Using Noncanonical Amino Acids in Computational Protein Design

P. Douglas Renfrew

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biochemistry and Biophysics (Program in Molecular and Cellular Biophysics).

Chapel Hill
2009

Approved by:

Advisor: Brian Kuhlman

Reader: Nikolay Dokholyan

Reader: Charles Carter

Reader: Jan Hermans

Reader: Richard Wolfenden

Abstract

P. Douglas Renfrew: Using Noncanonical Amino Acids in Computational Protein Design

(Under the direction of Brian Kuhlman)

The structure of noncanonical amino acid (NCAA) side chains allows them to explore conformations inaccessible to canonical amino acids (CAAs). Peptides made of the D-enantiomers of amino acid backbones are resistant to proteolysis. The long term goal of this research is to adapt the current tools of computational protein design to create functional molecules be they proteins or not. In this thesis we have attempted the first steps toward this longer goal. The increased sequence and conformation space accessible to a protein during a design simulation when NCAAs are included, allows us to design tighter protein-protein interactions, with a higher degree of specificity.

The computational protein design program Rosetta has been modified for compatibility with NCAAs. The use of knowledge-based potentials was the major hurdle as the potentials are based on statistics collected from known protein structures and few protein structures have been determined containing NCAAs.

Using quantum mechanics (QM) calculations of the amino acids valine and isoleucine, with a helical conformation, we found an even distribution of rotamer preference. When that was used in rotamer recovery benchmarks, outperformed the knowledge-based potential that was biased because of long-range interactions imposed by the α-helical secondary structure. QM, although accurate and compatible with NCAAs was found to be too computationally expensive.

We created a modified energy function that can evaluate the energy of both CAAs and NCAAs, where the knowledge-based energy potentials have been replaced with physically-based MM

potentials that performs comparable to the stock energy function. We have developed methods to create rotamer libraries for both CAAs and NCAAs that are comparable to knowledge-based rotamer libraries. We have used these tools to create rotamer libraries for 88 different NCAAs that can now be used within Rosetta.

The interface between calpain and the calpastatin peptide as well as the interface between HIV GP41 and the integration inhibitor, PIE12, developed by the Kay lab, has been redesigned using NCAAs to increase the binding affinity between the two pairs. The research has take protein design in a new direction and has enabled the development of novel protein interactions, and protein-like therapeutics.

Acknowledgements

More so than ever science is not done in a vacuum and a Ph.D. thesis is not a solitary work. There were many people who have helped me get to this point in my life. A gigantic thank you has to go to Brian Kuhlman for being a fantastic advisor, giving me an incredible project, and encouraging me despite my short comings (Sorry I took so long!). All members of the Kuhlman lab past and present have helped me in some way, but some have gone out of their way: Glenn Butterfoss helped me get started on my first project of graduate school and has continued to be a friend and collaborator, Eun Jung Choi spent considerable time helping me with experiments. A very special thanks need to be given to Sylvie Doublie and the Doublie and Rould Labs at the University of Vermont where I worked for 3 years doing research as an undergraduate. Sylvie's encouragement and the fun I had in her lab is the main reason I choose to pursue graduate school and a Ph.D. Thanks needs to go out to Michael Kay and his lab at the University of Utah for trying our designs on his HIV integration inhibitor. Thank you to my committee members for their advice over the years.

I am especially grateful for my parents, Paul and Sue Ellen Renfrew for their guidance, support, encouragement and cookies. I am also grateful and the love of my life Jennifer Cable.

Table of Contents

List of Tables

List of Figures

Introduction

All computational protein design programs attempt to solve what has been called the inverse protein folding problem: given some structural information (usually the protein backbone conformation) find the sequence with the lowest free energy for that structure [1]. There have been significant advances toward this goal. The earliest attempts came from simply looking at the structures of proteins and the commonalities between them; common targets were helical bundle proteins (reviewed by De Grado *et al.* [2] and Richardson *et al.* [3]). Work progressed to improve the packing of designed proteins using algorithms to find compatible side chain conformations with simple secondary structures [4] and eventually known protein backbones serving as scaffolds [5] including the first fully automated full sequence design of a protein by Dahiyat and Mayo [6]. There have been additional advances toward the *de novo* design of proteins with arbitrary shape and unseen topology such as the design of a novel α/β protein fold by Kuhlman *et al.* [7], the design of a 4-helix bundle by Summa *et al.* [8], and the design a β-sheet protein by Kraemer-Pecore *et al.* [9]. Recently protein design has progressed to the point where the design of enzymes is possible [10, 11].

The potential applications of the rational manipulation of proteins are staggering. The long term goal of this research is to adapt the tools of computational protein design to create functional molecules be they traditional proteins or not. The research conducted in this thesis takes the first step toward the goal of a general molecular design program that will have far reaching influence on the creation of therapeutics, biological tools and many other unforeseeable applications.

In this thesis, I modify the molecular modeling program Rosetta so that it can perform protein design simulations with noncanonical amino acids (NCAAs). In particular, I focus on the redesign of peptide-protein interactions with NCAAs incorporated into the peptide sequences. Designing peptides

and proteins is coupled to the extent with which we can sample conformational space [12, 13]. The use of NCAAs will allow us to explore an increased sequence and conformational space. I find that NCAAs with novel geometries allow me to fill voids that are inaccessible to the canonical amino acids (CAAs), while amino acids with novel polar groups allow for new hydrogen bond patterns. In addition, classes of NCAAs have properties, such as resistance to proteolysis, which could be advantageous in the design of effective protein-like therapeutics.

Computational protein design programs generally contain two major parts: an energy function to evaluate the fitness of the sequence for the structure, and a method to sample conformational space [14, 15]. Previously, Rosetta was only able to manipulate the 20 CAAs. It has been modified here and used to increase the binding affinity in peptide-protein interfaces.

Noncanonical Amino Acids

An understanding of how all life on Earth came to use the 20 canonical amino acids is a long standing question in biology. The number of amino acids and their relative abundance has been studied and found to span a variety of functional groups but may not be more diverse than a random sample of the potentially hundreds of pre-biotic amino acids available [16]. It is clear that nature often requires chemistry not available in the 20 CAAs to perform its various functions, as evidenced by the abundance of protein modifications that take place before, during and after translation [17-19]. Additional amino acids are also genetically encoded in special cases, selinocysteine [20] and pyrollysine[21], the 21st and 22nd amino acids, as well as the incorporation of N-formalmethionine (fMET).

The work conducted in this thesis describes the incorporation and use of NCAAs in computational protein design (*in silico*). Experimentally, there are several alternative strategies for incorporating NCAAs into engineered proteins. Solid phase peptide synthesis is the most powerful approach and can be used to build novel backbones as well as side chains. Solid phase synthesis is generally

limited to peptides below 50 amino acids, but can be combined with splicing techniques to create even larger chains [22]. Alternatively, methods have been developed to make recombinant proteins with NCAAs. These methods make use of orthogonal tRNA-codon pairs, orthogonal aminoacyl-tRNA synthetases, and novel codons to specify NCAAs [23, 24]. Here we focus on the redesign of peptides which can be readily made by solid phase synthesis with NCAAs.

NCAAs can be used to modify the function or biophysical properties of naturally occurring proteins. Some interesting and well known examples include the use of selinomethionine in protein crystal phasing [25], the modification of enhanced cyan fluorescent protein by incorporating 4-amino-tryptophan to change its fluorescent properties[26], and the use of hexafluoro-leucine in the cores of coiled-coil proteins to drive equilibrium of a mixture of wild-type and fluorinated from heterodimers to homodimers [27]. According to Wang *et al.* more than 110 NCAAs have been incorporated in to proteins using a variety of experimental methods [24].

NCAAs have been used in computational simulations before but this is the first time that they have been used to this extent in computational protein design. Datta *et al.* redesigned the phenylalanine amino-acyl-tRNA synthatase to use acetylate phenylalanine analogs. The analogs were modeled as if they were a phenylalanine. Ali *et al.* used the D-enantiomers of alanine and proline as well as L-amino-butyric acid when trying to design a 21 residue "miniprotein" [28]. The amino-butyric acid was modeled as a serine. Both of these results used physically based energy functions.

Peptide ligands often incorporate NCAAs and modelers must take this into account. These techniques used in these studies often involve the use of molecular mechanics force fields or the parameterization of torsional parameters based on small molecule structures [29, 30]. The techniques developed for designing these small ligands are only designed to handle small peptides. It is the use of computational protein design tools that make this work novel. The use of computational protein design affords us that ability to rapidly search through large regions of sequence space on large

molecules which is different than the single species that is typically used in molecular dynamics simulations.

Rosetta Modeling Suite

Rosetta is a suite of programs that share common search functions, energy functions and other algorithms. The Rosetta suite is maintained by approximately 100 developers at 15 universities, 2 corporations and 1 national lab. Parts of the Rosetta suite are used in commercial operations as well as hundreds of academic institutions around the world. We benefit from the continuing development efforts of this community and the work conducted here will benefit all users. The most recent version of Rosetta, with which this work was primarily conducted, consists of several libraries. The lower level libraries have routines for evaluating the energy of proteins and manipulating protein backbones and side chains. The higher level libraries use the low level functionality to create complex algorithms and protocols. A suite of applications have been built using these libraries that can perform many protein modeling applications from which Rosetta has been able to achieve its diverse success.

Rosetta Applications

The Rosetta suite, is perhaps best known for its ability to predict protein structures from only their sequences [12] and its success in the Critical Assessment of Techniques for Protein Structure Prediction (CASP, blind protein structure prediction) competition [31]. The structure prediction techniques have recently been extended to membrane proteins with helical trans-membrane domains [32]. The Rosetta structure prediction algorithms have even been applied to the entire Pfam-A database [33]. We incorporate some of the techniques developed for use in structure prediction to design protein-peptide interfaces (chapters 3 and 4).

Predicting how two proteins will bind each other is another area of research that has received significant attention [34, 35] because it impacts many other aspects of protein modeling. Rosetta has

been successful in the Critical Assessment of Predicted Interactions (CAPRI, blind protein-protein docking prediction) competition [36]. We incorporate some of the techniques developed for use in protein-protein docking during the protein-peptide interface designs (chapters 3 and 4).

Rosetta also has tools to manipulate and dock small molecules [37], model protein loops [38, 39], aid the solution of NMR [40] and electron cryomicroscopy structures[41], develop novel protein structure validation tools [42], and very recently has been able to design enzymes that carry out the retro-aldol [43] and Kemp-elimination [11] reactions. Additionally Rosetta is the backend behind the RosettaDesign protein design server [44], the RosettaDoc protein-protein docking server [45], the Robetta structure prediction server [46], the Rosetta@Home distributed computing project (http://boinc.bakerlab.org/rosetta), and a popular protein folding game called FoldIt (http://fold.it).

The focus of this thesis is on the protein design and protein interface design tools of Rosetta. The protein design aspects of Rosetta, specifically, has achieved success in redesigning the folding pathway of the protein G variants NuG1 and NuG2 [47, 48], redesigning a loop in protein L [49], the design of a globular protein (TOP7) with a novel fold [7], used to create a tool to study the ubiquitin pathway [50], to design a single amino acid sequence that can switch between a coiled-coil and a zinc finger [51], used as a tool to find mutations that can increase binding affinity [52], and redesign the loop of an all beta protein [53].

Rotamer Libraries in Computational Protein Design

The theoretical number of conformations a protein sequence can have is on the same scale as the number of atoms in the universe. In the majority of cases folded proteins are at the lowest free energy conformation possible. A close examination however reveals that not all individual components are at the lowest free energy conformation and we see a distribution of values for features like the side chain torsional angles. Since the first protein structures became available it has been seen that these

distributions are not continuous but discrete, with protein side chains being observed in a variety of local free energy minima know as rotamers (short for rotational isomers)[54].

Protein modelers take advantage of these discontinuities to sample the most probable side chain conformations when trying to find the set of side chain conformation with the lowest free energy for a given backbone fold, which dramatically reduces the conformational space that must be searched. Two approaches are generally taken: conformer libraries or rotamer libraries. Conformer libraries are lists of amino acid side chain atomic coordinates that have been taken from high resolution protein structures[15]. These coordinates are oriented on protein backbone conformations during a design simulation. Rotamer libraries are lists of common side chain χ angle values. X angles are the side chain torsional dihedral angles.

In Rosetta, the coordinates for the side chain at a position are built using the χ angles supplied by the rotamer library, assuming that the side chain will have ideal bond lengths, bond angles, and non-χ dihedral angles. The size and quality of rotamer libraries has increased as more and higher quality data is deposited in the protein data bank [54]. Rotamer libraries have been used extensively and achieved a great deal of success. There is some evidence however to suggest that conformer libraries are more accurate in certain situations [55].

One of the fundamental assumptions of the Rosetta suite, is that a protein can be represented simply as its sequence and backbone dihedral angles φ, ψ, ω, and assuming ideal bond lengths and angles [56], an accurate all-atom representation can be built. Sampling torsion space as opposed to Cartesian space reduces the degrees of freedom and mimics the way protein fold in nature. Protein conformations in Rosetta are sampled by changing torsional angles. Therefore, the quality of a predicted conformation is due in large part to the accuracy of the torsional terms of the energy function. The use of rotamer libraries is compatible with this assumption and Rosetta uses the Dunbrack rotamer library for the 20 CAAs [57].

