffer pH 5.5 using 9:1 $H_2O/^2H_2O$.

3.5 Bibliography

1. Dahiyat, B. I., and Mayo, S. L. Protein design automation. *Protein Sci* **5**, 895-903 (1996).

2. Street, A. G., Datta, D., Gordon, D. B., and Mayo, S. L. Designing protein beta-sheet surfaces by Z-score optimization. *Phys Rev Lett* **84**, 5010-5013 (2000).

3. Lazaridis, T., and Karplus, M. Effective energy function for proteins in solution. *Proteins* **35**, 133-152 (1999).

4. Mendes, J., Guerois, R., and Serrano, L. Energy estimation in protein design. *Curr Opin Struct Biol* **12**, 441-446 (2002).

5. Dunbrack, R. L., Jr., and Cohen, F. E. Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Sci* **6**, 1661-1681 (1997).

6. Shetty, R. P., De Bakker, P. I., DePristo, M. A., and Blundell, T. L. Advantages of fine-grained side chain conformer libraries. *Protein Eng* **16**, 963-969 (2003).

7. Chiu, T. L., and Goldstein, R. A. Optimizing potentials for the inverse protein folding problem. *Protein Eng* **11**, 749-752 (1998).

8. Gordon, D. B., Marshall, S. A., and Mayo, S. L. Energy functions for protein design. *Curr Opin Struct Biol* **9**, 509-513 (1999).

9. Lassila, J. K., Privett, H. K., Allen, B. D., and Mayo, S. L. Combinatorial methods for small-molecule placement in computational enzyme design. *Proc Natl Acad Sci USA* **103**, 16710-16715 (2006).

10. Hom, G. K., and Mayo, S. L. A search algorithm for fixed-composition protein design. *J Comput Chem* **27**, 375-378 (2006).

11. Dahiyat, B. I., and Mayo, S. L. De novo protein design: fully automated sequence selection. *Science* **278**, 82-87 (1997).

12. Dahiyat, B. I., Gordon, D. B., and Mayo, S. L. Automated design of the surface positions of protein helices. *Protein Sci* **6**, 1333-1337 (1997).

13. Dahiyat, B. I., and Mayo, S. L. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci USA* **94**, 10172-10177 (1997).

14. Dahiyat, B. I., Sarisky, C. A., and Mayo, S. L. De novo protein design: towards fully automated sequence selection. *J Mol Biol* **273**, 789-796 (1997).

15. Dunbrack, R. L., Jr., and Karplus, M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol* **230**, 543-574 (1993).

Table 3-1: Percent Wild-Type Sequence Identity of Gβ1 Predicted Fixed Composition Sequences

		Percent Sequence Identity*			
	Parameters	Total	Core	Boundary	Surface
A	Initial				
	sbias0.0	16	20	8	17
	sbias1.0	35	60	17	34
	sbias2.0	55	70	33	59
	sbias3.0	69	70	33	83
	sbias4.0	84	80	67	93
	sbias5.0	92	90	92	93
	sbias6.0	100	100	100	100
B	Solvent Exclusion-Based Solvation Model				
	sbias0.0	32	50	33	24
	sbias1.0	59	90	42	55
	sbias2.0	82	90	83	79
	sbias3.0	90	90	83	93
	sbias4.0	100	100	100	100
C	Larger Rotamer Library and No Surface Side-Chain Side-Chain HB				
	sbias0.0	25	50	25	17
	sbias1.0	82	100	50	90
	sbias2.0	100	100	100	100
D	Rotamer Probability Scale Factor Set to 0.3				
	sbias0.0	33	70	17	28
	sbias0.5	49	70	25	52
	sbias1.0	75	90	58	76
	sbias2.0	100	100	100	100
E	Decreased Hydrogen Bond Well Depth to 4.0				
	sbias0.0	31	70	17	24
	sbias0.5	63	90	42	62
	sbias1.0	84	100	67	86
	sbias1.5	96	100	92	97
	sbias2.0	100	100	100	100
F	Set Polar Burial to 0.6				
	sbias0.0	41	90	17	34
	sbias0.5	71	90	50	72
	sbias1.0	96	100	92	97
	sbias1.5	96	100	92	97
	sbias2.0	100	100	100	100
G	Reduced Benefit for H-Bond Between Immediate Neighbors				
	sbias0.0	37	80	25	28
	sbias0.5	76	90	58	70
	sbias1.0	96	100	92	97
	sbias1.5	96	100	92	97
	sbias2.0	100	100	100	100
H	Included Position 7's Crystallographic Rotamer				
	sbias0.0	53	100	50	38
	sbias0.5	82	100	58	86
	sbias1.0	96	100	92	97
	sbias1.5	96	100	92	97
	sbias2.0	100	100	100	100
I	Used a Conformer Library				
	sbias0.0	49	100	42	34
	sbias0.5	71	100	50	69
	sbias1.0	92	100	83	93
	sbias1.5	96	100	92	97
	sbias2.0	100	100	100	100
J	Dielectric Constant Decreased to 20				
	sbias0.0	53	100	50	38
	sbias0.5	76	100	50	79
	sbias1.0	92	100	83	93
	sbias1.5	96	100	92	97
	sbias2.0	100	100	100	100

* Wild-type sequence identity was determined using the 51 Gβ1 designed positions. Values are rounded to the nearest integer.

Table 3-2: RMSDs Following Side-Chain Placement of Wild-Type Sequence on Gβ1 Scaffold Using Different Rotamer Probability Scale Factors

Rotamer Probability Scale Factor	RMSD (Å)*			
	Core	Boundary	Surface	Total†
0.0	**0.31**	1.23	2.10	1.69
0.1	0.50	0.95	1.97	1.54
0.2	0.50	0.95	1.90	1.49
0.4	0.50	0.95	1.94	1.52
0.6	0.50	0.95	1.79	1.42
0.8	0.48	**0.89**	**1.74**	**1.38**
1.0	0.48	0.89	1.74	1.38

* In each column, first appearance of lowest RMSD is in bold.

† RMSD for entire protein.

Table 3-3: Percent Wild-Type Sequence Recovery for Unbiased Fixed Composition Designs Using Different Rotamer Probability Scale Factors

Rotamer Probability Scale Factor	Percent Sequence Identity*			
	Core	Boundary	Surface	Total†
0.0	50	25	17	28
0.1	40	25	35	33
0.2	70	17	28	33
0.3	70	17	28	33
0.4	50	17	28	29
0.5	40	17	24	26

*Wild-type sequence identity was determined using the 51 positions in the Gβ1 unbaised design. Values are rounded off to the nearest integer.

Table 3-4: Percent Wild-Type Sequence Identity of Lβ1 and ENH Predicted Fixed Composition Sequences

	Percent Sequence Identity*			
Parameters	Total	Core	Boundary	Surface
Lβ1				
sbias0.0	31	60	39	20
sbias0.5	74	100	62	71
sbias1.0	100	100	100	100
ENH				
sbias0.0	32	70	27	21
sbias0.5	72	100	64	66
sbias1.0	96	100	100	93
sbias1.5	100	100	100	100

* Wild-type sequence identity was determined using only the designed positions. Values are rounded to the nearest integer.

Fig. 3-1: Force Field Optimization Flow Chart

Force field optimization was carried out in an iterative procedure that included fixed composition designs followed by sequence analysis and experimental verification of sequence predictions. A hypothesis was formed after sequence and experimental analysis and used to make adjustments to the force field. A correct hypothesis would lead to improved fixed composition design sequences. In cases where sequence analysis revealed clear violations in protein stability, the experimental step was skipped.

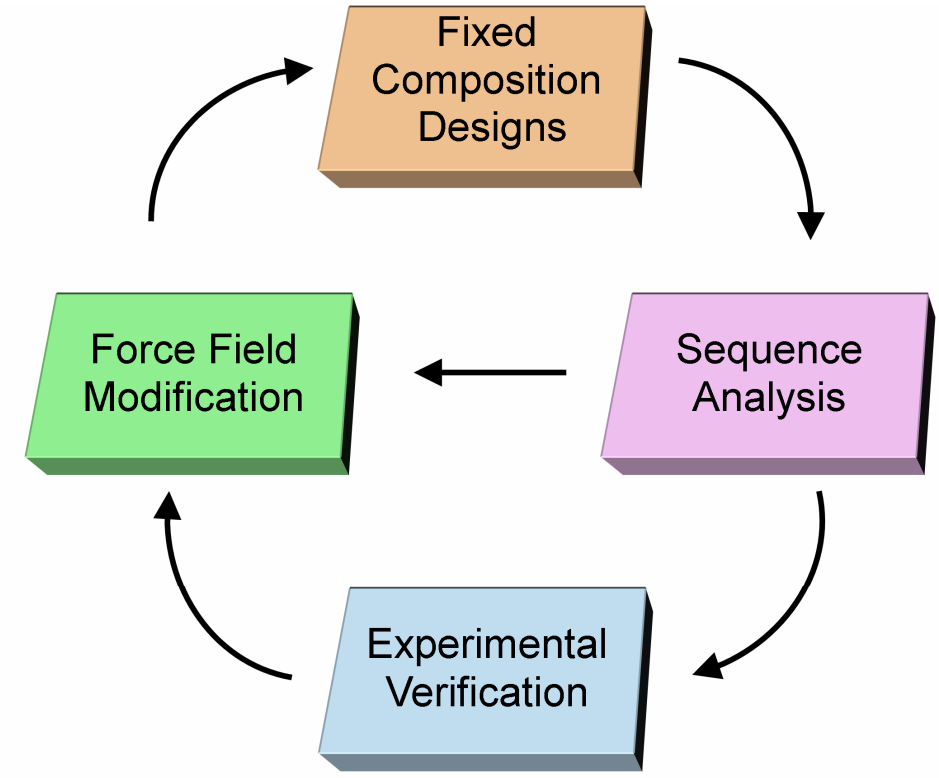
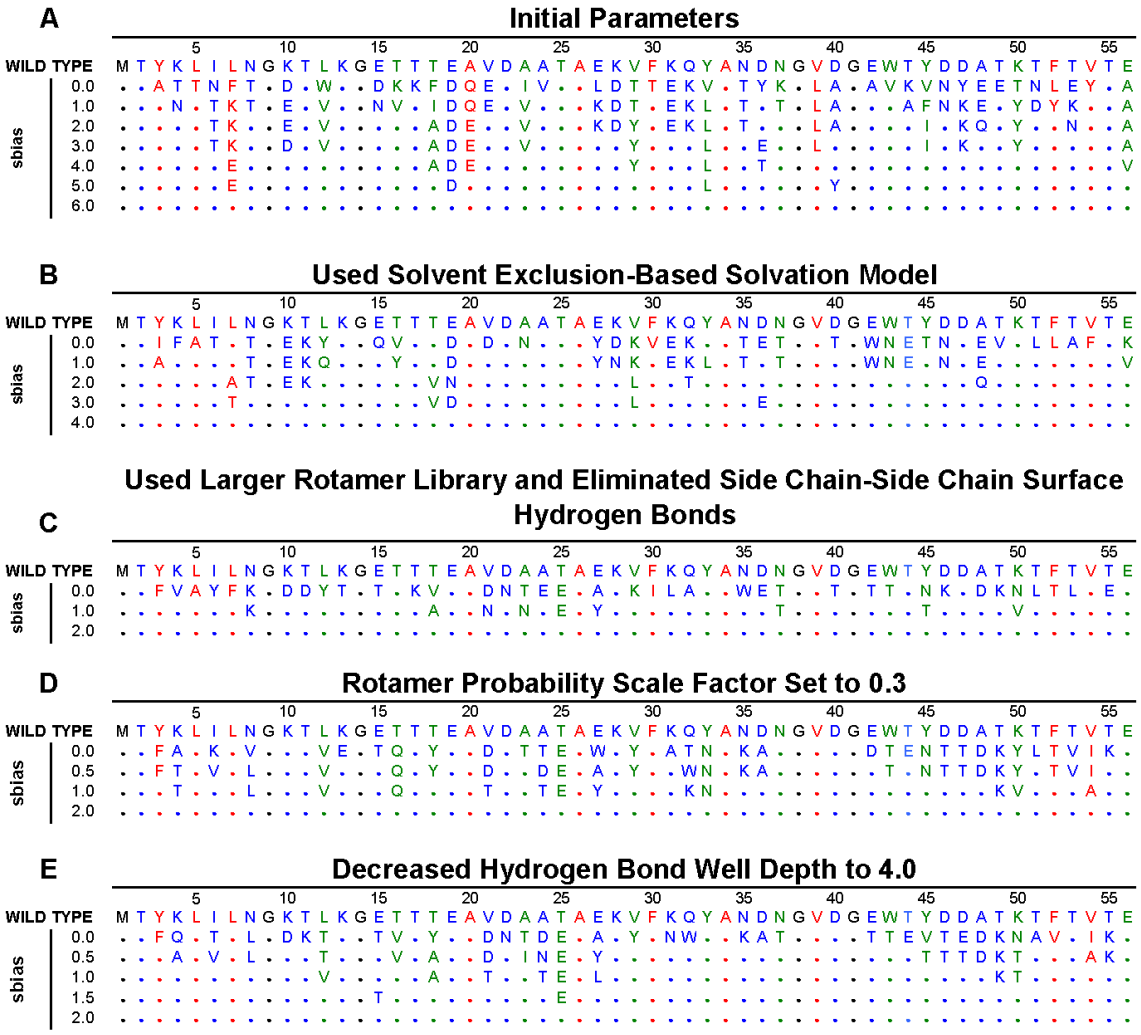


Fig. 3-2: Fixed Composition Predicted Sequences

Predicted sequences for Gβ1 fixed composition designs. Designed core, boundary, and surface positions are shown in red, green, and blue, respectively. The wild-type sequence is shown followed by the sequences predicted at increasing sequence bias values (kcal/mol/position). Dots represent wild-type amino acids. The title for each set of sequences states the change made to the force field; all other parameters are identical to those in the preceding set of sequences.



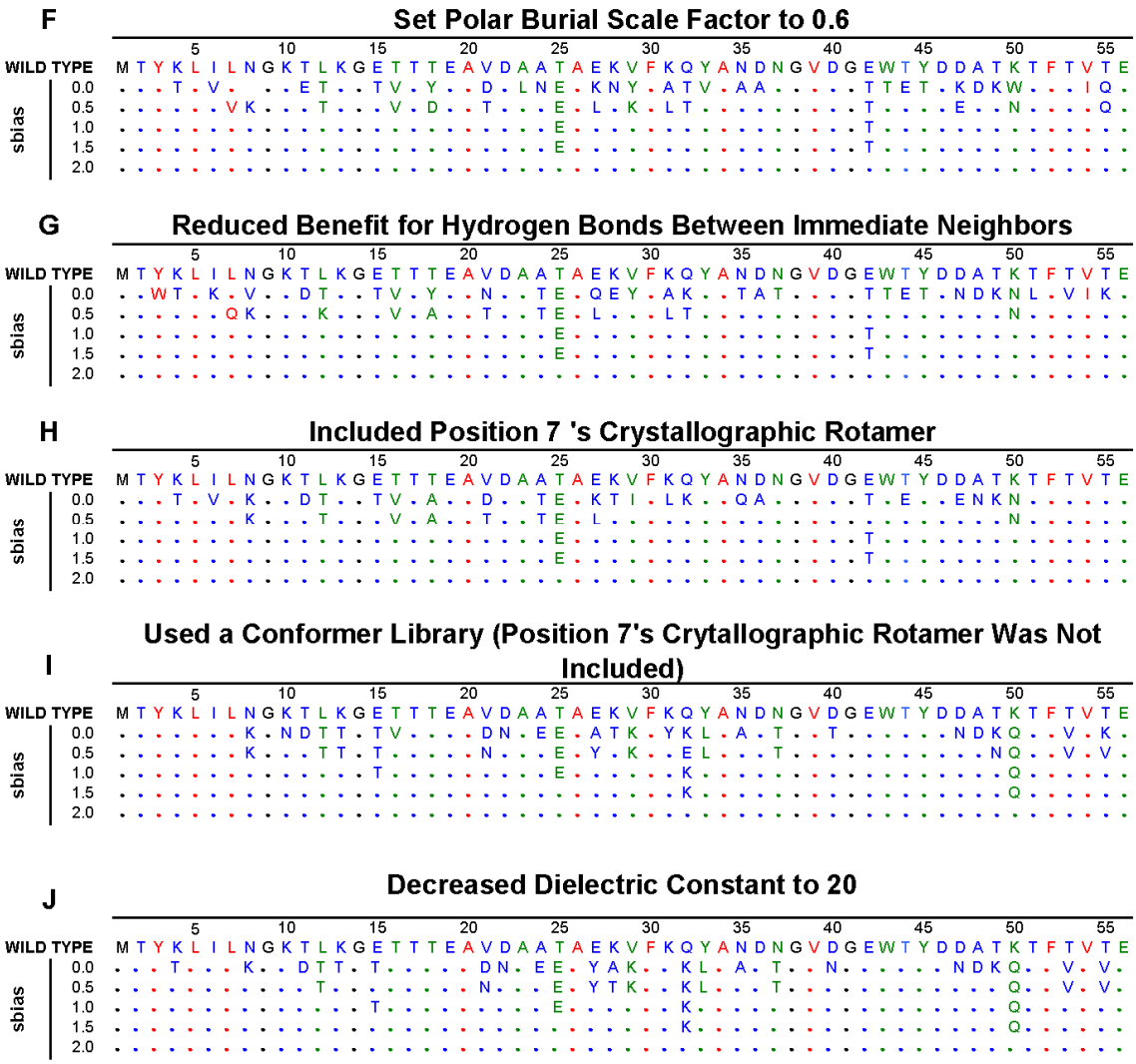

F
Set Polar Burial Scale Factor to 0.6
G
Reduced Benefit for Hydrogen Bonds Between Immediate Neighbors
H
Included Position 7 's Crystallographic Rotamer
I
Used a Conformer Library (Position 7's Crytallographic Rotamer Was Not Included)
J
Decreased Dielectric Constant to 20
WILD TYPE
sbias
0.0
0.5
1.0
1.5
2.0
5
10
15
20
25
30
35
40
45
50
55
M T Y K L I L N G K T L K G E T T T E A V D A A T A E K V F K Q Y A N D N G V D G E W T Y D D A T K T F T V T E

Fig. 3-3: Thermodynamic Data on E19D/D36E Mutant

(*A*) The two-point mutant, E19D/D36E, is predicted to form the two hydrogen bonds shown. (*B*) Temperature denaturations of the two-point mutant and the wild-type protein show overlapping data points. The T_m for both molecules was calculated to be 88°C. (*C*) Chemical denaturation with guanidinium hydrochloride shows no significant difference between the free energy of folding of the two-point mutant and the wild-type protein. A ΔG of 5.07 kcal/mol and 5.23 kcal/mol was calculated for the two-point mutant and the wild-type protein, respectively.

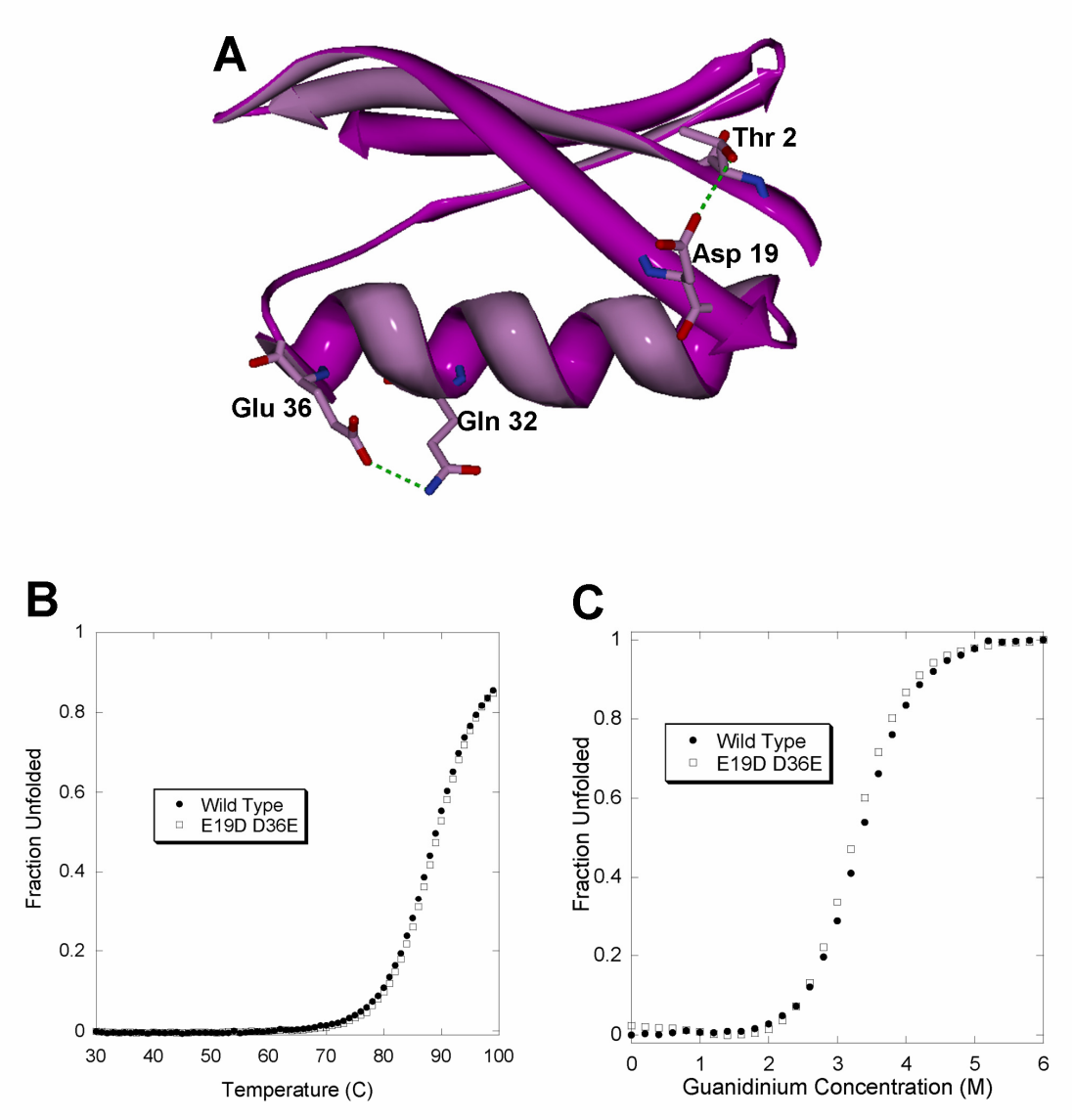


Fig. 3-4: Thermodynamic Data for Sequences Obtained After Reducing Polar Burial Benefit

Sequences of mutants are given in Fig. 3-2 *F*. Thermodynamic data is shown for mutants predicted after reducing the rotamer polar burial scale factor to 0.6. (*A*) T_ms of 88, 89, and 81.8˚C were calculated from the temperature denaturation of the wild-type, sbias1.0 and sbias0.5 proteins, respectively. (*B*) Free energy of folding (ΔG) was calculated to be 5.2, 5.3, and 3.3 kcal/mol for the wild-type, sbias1.0, and sbias0.5 proteins, respectively. (*C*) Residue 55 and 18 are shown hydrogen bonding with their respective +1 position in the predicted conformation for sbias0.5.

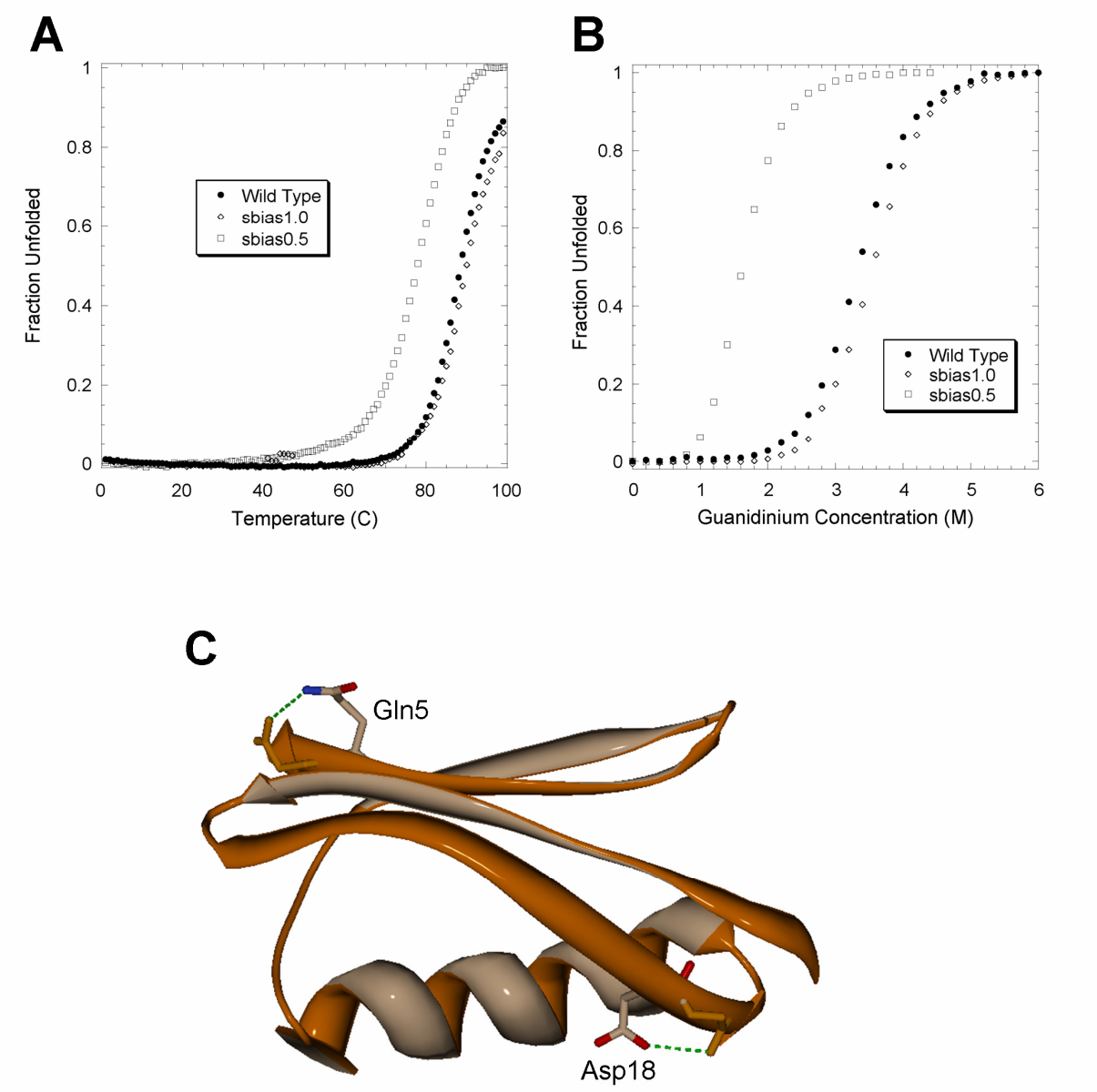


Fig. 3-5: Thermodynamic Data for Sequence Obtained After Including Crystallographic Rotamer at Position 7

Sequences of mutants are given in Fig. 3-2 *G*. Thermodynamic data is shown for mutants predicted after addition of the crystallographic rotamer at position 7.