In addition to decreasing the search space, the relative probabilities between rotamers can be used to compute a pseudo-energy (see below). The use of the log probabilities of rotamers is common in computational protein design and is used by Rosetta, as well as several other computational protein design groups [54]. There are two major problems using protein statistics in this manner. First, there are not sufficient protein structures containing NCAAs to generate statistics and the method is therefore incompatible with NCAAs. Second, it can lead to rotamer selection bias. Selection bias arises as a result of the sampling of protein structures and the way they are used in Rosetta. In a helix, for example, the conformation of the amino acid side chain is often constrained to a particular rotamer by the steric clashes with atoms in the neighboring helical turns. The statistics collected for helical φ and ψ not only contain information about the internal energy of the conformation of the side chain but information that is explicitly taken into account in other terms in the energy function. The combination of the physical and knowledge-based potentials in Rosetta's energy function leads to double counting, and bias in the selection of rotamers.

For example if the region of a protein being designed has helical φ and ψ but is not in a helix, Rosetta retains the conformational preferences of the helix. Additionally, in cases where the backbone coordinates are not known with accuracy, such as a disordered region in a protein crystal structure, using the statistically most likely rotamer is the most logical choice. The problems with rotamer libraries are addressed in chapters 1 and 2. In chapter 1 we use quantum mechanics to compute the energies of amino acid dipeptides [58, 59] and use these energies in place of the probabilities of the knowledge-based rotamer library. In chapter 2 we develop and describe methods to create rotamer libraries for NCAAs.

Rosetta Energy Function

An important part of protein design is the ability to differentiate between sequence designs with high and low free energies for a given structure, which is accomplished by the energy or scoring function. The Rosetta suite contains two main energy functions, an all-atom energy function and an energy

function that represents each amino acid side chain as a single sphere called a centroid[60]. The centroid based energy function is primarily used in the protein structure prediction component of Rosetta, and was not used in this thesis, the all-atom energy function is the energy function used primarily in protein design (although both were required for the *de novo* design of TOP7)[7]. The stock Rosetta energy function cannot evaluate the quality of designs involving NCAAs because the energy function includes knowledge-based energy potentials, terms that are based statistics collected from known protein structures.

The energy function used in Rosetta simulations is comprised of both physical and knowledge-based terms as follows: a Leonard-Jones potential [61], a Lazaridis-Karplus solvation term [62], a hydrogen-bonding term [63], a pair interaction term [7], side chain torsional term, amino acid backbone torsional preference, and a reference term that represents the energy of an amino acid in the unfolded state [7, 12].

The energy that comes from the pair term, and amino acid backbone and side chain torsional preference are knowledge-based potentials derived from known protein structures and are incompatible with NCAAs. Ideally an energy function should be able to score both CAAs and NCAAs. In chapter 2 we describe the development of a modified version of the Rosetta energy function that is compatible with both CAAs and NCAAs. Knowledge-based torsional terms are replaced with molecular mechanics torsional and Lennard-Jones terms, and amino acid reference energies are replaced with explicit calculations of amino acids in the unfolded state. We have re-weighted the energy function based on its ability to reproduce native sequences of CAAs.

Lennard-Jones Interactions Energy Term

The Lennard-Jones term is a standard 12-6 potential. The well depths are taken from the CHARMM19 parameter set.

$$E_{LJ} = \sum_{i} \sum_{j>i} e_{ij} \left(\left(\frac{r_{ij}}{d_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}}{d_{ij}} \right)^{6} \right)$$

Where *i* and *j* are atom indices, *d* is the inter atomic distance, *e* is the geometric mean of atomic well depths, and *r* is the summed van der Waals radii. After a given cutoff value, the function is evaluated linearly to adjust for the discrete side chain conformations. Evaluating packing of the backbone and side chains is crucial to successful protein design. This term is calculated on an atomic basis (as opposed to a residue) and is therefore compatible with both NCAAs and CAAs.

Hydrogen bonding Energy Term

Hydrogen bonding is important in stabilizing secondary structure and protein-protein interface interactions. The orientation-dependent hydrogen bonding term is calculated by looking at distances and angles in known protein structures.

$$E_{HB} = \sum_{i} \sum_{j} \left(- \ln \left(P\left(d_{ij} \left| \mathrm{h_j} ss_{ij} \right. \right) \right) + - \ln \left(P\left(\cos \theta_{ij} \left| d_{ij} \, \mathrm{h_j} ss_{ij} \right. \right) \right) + - \ln \left(P\left(cos \, \varphi_{ij} \left| d_{ij} \, \mathrm{h_j} ss_{ij} \right. \right) \right) \right)$$

Where *i* is the donor residue index, *j* is the acceptor residue index, *d* is the acceptor-proton inter atomic distance, *h* is the hybridization (sp2, sp3), θ is the proton-acceptor-acceptor base bond angle, and φ is the donor-proton-acceptor base bond angle. Although the hydrogen bonding potential is parameterized based on crystal structures, it is evaluated on an atomistic level and only dependent on the type of atom, the hybridization, and the angle of the hydrogen bond. The hydrogen bond donors and acceptors on the NCAAs we have added have the same or similar hybridization as those in the CAAs. This term is therefore compatible with NCAAs.

Lazaridis-Karplus Solvation Energy Term

The solvation energy of a design is evaluated based on the implicit solvation model of Lazaridis and Karplus [62].

$$E_{SOLV} = \sum_i \left(\Delta G_i^{ref} - \sum_j \left[\left(\frac{2\Delta G_i^{free}}{4\pi^{3/2}\gamma_j r_{ij}^2} \right) e^{-d_{ij}^2 V_j} + \left(\frac{2\Delta G_i^{free}}{4\pi^{3/2}\gamma_j r_{ij}^2} \right) e^{-d_{ij}^2 V_i} \right] \right)$$

Where *i* and *j* are atom indices, d is the inter atomic distance, *r* is the summed van der Waals radii, γ is the correlation length, *V* is the atomic volume, and ΔG^{ref} and ΔG^{free} are the energy of a fully solvated atom. The solvation term is atomistic and there for compatible with both types of NCAAs. In our model the ΔG^{free} values of several of the atoms have been modified from the original reference to better replicate how often a residue is seen on the surface or buried [7].

Ramachandrin Torsional Energy Term

The energy of the Ramachandrin torsional preferences is the log probability of seeing a residue given the amino acid type and secondary structure.

$$E_{RAMA} = \sum_i -\ln\left(P(\phi_i, \psi_i | aa_i, ss_i)\right)$$

Where *i* is the residue index, φ/ψ are the backbone torsion angles (in 36 degree bins), *aa* is the amino acid type, and *ss* is the secondary structure type. Secondary structure are calculated by Rosetta using the DSSP algorithm [64]. This is one of the knowledge-based terms in the energy function. During a design simulation where the backbone is allowed to move the change in the Rama energy is significant in deciding if those moves should be accepted or rejected. This term was derived by observing the frequency of φ/ψ pairs based on 3 secondary structure regions (helix, sheet, other) for each amino acid.

Residue-pair Interaction Energy Term

$$E_{PAIR} = \sum_i \sum_{j>i} -\ln\left(\frac{P\left(aa_i, aa_j \middle| d_{ij}, env_i, env_j\right)}{P\left(aa_i \middle| d_{ij}\, env_i\right) P\left(aa_j \middle| d_{ij}\, env_j\right)} \right)$$

Where *i* and *j* are the residue indices, *d* is the distance between residues, *aa* is the amino acid type, and *env* is the environment of residue based on the number of neighbors. The residue pair interaction term is based on the frequency of seeing two residues of a given type, a certain distance (distances are binned and are calculated based on the β-carbon from each other in the protein, and given a certain environment (buried or surface) in the protein. This term is only evaluated for polar residues. This term is meant to represent the electrostatic and disulfide bonds formation preferences. This term has been omitted when evaluating NCAAs.

Rotamer Self-Energy Term

$$E_{ROT} = \sum_{i} -\ln\left(\frac{P(rot_i|\phi_i,\psi_i)P(aa_i|\phi_i,\psi_i)}{P(aa_i)}\right)$$

Where *i* is the residue index, *rot* is the Dunbrack backbone-dependant rotamer, *aa* is the amino acid type, and φ and ψ are the backbone torsion angles. The rotamer self energy is dependent upon the probability of seeing a particular amino acid with a particular rotamer given its backbone dihedral angles, and the frequency of an amino acid. The rotamer probabilities came from the rotamer library of Dunbrack and Cohen [57]. In Rosetta it is a measure of the internal energy of the of a side chain. This energy takes into account mainly the torsional preferences, but also the energies contributed by the bond lengths and angles of the side chain when it was sampled in the pdb. This term is knowledge-based and is not compatible with NCAAs and has been replaced with a molecular mechanics torsional and Lennard-Jones potential, discussed in chapter 2.

Unfolded State Reference Energy Term

$$E_{REF} = \sum_{aa} n_{aa}$$

Where *aa* is the amino acid type, and *n* is the number of residues. The energies produced by computational protein design potentials are intended to be a measure of free energy of folding, the energy of the unfolded state is an important factor. For each amino acid, a empirical reference energy

is applied during the weighting of the energy function. This energy represents the energy of the residue in the unfolded protein. We have developed methods, discussed in chapter 2, to evaluate the unfolded energy of NCAAs.

Rosetta Search Function

The purpose of the search function is to explore conformational space and to direct the simulation toward low energy conformations. Search functions for protein design generally fall in to two categories: stochastic search functions such as Monte Carlo and genetic algorithms, or deterministic search functions such as dead end elimination or self-consistent mean field algorithms [65]. A compromise is made in the choice of search algorithm between speed and accuracy. For example with dead end elimination, if the search converges it is guaranteed to be the global energy minimum. A Monte Carlo search algorithm does not necessarily find the global minimum however it can be considerably faster. Rosetta uses a Monte Carlo search algorithm because of its speed.

Given a backbone template a typical sequence design simulation in Rosetta proceeds as follows. The residues to be designed are selected as well as the set of amino acids each position may mutate too (this could be a full sequence design allowing every amino acid at all positions or a partial allowing only a subset). A set of probable side chain conformations are selected for each sequence position from the rotamer library of Dunbrack and Cohen [57]. In most cases, the energies of all pairs of rotamer combinations are calculated using the pair-wise decomposable energy function described above. The search function then randomly changes the rotamer or amino acid identity at a residue and evaluates the new energy and either accepts or rejects the change based on the Metropolis criterion[66]. The Metropolis criterion automatically accepts changes that lower the energy and uses a Boltzmann probability to evaluate changes that raise the energy (see below). P is the probability of accepting the change, E_{old} and E_{new} are the energy of the protein before and after the change respectfully, kB is the Boltzmann constant, and T is a temperature.

$$P_{ACCEPT} = e^{-\left(\frac{E_{new} - E_{old}}{k_B T}\right)}$$

The search function further also seeks the global minimum though a technique called simulated annealing where the temperature is set high to begin and lowered as the search progresses. This allows a large conformational search which eventually narrows to a local minimum as changes that have a large affect on the energy become less likely.

This Monte Carlo simulated annealing procedure is all that is used for fixed backbone design. For flexible backbone design, we iterate between the above protocol and a step where a conformational perturbation is made to the backbones. There are several different chemically relevant perturbations that are used to search through backbone conformational space. "Small" perturbations are 1-3 degree rotations about φ or ψ. "Shear" perturbations are when ψ is changed and then φ is changed by the same amount in the opposite direction. "Wobble" perturbations are a section of 1 to 3 residue is swapped out for a another section from a fragment library made from know protein structures [60]. Wobble moves are not compatible with NCAAs as with the knowledge-based scoring terms, there are not enough structures that contain NCAAs to generate a fragment library. "Backrub" perturbations are a combination of a rotation about 2 backbone torsional angles and a compensating bending of the bond angles formed between the N, CA, and CB [67]. Two other perturbations involved making a break in the backbone of the protein, varying the backbone dihedral angles near the break and then relinking the chain in a new conformation using a cyclic coordinate descent algorithm or kinematic loop closure algorithm [38]. Additionally backbones of different protein chains can be rotated or translated relative to each other.

Peptide/Protein Design Model Systems

I have used two peptide/protein systems to test the modifications made to Rosetta to enable it to use NCAAs. Both systems involved the redesign of a peptide/protein interface by incorporating NCAAs in the peptide.

Calpain/Calpastatin System

Calpain is a ubiquitous cysteine protease[68]. Calpain functions as a heterodimer made up of two subunits. An 80 kd subunit comprises the first 4 domains (DI-DIV) and includes the catalytic domain. A 30 kd subunit comprised of last two domains (DV and DVI) helps regulate the protease. Upon calcium binding by domains DII, DIV, and DVI, calpain under goes a conformational change that activates the enzyme. The conformational change also allows calpains inhibitor, calpastatin, to bind[69, 70]. Todd *et al.* determined the structure of a 19 residue subdomain of calpastatin binding to domain DVI of calpain[71]. This is an ideal system to test the use of NCAAs because of the small peptide size and the large hydrophobic pocket on DVI. Rosetta predicts several mutations that could increase the binding affinity of the peptide for the protein. Experimental validation of the predictions is ongoing. Results obtained indicate that designs predicted by Rosetta are have lower disassociation constant than the native sequence. The calpain/calpastatin system, the design methodology, and experimental validation are discussed in chapter 3.

HIV GP41/PIE12 System

The HIV gp41 protein is responsible for bringing the HIV virus membrane in proximity to the host cell membrane allowing for membrane fusion and viral entry[72]. The integration process is starts by the HIV gp120 protein binding to 2 receptors on the host cell, CD4 and member of the chemokine family. Upon receptor binding, conformational changes in gp120 induce conformational changes in gp41 that cause it to extend and its transmembrane domain to penetrate the host cells membrane, linking the virus to the host cell and forming what is called the pre-hairpin complex. For fusion to occur, gp41 must undergo a second conformational change where it folds back on itself, forming a hairpin, and pulling the virus and host cell membrane together. The conformational change and resulting membrane fusion has been inhibited by molecules that bind to a conserved region of gp41 that is exposed in the pre-hairpin complex[73]. Recently the Kay lab at the University of Utah has designed inhibitors made out of the D-enatiomers of the canonical amino acids[74]. D-peptides were

used primarily because they are resistant to protolysis[75] which is a problem for protein therapeutics made of the L-enantiomers[76]. We are collaborating with the Kay lab to use the D-enantiomers of the NCAAs added to Rosetta to try to increase the binding affinity of the inhibitory peptide for gp41. Experimental validation of the designs predicted to increase the binding affinity is currently ongoing. The gp41/PIE system, the design methodology, and the predicted designs are discussed in chapter 4.