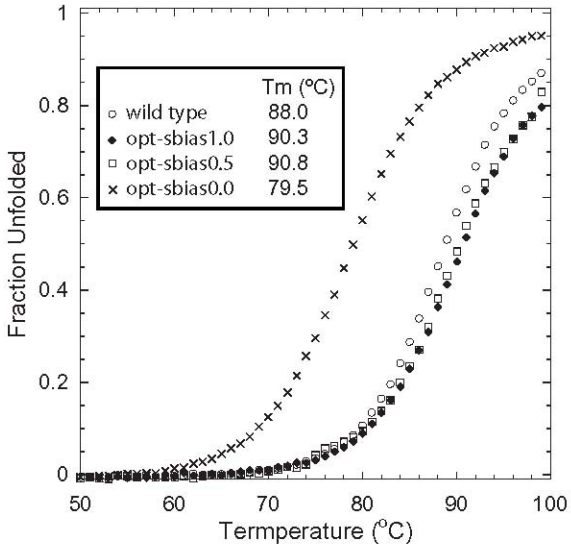
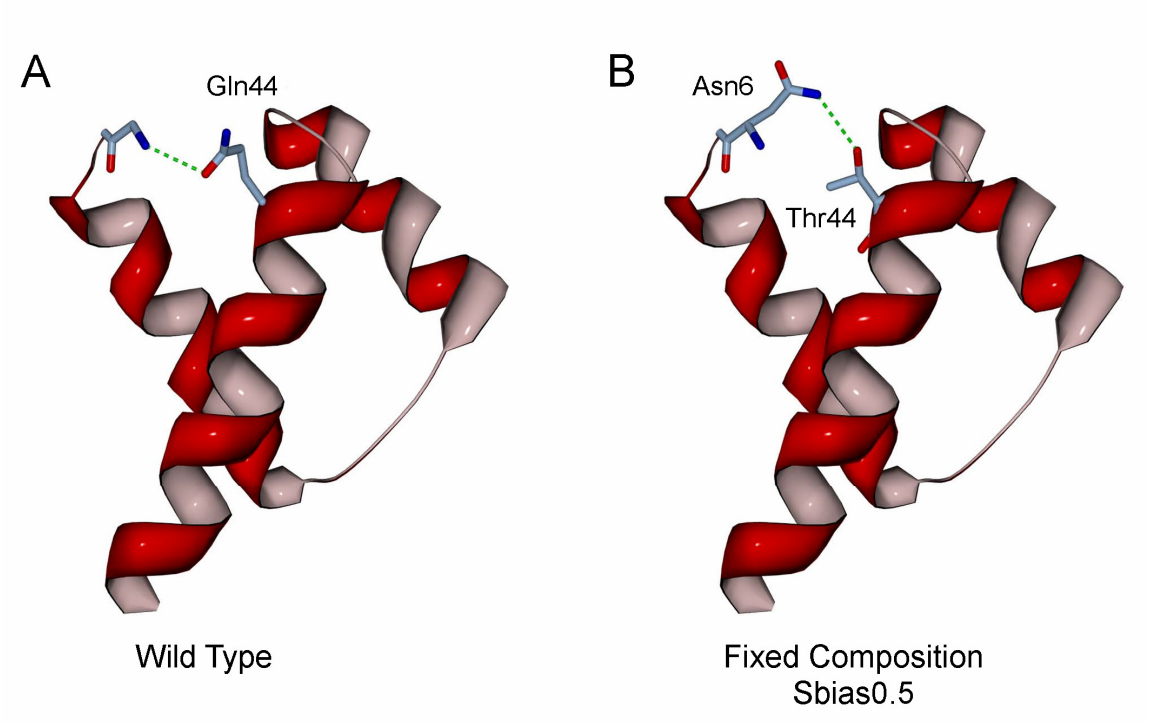


Fig. 3-6: Conformational Comparison of Position 44 in the Engrailed Scaffold

(*A*) In the wild-type sequence, Gln44 is shown hydrogen bonding to the backbone at position 6 (first residue in the crystal structure). (*B*) In the unbiased fixed composition design, Thr44 forms a similar hydrogen bond as in the wild-type sequence with position 6. Instead of hydrogen bonding to the backbone, Thr44 hydrogen bonds to the side chain of Asn6.



Chapter 4

Collision-Induced Dissociation of Gβ1

4.1 Mass Spectrometry Background

Mass spectrometry (MS) can be broadly defined as the use of magnetic or electric fields to isolate and identify molecules with distinct mass to charge ratios (m/z). Some of the first work on this topic was carried out in the late 19th century by Sir Joseph John Thomson, who is credited for inventing the first mass spectrometer. Since then, multiple Nobel Prizes have been awarded to scientists who developed techniques that have advanced the field of mass spectrometry. The latest was awarded in 2002 to John Fenn and Koichi Tanaka for developing methods that allow mass analysis of biological macromolecules.

Currently, several instruments can be used to identify the m/z of a sample (linear quadruple ion trap, orbit trap, quadruple, quadruple ion-trap, sector, and time-of-flight), each with its specific strengths and weaknesses [1, 2]. Time-of-flight (TOF) is a commonly used method for ion detection and is the most conceptually straightforward.

TOF relies on the fact that the potential energy (E_p) of a charged particle can be equated to its kinetic energy (E_k) when it is accelerated using an electric field (Eq. 4-3). The resulting equation reveals that m/z is dependent on the distance of the chamber (d), the electric field charge strength (V), the elementary charge constant (e = 1.602 x 10^{-19}C) and the time the ion took to reach the detector (t) (Eq. 4-4). The larger the molecule, the smaller the velocity, and hence the longer the time required to reach the detector.

$$E_p = zeV \quad \text{(Eq. 4-1)}$$

$$E_k = \frac{1}{2}m\left(\frac{d}{t}\right)^2 \quad \text{(Eq. 4-2)}$$

$$\frac{1}{2}m\left(\frac{d}{t}\right)^2 = zeV \quad \text{(Eq. 4-3)}$$

$$\frac{m}{z} = 2eV\left(\frac{t}{d}\right)^2 \quad \text{(Eq. 4-4)}$$

One of the benefits of TOF is its compatibility with matrix-assisted laser desorption/ionization (MALDI) [3-5]. MALDI is a gentle method used to ionize and analyze large biomolecules that are easily fragmented, such as proteins and polysaccharides. The matrix is composed of a small acid that co-crystallizes with the biomolecule and is able to protect the delicate analyte during desorption by a nitrogen laser beam. A good alternative to MALDI is electro-spray ionization (ESI), which uses a charge capillary to vaporize the sample [6]. All the solvent in the sample is evaporated off before redirecting the charged biomolecules towards the mass analyzing chamber.

ESI is extremely useful when coupled with a high-performance liquid chromatography (HPLC) system. LC-MS (liquid chromatography–mass spectrometry) is often used when analyzing samples containing multiple species. Before analyzing the

mass distribution of the sample, the components are separated using HPLC, and MS is used to analyze specific fractions. The pharmacokinetics field routinely uses this technique when identifying trace amounts of doping agents in an individual's blood. In the last decade, the proteomics field has also greatly benefited from the use of LC-MS in conjunction with tandem mass spectrometry (MS/MS) [7, 8].

Proteomics is a quickly growing field that focuses on the large-scale study of protein expression, structure, and function. The field is largely driven by the fact that genomics, the study of genes and gene expression, drastically underestimates protein diversity in a living cell. In order to obtain information on protein expression levels, scientists are developing high-throughput methods to identify the presence or absence of specific proteins in cell extracts. Innovations in mass spectrometry technology are leading the way by providing highly accurate, reproducible and rapid data on a cell's proteome [8, 9]. The first step in the process is always enzymatic digest of a cell extract to yield small peptide fragments that can be easily analyzed by LC-MS and quickly followed by tandem MS/MS (Fig. 4-1).

MS/MS is the process of determining the mass of a sample before and after collision-induced dissociation (CID) [10]. The initial mass determination is used to identify all the species in solution and to isolate ion species with the target mass. The isolated ions are re-directed to a chamber where they are forced to collide with an inert gas, usually nitrogen or helium. The high-energy collision results in fragmentation of the peptides, which are used for a second mass determination. The results yield a distribution of masses that can serve as a fingerprint to identify specific proteins in the original sample [11, 12].

Two types of software are used to identify protein fragments. The first type is referred to as the top-down method and requires the use of a database of known protein fragments to identify the protein in question [11, 13]. Success using this method is highly dependent on the size of the database and the quality of the data. Low-expressing proteins are less likely to be characterized and are thus unlikely to yield a match. In such cases, the second method, referred to as the bottom-up approach, can be of great use. This method attempts to piece together an amino acid sequence from the data provided [14]. The bottom-up approach uses known CID fragmentation trends to identify sequences de novo [15, 16].

The work presented in this chapter attempts to use CID to analyze Gβ1 variants with identical amino acid composition and tertiary structure. Since all the variant sequences have identical amino acid composition, they all have identical mass and are indistinguishable by standard MS. CID in combination with analysis using a bottom-up approach was expected to successfully distinguish between them. The resulting spectra for each of the sequences were analyzed and were shown to be consistent with previously observed CID trends. The observations made in this study helped us identify a straightforward method of distinguishing between Gβ1 variants with identical amino acid composition.

4.2 Results and Discussion

The work presented in Chapters 2 and 3 relied on the use of sequences with fixed amino acid composition. Given that all the fixed composition sequences have identical mass, alternate methods to standard MS were required for sequence identification. The shotgun method used in proteome screening provides insight into potential alternatives

(Fig. 4-1) [8]. In our case, the degree of complexity was significantly less, since we were dealing with a single protein sequence instead of an entire proteome. As a result, only partial implementation of the procedure was sufficient to specifically identify each protein.

The procedure laid out in Fig. 4-1 requires the fragmentation of the sample at two distinct stages. The first is obtained from enzymatic cleavage by a protease and the second by CID. Each fragmentation step can be used independently for sequence specific analysis of fixed composition proteins (Fig. 4-2). Although each intact fixed composition sequence has identical mass, protein fragmentation results in smaller peptides with distinct amino acid compositions and masses that can be traced back to the original protein sequence.

Trypsin is a reliable protease commonly used for fragmentation since it is highly specific for the positively charged amino acids, Arg and Lys. The Gβ1 protein contains no Arg and seven Lys, making it ideal for trypsin fragmentation (Fig. 4-3). Three fixed composition sequences were analyzed using the trypsin digest method. In all cases, the resulting MS of the fragmentation reactions were convoluted with side products that were unidentifiable. LC-MS was run on the digested samples and clearly shows that the reactions yield a number of fragments; most were not consistent with the expected masses of peptides resulting from the C-terminal cleavage of a lysine (Fig. 4-4).

Rot-sbias0.0 and rot-sbias0.5 (sequences pertain to those in Fig 3-2 *G* and were obtained with the use of a rotamer library and a sequence bias of 0.0 and 0.5, respectively) were the only two sequences that produced unique peptides that were successfully traced back to their original sequences. The wild-type sequence and rot-

sbias1.0 were too similar to be distinguishable using this method. The procedure required cleavage after Lys28 or Lys31 and successful identification of either the C-terminal or N-terminal fragment to distinguish these two sequences; however, none of the required peptides were observed. Given the poor performance of trypsin fragmentation, CID of fixed composition sequences was attempted.

CID was carried out on the five fixed composition sequences predicted by the final force field discussed in Chapters 2 and 3 (Fig. 2-2). Similar to the previous set of sequences, all of these were shown to be folded and well behaved. The wild-type sequence identity for the sequences ranged from 53% to 96%. Initial MS of all the samples showed a nice distribution of charged states confirming the purity and ionizability of the proteins. The +4 charge state was selected for CID since it exhibited the largest signal to noise and had an m/z of 1550.6, approximately halfway between 1000 and 2000, the limit of the instrument.

CID of the selected fixed composition sequences resulted in a clean distribution of fragments; most of the predominant species were successfully assigned (Fig. 4-5). All of the identified fragments were produced from backbone dissociation at the amide bond. From the resulting products, the vast majority were assigned to the N-terminus peptide (the b fragments from Fig. 4-6). Due to the large sequence variation between the proteins, a random fragmentation pattern was expected. However, close analysis of the resulting spectra suggested that dissociation along the protein backbone occurred in identical locations.

Table 4-1 lists some of the observed ions for the five proteins analyzed; four of the five peptides produced seven identical fragments (b_{22}^{+2}, b_{36}^{+3}, b_{40}^{+3}, b_{46}^{+3}, b_{47}^{+3}, b_{54}^{+4},

and b_{55}^{+4}). Con-sbias0.0 (sequence pertains to that in Fig 3-2 *J* and was obtained with the use of a conformer library and a sequence bias of 0.0) produced only four of the seven fragments common to the other sequences; the b_{22}^{+2}, b_{40}^{+3}, and b_{47}^{+3} ions were not observed. The absence of three out of the seven fragments could result from either low detection or unsuccessful dissociation at the required loci.

Fig. 4-7 *A* highlights fragmentation sites along the Gβ1 scaffold. Dissociation of the proteins seemed to cluster around surface exposed turns and loops. As a result, it was initially hypothesized that conserved backbone dissociation was dependent on the tertiary structure of the proteins [17]. To establish a solid connection, CD wavelength scans were carried out under conditions similar to those used for mass analysis of the two sequences with the lowest thermostability, con-sbias0.0 and con-sbias0.5. The wavelength scans overlapped nicely with the wild-type sequence and confirmed the presence of folded protein (Fig. 4-8). In addition, two other fixed composition sequences obtained at low sequence bias, rot-sbias0.0 and rot-sbias0.5, were also shown to be folded under these stringent conditions (Fig. 4-8).

Sequence comparison between the wild type, rot-sbias0.0, rot-sbias0.5, con-sbias0.0, and con-sbias0.5 showed sequences with 57.1–83.9% sequence similarity (Table 4-2). If CID fragmentation patterns were dependent on tertiary structure, CID on rot-sbias0.0 and rot-sbias0.5 would be expected to exhibit similar spectra to those observed for con-sbias0.0, con-sbias0.5, and the wild-type. Unfortunately, only three out of the seven fragments were conserved throughout the five sequences, suggesting that dissociation is independent of tertiary structure (Table 4-3).

However, closer inspection of the wild-type sequence revealed that five out of the seven cleavages occurred after an Asp (Fig. 4-3). In fact, variant sequences that did not contain an Asp immediately preceding a cleavage site were not fragmented at the expected sites. The b_{40}^{+3} fragment, for example, is observed in all the sequences except in con-sbias0.0, since it contains Asn instead of Asp at position 40.

Previously published work reported a similar phenomenon in Arg-containing sequences [18, 19]. Gβ1, however, contains no Arg but does have five Lys, another basic amino acid. It is believed that successful backbone cleavage requires protonation of the amide nitrogen. In the presence of excess protons, cleavage occurs randomly along the backbone [20]. However, when the number of protons is equal to or less than the number of Arg (or in our case, Lys) in the sequence, there are no free protons to bind to the amide nitrogen. As a result, cleavage occurs specifically after Asp residues.

There are a few proposed mechanisms that attempt to explain the sequence of events [21]. All of them require a proton transfer from the aspartate to the amide nitrogen, forming two charged species (negatively charged aspartic acid and positively charged backbone amide). The main difference in the mechanisms is in the role of Arg. In some mechanisms, the Arg is predicted to form a salt bridge with the aspartic acid, while in others it ideally stands by (Fig. 4-9) [22, 23]. After proton transfer, the carbon center of the protonated amide bond is either attacked by the aspartate's carboxyl oxygen or by the N-terminal neighbor amide oxygen, resulting in a cyclic anhydride or an oxazolone, respectively.

Insight into the potential mechanism can be obtained by noting the locations of all the Asp and Lys in the folded structure of Gβ1. A Lys must be physically close to an

Asp to form the salt bridge. However, in some instances, there are Asp residues with no Lys residues in close proximity in the folded structure (Fig. 4-7 *B*). This observation supports the idea that a Lys salt bridge is unnecessary for successful backbone cleavage by an Asp. In order for the previous statement to be true, it must be shown that the sequences are folded during the fragmentation step. CD wavelength scans show the sequences to be folded in solvent conditions used for MS analysis (Fig 4-8), but fail to prove that the proteins are still properly folded after solvent evaporation. A sequence could easily unfold and possibly aggregate as soon as it goes into vacuum, allowing for formation of the salt bridge. However, the low protein concentration required for MS analysis and studies on folded proteins in vacuo support the idea that the proteins retain their tertiary structure prior to fragmentation [24, 25].