The incorporation of NCAAs in to our computational protein design program Rosetta is the first logical step towards the goal of generalized molecular design. NCAAs are a powerful tool that will enable us to design new biological tools, strengthen protein-protein interfaces, and design or improve protein therapeutics.

Bibliography

1. Drexler, K.E., *Molecular engineering: An approach to the development of general capabilities for molecular manipulation.* Proceedings of the National Academy of Sciences of the United States of America, 1981. **78**(9): p. 5275-5278.

2. DeGrado, W.F., Z.R. Wasserman, and J.D. Lear, *Protein design, a minimalist approach.* Science, 1989. **243**(4891): p. 622-628.

3. Richardson, J.S. and D.C. Richardson, *The de novo design of protein structures.* Trends in Biochemical Sciences, 1989. **14**(7): p. 304-309.

4. Schneider, J.P., A. Lombardi, and W.F. DeGrado, *Analysis and design of three-stranded coiled coils and three-helix bundles.* Folding & Design, 1998. **3**(2): p. R29-40-R29-40.

5. Malakauskas, S.M. and S.L. Mayo, *Design, structure and stability of a hyperthermophilic protein variant.* Nat Struct Mol Biol, 1998. **5**(6): p. 470-475.

6. Dahiyat, B.I. and S.L. Mayo, *De Novo Protein Design: Fully Automated Sequence Selection.* Science, 1997. **278**(5335): p. 82-87.

7. Kuhlman, B., et al., *Design of a Novel Globular Protein Fold with Atomic-Level Accuracy.* Science, 2003. **302**(5649): p. 1364-1368.

8. Summa, C.M., et al., *Computational de novo Design, and Characterization of an A2B2 Diiron Protein.* Journal of Molecular Biology, 2002. **321**(5): p. 923-938.

9. Kraemer-Pecore, C.M., J.T.J. Lecomte, and J.R. Desjarlais, *A de novo redesign of the WW domain.* Protein Science, 2003. **12**(10): p. 2194-2205.

10. Jain, T., *Configurational-bias sampling technique for predicting side-chain conformations in proteins.* Protein Science, 2006. **15**(9): p. 2029-2039.

11. Rothlisberger, D., et al., *Kemp elimination catalysts by computational enzyme design.* Nature, 2008. **453**(7192): p. 190-5.

12. Bradley, P., K.M.S. Misura, and D. Baker, *Toward High-Resolution de Novo Structure Prediction for Small Proteins.* Science, 2005. **309**(5742): p. 1868-1871.

13. Butterfoss, G.L. and B. Kuhlman, *COMPUTER-BASED DESIGN OF NOVEL PROTEIN STRUCTURES.* Annual Review of Biophysics and Biomolecular Structure, 2006. **35**(1): p. 49-65.

14. Boas, F.E. and P.B. Harbury, *Potential energy functions for protein design.* Curr Opin Struct Biol, 2007. **17**(2): p. 199-204.

15. Gordon, D.B., S.A. Marshall, and S.L. Mayo, *Energy functions for protein design.* Curr Opin Struct Biol, 1999. **9**(4): p. 509-13.

16. Lu, Y. and S.J. Freeland, *A quantitative investigation of the chemical space surrounding amino acid alphabet formation.* Journal of Theoretical Biology, 2008. **250**(2): p. 349 - 361-349 - 361.

17. Creasy, D.M. and J.S. Cottrell, *Unimod: Protein modifications for mass spectrometry.* PROTEOMICS, 2004. **4**(6): p. 1534-1536.

18. Garavelli, J.S., *The RESID Database of Protein Modifications as a resource and annotation tool.* PROTEOMICS, 2004. **4**(6): p. 1527-1533.

19. Hornbeck, P.V., et al., *PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation.* PROTEOMICS, 2004. **4**(6): p. 1551-1561.

20. Bock, A., et al., *Selenocysteine: the 21st amino acid.* Mol Microbiol, 1991. **5**(3): p. 515-20.

21. Srinivasan, G., C.M. James, and J.A. Krzycki, *Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA.* Science, 2002. **296**(5572): p. 1459-62.

22. Muir, T.W., *Semisynthesis of proteins by expressed protein ligation.* Annu Rev Biochem, 2003. **72**: p. 249-89.

23. Hendrickson, T.L., V.d. Crecy-Lagard, and P. Schimmel, *INCORPORATION OF NONNATURAL AMINO ACIDS INTO PROTEINS.* Annual Review of Biochemistry, 2004. **73**(1): p. 147-176.

24. Wang, L., J. Xie, and P.G. Schultz, *EXPANDING THE GENETIC CODE.* Annual Review of Biophysics and Biomolecular Structure, 2006. **35**(1): p. 225-249.

25. Hendrickson, W.A., J.R. Horton, and D.M. LeMaster, *Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure.* EMBO J, 1990. **9**(5): p. 1665-72.

26. Bae, J.H., et al., *Expansion of the genetic code enables design of a novel "gold" class of green fluorescent proteins.* J Mol Biol, 2003. **328**(5): p. 1071-81.

27. Bilgiçer, B. and K. Kumar, *Synthesis and thermodynamic characterization of self-sorting coiled coils.* Tetrahedron, 2002. **58**(20): p. 4105-4112.

28. Ali, M.H., et al., *Design of a Heterospecific, Tetrameric, 21-Residue Miniprotein with Mixed [alpha]/[beta] Structure.* Structure, 2005. **13**(2): p. 225-234.

29. Bohm, H.J., *Towards the automatic design of synthetically accessible protein ligands: peptides, amides and peptidomimetics.* J Comput Aided Mol Des, 1996. **10**(4): p. 265-72.

30. Klebe, G. and T. Mietzner, *A fast and efficient method to generate biologically relevant conformations.* J Comput Aided Mol Des, 1994. **8**(5): p. 583-606.

31. Lange, O., et al., *Structure prediction for CASP8 with all-atom refinement using Rosetta.* Proteins: Structure, Function, and Bioinformatics, 2009. **9999**(9999): p. NA-NA.

32. Barth, P., B. Wallner, and D. Baker, *Prediction of membrane protein structures with complex topologies using limited constraints.* Proceedings of the National Academy of Sciences, 2009. **106**(5): p. 1409-1414.

33. Bonneau, R., et al., *De Novo Prediction of Three-dimensional Structures for Major Protein Families.* Journal of Molecular Biology, 2002. **322**(1): p. 65-78.

34. Gray, J.J., *High-resolution protein-protein docking.* Current Opinion in Structural Biology, 2006. **16**(2): p. 183-193.

35. Wang, C., P. Bradley, and D. Baker, *Protein-Protein Docking with Backbone Flexibility.* Journal of Molecular Biology, 2007. **373**(2): p. 503-519.

36. Wang, C., et al., *RosettaDock in CAPRI rounds 6-12.* Proteins: Structure, Function, and Bioinformatics, 2007. **69**(4): p. 758-763.

37. Meiler, J. and D. Baker, *ROSETTALIGAND: Protein-small molecule docking with full side-chain flexibility.* Proteins: Structure, Function, and Bioinformatics, 2006. **65**(3): p. 538-548.

38. Mandell, D.J., E.A. Coutsias, and T. Kortemme, *Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling.* Nat Meth, 2009. **6**(8): p. 551-552.

39. Rohl, C.A., et al., *Modeling structurally variable regions in homologous proteins with rosetta.* Proteins: Structure, Function, and Bioinformatics, 2004. **55**(3): p. 656-677.

40. Rohl, C.A. and T.L. James, *Protein Structure Estimation from Minimal Restraints Using Rosetta*, in *Nuclear Magnetic Resonance of Biological Macromolecules*. 2005, Academic Press. p. 244 - 260-244 - 260.

41. DiMaio, F., et al., *Refinement of Protein Structures into Low-Resolution Density Maps Using Rosetta*. Journal of Molecular Biology, 2009. **392**(1): p. 181-190.

42. Sheffler, W. and D. Baker, *RosettaHoles: Rapid assessment of protein core packing for structure prediction, refinement, design, and validation.* Protein Science, 2009. **18**(1): p. 229-239.

43. Barbas, C.F., et al., *De Novo Computational Design of Retro-Aldol Enzymes.* Science, 2008. **319**(5868): p. 1387-1391.

44. Liu, Y. and B. Kuhlman, *RosettaDesign server for protein design.* Nucl. Acids Res., 2006. **34**(suppl_2): p. W235-238-W235-238.

45. Lyskov, S. and J.J. Gray, *The RosettaDock server for local protein-protein docking*. Nucl. Acids Res., 2008. **36**(suppl_2): p. W233-238-W233-238.

46. Chivian, D., et al., *Prediction of CASP6 structures using automated robetta protocols.* Proteins: Structure, Function, and Bioinformatics, 2005. **61**(S7): p. 157-166.

47. Nauli, S., B. Kuhlman, and D. Baker, *Computer-based redesign of a protein folding pathway*. Nat Struct Mol Biol, 2001. **8**(7): p. 602-605.

48. Nauli, S., et al., *Crystal structures and increased stabilization of the protein G variants with switched folding pathways NuG1 and NuG2.* Protein Science, 2002. **11**(12): p. 2924-2931.

49. Kuhlman, B., et al., *Conversion of monomeric protein L to an obligate dimer by computational protein design.* Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10687-91.

50. Eletr, Z.M., et al., *E2 conjugating enzymes must disengage from their E1 enzymes before E3-dependent ubiquitin and ubiquitin-like transfer*. Nat Struct Mol Biol, 2005. **12**(10): p. 933-934.

51. Ambroggio, X.I. and B. Kuhlman, *Computational Design of a Single Amino Acid Sequence that Can Switch between Two Distinct Protein Folds.* Journal of the American Chemical Society, 2006. **128**(4): p. 1154-1161.

52. Sammond, D.W., et al., *Structure-based protocol for identifying mutations that enhance protein-protein binding affinities.* J Mol Biol, 2007. **371**(5): p. 1392-404.

53. Hu, X., et al., *High-resolution design of a protein loop.* Proc Natl Acad Sci U S A, 2007. **104**(45): p. 17668-73.

54. Dunbrack, R.L., *Rotamer Libraries in the 21st Century.* Current Opinion in Structural Biology, 2002. **12**(4): p. 431-440.

55. Shetty, R.P., et al., *Advantages of fine-grained side chain conformer libraries.* Protein Eng., 2003. **16**(12): p. 963-969.

56. Engh, R.A. and R. Huber, *Accurate bond and angle parameters for X-ray protein structure refinement.* Acta Crystallographica Section A, 1991. **47**(4): p. 392-400.

57. Jr, R.L.D. and F.E. Cohen, *Bayesian statistical analysis of protein side-chain rotamer preferences.* Protein Science : A Publication of the Protein Society, 1997. **6**(8): p. 1661–1681-1661–1681.

58. Butterfoss, G.L. and J. Hermans, *Boltzmann-type distribution of side-chain conformation in proteins.* Protein Science, 2003. **12**(12): p. 2719-2731.

59. Butterfoss, G.L., J.S. Richardson, and J. Hermans, *Protein imperfections: separating intrinsic from extrinsic variation of torsion angles.* Acta Crystallographica Section D, 2005. **61**(1): p. 88-98.

60. Rohl, C.A., et al., *Protein Structure Prediction Using Rosetta*, in *Numerical Computer Methods, Part D*. 2004, Academic Press. p. 66 - 93-66 - 93.

61. Neria, E., S. Fischer, and M. Karplus, *Simulation of activation free energies in molecular systems.* The Journal of Chemical Physics, 1996. **105**(5): p. 1902-1921.

62. Lazaridis, T. and M. Karplus, *Effective energy function for proteins in solution.* Proteins: Structure, Function, and Genetics, 1999. **35**(2): p. 133-152.

63. Kortemme, T., A.V. Morozov, and D. Baker, *An Orientation-dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein-Protein Complexes.* Journal of Molecular Biology, 2003. **326**(4): p. 1239-1259.

64. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-637.

65. Voigt, C.A., D.B. Gordon, and S.L. Mayo, *Trading accuracy for speed: a quantitative comparison of search algorithms in protein sequence design.* Journal of Molecular Biology, 2000. **299**(3): p. 789-803.

66. Metropolis, N., et al., *Equation of State Calculations by Fast Computing Machines.* The Journal of Chemical Physics, 1953. **21**(6): p. 1087-1092.

67. Smith, C.A. and T. Kortemme, *Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction.* J Mol Biol, 2008. **380**(4): p. 742-56.

68. Goll, D.E., et al., *The calpain system.* Physiol Rev, 2003. **83**(3): p. 731-801.

69. Hanna, R.A., R.L. Campbell, and P.L. Davies, *Calcium-bound structure of calpain and its mechanism of inhibition by calpastatin.* Nature, 2008. **456**(7220): p. 409-12.

70. Moldoveanu, T., K. Gehring, and D.R. Green, *Concerted multi-pronged attack by calpastatin to occlude the catalytic cleft of heterodimeric calpains.* Nature, 2008. **456**(7220): p. 404-8.