Most of the fragmentation patterns of the fixed composition sequences can be explained by the Asp effect. However, fragmentation after position 54 and 55 cannot, since Asp is absent at those positions. Yet, cleavage at the C-terminus is often observed when Glu is the last residue in the sequence [26]. The predicted mechanism involves Glu folding back and carrying out a nucleophilic attack on the carbonyl carbon of an adjacent residue (Fig. 4-10). In the case of Gβ1, the Glu attacks either the -1 or -2 position, resulting in the b_{54}^{+4} or b_{55}^{+4} ions.

4.3 Conclusion

Specific cleavage of Gβ1 fixed composition sequences during CID can be explained by either the Glu or Asp effect. Fragmentation after Asp residues was shown to be successful in sequences free of Arg. The presence of Lys, another positively charged basic amino acid, is believed to serve as a good substitute for Arg. The absence

of a basic residue within salt bridging distance of a specific Asp suggests that basic amino acids are required to be present for the Asp effect, but are not directly involved in the fragmentation of the peptide backbone. The presence of a basic amino acid prevents indiscriminate fragmentation by capturing the acidic protons that would otherwise initiate random backbone cleavage. These observations suggest that the mechanism shown in Fig 4-9 *A* is correct [22].

Segmentation after Asp residues provides a reliable method for specifically identifying fixed composition sequences. CID is a clean and rapid method for sequence fragmentation and identification. Trypsin digest, while successful in identifying some fixed composition sequences, is significantly more time consuming and less reliable. One drawback with both methods is their inability to identify contamination by a second fixed composition sequence. The methods confirm the presence or absence of sequences by successfully matching expected masses with those observed. As a result, the amino acid sequence must be known prior to analysis. In the case of a contaminating sequence, the user may have no prior knowledge of the sequence of the contaminant and will be unsuccessful in identifying its fragments.

In a heterogeneous mixture, the contaminant is expected to be one of the fixed composition sequences being investigated, since an arbitrary fixed composition sequence is highly unlikely to randomly arise. As a result, a database can be easily generated that contains all the fragmentation patterns for every fixed composition sequence in the study. The resulting CID fragmentation pattern can then be cross-referenced with the database to confirm the purity of the sample.

4.4 Materials and Methods

4.4.1 Fixed Composition Sequences

All fixed composition sequences were generated using ORBIT. Details on potential functions and parameters can be found in the Material and Methods section of Chapters 2 and 3. The genes were obtained from BlueHeron and expressed in BL21 (DE3) cells at an OD600 of 1 at 37^{o}C for three hr after IPTG induction. Cells were lysed by sonication and pelleted. An equal volume of acetonitrile was added to the supernatant and centrifuged. The supernatant from the resulting sample was purified by HPLC. Pure samples were flash frozen and lyophilized.

4.4.2 Circular Dichroism

Lyophilized samples were taken up in 8 M guanidinium and refolded in water. Samples were diluted down to a final concentration of 50 uM in 50% ethanol and 0.1% acetic acid. Wavelength scans were carried out on a Aviv 62A DS spectropolarimeter with a thermoelectric cell holder set to 25˚C. Acquired data was averaged for 1 sec every 1 nm from 200 nm to 250 nm.

4.4.3 Mass Spectrometry

Samples were diluted down to a final concentration of 5 μM in 50% ethanol and 0.1% acetic acid. CID was carried out on a ThermoFinnigan LCQ Deca XP quadrupole ion trap mass spectrometer with normalized collision energy of 23. The +4 ion with m/z of 1550.6 was isolated and used for dissociation. Protein prospector was used to identify the observed ions.

4.5 Bibliography

1. Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M., and Graham Cooks, R. The Orbitrap: a new mass spectrometer. *J Mass Spectrom* **40**, 430-443 (2005).

2. McLuckey, S. A., and Wells, J. M. Mass analysis at the advent of the 21st century. *Chem Rev* **101**, 571-606 (2001).

3. Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T., and Matsuo, T. Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. *Rapid Commun Mass Spectrom* **2**, 151 - 153 (1988).

4. Karas, M., and Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal Chem* **60**, 2299-2301 (1988).

5. Kaufmann, R. Matrix-assisted laser-desorption ionization (Maldi) mass-spectrometry - a novel analytical tool in molecular-biology and biotechnology. *J Biotechnol* **41**, 155-175 (1995).

6. Meng, C. K., Mann, M., and Fenn, J. B. Of protons or proteins. *Z Phys D Atom Mol Cl* **10**, 361-368 (1988).

7. Mann, M., Hendrickson, R. C., and Pandey, A. Analysis of proteins and proteomes by mass spectrometry. *Annu Rev Biochem* **70**, 437-473 (2001).

8. Purvine, S., Eppel, J. T., Yi, E. C., and Goodlett, D. R. Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. *Proteomics* **3**, 847-850 (2003).

9. Sadygov, R. G., Cociorva, D., and Yates, J. R., 3rd. Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. *Nat Methods* **1**, 195-202 (2004).

10. Wells, J. M., and McLuckey, S. A. Collision-induced dissociation (CID) of peptides and proteins. *Methods Enzymol* **402**, 148-185 (2005).

11. Pappin, D. J. C., Hojrup, P., and Bleasby, A. J. Rapid Identification of Proteins by Peptide-Mass Fingerprinting. *Curr Bio* **3**, 327-332 (1993).

12. Griffin, P. R., MacCoss, M. J., Eng, J. K., Blevins, R. A., Aaronson, J. S., and Yates, J. R., 3rd. Direct database searching with MALDI-PSD spectra of peptides. *Rapid Commun Mass Spectrom* **9**, 1546-1551 (1995).

13. Eng, J. K., Mccormack, A. L., and Yates, J. R. An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database. *J Am Soc Mass Spectrom* **5**, 976-989 (1994).

14. Steen, H., and Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat Rev Mol Cell Biol* **5**, 699-711 (2004).

15. Ma, B., Zhang, K. Z., Hendrie, C., Liang, C. Z., Li, M., Doherty-Kirby, A., and Lajoie, G. PEAKS: Powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* **17**, 2337-2342 (2003).

16. Johnson, R. S., and Taylor, J. A. Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry. *Mol Biotechnol* **22**, 301-315 (2002).

17. Tsaprailis, G., Nair, H., Somogyi, A., Wysocki, V. H., Zhong, W. Q., Futrell, J. H., Summerfield, S. G., and Gaskell, S. J. Influence of secondary structure on the fragmentation of protonated peptides. *J Am Chem Soc* **121**, 5142-5154 (1999).

18. Newton, K. A., Pitteri, S. J., Laskowski, M., Jr., and McLuckey, S. A. Effects of single amino acid substitution on the collision-induced dissociation of intact protein ions: Turkey ovomucoid third domain. *J Proteome Res* **3**, 1033-1041 (2004).

19. Xia, Y., Liang, X., and McLuckey, S. A. Ion trap versus low-energy beam-type collision-induced dissociation of protonated ubiquitin ions. *Anal Chem* **78**, 1218-1227 (2006).

20. Gu, C., Tsaprailis, G., Breci, L., and Wysocki, V. H. Selective gas-phase cleavage at the peptide bond C-terminal to aspartic acid in fixed-charge derivatives of Asp-containing peptides. *Anal Chem* **72**, 5804-5813 (2000).

21. Paizs, B., and Suhai, S. Fragmentation pathways of protonated peptides. *Mass Spectrom Rev* **24**, 508-548 (2005).

22. Yu, W., Vath, J. E., Huberty, M. C., and Martin, S. A. Identification of the facile gas-phase cleavage of the Asp-Pro and Asp-Xxx peptide bonds in matrix-assisted laser desorption time-of-flight mass spectrometry. *Anal Chem* **65**, 3015-3023 (1993).

23. Paizs, B., and Suhai, S. Towards understanding the tandem mass spectra of protonated oligopeptides. 1: Mechanism of amide bond cleavage. *J Am Soc Mass Spectrom* **15**, 103-113 (2004).

24. Wolynes, P. G. Biomolecular folding in vacuo!!!(?). *Proc Natl Acad Sci USA* **92**, 2426-2427 (1995).

25. Wood, T. D., Chorush, R. A., Wampler, F. M., 3rd, Little, D. P., O'Connor, P. B., and McLafferty, F. W. Gas-phase folding and unfolding of cytochrome C cations. *Proc Natl Acad Sci USA* **92**, 2451-2454 (1995).

26. Li, Z., Yalcin, T., and Cassady, C. J. C-terminal amino acid residue loss for deprotonated peptide ions containing glutamic acid, aspartic acid, or serine residues at the C-terminus. *J Mass Spectrom* **41**, 939-949 (2006).

Table 4-1: Successfully Identified Fragmentation Ions From Fixed Composition Gβ1 Designed Sequences

	b_{22}^{+2}	b_{36}^{+3}	b_{40}^{+3}	b_{46}^{+3}	b_{47}^{+3}	b_{54}^{+4}	b_{55}^{+4}
Wild Type	1205.3	1316.6	1444.9	1695.5	1734.1	1488.1	1513.2
con-sbias0.0	-	1298.7	-	1673.0	-	1488.2	1513.2
con-sbias0.5	1207.1	1322.5	1446.1	1679.0	1735.0	1488.2	1513.2
con-sbias1.0	1190.5	1316.6	1444.7	1695.3	1733.7	1487.8	1513.0
con-sbias1.5	1205.1	1316.4	1444.9	1695.0	1733.6	1487.7	1513.1

Table 4-2: Percent Sequence Similarity of Predicted Fixed Composition Gβ1 Variants

	Wild Type	rot-sbias0.0	rot-sbias0.5	con-sbias0.0	con-sbias0.5
Wild Type	100.0				
rot-sbias0.0	57.1	100.0			
rot-sbias0.5	83.9	69.6	100.0		
con-sibas0.0	57.1	60.7	58.9	100.0	
con-sbias0.5	78.6	57.1	75.0	75.0	100.0

Table 4-3: Successfully Identified Fragmentation Ions From Fixed Composition Gβ1 Designed Sequences Obtained at a Low Sequence Bias

	b_{22}^{+2}	b_{36}^{+3}	b_{40}^{+3}	b_{46}^{+3}	b_{47}^{+3}	b_{54}^{+4}	b_{55}^{+4}
Wild Type	1204.7	1316.6	1444.9	1695.5	1734.1	1488.1	1513.2
con-sbias0.0	-	1298.7	-	1673.0	-	1488.2	1513.2
con-sbias0.5	1207.1	1322.5	1446.1	1967.0	1735.0	1488.2	1513.2
rot-sbias0.0	1170.7	-	1421.3	1671.9	-	1487.7	1513.2
rot-sbias0.5	-	1321.0	1449.9	1699.8	1738.3	1488.1	1513.2

Fig. 4-1: The Shotgun Method: Using LC-MS and CID for High-Throughput Proteome Screening

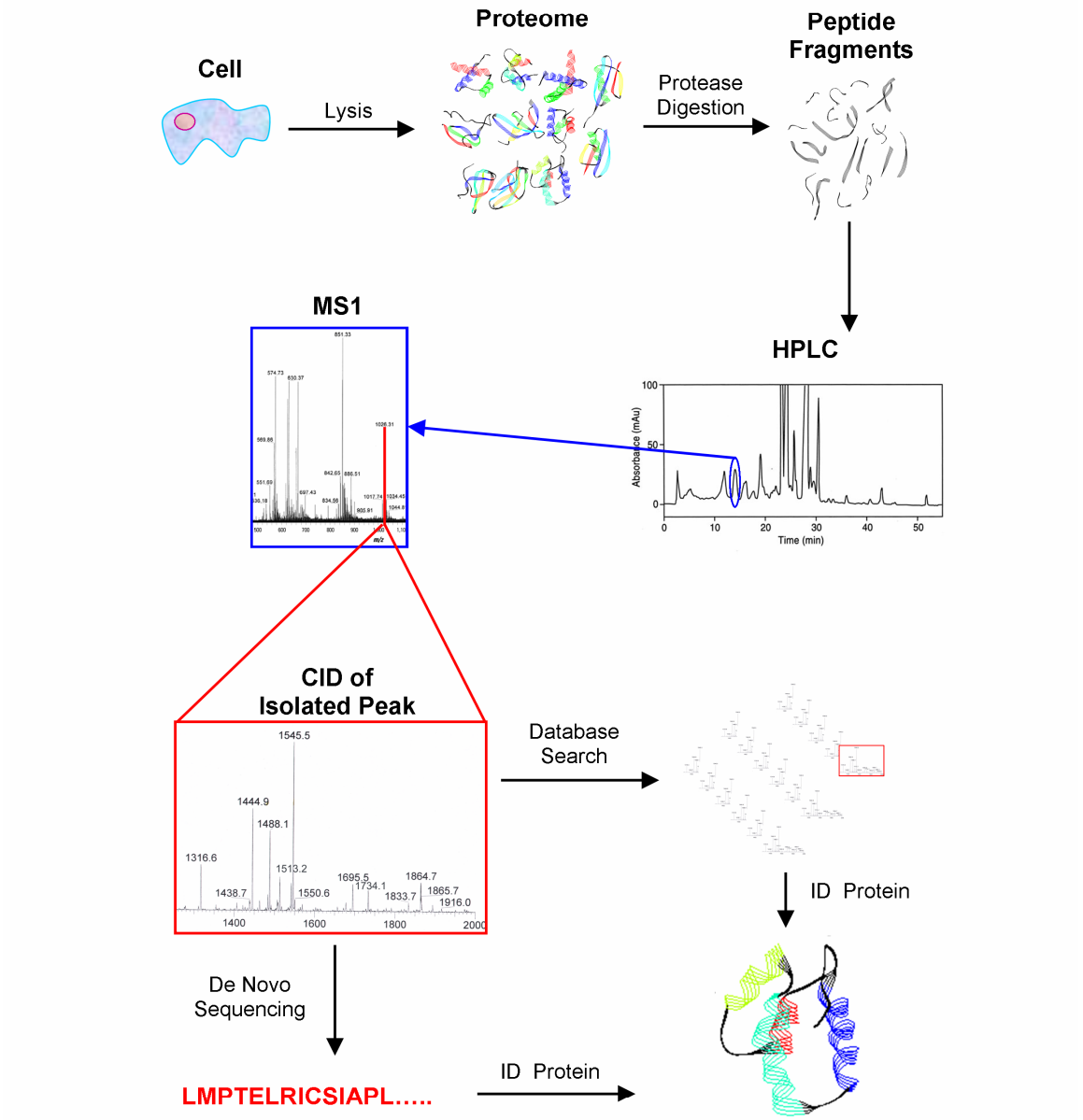


Fig. 4-2: Fragmentation Options for Gβ1

Two fragmentation options are shown, chemical and mechanical. Chemical fragmentation includes trypsin digest, HPLC of fragmented peptides and mass spectrometry of HPLC fractions. Mechanical fragmentation requires isolating an ion of the intact protein followed by CID.