71. Todd, B., et al., *A Structural Model for the Inhibition of Calpain by Calpastatin: Crystal Structures of the Native Domain VI of Calpain and its Complexes with Calpastatin Peptide and a Small Molecule Inhibitor.* Journal of Molecular Biology, 2003. **328**(1): p. 131-146.

72. Eckert, D.M. and P.S. Kim, *Mechanisms of viral membrane fusion and its inhibition.* Annu Rev Biochem, 2001. **70**: p. 777-810.

73. Leonard, J.T. and K. Roy, *The HIV entry inhibitors revisited.* Curr Med Chem, 2006. **13**(8): p. 911-34.

74. Welch, B.D., et al., *Potent D-peptide inhibitors of HIV-1 entry*. Proc Natl Acad Sci U S A, 2007. **104**(43): p. 16828-33.

75. Milton, R.C., S.C. Milton, and S.B. Kent, *Total chemical synthesis of a D-enzyme: the enantiomers of HIV-1 protease show reciprocal chiral substrate specificity [corrected]*. Science, 1992. **256**(5062): p. 1445-8.

76. Fung, H.B. and Y. Guo, *Enfuvirtide: a fusion inhibitor for the treatment of HIV infection.* Clin Ther, 2004. **26**(3): p. 352-78.

Using Quantum Mechanics to Improve Estimates of Amino Acid Side Chain Rotamer Energies

P. Douglas Renfrew, Glenn Butterfoss, Brian Kuhlman

Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27516

This work was published in Proteins: Structure, Function and Bioinformatics 2008 Jun;71(4):1637-46

Abstract

Amino acid side chains adopt a discrete set of favorable conformations typically referred to as rotamers. The relative energies of rotamers partially determine which side chain conformations are more often observed in protein structures and accurate estimates of these energies are important for predicting protein structure and designing new proteins. Protein modelers typically calculate side chain rotamer energies by using molecular mechanics (MM) potentials or by converting rotamer probabilities from the protein database (PDB) into relative free energies. One limitation of the knowledge-based energies is that rotamer preferences observed in the PDB can reflect internal side chain energies as well as longer-range interactions with the rest of the protein. Here, we test an alternative approach for calculating rotamer energies. We use three different quantum mechanics (QM) methods (second order Moller-Plesset (MP2), density functional theory (DFT) energy calculation using the B3LYP functional, and Hartree-Fock) to calculate the energy of amino acid rotamers in a dipeptide model system, and then use these pre-calculated values in side chain placement simulations. Energies were calculated for over 35,000 different conformations of leucine, isoleucine and valine dipeptides with backbone torsion angles from the helical and strand regions of the Ramachandran plot. In a subset of cases these energies differ significantly from those calculated

with standard molecular mechanics potentials or those derived from PDB statistics. We find that in these cases the energies from the QM methods result in more accurate placement of amino acid side chains in structure prediction tests.

Introduction

Amino acid side chains adopt a variety of conformations. An accurate estimate of the relative energies of different side chain conformations is essential for high resolution structure prediction, protein design and modeling protein dynamics. These energies are generally calculated by using molecular mechanics (MM) potentials or by deriving energies from the probability of observing a particular side chain conformation in the PDB [1] [2]. Most MM potentials use empirically derived functions to model the energetics of bond stretching, bending and torsion angle perturbation [3]. Non-bonded interactions are generally modeled with a Lennard-Jones potential and a form of Coulomb's potential to model electrostatics. MM potentials are often parameterized to match results from quantum mechanics (QM) calculations on model compounds. The advantage of MM potentials is that they are generalizable to a variety of atom types, they are fast to evaluate and the same force field can be applied throughout a molecule. For instance, the same MM expressions can be used to model energetics within an amino acid side chain as between side chains. A limitation of MM potentials is that the calculated energies are sensitive to the model systems used to parameterize them [4]. Different MM potentials often give different answers when evaluating the same set of molecules, for instance, producing a MM potential that accurately represents the torsional preferences of a peptide backbone has proven to be difficult [5-7].

A common alternative to MM potentials are knowledge-based energy functions. Comparative analysis of amino acid side chains in protein structures has shown that most side chains only adopt a limited set of conformations, typically referred to as rotamers [8]. Additionally, some rotamers of an amino acid are observed more often than others, suggesting that the internal energies of the various rotamers are not equal. Using protein structures from the PDB, databases have been constructed,

commonly referred to as rotamer libraries, that specify the most commonly observed torsion angles associated with each rotamer of an amino acid, and the frequency that the various rotamers are observed in the protein database, most recently reviewed by Dunbrack[9]. Because rotamer probabilities depend on the local environment of a side chain, the probabilities are often measured as a function of the backbone dihedral angles of a residue, or as a function of secondary structure [10]. Rotamer probabilities are typically converted to energy by assuming Boltzmann sampling and taking the logarithm of the probability[11]. This assumption was supported in one case by showing that the relative favorability of different methionine rotamers as determined by high level quantum mechanics simulations matches the preference of methionine to adapt a particular rotamer in the PDB[12]. Knowledge-based torsional preferences have been used with good success to predict the conformations of amino acid side chains and to design new protein structures and functions [13, 14].

However, there are situations in which a knowledge-based approach may lead to an inaccurate estimate of protein energy. Particularly challenging is making sure that the knowledge-based term does not represent energies that are included in other terms in the energy function. For example, many modeling programs use a Lennard-Jones (LJ) potential evaluated between pairs of atoms to model van der Waals forces and steric repulsion. If the LJ potential is evaluated between pairs of atoms that contribute to rotamer probabilities observed in the PDB, then there will be double counting. In some cases it is clear which atom pairs to ignore to prevent double counting; if rotamer statistics are being used to evaluate side chain preferences than atom pairs within a side chain should not be considered. It is less clear if atom pair energies should be considered with backbone atoms in the neighboring residue. It will depend in part if the rotamer statistics are compiled as a function of the protein backbone dihedral angles. Because the amide group of the following residue (i+1) and the carbonyl group of the preceding residue (I -1) are determined by the phi and ψ angles of the central residue (i), it can be argued that atom pair energies should not be calculated between these groups and the side chain of i. Potentially even more subtle are longer range interactions commonly observed in

protein secondary structure. In a helix, the preferred side chain conformation at residue i is determined in part by interactions with the residue at positions i-3 and i-4, and therefore, the energetics of this interaction is folded into rotamer statistics of helical residues from the PDB.

Instead of using MM potentials or knowledge-based potentials to calculate protein energetics, an alternative approach is to use direct quantum mechanics (QM) calculations. Energies from QM calculations have been shown to more accurately reproduce backbone and side chain dihedral preferences in the PDB[5, 12]. A crucial limitation of QM is that in general it can not be applied to full-sized proteins, and even for a single amino acid most QM simulations require on the order of minutes to perform. For this reason it is not feasible to perform a QM simulation on every structure that is created during a protein design simulation or a molecular dynamics simulation. However, because only a limited set of side chain conformations are observed during a protein simulation, it is possible to precompute the energy of a side chain in various conformations with QM, and then use these energies during protein simulations. Here, we explore this approach by precomputing energies of ~35,000 conformations of valine, isoleucine and leucine with QM calculations, and then test these energies in side chain prediction tests on full-size proteins. We find that in situations where knowledge-based potentials are more likely to double count or miscount interactions, that the QM energies provide more accurate side chain predictions.

Materials and Methods

Dipeptides

To calculate the internal energies of amino acid side chain rotamers QM and MM calculations were performed on amino acid dipeptides (ACE-X-NME, where X is the amino acid being tested) (figure 1). The dipeptide is commonly used to probe side chain energetics because the relative positions of all the atoms in a dipeptide are primarily determined by phi/psi and the side chain chi angle of a single residue. Backbone and side chain dihedral angles were fixed to their desired values during the

calculations. Backbone dihedral angles were sampled combinatorially in regions of phi/psi space that correspond to α-helical and β-strand conformations (α: phi = -70 to -40 and psi = -50 to -20, β: phi = -110 to -160 and psi = 110 to 160) in ten degree intervals. Chi angles were sampled at their canonical angles (-60, 60, 180) and ±10, ±20, and ±30 degrees; resulting in 336 valine-α structures, 7056 isoleucine-α/leucine-α structures, 504 valine-β structures, and 10584 isoleucine-β/leucine-β structures. When referring to the various rotamers we use the nomenclature established in Lovell *et al*. [10] where "m" is minus gauche (-g, ~-60), "p" is plus gauche (+g, ~+60), and "t" is trans (t, ~180).

MM

Molecular mechanics simulations on the constrained dipeptides were carried out using the CHARMM force field (version 22) [15] and Cedar molecular mechanics force fields as implemented in the molecular mechanics package Sigma [16, 17]. The phi, psi, and chi dihedral angles of each dipeptide were constrained using a 1000 kcal / mol force. To optimize bond angles, bond lengths, and unconstrained dihedrals, structures were put through 2000 rounds of conjugate gradient minimization. Amber version 9 with the FF99 force field was also used to evaluate the energies of the dipeptides. As with the Cedar and CHARMM methods, the phi, psi, and chi dihedrals angles were constrained using a 5000 kcal / mol force and structure optimization was done with 5000 cycles of conjugate gradient minimization. The dihedral constraint energy was not included in the final calculated energies.

QM

Quantum mechanics calculations were carried out using Gaussian03 from Gaussian Inc.[18]. Energies were calculated by first performing a Hartree Fock (HF) minimization followed by a second order Moller-Plesset (MP2) energy calculation and a density functional theory (DFT) energy calculation using the B3LYP functional. In addition to the MP2 and DFT energies the final energy from the HF minimization was also used in the tests described below. All calculations were

performed with the 6-31G(d) basis set except where noted. The HF minimization is the slowest step in this process. To shorten the time of the calculations and to prevent large clashes that can occur when the dihedral angles of a starting structure are rotated and fixed, the starting structure used for each minimization varied by only ten degrees in either phi, psi or one of the chi dihedrals from the target set of angles. For example a minimized valine dipeptide with phi = -60, psi = -40, chi1 = -60, would be allowed to serve as a starting structure for ((-70 or -50), -40, -60), (-60, (-50 or -30), -60), and (-60, -40, (-70 or -50)).

Each class of calculations (valine-α, valine-β, isoleucine-α, isoleucine-β, leucine-α, leucine-β) had one set of phi/psi values tested with a larger basis (6-31+G(d), 6-311+G(d)) set to see if increasing the size of the basis set lead to improvements in the rotamer prediction benchmarks. Only the leucine α class of dipeptides showed improvement with an increased basis set (6-31+G(d)) and the entire phi/psi range was rerun using this larger basis set.

Calculations were performed on either a IBM P690 Model 681 running AIX or an SGI Altix 3700bx2 running RedHat Enterprise Linux 3 maintained by the UNC Information Technology Services (http://its.unc.edu) (both) or the National Center for Supercomputing Applications (IBM P690). Calculations on either machine take ~1 hour of CPU time per structure with the 6-31G(d) basis set and ~3 hours using the 6-31+G(d) basis set.

Knowledge-based rotamer energies

Knowledge-based rotamer energies were computed using the protein modeling program Rosetta:

$$E_{rotamer}(rot, aa, \phi, \psi, \vec{\chi}) = -RT \ln\left[P_{chi}(\vec{\chi} \mid rot, aa, \phi, \psi) * P_{rot}(rot \mid aa, \phi, \psi)\right] \qquad (1)$$

where P_{chi} is the probability that a particular rotamer will have a certain set of chi angles ($\vec{\chi}$), P_{rot} is the probability that, given phi and psi, a particular amino acid (*aa*) will adopt a particular rotamer (*rot*). P_{rot} is taken directly from Dunbrack's most recent rotamer library (http://dunbrack.fccc.edu/).

P_{chi} is determined using the standard deviations included in Dunbrack's library assuming each side chain torsion angle ($\chi 1$, $\chi 2$, …) is independent of the other torsion angles:

$$P_{chi}(\vec{\chi}, rot, aa, \phi, \psi) = \prod_{i}^{\max\_\chi} P(\chi(i) \mid rot, aa, \phi, \psi) \qquad (2)$$

$$P(\chi(i) \mid rot, aa, \phi, \psi) = \left(\frac{1}{\sqrt{2\pi}\sigma_{\chi}} \right) \exp\left(\frac{\left(\chi(i) - \bar{\chi}(i, rot, aa, \phi, \psi)\right)^2}{2 * \sigma_{\chi}^2} \right) \qquad (3)$$

Where $\chi(i)$ is the torsion angle for the ith chi angle, $\bar{\chi}$ is the average value for that chi angle for a particular rotamer, and $\sigma_{\chi}$is the standard deviation for that chi angle for the same rotamer.

Side chain prediction tests

To determine the usefulness of the different methods for calculating the relative energies of amino acid rotamers, we tested them to see how accurately they could reproduce native side chain conformations from a set of ~2800 protein structures with a resolution not higher than 2.0 angstroms [19]. In these tests the side chain of a residue was removed and rebuilt with Rosetta in the context of the whole protein. Neighboring residues were held fixed. The energy of each rotamer in the context of the whole protein was calculated by adding the intrinsic energy of the rotamer, as determined by the theoretical calculations on the dipeptides, to the standard Rosetta energy. Because the theoretical calculations were performed for 10 degree increments of phi, psi and chi angles, linear interpolation was used to estimate the energy for a specific set of torsion angles. The knowledge-based term usually used to evaluate internal rotamer preferences was removed from the Rosetta energy function except in the cases in which it was being tested. The lowest energy rotamer in the context of the whole protein was taken as the Rosetta prediction. The test was performed for all valine, isoleucine and leucine residues with phi and psi angles in the range covered by the QM simulations. Each side chain was sampled at its most probable chi angle (as given by Dunbrack's backbone dependent rotamer library) as well as chi angles that varied ±0.5, ±1, ±1.5, and ±2 standard deviations away

from the mean (again as given by Dunbrack's backbone dependent library). This results in 27 rotamers for each valine and 729 rotamers for isoleucine and leucine.