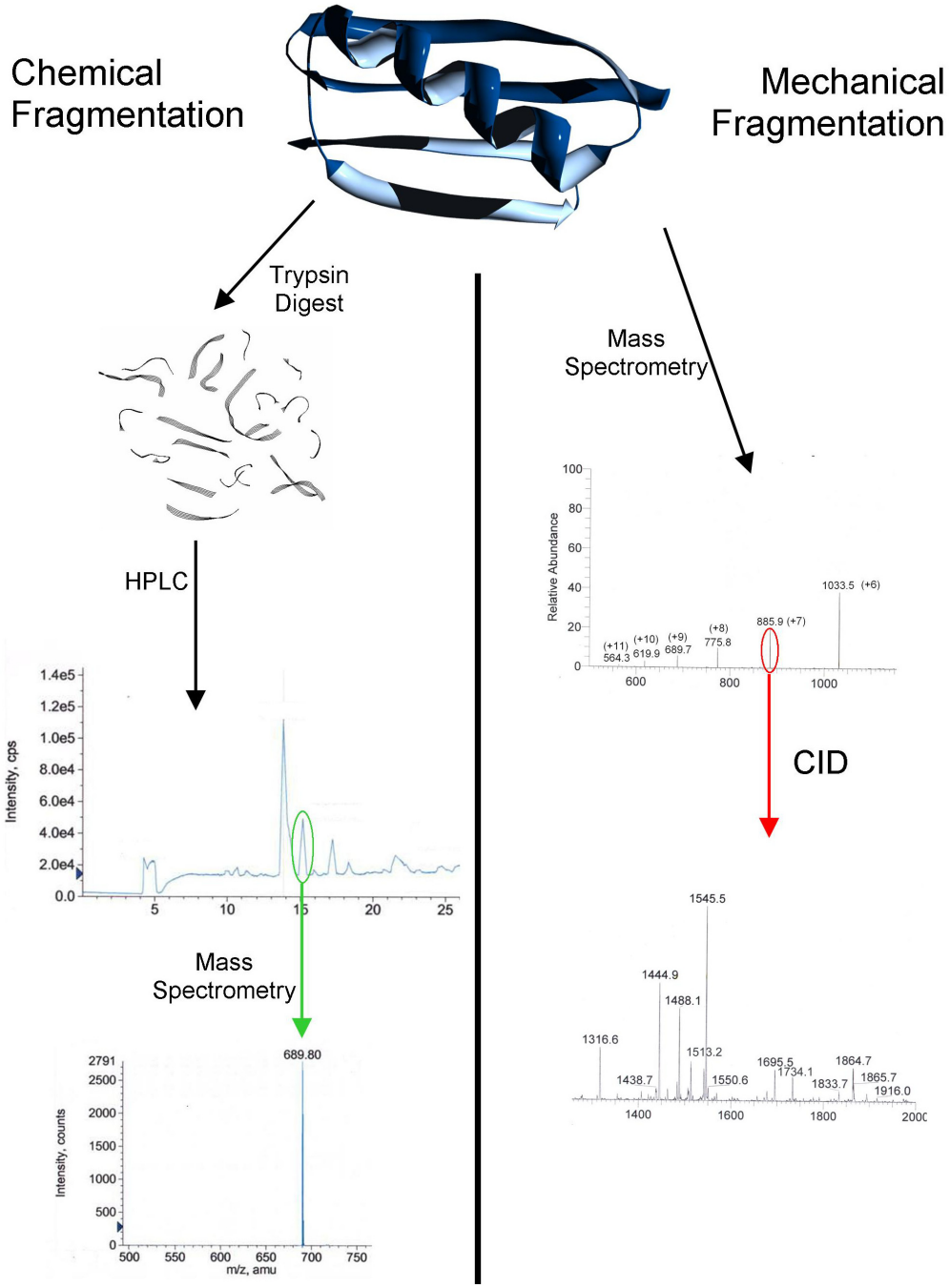


Fig. 4-3: Fixed Composition Sequences Used in CID

Fixed composition Gβ1 sequences used for MS analysis are listed below. Positions with wild-type amino acids are represented by a dot. Vertical dashed lines show observed cleavage after Asp residues in the wild-type sequence. No cleavage is observed after residues highlighted in red since they replaced the wild-type Asp. Vertical solid lines show non-selective C-terminus cleavage observed for all the fixed composition sequences analyzed. The five Lys in the wild-type sequence are shown in blue.

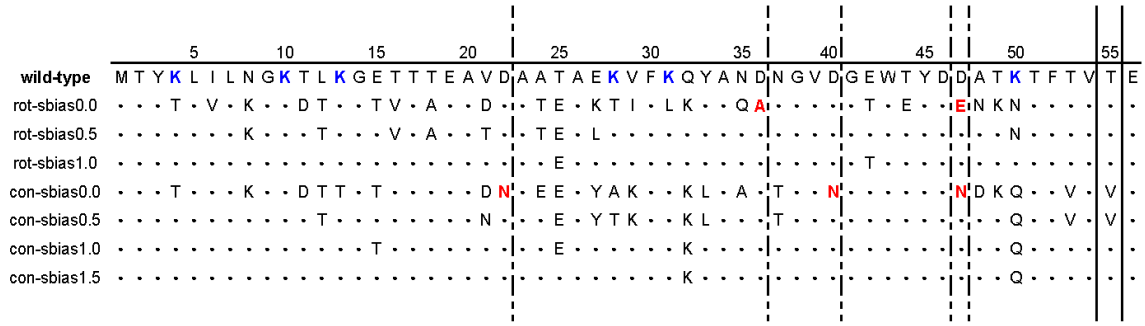


Fig. 4-4: LC-MS of Gβ1 Variant Rot-sbias0.0

The HPLC chromatograph from trypsin digest of rot-sbias0.0 is shown below. Masses pertaining to observed fragments are shown above each peak. Boxed masses belong to known trypsin digest product.

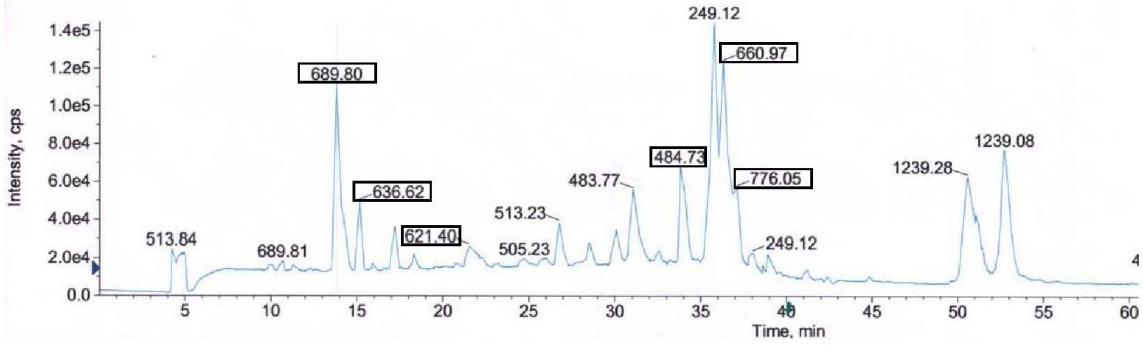


Fig. 4-5: CID on the Wild-Type Gβ1 Sequence

Assigned fragments are highlighted in color. Note that only b ions were successfully identified. The Gβ1 wild-type sequence spectrum is typical of other fixed composition sequences.

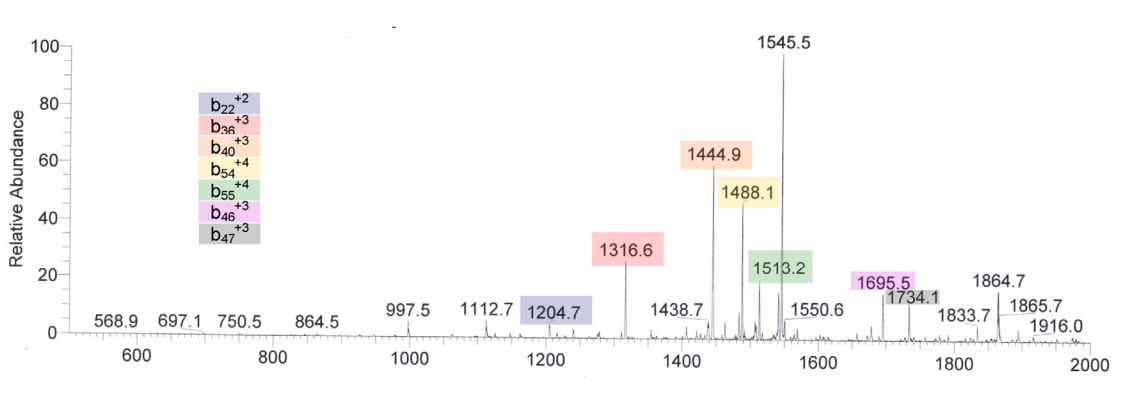


Fig. 4-6: Nomenclature for Backbone Fragmentation

CID preferentially fragments along the peptide backbone. Each of the vertical lines shows a potential fragmentation site. The labels above the vertical lines are assigned to the C-terminal fragment and those below the vertical lines are assigned to the N-terminal fragment. Fragmentation along the red vertical line results in two peptides. The C-terminal end is referred to as y_3 and the N-terminal end is referred to as b_2.

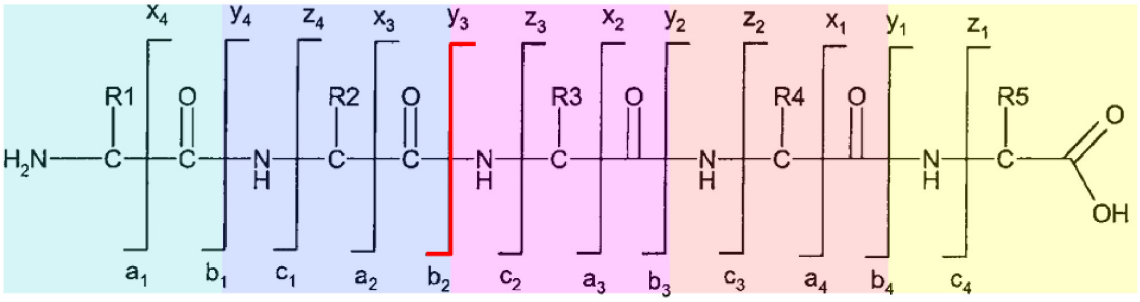


Fig. 4-7: Gβ1's Wild-Type Crystal Structure

(*A*) The crystal structure of Gβ1 is shown in orange with fragmentation sites along the backbone highlighted in green. (*B*) All the Lys residues are shown in cyan and Asp36 in orange. It is clear from (*B*) that none of the Lys residues are close enough for salt bridge formation with Asp36.

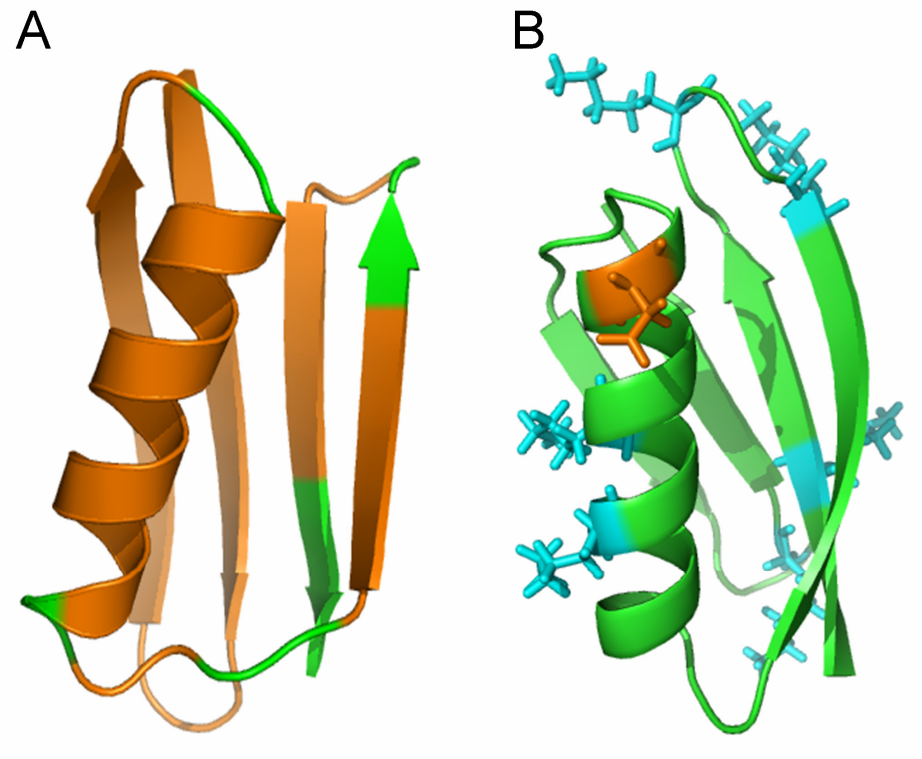


Fig. 4-8: CD Wavelength Scans of Fixed Compositions Sequences in Conditions Required for CID

Wavelength scans were carried out in 50% methanol and 0.1% acetic acid. All of the fixed composition sequences studied (wild-type, con-sbias0.0, con-sbias0.5, rot-sbias0.0, and rot-sbias0.5) showed CD wavelength scans typical for folded α/β proteins.