To insure that the position of the side chain was well-defined in the crystal structure, residues were only used for the side-chain replacement test if all atoms of the side chain in the crystal structure had B-factors less than 20. In the case of leucine rotamers, if the native conformation in the crystal structure was one of the commonly mistakenly assigned mp* and tt* rotamers [10], the position was omitted.

Analysis of side chain prediction results

A number of statistics were gathered to determine how well side chain conformations were predicted.

Percent Total correct: The percent of residue positions where the side chain conformation was correctly predicted. A prediction was considered correct if all chi angles in the predicted side chain were within the same torsional basin as the native side chain.

Percent Correct and Chi Free: The percent of residue positions where the rotamer was correctly predicted given that the position was "free." A residue is considered free if the preferred side chain conformation is not primarily determined by repulsive interactions with neighboring residues. We define a position to be free if the repulsive energy as computed by Rosetta between the side chain and neighboring residues is less than 0.5 kcal / mol for at least two alternate side chain conformations, where the conformations differ by more than 60 degrees in at least one of their chi angles. For Percent Correct and Chi1 Free, a position is only considered free if 2 conformations with low repulsive energies have chi1 angles that differ by more than 60 degrees.

Percent Minimum Energy and Closest Chi: The percent of positions where the dihedral angle with the lowest energy is the closest to that of the native angle of all the dihedrals tested.

Standard Rosetta Energy Function

The Rosetta energy function has been described previously [20]. Directly relevant to this study is the 12-6 Lennard-Jones potential that is used to evaluate van der Waals forces and steric repulsion. This potential is evaluated between most pairs of atoms in the protein. It is not evaluated between atoms within a residue. In addition it is not evaluated between the amide group and C_α of residue i+1 and the atoms in i, and it is not evaluated between the carbonyl group and C_α of the preceding residue (i-1) and the atoms in i. These interactions are left out because these interactions should be accounted for by backbone dependent rotamer energies derived from PDB statistics. These interactions with the neighboring backbone atoms will also contribute to the energies calculated for the dipeptides, and therefore it is appropriate that these energies are not included in the Rosetta Lennard-Jones calculations. Explicit hydrogens are modeled on all atoms, but they are only used to check for steric overlap and only contribute to the energy of the protein when they have Lennard-Jones energies that are greater than zero. The van der Waals radii and Lennard-Jones well depths have been described previously [20].

Results

QM and MM energy calculations were performed on dipeptides of valine, isoleucine and leucine with a variety of side chain conformations and phi and psi angles from either the α-helical or β-strand regions of the Ramachandran plot. In many cases, the QM energies from the final step of the HF minimization or the MP2 or DFT energy calculations were significantly different from those calculated with the CHARMM22, Cedar or Amber force fields. For example, for a valine dipeptide with a phi of -60° and a psi of -40° the m (chi1 ~ -60°) rotamer is predicted by the CHARMM22 force field to be -0.8 kcal / mol more favorable than the p rotamer (chi 1 ~ 60°) (figure 2). QM calculations at the MP2 level predict the opposite; the p rotamer is predicted to be -0.6 kcal / mol more favorable than the m rotamer. These are significant differences when one considers that proteins are only stable by a few kcals / mol. In some cases, the most preferred chi angle for each rotamer also differed between the QM and MM simulations. For the valine dipeptide the QM

calculations of the HF energy preferred a chi1 near 170° for the t rotamer while the CHARMM22 force field favors a chi1 near 190° (figure 2). A complete list of calculated energies is provided in the supplementary material.

Energy calculations with dipeptides that have different phi and psi angles highlight the importance of interactions between the side chain and the local backbone. The relative energies of the rotamers often shift dramatically with just small changes in one of the backbone dihedral angles: for valine with a phi -50° and a psi of -30° the QM calculations (MP2) predict that the t rotamer is 0.8 kcal / mol less favorable than the m rotamer, when psi is shifted to -50° the situation is reversed and the t rotamer is predicted to be 1.1 kcal / mol more favorable than the m rotamer (figure 3, table 1). This dramatic change with such a small change in psi reflects interactions between the backbone carbonyl oxygen and the side chain methyl groups on valine. In general when the backbone torsion angles are varied, the energies calculated with the 3 MM potentials follow the same trends observed with the 3 QM calculations. The strong dependence of rotamer energies on phi and psi indicates that if precomputed rotamer energies are to be used during protein simulations, they should be calculated as a function of phi and psi, and phi and psi should be sampled at least every 10 degrees.

To compare the QM and MM energies with rotamer statistics from the PDB, the energies were converted to rotamer probabilities assuming a Boltzmann distribution and a temperature of 298 K (figure 4). Overall agreement between two methods for a single amino acid and backbone conformation was measured by computing the root mean square deviation between the probabilities of observing each rotamer (table II). The biggest differences between the PDB statistics and the theoretical methods occur for valine and isoleucine with helical phi and psi angles. Unlike most amino acids, valine and isoleucine are β-branched, i.e. there are two non-hydrogen side chain atoms bonded to the C_β atom. When a valine or isoleucine is in a α-helix there is only one chi 1 rotamer it can adopt and avoid a clash between its $C_{\square}$ groups and the carbonyl oxygen on residue i-3. This restriction on chi 1 is evident in the PDB statistics (figure 4), but is absent from the theoretical

calculations that were performed in the context of a dipeptide. This provides a clear example of a case where double counting will occur if PDB statistics are used in combination with Lennard-Jones energies when calculating the energy of residues i and i-3. Tables showing the rotamer probabilities binned by phi and psi dihedral angles for all of the theoretical methods are provided in the supplementary material.

Aside from valine and isoleucine in the helical region, the energies calculated with QM and MM match reasonably with those derived from PDB statistics, although there are some specific cases where the MM potentials deviate significantly. The Amber potential favors the TP rotamer over the MT rotamer for leucine when it has helical torsion angles, but the MT rotamer is more commonly observed in the PDB. The Cedar potential strongly favors the trans rotamer for valine when it has α backbone angles, but this is the least common rotamer in the PDB. Not surprisingly, the aforementioned potentials perform poorly in side chain prediction tests for the regions of Ramachandran space in which they deviated from the QM and the PDB statistics.

Side chain prediction tests

We have shown several examples that demonstrate that the three different approaches, QM, MM and knowledge-based, give significantly different energies for many side chain rotamers. To determine which of these potentials more accurately represents the internal energy of amino acid residues, we performed side chain prediction tests with the Rosetta protein modeling program. In these tests a single side chain was removed from a residue in a protein, and Rosetta was used to predict the conformation of the removed side chain. The prediction was performed by cycling through all rotamers and sub-rotamers of the missing amino acid and choosing the one with the lowest energy. The energy function was a linear sum of the internal energy of the rotamer, as calculated by the QM, MM or knowledge-based potential, and long range interactions between the rotamer and its neighbors calculated with the standard Rosetta energy function. Neighboring residues were held fixed in this test because QM energies are only available for valine, isoleucine and leucine. As a control, tests

were also performed in which each rotamer of an amino acid was assumed to have equal internal energy (flat). The side chain prediction test was performed on 5360 valine-α, 6377 valine-β, 7569 leucine-α, 2546 leucine-β, 4278 isoleucine-α and 3928 isoleucine-β positions in over 2800 proteins. The predictions were analyzed to determine how often the correct rotamers were predicted (i.e. the correct torsional wells), and how close the chi angles were to the native chi angles as described in the materials and methods section.

Overall, all of the methods do well in the side chain prediction test; all of them predict the correct rotamer at more than 90% of the positions. This result was expected because at most sequence positions only one rotamer can fit without clashing with the neighboring residues, and the energy from a clash will overwhelm the internal energies of the amino acids. Indeed, in the tests without any internal energy for the side chain the correct rotamer was predicted over 85% of the time. This does not indicate that the internal rotamer energies are unimportant. This test is artificial in that we are keeping all the neighbors fixed as well as the protein backbone. In a full protein simulation all backbone positions and side chains are free to vary and changes in 1 kcal / mol as a side chain moves to a new rotamer are certainly important. To make the test more discriminatory, we focused on sequence positions at which the correct rotamer was not specified by simply looking for clashes with neighboring residues. If a side chain could adopt two rotamers that had a predicted clash score of less than 0.5 kcal / mol and differed by more than 60° at chi 1, than that position was included in our refined test. Because isoleucine, valine, and leucine are often found in the interior of a protein, this filter removed a large number of sequence positions from our test. The filter reduced the number of test positions to 118 valine-α, 549 valine-β, 842 leucine-α, 761 leucine-β, 2949 isoleucine-α and 477 isoleucine-β.

In the filtered side chain prediction test there are notable differences between the three methods, reflecting the different energies the methods give for the internal energy of rotamers. The largest differences are seen for isoleucines and valines with helical phi and psi angles. Rosetta's knowledge-

based potential, which is based on Dunbrack's backbone dependent rotamer library, only picks the correct rotamer 53% of the time for valine and 41% for isoleucine (table III). The QM calculations with the HF energy predict the correct rotamer 67% of the time for valine and isoleucine. The prediction accuracy with the MM potentials vary significantly; Cedar only places 35% of the valine side chains accurately while CHARMM22 places 55% correctly. These results confirm that for isoleucine and valine with helical torsion angles that the knowledge-based potential does not accurately reflect the internal energy of isoleucine and valine, but rather the potential is dominated by interactions that isoleucine and valine make with neighboring residues in a helix.

For residues with phi and psi angles in the β-strand region of the Ramachandran plot the QM and knowledge-based potentials do equally well. This suggests that for these residues that the knowledge-based potential is a fairly accurate measure of the internal energy of a side chain. The results with the MM potentials are more varied, and no single potential performs as well as the QM potential or the knowledge-based potential. The complete results table is available in the supplementary information.

Discussion

Accurate estimates for the relative energies of amino side chain conformations are important for protein structure prediction, protein design and drug design. Here, we have shown that various approaches for calculating these energies, molecular mechanics potentials, quantum mechanics calculations and knowledge-based potentials, can give significantly different results, in some cases on the order of 1 kcal / mol per side chain. In general, the QM and knowledge-based energies are more similar with each other than with the results from the molecular mechanics potentials. To evaluate which potentials were most accurate we performed side chain prediction tests. In particular, we examined residues in proteins for which the correct side chain conformation could not be predicted by searching for clashes with neighboring residues. In most scenarios the QM potentials and the knowledge-based potential performed equally well. The exceptions were valines and isoleucines with backbone torsion angles from the helical region of the Ramachandran plot. In these cases the QM

potential significantly outperformed the knowledge-based potential because the knowledge-based potential is not an accurate representation of the internal energy of the side chains in this situation, but rather also represents energetics terms derived from being in a helix.

The discrepancy between the QM and knowledge-based energies for β-branched amino acids with helical torsion angles, highlights one of the potential pitfalls of using knowledge-based potentials. The physical basis for preferences observed in the protein database may not always be cleanly assigned to a single energetic effect. For instance, the common hydrogen bond geometries and distances observed in the backbone of an α-helix represent more than the relative energy of different hydrogen bond configurations, they also reflect all the other energetic terms that go in to determining the optimal conformation for a helix. In other words, when knowledge-based potentials are combined with each other or with molecular mechanics potentials, there is a possibility of double counting.

The MM potentials gave fairly erratic results: performing well in some cases but poorly in others. The overall success of the QM energies in side chain prediction tests suggest that they could be used as a benchmark for improving the MM potentials [21, 22]. QM simulations on dipeptides have played extensive roles in the parameterization of molecular mechanics potentials from the beginning. Recently there have been attempts by the developers of the CHARMM (version 31) [4, 23, 24] and ECEPP (version 5) [25] suites to improve the modeling of the protein backbone using QM simulations similar to those conducted here. Both groups sampled either the complete or selected regions of phi/psi space of alanine, glycine, and proline dipeptides. The ECEPP group refit the parameters used to compute backbone torsional energy while the CHARMM group refit its torsional backbone parameters as well as created a 2D grid correction scheme. Both groups have shown improved modeling of the protein backbone [26].

In this study we have restricted our tests to hydrophobic amino acids that do not have the potential to form strong electrostatic interactions between the side chain and the polar atoms in the backbone. In

vacuum QM simulations with dipeptides will not be as useful for determining the rotamer preferences of polar side chains. An alternative approach is to perform QM/MM simulations where the dipeptide is treated by QM and explicit solvent is modeled with a MM forcefield. This type of approach has been used by Hermans and co-workers to map out the conformational preferences of solvated peptides [5]. The peptides intramolecular energies were calculated with the self-consistent charge density functional tight binding method (SCCDFTB) and the solvent was represented by either the SPC or TIP3P models. The distribution of backbone torsion angles obtained with the QM/MM approach more closely matched distributions from high-resolution protein structures than did distributions obtained using only MM potentials. Our results suggest that before performing computationally intensive QM/MM simulations with polar side chains, it will be prudent to test our knowledge-based potential in side chain prediction test with polar amino acids. The QM/MM simulations will be most useful for conformations for which the knowledge-based potential is not an accurate reflection of the internal energy of the residue, but rather reflects longer range interactions from the protein. In conclusion, our results indicate that calculating the relative energies of side chain rotamers is still a difficult problem, and combining QM calculations with knowledge-based scores may be the best way to generate an accurate potential.

Acknowledgements

UNC Information Technology Services

This work was partially supported by the National Center for Supercomputing Applications under grant MCB040053, and utilized the IBM pSeries 690 systems. This research was supported by an award from the W.M. Keck foundation and the grant GM073960 from the National Institutes of Health.

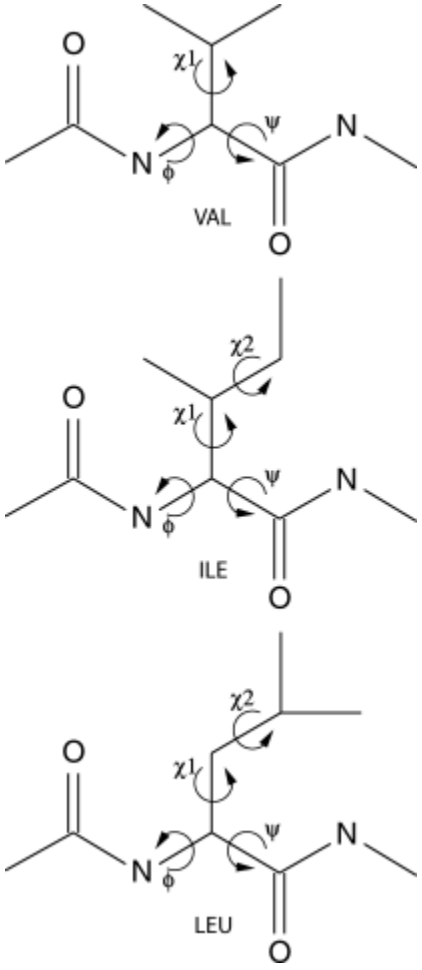


Figure 1.1 Diagrams of the (top) valine, (middle) isoleucine, and (bottom) leucine dipeptides showing backbone and side chain torsion angles.

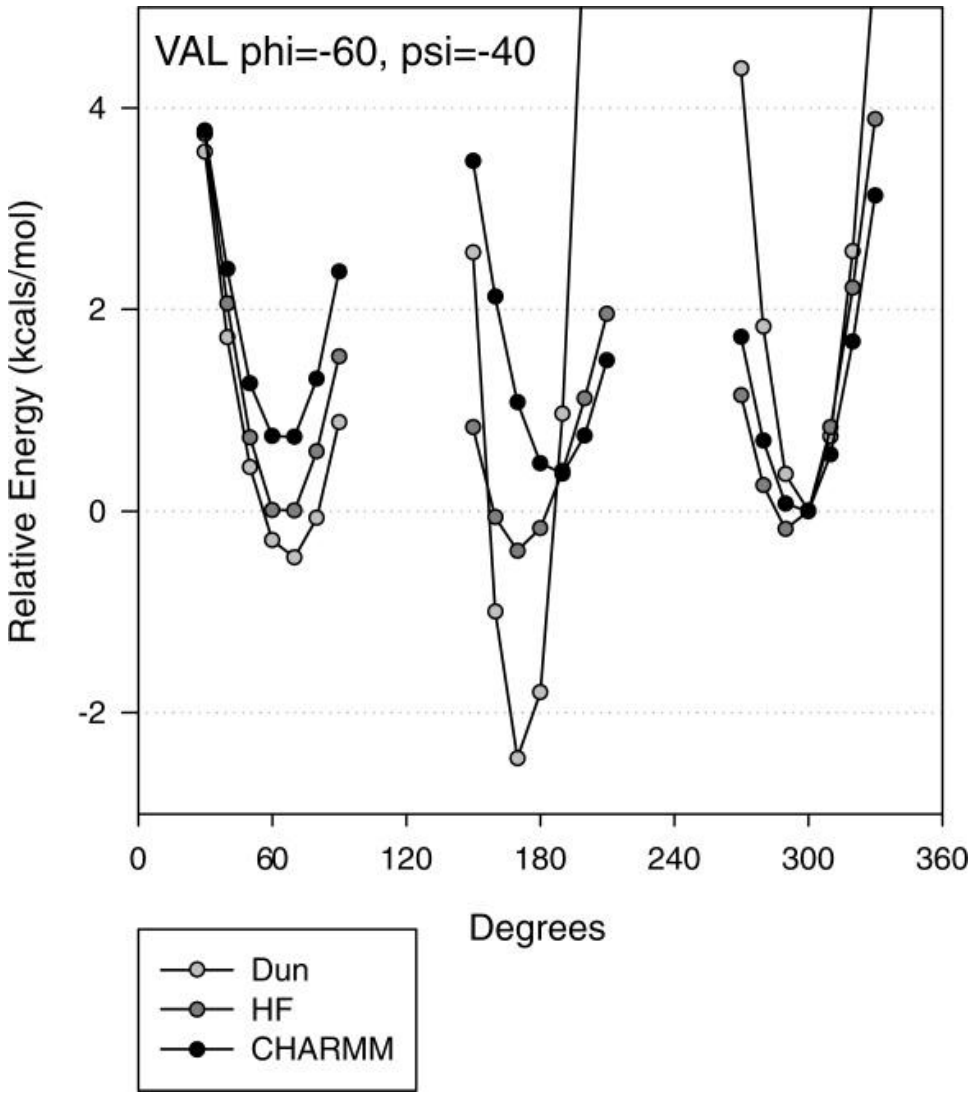


Figure 1.2 Comparison between the relative energy differences of (black) CHARMM22 MM potential, (dark grey) Dunbrack rotamer library, or (light grey) energies from the final step of HF minimization for valine in the α-helical region (phi = -60, psi = -40). Probabilities from the Dunbrack library were converted to energies using equations 2 and 3 from the text. Energies for each method were set equal to a value of 0 at a chi of -60 to allow for comparison.

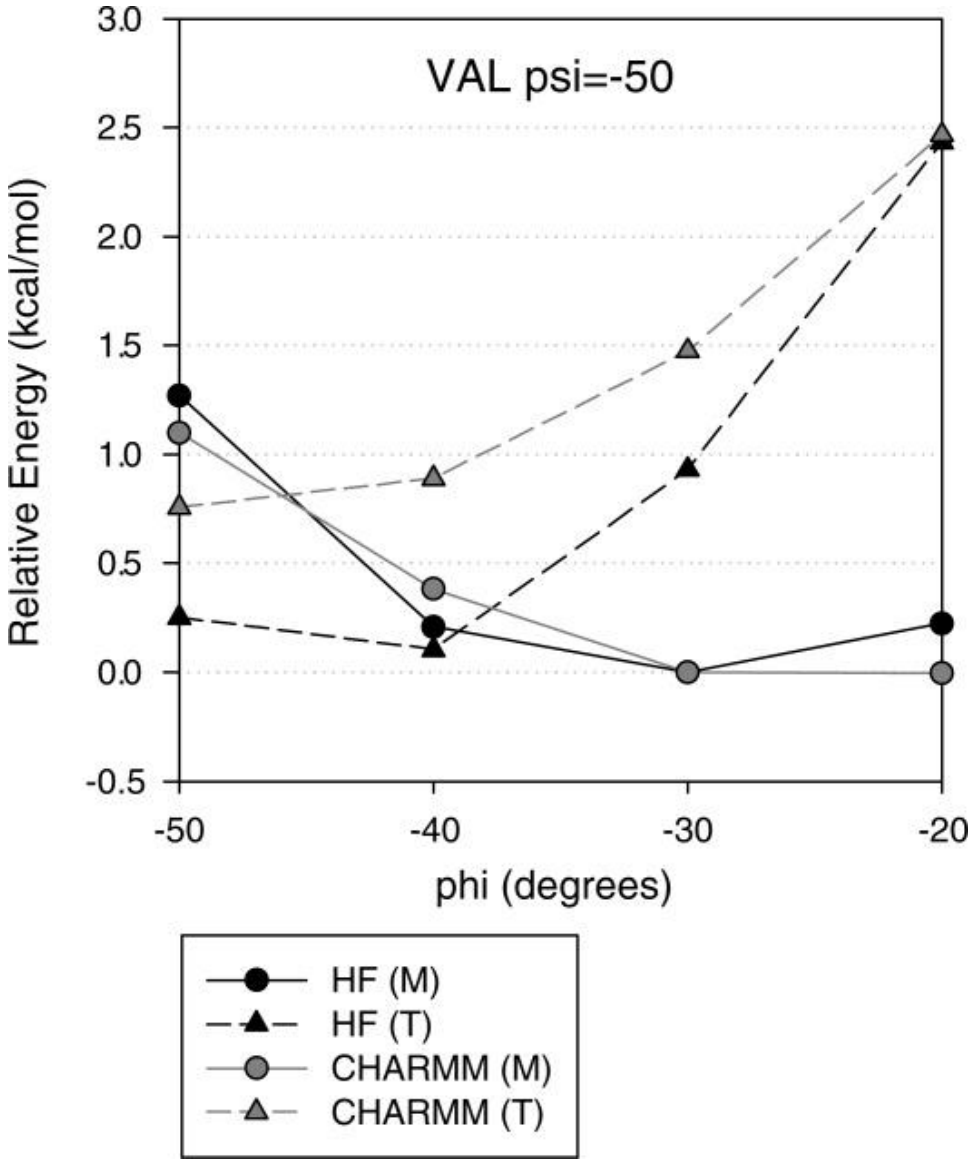


Figure 1.3 Relative energy versus psi angle for the M (solid line, circles) and T (dashed line, triangles) rotamers of valine dipeptides using the HF (black) and CHARMM (dark grey) methods. Energies shown are for the phi and psi angle shown and the chi angle that had the minimum energy for that rotamer bin relative to the calculated energy of the M rotamer minimum for each method at a phi of -50, and psi of -30. Psi and Chi angles are as follows HF (M): -50/-70, -40/-70, -30/-60, -20/-60; HF (T) -50/170, -40/170, -30/170, -20/170; CHARMM (M) -50/-70, -40/-60, -30/-60, -20/-60; CHARMM (T) -50/190, -40/190, -30/190, -20/190. See methods for rotamer labeling.

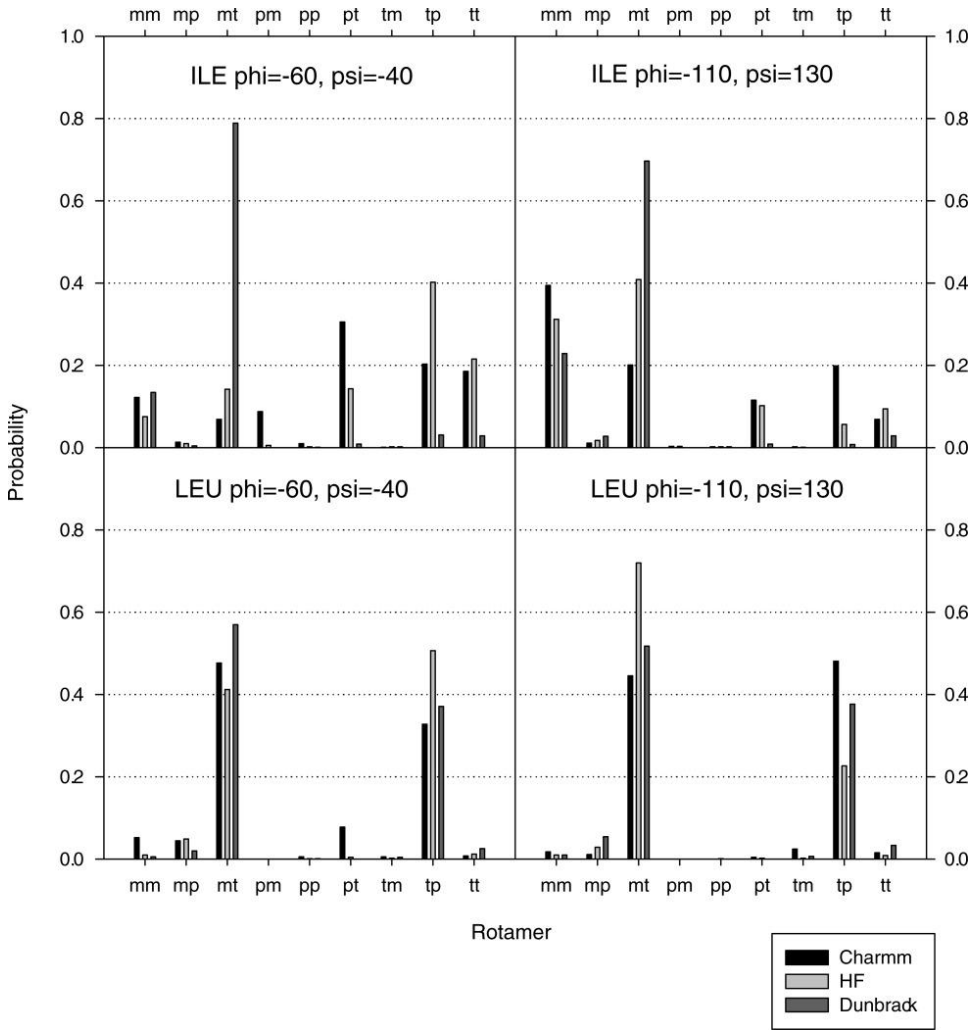


Figure 1.4 Probability of choosing a particular rotamer according to (black) CHARMM22 MM potential, (dark grey) Dunbrack rotamer library, or (light grey) HF QM potential for isoleucine and leucine in the canonical α-helical (phi = -60, psi = -40) and β-strand (phi = -110, psi = 130) region. Log probabilities were calculated from energies and normalized to 1 (P = exp(-E/RT)). See methods for rotamer labeling.

Bibliography

1. Gordon, D.B., S.A. Marshall, and S.L. Mayo, *Energy functions for protein design.* Curr Opin Struct Biol, 1999. **9**(4): p. 509-13.