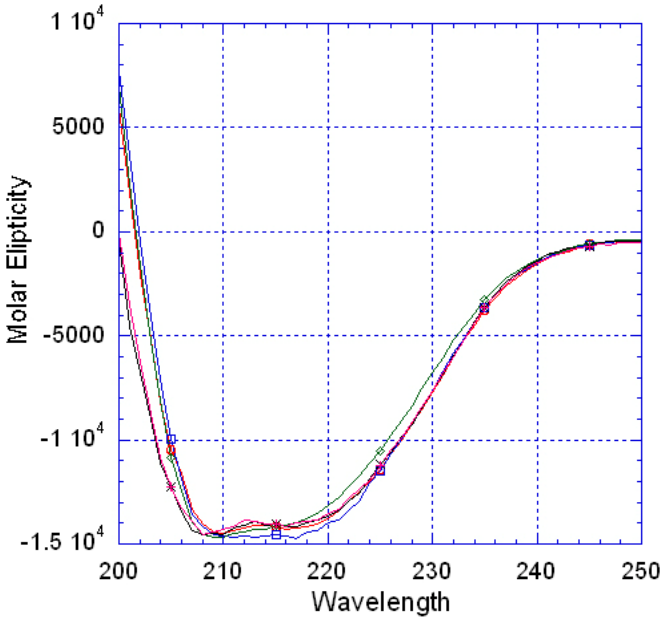


Fig. 4-9: Proposed Mechanisms for Asp Effect

Two possible mechanisms for fragmentation of the peptide backbone are shown below. In both cases, the presence of Arg is required to capture acidic protons; however, its role in the cleavage differs. In the first mechanism (*A*), Arg ideally stands by while the Asp carries out the cleavage. The resulting fragment contains a cyclic anhydride at the C-terminus. In the second mechanism (*B*), Arg forms a salt bridge with the deprotonated Asp following proton transfer. Backbone cleavage is initiated by the carbonyl oxygen of the preceding residue. The resulting fragment contains an oxazolone moiety at the C-terminus.

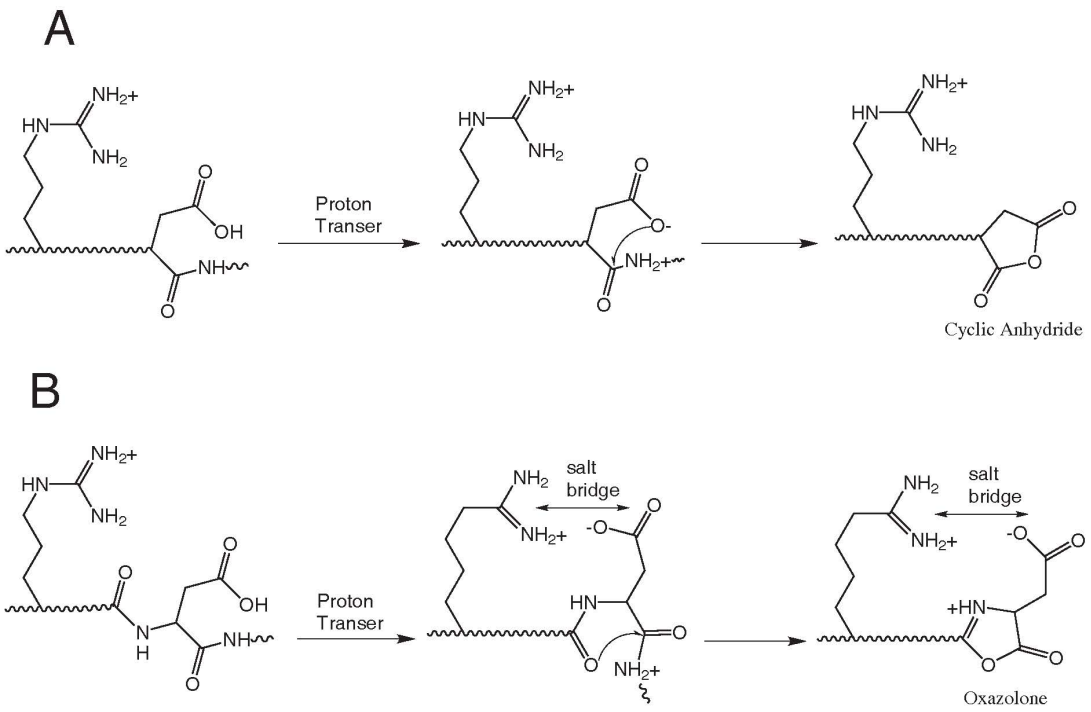
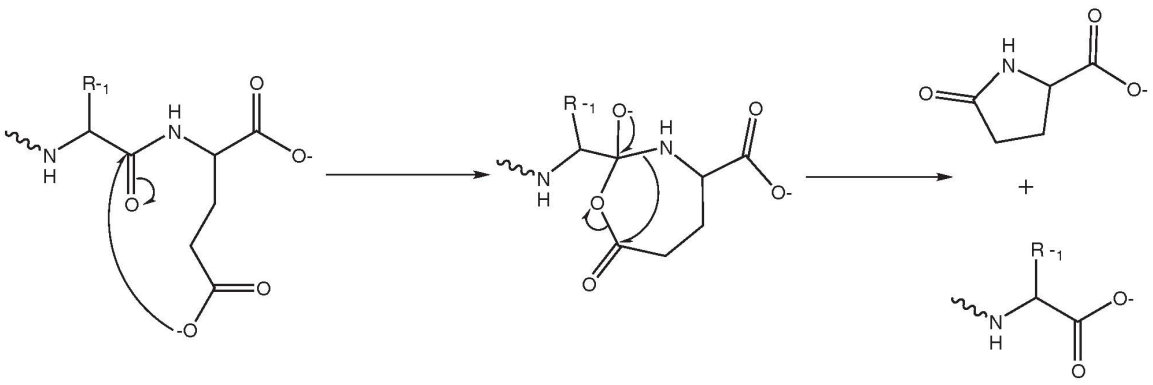


Fig. 4-10: Proposed Mechanisms for C-Terminal Glu Effect

C-terminal cleavage is commonly observed when Glu is the last residue in the sequence. The mechanism depicts the Glu folding back and carrying out a nucleophilic attack on the carbonyl carbon of a preceding residue.



Chapter 5

Optimizing HIV-1 Protease Specificity

5.1 Background on HIV Protease

HIV protease is a 99-residue homodimer essential for viral maturation. Its main function is to recognize and cleave sequences in the Gag and Pol polyproteins. Due to HIV protease's unique role, a major effort to engineer potent inhibitors has resulted in a number of successful therapeutics. Protease inhibitors are seen as crucial components in the multi-prong approach used to fight HIV infection. While there is currently no cure for HIV infection, state of the art therapeutics can keep the virus at bay for decades. The biggest threat to patients is the adaptability of the virus and its enzymes. HIV protease is notorious for mutating and developing resistance to inhibitors. There are countless examples of mutations along the substrate binding pocket that decrease the enzyme's specificity for inhibitors while retaining its ability to hydrolyze its native substrates.

Hydrolysis of the peptide backbone is carried out by two aspartic acids at the center of the substrate binding cleft. HIV protease was first categorized as belonging to

the aspartate protease family after successful inhibition by pepstatin, a nonspecific inhibitor of aspartate proteases [1]. There is currently no consensus on the detailed chemical mechanism of hydrolysis. However, it is widely believed that a water molecule serves as the nucleophile and that the aspartates serve as a general acid-base catalyst [2]. The aspartate residues are essential for hydrolysis; mutating them to asparagines results in a dead enzyme [3]. This fact was used by Schiffer and colleagues, who co-crystallized wild-type substrates with an inactive variant of HIV protease [4, 5].

The enzyme is of unique interest because its symmetrical binding region recognizes and cleaves asymmetrical substrates. While highly specific, HIV protease recognizes and hydrolyzes sequences that exhibit little sequence homology (Table 5-1). Studies on substrate binding have shown that specificity is driven by steric complementarity of the peptide side chains [5]. Hydrogen bonds to the side chains are rarely formed, and those that do form are not conserved between substrates [5]. All the conserved hydrogen bonds occur between the enzyme and the substrate backbone. An extensive network of hydrogen bonds is formed, which locks the substrate backbone into the proper conformation for peptide hydrolysis.

Hydrogen bonds are often seen as crucial for substrate specificity, and it has been suggested that HIV protease's lack of specific hydrogen bonds to peptide side chains allows it to bind to a broad range of sequences [5]. If this hypothesis is correct, then variants that can form specific hydrogen bonds to wild-type substrates should exhibit an increase in specificity [6]. The question of specificity has already been addressed in computational protein design for other proteins. However, the impact of negative design on specificity has yet to be fully addressed. Optimizing for hydrogen bonding can be

seen as a purely positive design approach, whereas emphasizing differences in electrostatics or van der Waals interactions can be used to incorporate negative design. Electrostatics can be important in discriminating between side chains with opposite charges, while van der Waals forces can be used to select for side chains of different dimensions.

Positive design proved to be a successful approach in the computational protein design of calmodulin (CaM) specificity [7, 8]. CaM is a protein that interacts with a large number of α-helical peptides. Upon binding, it undergoes a drastic conformational change to maximize contacts with its ligand. Designs on CaM used the bound compact structure to produce mutants with increased specificity towards a target peptide [7]. The substrates were either known wild-type CaM-binding sequences or sequences engineered to have increased specificity [7, 9]. The designs were limited to positions known to directly contact the peptide. A similar approach was also shown to be effective by Rein et al., who re-engineered PDZ domains to bind to novel sequence targets [10].

In addition to designing variants with increased specificity towards peptides, computational protein design has been used to engineer proteins with improved protein-protein specificities. Bolon et al. showed that focusing on positive design resulted in more stable complexes, whereas including negative design provided specificity at the cost of stability [11]. Negative design was also shown to be crucial in the design of coiled-coil interfaces. The computational algorithm used by Havranek and Harbury explicitly considered the aggregated, denatured, homo-dimeric and hetero-dimeric states [12]. While optimizing for one state, the algorithm designs against (predicts destabilizing mutations for) the other three states. Another example is the redesign of a protein

complex between colicin E7 DNAse and Im7 immunity protein. In this design, a new hydrogen bond network was predicted at the interface that provided increased specificity for the cognate dimer over the non-cognate dimer [13]. The success of the redesign was attributed to using positive design on an ensemble and designing against the native complex.

The work presented in this chapter attempts to increase the specificity of HIV protease by designing variants using a combination of positive and negative design strategies. Previous attempts to engineer protease specificity have largely focused on the rational design of trypsin [14-18]. Trypsin is a hydrolase that is highly specific for Lys- and Arg-containing peptides, while chymotrypsin favors peptides with aromatic residues such as Phe, Tyr, and Trp. In 1992, Headstrom et al. successfully engineered a trypsin mutant with specificity similar to that of chymotrypsin [19]. Since both enzymes have similar tertiary structures, replacing trypsin residues around the catalytic binding pocket and surface loops with those of chymotrypsin resulted in a preference for hydrophobic amino acids. The work done on trypsin focused on changing the specificity of the S1 binding pocket.

Our redesign of HIV protease considers the entire binding region, which is composed of eight binding pockets. In this work, we bridged the gap between the computational design of specificity and engineering for improved catalytic activity [20]. The computational approach aimed to re-engineer the binding pockets to have increased specificity for one of HIV protease's natural substrates. Unlike CaM or PDZ domains, HIV protease is catalytic, and the designs were expected to preserve its hydrolase activity in addition to increasing its specificity.

5.2 Results and Discussion

5.2.1 Design Calculations and Prediction of Mutants

Due to the large binding region of HIV protease, only side chains directly contacting the substrate were considered. In order to conserve function, the catalytic aspartates at positions 25 and 125 were kept in their crystallographic conformations. Given that HIV protease, a symmetrical dimer, binds to asymmetrical peptides, the two-fold symmetry of the binding region was not conserved. Three crystal structures of HIV protease, each bound to a different native substrate, were used in the designs (PDB files 1F7A, 1KJG, and 1KJ7). Early design calculations indicated that the RT-RH bound structure (1KJG) showed the most promise, so it was selected for further optimization.

Due to HIV protease's large binding region, the design was initially divided into eight small calculations, one for each binding pocket. For each pocket, residues within 4.2 Å of a substrate side chain were defined as being in the 1st shell, and residues within 4.2 Å of any 1st shell residue were defined as being in the 2nd shell. 1st shell residues were allowed to mutate, whereas 2nd shell residues and the substrate's side chains were floated (their conformations were allowed to change, but their amino acid identities were fixed to wild-type). Optimization of interactions within the binding region was expected to increase specificity [21].

The individual pocket designs were carried out in the following sequential order:

$$S1 \rightarrow S1' \rightarrow S2 \rightarrow S2' \rightarrow S3 \rightarrow S3' \rightarrow S4 \rightarrow S4'$$

Mutations predicted from preceding calculations were carried over to the next design. As a result, the final design on the S4′ binding pocket contained all the mutations from preceding designs.

One of the major drawbacks with designing individual pockets is their discrete nature. As implemented thus far, the optimization procedure had no means of predicting inter-pocket mutations that might be beneficial. To mitigate this problem, positions that predicted reasonable mutations in the individual pocket calculations were simultaneously designed in the context of the entire binding region. In the case of the 1KJG crystal structure, four positions (30, 48, 82, and 130) were selected for simultaneous design. All the other positions that had been considered in individual binding pocket calculations were floated. Not surprisingly, the design predicted mutations at all four of the selected positions (Fig. 5-1). Position 30 was mutated to Phe due to improved van der Waals interactions with AlaP4, and position 48 replaced Gly with Arg to form a salt bridge with GluP3. The mutations predicted at positions 82 (V→I) and 130 (D→N) were conservative, and were expected to have little impact on specificity for the target substrate, RT-RH.

Changing force field parameters can be helpful in revealing promising new mutations. We switched the solvation model from a solvent exclusion-based one to a surface area-based one and repeated the individual pocket designs. The resulting sequence predicted an Ala to Ser mutation at position 28 that forms a hydrogen bond with the P2 Thr of the substrate (Fig. 5-2). After considering the mutations from all the preceding designs, a four-fold mutant (A28S/D30F/G48R/V82I) was selected for further evaluation. Side-chain placement calculations were carried out on this mutant and on the wild type using three substrate-bound crystal structures (1KJG, 1F7A, and 1KJ7) in which the protease was bound to the RT-RH, CA-P2, or P2-NC peptide, respectively. Energy analysis of the side-chain placement results indicated that the four-point mutant

would stabilize the RT-RH substrate by 7.8 kcal/mol relative to the wild-type sequence (Table 5-2). In contrast, a much smaller increase in stability (1.96 kcal/mol) was predicted for the P2-NC peptide, and a decrease in stability was predicted for the CA-P2 peptide. The bulk of the stability for RT-RH arises from the new hydrogen bond and salt bridge that are predicted to form as a result of the A28S and G48R mutations.

The G48R mutation appears to stabilize interactions with RT-RH, while disfavoring binding of CA-P2 and P2-NC. Arg at position 48 is able to form a nice salt bridge with Glu at the P3 position on RT-RH (Fig. 5-2). The absence of Glu at P3 in the other two substrates clearly prevents a similar interaction (Table 5-1). Position P3 in P2-NC is a Thr, a residue too small to accommodate even a hydrogen bond. CA-P2 contains an Arg at P3, forcing an interaction between two side chains with similar charges. The unfavorable electrostatic contact would normally be expected to significantly destabilize binding of CA-P2. However, due to the two-fold symmetry in the binding region of HIV protease, substrates can bind in either of two orientations. To avoid interacting with Arg48, the CA-P2 substrate is likely to bind in the opposite orientation where P3Arg can interact with Ala148 instead. Loss in specificity would still be observed, since there would be limited binding in one of the two possible binding orientations. The same argument can be made about substrate P2-NC; although there are no unfavorable interactions predicted between Arg48 and P3Thr, binding in the alternate conformation would result in Arg48 having direct contact with P3′Arg (Table 5-1). As a result, the mutation at position 48 to an Arg can be said to contain negative design features in addition to its positive design attributes.

5.2.2 Kinetic Experiments

Experimental results support the idea that Arg at position 48 results in a change in specificity of the S3 and the S3′ binding pockets [22]. Replacing Gly48 with Arg in the wild-type protein has previously been shown to decrease affinity for negatively charged residues at the P3 and the P3′ positions [22]. Since our designs allowed asymmetrical mutations, in constructing the wild type and mutant proteins we used a tethered dimer to ensure a heterodimer complex. Given that the crystal structure of the dimer shows that the N-terminus of one monomer is adjacent to the C-terminus of the second monomer, using a linker to covalently attach the monomers seemed reasonable.

5.2.3 Positive Design Results

After evaluating all the positive design results, we constructed an asymmetrical mutant, HIVpr-positive, which included three of the mutations predicted to stabilize RT-RH binding from the positive design calculations (A28S, D30F, and G48R) (Fig. 5-2). Kinetic experiments for the tethered wild-type protein and for HIVpr-positive were carried out and kinetic parameters were compared. Due to minimal peptide solubility, limited data was obtained for substrates with K_ms larger than 50 μM. Nevertheless, V_{max}/K_m values were successfully obtained using hydrolysis rates at low substrate concentrations. The wild-type protein showed a preference for the CA-P2 peptide: V_{max}/K_m for CA-P2 was 1.44 times larger than for RT-RH (Table 5-3). The P2-NC peptide was the least efficiently hydrolyzed of the three substrates, with a V_{max}/K_m slightly lower than the RT-RH value. The three-point mutant, HIVpr-positive, exhibited a significantly different specificity profile. It was most efficient in hydrolyzing RT-RH, followed by the other two substrates, which were hydrolyzed at about one-third the

RT-RH rate. The normalized values show that, relative to wild type, specificity towards the RT-RH substrate increased three-fold and four-fold over the P2-NC and CA-P2 substrates, respectively.

The experimental results obtained from the three-point variant support the idea that using only positive design to optimize for the target structure is an effective way to increase substrate specificity. Positive design is extremely effective in optimizing hydrogen bonds and salt bridges. In the case of HIV protease, new hydrogen bonds are especially important for improving specificity because substrate side chains in the wild-type protein contain buried unsatisfied hydrogen bond acceptor and donor atoms. As observed in the optimization of the RT-RH substrate, the incorporation of salt bridges can have the additional benefit of incorporating unintentional negative design features in the predicted mutants.

5.2.4 Incorporating Negative Design

In an attempt to increase the specificity exhibited by the three-point mutant, an explicit negative design approach was implemented. The calculation required the design of all three crystal structures, 1KJG, 1KJ7, and 1F7A, in parallel. A scoring function was used that benefited mutations having favorable interactions with RT-RH in 1KJG and unfavorable interactions with the CA-P2 and P2-NC substrates in 1F7A and 1KJ7, respectively. Initial designs used the following simple scoring function:

$$Score = E_{1KJG} - E_{1KJ7} - E_{1F7A} \qquad \text{(Eq. 5-1)}$$

Only sequences that exhibited energies within 20% of the global energy minimum conformation in the 1KJG structure were considered.

$$E_{1KJG} \leq \left(E_{1KJG}^{\min} \times 0.8\right) \qquad \text{(Eq. 5-2)}$$

Sequences predicted to have energies above 200 kcal/mol in the 1KJ7 and 1F7A structures were assigned an energy of 200 kcal/mol. This hard energy ceiling caps the large unfavorable energies that result from overlapping atoms. A successful negative design will predict mutations that pack nicely into the 1KJG structure, but cause at least one van der Waals clash in both the 1F7A and the 1KJ7 crystal structures. The extent of the atomic clash is not crucial and there is little benefit in discriminating between them. Any sequence with a reasonable clash will hit the hard energy ceiling of 200 kcal/mol, at which point the best sequence is the one with the lowest energy in the 1KJG crystal structure (Fig. 5-3).

This negative design approach was carried out on individual pockets in the hopes of identifying positions predicted to clash with the bound substrate in the 1KJ7 and 1F7A structures. Unfortunately, only three of the eight designs predicted sequences with clashing residues in the desired structures (Table 5-4). Energy analysis of the predicted sequences revealed that only P1 pocket mutations resulted in extremely unfavorable interactions in both the 1F7A and 1KJ7 structures. The P3′ pocket design predicted a sequence with unfavorable energies in the 1KJ7 structure, but reasonable energies in the 1F7A and the 1KJG structures. The fact that only one of eight pocket designs predicted clashing mutations in both 1F7A and 1KJ7, and not in 1KJG, supports the idea that the binding pockets are optimized for steric complementarity for all three substrates. Mutations that clash in the 1F7A and 1KJ7 scaffolds are likely to also clash in the 1KJG structure and, thus, are not predicted.

Closer inspection of the ten best-scoring sequences for the P1 pocket design showed position 180 was consistently mutated to an Ala and position 184 was mutated to

a large hydrophobic amino acid (Phe, Trp, or Tyr). Since it is unlikely that replacing a Thr at position 180 with an Ala would result in large clashes, we focused our attention on mutations at position 184. The large hydrophobic residues selected at position 184 were predicted to cause large van der Waals clashes with Met and Leu at the P1 position in the P2-NC and CA-P2 peptides. Interestingly, position P1 in the RT-RH substrate is a Phe, a residue much bulkier then either Met or Leu. Phe at P1 was predicted to take on a conformation that allows it to make good π-stacking interactions with the hydrophobic aromatic residues predicted at 180. Met and Leu, on the other hand, were unable to accommodate the change at position 184.

Experimental analysis was carried out on a four-point mutant, HIVpr-negative, which included a I184F mutation in addition to A28S, D30F, and G48R mutations. Relative to the results observed for HIVpr-positive, the four-point mutant showed decreased specificity (Table 5-3). Specificity for CA-P2 reverted back to a value closer to that observed for the wild-type protein: 1.63 compared to 1.44 for the wild-type protein. Mutating position 184 to Phe apparently did not increase specificity for the target peptide; instead, it increased specificity for CA-P2.

It is difficult to predict the true impact of van der Waals clashes in a computational protein design procedure that requires the use of a fixed backbone. Proteins are intrinsically dynamic and can adapt to mutations in order to prevent van der Waals violations. With the fixed backbone restriction, there is no means of modeling protein motions that might accommodate unfavorable van der Waals energies. In addition, the discrete nature of the rotamer library can exclude side-chain conformations that might prevent atomic clashes. One possible solution to the problem is minimization,

in which the protein backbone and predicted side-chain conformations are relaxed. Minimization, however, is computationally expensive, since it would have to be performed at every step in the sequence search. If minimization had been incorporated into the negative design procedure, the mutation at position 184 would not have been selected, since the clash would have been alleviated in all the scaffolds. A good alternative to minimization is the parallel design of multiple static backbone structures that represent a protein's dynamic range. Future studies in computational protein design will most likely incorporate methods that include protein backbone motion.

5.3 Conclusions

Positive design proved to be an effective way to alter protein specificity. Positive design can easily identify stabilizing electrostatic interactions and hydrogen bonds that will increase substrate specificity. Negative design was shown to successfully recover mutations predicted using positive design alone and to predict mutations that clash in alternate substrates. Experimental results support the idea that hydrogen bonds and salt bridges increase specificity for the target substrate. Steric complementarity is known to be crucial for specificity; however, designing optimal steric complementarity for one substrate and destabilizing another is complicated by protein dynamics.

5.4 Materials and Methods

5.4.1 Computational Positive Design

The crystal structure of HIV protease bound to the RT-RH substrate (PDB code 1KJG) was used as the positive design scaffold. The crystal structure was put through 50 steps of minimization to relax van der Waals interactions, atomic bonds, and angles.

Residues within 4.2 Å of a substrate side chain were selected for design (1st shell residues); residues within 8.4 Å not selected to be in the 1st shell (2nd shell residues), and substrate side chains were floated (allowed to change conformation but not amino acid identity). Five conserved water molecules, residues hydrogen bonding to the waters (8, 29, 87, 108, 129, and 187) and catalytic residues (25 and 125) were fixed in their crystallographic conformations. In addition, proline-containing positions (81 and 181) were only allowed to change conformation.

A Dunbrack and Cohen-based backbone dependent rotamer library was used for side-chain optimization. χ_1 and χ_2 values were expanded 1 standard deviation for all amino acids. In addition, the crystallographic rotamer at every design position was included. Either a solvent exclusion-based or an atomic surface area-based solvation potential was used. A rotamer probability scale factor of 0.3 proportionally penalized side-chain conformations based on their pre-calculated probabilities. All other parameters and potential functions have been described in previous Mayo lab publications [23-26]. An optimization algorithm based on the Dead-End Elimination (DEE) theorem was used in the design of individual pockets [27]. Designs on the entire binding region required the use of the FASTER algorithm to achieve convergence [28]. A combination of energy analysis and visual inspection of the predicted low-energy conformations was used to identify promising mutations.

5.4.2 Computational Negative Design

Crystal structures of HIV protease bound to the CA-P2 and P2-NC peptides (PDB codes 1F7A and 1KJ7, respectively) were used as negative design scaffolds. Both structures were minimized in the same fashion as the 1KJG crystal structure. All force

field parameters were identical to those used in the positive design procedure. Computational negative design, however, carried out sequence optimization using all three scaffolds simultaneously. The search algorithm selects an amino acid mutation and determines the lowest energy conformation for that amino acid on each of the scaffolds. The scoring function then subtracts the predicted energies on the negative scaffolds from that of the positive scaffold (Eq. 5-1) to yield a fitness value representative of the final objective of the calculation.

Design of specificity requires a scoring function that favors stabilizing mutations in the target substrate while destabilizing alternate substrates. Equation 5-1 is simple but effective when adequately restricted. In order to ensure reasonable sequences for the target substrate, a DEE calculation is carried out to identify the global minimum energy conformation (GMEC) in the target's scaffold. Only sequences within 20% of the GMEC energy are evaluated, ensuring that the predicted sequences are reasonable. In addition, scoring for the undesired substrate is capped at a cutoff of 200 kcal/mol.

5.4.3 Protein Kinetics

HIV protease variants were expressed from a Pet11A *E. coli* plasmid vector. The gene construct coded for two copies of the HIV monomer linked by the nucleotide sequence that codes for Gly-Gly-Ser-Ser-Gly. The nucleotide sequence that coded for each monomer was unique; the use of two different sequences allowed for site-directed mutagenesis to be targeted to a specific monomer. Cysteines at position 67, 95, 167, and 195 were mutated to Leu, Met, Leu, and Met, respectively.

HIV protease variants were expressed in two liter cultures of *E. coli* BL21(pLys) cells at 37˚C. Protein expression was initiated at an OD_{600} of 0.6 by adding IPTG. Cells

were harvested after three hr and lysed using an emulsiflex. The inclusion bodies were isolated and resuspended in 66% acetic acid and diluted ten-fold in water. The soluble fraction was isolated after centrifugation and dialyzed overnight against 100% water. Any precipitate was removed and the resulting sample was purified using cation exchange chromatography. The pure sample was desalted and lyophilized. Active enzyme was produced by taking the lypholized sample in 8 M guanidinium and refolding the protein at 0.6 mg/ml in 55 mM Tris pH 8.2, 10.56 mM NaCl, 0.44 mM KCl, 0.055% PEG 3350, 550 mM guanidine HCl, 1.1mM EDTA, 440 mM sucrose, and 1mM DTT at 0˚C. Kinetics were determined using three DABCYL/EDANS substrates: NH_2-D(Edans)-KARVLAEAM-K(Dabcyl)-R-COOH, NH_2-D(Edans)-ATIMMQRGN-K(Dabcyl)-R-COOH, and NH_2-D(Edans)-AETFYVDGA-K(Dabcyl)-R-COOH, which are derivatives of CA-P2, P2-NC, and RT-RH substrates, respectively [29]. Hydrolysis was monitored at 490 nm while exciting at 340 nm in a PTI fluorimeter. The reaction buffer was composed of 0.1 M sodium acetate, 1 M NaCl, 1 mM EDTA, 1 mM DTT, 1 mg/ml BSA, and 10% DMSO at a pH of 4.7.

5.5 Bibliography

1. Vonderhelm, K., Seelmeier, S., Schmidt, H., and Juncker, U. Hiv has an aspartic protease and can be inhibited by pepstatin. *Biol Chem H-S* **369**, 835-835 (1988).

2. Brik, A., and Wong, C. H. HIV-1 protease: Mechanism and drug discovery. *Org Biomol Chem* **1**, 5-14 (2003).

3. Darke, P. L., Leu, C. T., Davis, L. J., Heimbach, J. C., Diehl, R. E., Hill, W. S., Dixon, R. A., and Sigal, I. S. Human immunodeficiency virus protease. Bacterial expression and characterization of the purified aspartic protease. *J Biol Chem* **264**, 2307-2312 (1989).

4. Prabu-Jeyabalan, M., Nalivaika, E., and Schiffer, C. A. How does a symmetric dimer recognize an asymmetric substrate? A substrate complex of HIV-1 protease. *J Mol Biol* **301**, 1207-1220 (2000).

5. Prabu-Jeyabalan, M., Nalivaika, E., and Schiffer, C. A. Substrate shape determines specificity of recognition for HIV-1 protease: Analysis of crystal structures of six substrate complexes. *Structure* **10**, 369-381 (2002).

6. Wells, J. A., Powers, D. B., Bott, R. R., Graycar, T. P., and Estell, D. A. Designing substrate specificity by protein engineering of electrostatic interactions. *Proc Natl Acad Sci USA* **84**, 1219-1223 (1987).

7. Shifman, J. M., and Mayo, S. L. Modulating calmodulin binding specificity through computational protein design. *J Mol Biol* **323**, 417-423 (2002).

8. Shifman, J. M., and Mayo, S. L. Exploring the origins of binding specificity through the computational redesign of calmodulin. *Proc Natl Acad Sci USA* **100**, 13274-13279 (2003).