2. Moult, J., *Comparison of database potentials and molecular mechanics force fields.* Curr Opin Struct Biol, 1997. **7**(2): p. 194-9.

3. McCammon, J.A. and S.C. Harvey, *Dynamics of Proteins and Nucleic Acids*. 1987, Cambridge, UK: Cambridge University Press.

4. Mackerell, A.D., *Empirical force fields for biological macromolecules: Overview and issues.* Journal of Computational Chemistry, 2004. **25**(13): p. 1584-1604.

5. Hu, H., M. Elstner, and J. Hermans, *Comparison of a QM/MM force field and molecular mechanics force fields in simulations of alanine and glycine "dipeptides" (Ace-Ala-Nme and Ace-Gly-Nme) in water in relation to the problem of modeling the unfolded peptide backbone in solution.* Proteins, 2003. **50**(3): p. 451-63.

6. Feig, M., A.D. MacKerell, and C.L. Brooks, *Force field influence on the observation of pi-helical protein structures in molecular dynamics simulations.* Journal of Physical Chemistry B, 2003. **107**(12): p. 2831-2836.

7. Beachy, M.D., et al., *Accurate ab initio quantum chemical determination of the relative energetics of peptide conformations and assessment of empirical force fields.* Journal of the American Chemical Society, 1997. **119**(25): p. 5908-5920.

8. Ponder, J.W. and F.M. Richards, *Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes.* J Mol Biol, 1987. **193**(4): p. 775-91.

9. Dunbrack, R.L., Jr., *Rotamer libraries in the 21st century.* Curr Opin Struct Biol, 2002. **12**(4): p. 431-40.

10. Lovell, S.C., et al., *The penultimate rotamer library.* Proteins-Structure Function and Genetics, 2000. **40**(3): p. 389-408.

11. Pohl, F.M., *Empirical Protein Energy Maps.* Nature-New Biology, 1971. **234**(52): p. 277-&.

12. Butterfoss, G.L. and J. Hermans, *Boltzmann-type distribution of side-chain conformation in proteins.* Protein Sci, 2003. **12**(12): p. 2719-31.

13. Butterfoss, G.L. and B. Kuhlman, *Computer-based design of novel protein structures.* Annu Rev Biophys Biomol Struct, 2006. **35**: p. 49-65.

14. Dunbrack, R.L., Jr., *Comparative modeling of CASP3 targets using PSI-BLAST and SCWRL.* Proteins, 1999. **Suppl 3**: p. 81-7.

15. Brooks, B.R., et al., *Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations.* Journal of Computational Chemistry, 1983. **4**(2): p. 187-217.

16. Hermans, J., et al., *A Consistent Empirical Potential for Water-Protein Interactions.* Biopolymers, 1984. **23**(8): p. 1513-1518.

17. Ferro, D.R., et al., *Energy Minimizations of Rubredoxin.* Journal of Molecular Biology, 1980. **136**(1): p. 1-18.

18. M. J. Frisch, G.W.T., H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople, *Gaussian 03, Revision C.02*. 2004, Gaussian, Inc.: Wallingford CT.

19. Wang, G.L. and R.L. Dunbrack, *PISCES: a protein sequence culling server.* Bioinformatics, 2003. **19**(12): p. 1589-1591.

20. Kuhlman, B., et al., *Design of a novel globular protein fold with atomic-level accuracy.* Science, 2003. **302**(5649): p. 1364-8.

21. Petrella, R.J., T. Lazaridis, and M. Karplus, *Protein sidechain conformer prediction: a test of the energy function.* Folding & Design, 1998. **3**(5): p. 353-377.

22. Jacobson, M.P., et al., *Force field validation using protein side chain prediction.* Journal of Physical Chemistry B, 2002. **106**(44): p. 11673-11680.

23. Mackerell, A.D., M. Feig, and C.L. Brooks, *Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations.* Journal of Computational Chemistry, 2004. **25**(11): p. 1400-1415.

24. MacKerell, A.D., M. Feig, and C.L. Brooks, *Improved treatment of the protein backbone in empirical force fields.* Journal of the American Chemical Society, 2004. **126**(3): p. 698-699.

25. Arnautova, Y.A., A. Jagielska, and H.A. Scheraga, *A new force field (ECEPP-05) for peptides, proteins, and organic molecules.* Journal of Physical Chemistry B, 2006. **110**(10): p. 5025-5044.

26. Buck, M., et al., *Importance of the CMAP correction to the CHARMM22 protein force field: Dynamics of hen lysozyme.* Biophysical Journal, 2006. **90**(4): p. L36-L38.

Incorporating Noncanonical Amino Acids into Rosetta

Introduction

From the original full automated sequence design of Dahiyat and Mayo [1] to the recently designed enzymes designed by Jain *et al.* and Rothlisberger *et al.* [2, 3], computational protein design has become an increasingly powerful tool for protein modelers. A common theme amongst all of the programs used to produce these results is that they are designed to primarily work with the twenty canonical amino acids (CAAs) found in humans. The ability to apply the tools and techniques that have been developed to design proteins to design other protein-like polymers could allow for the creation of new therapeutics and biological tools. The first logical step towards this goal is the incorporation of noncanonical amino acids (NCAAs) in to computational protein design software. The use of NCAAs in protein design programs has advantages both biologically and computationally.

Simply changing the chirality of a protein by constructing it out of the D-enantiomers of its amino acids has been shown to provide proteolytic resistance[4], an issue which has been a problem for protein therapeutics[5]. Protein stability has been increased without significantly disturbing the protein structure by replacing common hydrophobic residues with fluorinated derivatives[6]. Numerous protein crystal structures have be solved with the aid of selino-methionine phasing[7]. Chemically restrained amino acids that have particular φ and ψ angle preferences have been used to promote helix formation[8]. These results have been obtained without the use of computational modeling and were limited in the scope of what they could design by similarity to the CAAs.

NCAAs can increase the number of sequences and therefore the number of conformations that can be sampled during a design simulation which has been shown to increase the accuracy of the simulation[9, 10]. Additional conformations will allow us to find more optimum packing

conformations to form additional hydrogen bonds. In contrast incorporating amino acids with torsional constraints will restrict the number of conformations the protein could have reducing the conformational entropy.

The term “nonnatural amino acid” is often used to describe NCAAs but its use is a misnomer as amino acids that differ from the canonical twenty are frequently found in nature. The most common NCAAs are residues with pre-/co-/post-translational modifications that provide them with additional functionality [11-13]. Eukaryotes, prokaryotes, and archea have all been found to have selenocysteine residues which are genetically encoded indirectly by overloading the UGA stop codon in conjunction with a selenocysteine insertion sequence element[14]. Additionally some methanogenic archaea genetically encode pyrrolysine indirectly by overloading the UAG stop codon in conjunction with a pyrrolysine insertion sequence element [15].

Computational protein design programs typically contain two major parts: an energy or scoring function to evaluate how well a particular amino acid sequence fits a given scaffold and a search function that samples sequences as well as backbone and side chain conformations. Energy functions often contain a combination of physically based and knowledge-based terms. Knowledge-based terms are generated from protein structures and compiling statistics from which probabilistic pseudo-energies can be calculated. They are information rich and generally quick to evaluate, but care must be taken to avoid double counting[16]. Unfortunately knowledge-based potentials are not compatible with NCAAs because there are not enough structures that contain NCAAs to derive meaningful statistics and the structures of many NCAA side chains have never been solved in a protein context. We have modified the energy function of Rosetta, the computational protein modeling suite developed in our lab, by removing knowledge-based terms incompatible with NCAAs and replaced them physically-based terms to create an energy function that can be used to evaluate the energy of both canonical and noncanonical amino acids.

Conformational searches of the backbone degrees of freedom is typically done using small perturbations to the backbone dihedral angles, fragment insertion, backrub movements, or using robotic loop-closure algorithms[17]. Conformational searches of the side chain degrees of freedom are done with the aid rotamer libraries. Rotamer libraries are lists of commonly seen side chain dihedral angles[18]. In Rosetta, the side chain rotamer coordinates are determined from dihedral angles from the rotamer library and idealized bond lengths, bond angles, and non-χ dihedrals[19]. Additionally rotamer libraries include the probability with which set of side chain dihedrals was observed in the set of proteins it was trained on. Amino acid rotamers are not observed with equal frequency suggesting that the internal energy of the conformation is different. These probabilities can be used to compute a pseudo-energy that represents the internal energy of the amino acid. To compute the internal energy of a side chain conformation, Rosetta assumes a Boltzmann distribution and uses the log of the probability of seeing a given rotamer with particular φ and ψ backbone dihedral angles as a measure of the energy as shown bellow.

$$E_{rot\ i} = -\ln\left(P(rot_i|\phi_i,\psi_i)\right)$$

Where E_{roti} is the energy of rotamer i, φ_i and ψ_i are the φ and ψ backbone dihedral angles at position i, and P_{roti} is the probability of seeing rotamer i when the backbone dihedral are φ_i and ψ_i. The probabilities in this equation come from the Dunbrack rotamer library[20]. The frequency of rotamers also provides a way of limiting the conformational search to the statistically most likely conformation. As with the knowledge-based potentials, the use of rotamers libraries to provide common side chain coordinates for use in side chain packing is incompatible with NCAAs because there are not enough protein structures to compute statistics. We have developed a method to create rotamer libraries for NCAAs that can reproduce the rotamers seen in CAA.

The modifications we have made to the energy function that allow for the scoring of NCAAs and the ability to create rotamers libraries allows us to now be able to use NCAAs in the computational

protein design program Rosetta. Currently we have incorporated an additional 88 amino acids in to Rosetta.

Materials and Methods

Modification of the Rosetta Energy Function

The Rosetta energy function is a sum of individually weighted terms and is shown bellow. It contains a physically-based inter-residue Lennard-Jones term split into repulsive and attractive components ($E_{inter\ rep}$ and $E_{inter\ atr}$)[21], a implicit solvation term implemented as described by Lazarids and Karplus ($E_{solvation}$)[22], knowledge-based reside pair electrostatics term (E_{pair}), orientation dependent hydrogen bonding term ($E_{bb:sc\ HB}$, $E_{bb:bb\ HB}$, and $E_{sc:sc\ HB}$)[23], a knowledge-based term that measures the internal energy of an amino acid based on probabilities from rotamer libraries ($E_{dunbrack}$), a knowledge-based term that measures Ramachandrin backbone torsion preferences of a position (E_{rama}), and a reference energy term that represents the energy of the unfolded state of a protein ($E_{reference}$)[9, 24].

$$E_{protein} = W_{inter\ rep}E_{inter\ rep} + W_{inter\ atr}E_{inter\ atr} + W_{solvation}E_{solvation} + W_{pair}E_{pair} + W_{bb:bb\ HB}E_{bb:bb\ HB} + W_{bb:sc\ HB}E_{bb:sc\ HB} + W_{sc:sc\ HB}E_{sc:sc\ HB} + W_{Dunbrack}E_{Dunbrack} + W_{rama}E_{rama} + W_{reference}E_{reference}$$

The inter-residue attractive and repulsive terms are physically based and are compatible with NCAAs.

The solvation term and the hydrogen bonding terms were trained on canonical protein data but are evaluated on atom-atom pairs. The atom types found in CAAs are the same or similar to the atom types in the NCAAs and these terms are therefore compatible with NCAAs. The internal energy term, the rama term and the pair term are knowledge-based, evaluated based in part on residue identity and are not compatible with NCAAs. To replace the internal energy term and the rama term we have implemented a intra-residue molecular mechanics Lennard-Jones term and a matching molecular

mechanics torsion term, both described below. The reference energy term has been replaced with a term that uses an explicit unfolded state model described below. The pair electrostatic term has been omitted. The modified energy function used for scoring CAAs and NCAAs is shown in equation bellow.

$$E_{protein} = W_{inter\ rep} E_{inter\ rep} + W_{inter\ atr} E_{inter\ atr} + W_{solvation} E_{solvation} + W_{torsion} E_{torsion} + W_{bb:bb\ HB} E_{bb:bb\ HB} + W_{bb:sc\ HB} E_{bb:sc\ HB} + W_{sc:sc\ HB} E_{sc:sc\ HB} + W_{intra\ rep} E_{intra\ rep} + W_{intra\ atr} E_{intra\ atr} + W_{unfolded} E_{unfolded}$$

Molecular mechanics energy terms functions are commonly used in computational protein design programs[25]. In contrast to molecular mechanics which often views proteins as a fixed set of atoms, bonds, bond angles, and dihedral angles, the energy functions used by computational protein design programs must be able rapidly handle changes to the protein amino acid sequence. This is achieved by decomposing the energy function in to terms that can be evaluated between pairs of prospective amino acid rotamers. Energy terms that that can be evaluated without information about the surrounding rotamers are called one-body terms (ie. $E_{Dunbrack}$). Energy terms that require information about the surrounding rotamers are called two-body terms (ie. $E_{inter\ rep}$). The combination of the molecular mechanics torsion and intra-residue Lennard-Jones terms can accurately describe the rotation about a bond in a protein design scenario using fixed bond lengths and angles[26].

Implementation of the CHARMM Torsion Potential in Rosetta

We have implemented a molecular mechanics torsion term of the form shown bellow using the CHARMM27 parameter set[27].

$$E_{dihe,ijkl} = K_{ijkl}\left(1 + \cos\left(n\chi_{ijkl} - \theta_{ijkl}\right)\right)$$

Where for four atoms *i*, *j*, *k*, and *l* that comprise the dihedral angle, *K* is a constant, *n* is the multiplicity, χ is the value of the dihedral, and θ is the offset. Note that a single chemical bond may have more than one of these terms such that the sum is expressed as a Fourier series. The torsion term is evaluated for all sets for 4 connected atoms in a protein. The energy from dihedral angles comprised entirely of a set of atoms from one rotamer is calculated and stored as a one-body energy at full weight. The energy from dihedral angles comprised of a set of atoms from two rotamers is calculated and stored as a two-body energy with half of the energy being stored in each rotamer. The sum of the one-body and two-body components for a set of rotamer is the full torsion energy of the protein.