9. Green, D. F., Dennis, A. T., Fam, P. S., Tidor, B., and Jasanoff, A. Rational design of new binding specificity by simultaneous mutagenesis of calmodulin and a target peptide. *Biochemistry* **45**, 12547-12559 (2006).

10. Reina, J., Lacroix, E., Hobson, S. D., Fernandez-Ballester, G., Rybin, V., Schwab, M. S., Serrano, L., and Gonzalez, C. Computer-aided design of a PDZ domain to recognize new target sequences. *Nat Struct Biol* **9**, 621-627 (2002).

11. Bolon, D. N., Grant, R. A., Baker, T. A., and Sauer, R. T. Specificity versus stability in computational protein design. *Proc Natl Acad Sci USA* **102**, 12724-12729 (2005).

12. Havranek, J. J., and Harbury, P. B. Automated design of specificity in molecular recognition. *Nat Struct Biol* **10**, 45-52 (2003).

13. Joachimiak, L. A., Kortemme, T., Stoddard, B. L., and Baker, D. Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface. *J Mol Biol* **361**, 195-208 (2006).

14. Hung, S. H., and Hedstrom, L. Converting trypsin into elastase. *Biophys J* **72**, Mp295-Mp295 (1997).

15. Hung, S. H., and Hedstrom, L. Converting trypsin to elastase: Substitution of the S1 site and adjacent loops reconstitutes esterase specificity but not amidase activity. *Protein Eng* **11**, 669-673 (1998).

16. Kurth, T., Grahn, S., Thormann, M., Ullmann, D., Hofmann, H. J., Jakubke, H. D., and Hedstrom, L. Engineering the S1' subsite of trypsin: Design of a protease which cleaves between dibasic residues. *Biochemistry* **37**, 11434-11440 (1998).

17. Page, M. J., Wong, S. L., Hewitt, J., Strynadka, N. C., and MacGillivray, R. T. Engineering the primary substrate specificity of Streptomyces griseus trypsin. *Biochemistry* **42**, 9060-9066 (2003).

18. Tanaka, T., and Yada, R. Y. Redesign of catalytic center of an enzyme: Aspartic to serine proteinase. *Biochem Biophys Res Commun* **323**, 947-953 (2004).

19. Hedstrom, L., Szilagyi, L., and Rutter, W. J. Converting trypsin to chymotrypsin - the role of surface loops. *Science* **255**, 1249-1253 (1992).

20. Wilson, C., Mace, J. E., and Agard, D. A. Computational method for the design of enzymes with altered substrate specificity. *J Mol Biol* **220**, 495-506 (1991).

21. Park, S., Morley, K. L., Horsman, G. P., Holmquist, M., Hult, K., and Kazlauskas, R. J. Focusing mutations into the P. fluorescens esterase binding site increases enantioselectivity more effectively than distant mutations. *Chem Biol* **12**, 45-54 (2005).

22. Moody, M. D., Pettit, S. C., Shao, W., Everitt, L., Loeb, D. D., Hutchison, C. A., 3rd, and Swanstrom, R. A side chain at position 48 of the human immunodeficiency virus type-1 protease flap provides an additional specificity determinant. *Virology* **207**, 475-485 (1995).

23. Dahiyat, B. I., Gordon, D. B., and Mayo, S. L. Automated design of the surface positions of protein helices. *Protein Sci* **6**, 1333-1337 (1997).

24. Dahiyat, B. I., and Mayo, S. L. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci USA* **94**, 10172-10177 (1997).

25. Dahiyat, B. I., and Mayo, S. L. De novo protein design: Fully automated sequence selection. *Science* **278**, 82-87 (1997).

26. Dahiyat, B. I., Sarisky, C. A., and Mayo, S. L. De novo protein design: Towards fully automated sequence selection. *J Mol Biol* **273**, 789-796 (1997).

27. Gordon, D. B., Hom, G. K., Mayo, S. L., and Pierce, N. A. Exact rotamer optimization for protein design. *J Comput Chem* **24**, 232-243 (2003).

28. Desmet, J., Spriet, J., and Lasters, I. Fast and accurate side-chain topology and energy refinement (FASTER) as a new method for protein structure optimization. *Proteins* **48**, 31-43 (2002).

29. Matayoshi, E. D., Wang, G. T., Krafft, G. A., and Erickson, J. Novel fluorogenic substrates for assaying retroviral proteases by resonance energy transfer. *Science* **247**, 954-958 (1990).

Table 5-1: Sequences of Peptide Substrates Hydrolyzed by Wild-Type HIV-1 Protease

	Residue Position							
Substrate	**P4**	**P3**	**P2**	**P1**	**P1'**	**P2'**	**P3'**	**P4'**
MA-CA	SER	GLN	ASN	TYR	PRO	ILE	VAL	GLN
CA-P2*	ALA	ARG	VAL	LEU	ALA	GLU	ALA	MET
P2-NC*	ALA	THR	ILE	MET	MET	GLN	ARG	GLN
NC-P6	PRO	GLN	ASN	PHE	LEU	GLN	SER	ARG
in P6	LYS	GLU	LEU	TYR	PRO	LEU	THR	SER
TF-PR	SER	PHE	ASN	PHE	PRO	GLN	ILE	THR
PR-RT	THR	LEU	ASN	PHE	PRO	ILE	SER	PRO
RT-RH†	ALA	GLU	THR	PEH	TYR	VAL	ASP	GLY
RT-IN	ARG	LYS	ILE	LEU	PHE	LEU	ASP	GLY

The red line running through the center of the table depicts the bond that is hydrolyzed by HIV protease.

* Peptides used for negative design.

† Peptide used for positive design.

Table 5-2: Energy Scores of Side-Chain Placement Calculation on the Binding Region of Wild-Type HIV Protease and a Predicted Four-Point Mutant

Substrate (PDB)*	Wild Type (kcal/mol)	A28S D30F G48R V82I (kcal/mol)	Δ (kcal/mol)
RT-RH (1KJG)	-282.18	-289.98	-7.80
P2-NC (1KJ7)	-276.04	-278.00	-1.96
CA-P2 (1F7A)	-275.31	-274.35	0.97

* The crystal structure file used for the side-chain placement designs.

Table 5-3: Experimental Kinetic Values for HIV Protease and Two Variants Using Three Peptide Substrates

	RT-RH (1KJG)		P2-NC (1KJ7)		CA-P2 (1F7A)	
Protease	**V_{max}/K_m (s^{-1})**	**Normalized**	**V_{max}/K_m (s^{-1})**	**Normalized**	**V_{max}/K_m (s^{-1})**	**Normalized**
Wild Type	1.54E-03	1.00	1.40E-03	0.91	2.22E-03	1.44
HIVpr-positive*	1.03E-03	1.00	3.03E-04	0.29	3.38E-04	0.33
HIVpr-negative†	1.17E-04	1.00	4.05E-05	0.35	1.91E-04	1.63

* Mutations: A28S D30F G48R

† Mutations: A28S D30F G48R I184F

Table 5-4: Energies and Scores for Individual Pocket Negative Designs of Three HIV Protease Structures

Pocket	Score*	1KJG† Energy (kcal/mol)	1F7A† Energy (kcal/mol)	1KJ7† Energy (kcal/mol)
P1	-496.32	-96.32	6053.00	5327.00
P1'	-245.33	-73.65	-28.32	6926.84
P2	5.58	-73.97	-35.93	-43.62
P2'	-141.15	-67.80	81.48	-8.13
P3	33.88	-44.83	-32.89	-45.81
P3'	-196.40	-41.05	-44.62	6494.00
P4	16.80	-38.50	-17.59	-38.11
P4'	23.77	-37.08	-14.61	-46.24

* The sequence score is calculated according to Eq. 5-1.

† All three crystal structures were designed in parallel.

Fig. 5-1: Predicted Conformation of Four-Fold HIV Protease Mutant: D30F/G48R/V82I/D130N

Peptide substrate, RT-RH, is shown in orange and HIV protease positions are shown in green. Hydrogen bonds are depicted by a dashed green line. The G48R mutation is predicted to form a nice salt bridge with Glu at position P3. The wild-type hydrogen bond between residue 130 and 145 is preserved due to the conservative mutation from Asp to Asn at position 130. On the other hand, Phe at position 30 is predicted to form improved van der Waals interactions with Ala at P4, which might significantly increase substrate specificity. A conservative mutation from Val to Ile at position 82 is unlikely to significantly impact specificity.

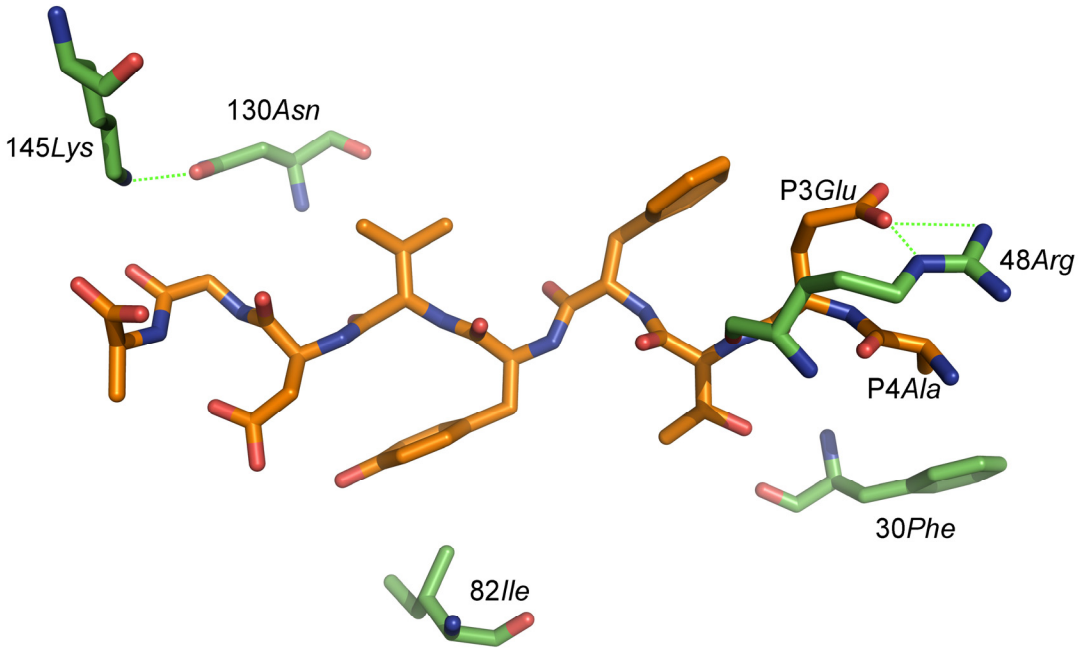


Fig. 5-2: Predicted Conformation for HIV-positive Mutant in the RT-RH Bound Scaffold

The Arg at position 48 is predicted to form a nice salt bridge with P3 Glu of RT-RH. A hydrogen bond is predicted to form between the Ser at position 28 and a Thr at P2. The mutation at position 30 is selected for its improved van der Waals interactions with P3 Ala.

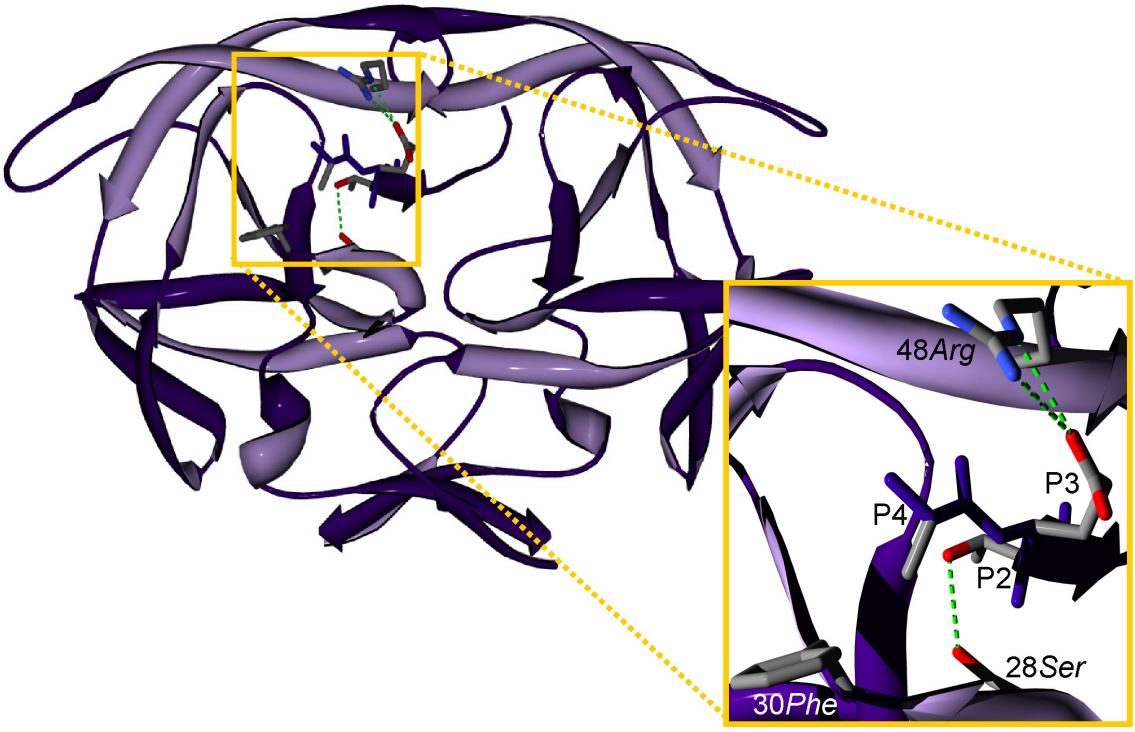


Fig. 5-3: Schematic Representation of Negative Design Sequence Optimization Procedure

The optimization procedure is represented below by a handful of time points that proceed from left to right. The sequence search is limited to sequences with energies within 20% of the global minimum energy conformation (highlighted in yellow) of the positive design scaffold, 1KJG. For the negative design scaffolds, 1F7A and 1KJ7, sequences with energies greater than 200 kcal/mol automatically receive a score of 200. The GMEC for the positive design scaffold is used as the starting sequence; the energy in the alternative scaffolds tends to be fairly optimal (*A*). As the negative design optimization proceeds, sequences that exhibit unfavorable energies in the negative design scaffolds are chosen until they hit the hard energy ceiling of 200 kcal/mol (*B, C,* and *D*). Once the ceiling is reached, the sequence score is only improved by optimizing the positive design structure represented in blue (*D* and *E*).