Implementation of the CHARMM Lennard-Jones Potential in Rosetta

We have matched the molecular mechanics torsion term with a matching molecular mechanics Lennard-Jones term of the form shown bellow also using the CHARMM27 parameter set[27].

$$E_{LJ,ij} = \sqrt{\varepsilon_i \varepsilon_j}\left(\left(\frac{R_{min,ij}}{R_{ij}}\right)^{12} - 2\left(\frac{R_{min,ij}}{R_{ij}}\right)^{6}\right)$$

Where for two atoms of types *i* and *j*, $\sqrt{\varepsilon_i \varepsilon_j}$ is the well depth, $R_{min,ij}$ is the distance at which atoms of type *i* and *j* are at an energetic minimum, and R_{ij} is the distance between the two atoms. The term is evaluated between all pairs of atoms within an amino acid rotamer that are separated by three or more chemical bonds. The current Lennard-Jones term in Rosetta has the same form but is only evaluated between atoms in different rotamers and at separations of 4 or more chemical bonds. The parameters of the inter residue Lennard-Jones term are based on CHARMM but have been adjusted to more closely reflect atom/atom distances seen between residues in proteins and are not appropriate for evaluating intra-residue energies. Like the inter-residue Lennard-Jones term, the intra-residue term is artificially split into attractive and repulsive components at the energy minimum.

Explicit Unfolded Energy Term

The reference energy term in Rosetta represents the unfolded energy of the protein. The individual values for each CAA are variable degrees of freedom in the energy function weight fitting procedure. Weight fitting is done on a training set of proteins that contain only CAAs and the reference energy is therefore incompatible with NCAAs. We have implemented a term to replace the reference energy term that uses an explicit unfolded state model and is compatible with both CAAs and NCAAs. To calculate the unfolded energy of an amino acid we break a set of ~2000 high resolution, low redundancy, protein structures into randomly chosen 5-mer fragments. The list of structures was generated from the culled pdb[28]. The central residue of each fragment is mutated, and allowed to repack. The unweighted energies of each energy term for each central residue are averaged and stored. When scoring a particular position, the averaged unweighted residue-based energies are multiplied by the weight from the respective energy term as shown in bellow. 5-mer fragments are used over longer fragments due to cases that happen when mutating a position to an amino acid for which the backbone and surrounding side chains are not optimized.

$$E_{unfolded\ ,i} = W_{inter\ rep}\overline{E_{inter\ rep,i}} + W_{inter\ atr}\overline{E_{inter\ atr,i}} + W_{solvation}\overline{E_{solvation\ ,i}} + W_{torsion}\overline{E_{torsion\ ,i}} + W_{bb:bb\ HB}\overline{E_{bb:bb\ HB,i}} + W_{bb:sc\ HB}\overline{E_{bb:sc\ HB,i}} + W_{sc:sc\ HB}\overline{E_{sc:sc\ HB,i}} + W_{intra\ rep}\overline{E_{intra\ rep,i}} + W_{intra\ atr}\overline{E_{intra\ atr,i}}$$

Where $\overline{E_{j,i}}$ is the average unweighted energy for energy term j and residue type i.

Energy Function Training

The Rosetta energy function is the sum of individual weighted energy terms as show above. Substantial changes to the terms in the energy function require a re-optimization of the weights on the individual terms. The weights are trained to maximize the probability of seeing the native amino acid at each position in a set of high resolution protein structures during a complete sequence redesign. The weights on certain terms can be kept fixed or allowed to be free to change. The weight fitting is done by first calculating the unweighted energies for all rotamers at all positions in all of the

structures. Second, the weights on the free terms are optimized using a combination of particle swarm optimization and David-Fletcher-Powell minimization routines to maximize a fitness function. The fitness function used is designed to maximize the probability of the native amino acid having a lower energy than all other amino acids and is shown bellow. Third, the new set of weights is used to redesign the set of training proteins and the sequence recovery is tested. If the sequence recovery increases, the new set of weights is accepted. If the sequence recovery decreases the new weight set is averaged with the previous weight set. These three steps are repeated 10 times. The fitness function, *F*, that is maximized during the optimization is shown bellow.

$$F = \sum_{\substack{all\ proteins \\ all\ positions}} -\ln\left(\frac{e^{\left(\frac{-E_{native\ \ AA}}{k_B T}\right)}}{\sum_{all\ AA} e^{\left(\frac{-E_{AA}}{k_B T}\right)}}\right)$$

Where *E* is the Rosseta energy, *kB* is the Boltzmann constant, and *T* is the temperature.

CAA Sequence and Rotamer Recovery Benchmarks

The modified energy function is tested on its ability to score CAAs using two benchmarks: a rotamer recovery benchmark, and a sequence recovery benchmark. In the benchmarks a side chain optimization procedure is performed on a set of high resolution protein structures. In the rotamer recovery benchmark, the rotamers used are limited to the rotamers of the native amino acid and the percent of native rotamer recovered is recorded. In the sequence recovery benchmark, the rotamers of all CAAs are allowed at each position and the sequence identity is recorded.

Rotamer Library Creation

Rotamer Library Creation Protocol

We have developed a simple protocol, called MakeRotLib, which can create backbone dependent amino acid rotamer libraries for both CAAs and NCAAs as shown in figure 1. The rotamer calculations are done using an amino acid dipeptide model system, a single residue with an acetylated

N-terminus and an N-methylated C-terminus. The dipeptide system mimics all the interactions that a side chain would have with the surrounding protein backbone. φ and ψ backbone dihedrals are combinatorial sampled in 10 degree intervals creating 1296 φ/ψ bins. For each of these φ/ψ bins, a set of amino acid dipeptides are created where the side chain χ dihedrals are combinatorial sampled in varying size intervals depending on the number of χ angles, the composition of the side chain, and the expected number of rotamers. Figure 1A shows the starting side chain dihedral angles for leucine with α-helical backbone dihedrals (φ=-60 and ψ=-40), with both side chain χ angles sampled at 5 degree intervals.

Each dipeptide is minimized with 25 steps of linear-gradient minimization to the closest local minimum.

The φ and ψ backbone dihedrals as well as non-χ side chain dihedrals are kept fixed during minimization. Linear minimization was chosen over other forms of minimization because it causes the side chain to move to the closest local minimum and will not jump out of the local energy well. The rotamers of amino acids side chain are simply the local minimum in the side chain energy landscape. The set of minimized side chain dihedral angles for leucine with α-helical backbone dihedrals (φ=-60 and ψ=-40) is shown in figure 1B.

Following minimization, the set of minimized side chain dihedral angles is clustered using a K-means clustering algorithm. The K-means algorithm works by first calculating the root mean squared distance between each set of side chain dihedral angles and each member of a set of cluster centroids. The set of side chain dihedrals becomes assigned to the cluster it is closest to. Second, the cluster centroids are recalculated to be the geometric mean of the members of that cluster. The algorithm iterates between these two steps until no side chain dihedral sets change clusters or 500 iterations. The minimized angles are shown colored by cluster and with the final centroid positions for leucine with α-helical backbone dihedrals (φ=-60 and ψ=-40) in figure 1C.

We do not predefine limits or bins in which rotamers can exist. A draw back however of the algorithm is that it requires knowing the number of clusters and an estimate of the starting positions of the cluster centroids before hand. However, Rosetta calculates side chain dihedral angles for positions with φ and ψ that fall in between the predefined bins using a linear interpolation between φ and ψ bins but also between rotamer bins. In order to properly interpolate between rotamer bins the number of rotamers for each φ/ψ bin must be equal. The number of rotamer bins for each amino acid and the starting values of the cluster centroid positions are determined using test runs and expected results based on previous rotamer libraries. The set side chain dihedral angles to be used as the angles for each rotamer is the lowest energy set of angles in each cluster after the iterative clustering procedure. The final rotamers for leucine with α-helical backbone dihedrals (φ=-60 and ψ=-40) are shown in figure 1D.

The Dunbrack rotamer library assumes that side chains are rotameric and can be fit to a Gaussian distribution. Dunbrack provides standard deviations in addition to the mean angles in its library. Rosetta uses these standard deviations to calculate off rotamer side chain conformations that increase the number of rotamers sampled. To calculate standard deviations we sample around each side chain χ angle until the energy increases by 0.5 kcals/mol.

Rosetta makes use of the probabilities of a given rotamer listed in the Dunbrack rotamer library for determining the internal energy but also a way to screen bad rotamers. Rosetta only uses the top 95% of rotamers for each φ/ψ bin during side chain optimization. The rotamer libraries generated here are not used for energy evaluation but only as starting points for the side chain packing. However the removal of high energy rotamers speeds up side chain optimization. We therefore convert the energies to probabilities using the equation bellow.

$$P = e^{\left(\frac{-E}{k_B T}\right)}$$

Where *P* is the probability, *E* is the energy of the rotamer, and *kBT* is the Boltzmann constant. The probabilities are normalized to sum to 100% for each φ/ψ bin.

NCAA Rotamer Libraries

NCAAs were chosen based on what is commercially available, could be modeled using the existing CHARMM torsion and Lennard-Jones parameters, and has four or fewer heavy atom side chain χ angles. Some conformers of NCAAs are difficult to model using rotamer libraries because they involve coordinated movements of multiple torsion angles (ie. The transition between cyclohexo ring conformers). In these cases the different conformers were modeled as independent types residue types.

We have added the following list of NCAAs to Rosetta and generated rotamer libraries for the ones with rotatable side chain dihedrals: 1-amino-cyclopentane-carboxylic acid (2 conformers), 2.4-dimethyl-phenylalanine, 2-allyl-glycine, 2-amino-2-phenylbutyric acid, 2-amino-5-phenyl-pentanoic acid, 2-amino-heptanoic acid, 2-aminomethyl-phenylalanine, 2-hydroxy-phenylalanine, 2-indanyl-glycine (2 conformers), 2-methyl-phenylalanine, 3-aminomethyl-phenylalanine, 3-amino-tyrosine, 3-hydroxy-phenylalanine, 3-hydroxy-tyrosine, 3-methyl-phenylalanine, 4.5-dehydro-leucine, 4.5-dehydro-lysine, 4-aminomethyl-phenylalanine, 4-carboxy-phenylalanine, 4-fluoro-proline (2 conformers), 4-fluoro-tryptophan, 4-hydroxy-phenylglycine, 4-methyl-phenylalanine, 4-methyl-tryptophan, 4-phenyl-phenylalanine, 4-tert-butyl-phenylalanine, 5-bromo-tryptophan, 5-chloro-tryptophan, 5-fluoro-tryptophan, 5-hydroxy-tryptophan, 5-methyl-tryptophan, 6-bromo-tryptophan, 6-chloro-tryptophan, 6-fluoro-tryptophan, 6-methyl-tryptophan, 7-azatryptophan, 7-bromo-tryptophan, 7-methyl-tryptophan, 9-anthryl-alanine, allo-isoleucine, allo-threonine, α-aminoadipic acid, α-amino-glycine, α.β-diaminoproionic acid, α.gama-diaminobutyric acid, α-methyl-3-hydroxy-tyrosine, α-methyl-histidine, α-methyl-leucine, α-methyl-phenylalanine, α-methyl-proline, α-methyl-tryptophan, α-methyl-tyrosine, α-methyl-valine, β-(1-naphthyl)-alanine, β-(2-naphthyl)-alanine, β-β-dicyclohexyl-alanine (4 conformers), β.β-diphenyl-alanine, β-cyclohexyl-alanine (2 conformers), β-cyclopentyl-

alanine (2 conformers), β-hydroxy-norvaline, cyclohexyl-glycine (2 conformers), diphenylglycine, dipropyl-glycine, ethionine, 2-fluoro-leucine, 2'-fluoro-leucine, hexafluoro-leucine, homocysteine, homophenylalanine, homoserine, n-in-methyl-tryptophan, ornithine, penicillamine, phenylglycine, phenyl-serine, tert-butyl-alanine, tert-butyl-cysteine, tert-butyl-glycine, 2,2,2-trifluoro-leucine, 2',2',2'-trifluoro-leucine, 1-methyl-histidine, 1-methyl-histidine prot, 2-amino-4-bromo-4-pentenoic acid, 3-methyl-histidine, 3-methyl-histidine prot, 4-amino-piperidine-4-carboxylic-acid (4 conformers), 4-amino-tetrahydropyran-4-carboxylic acid (4 conformers), 4-amino-tetrahydrothiopyran-4-carboxylic acid (4 conformers), amino-ethyl-cysteine, β-chloro-alanine, β-fluoro-alanine, β-iodo-alanine, and trifluoro-alanine.

Comparison to Knowledge-Based Rotamer Libraries

To test the MakeRotLib protocol we compared its ability to reproduce the rotamers of the CAAs. Rosetta uses only the top 95% of rotamers given by the Dunbrack rotamer library for each φ/ψ bin during side chain optimization. We therefore compare the percent overlap in rotamer identity between the top 95% of rotamers predicted by the MakeRotLib protocol and the top 95% of Dunbrack rotamers for each φ/ψ bin and for all amino acids except gylcine and proline. For each φ /ψ bin where the Dunbrack rotamer library has more than 10 observations for a particular amino acid, we compare the percent overlap between the identities of the rotamers bins. Additionally, for matching rotamer bins, we compute the root mean squared distance between side chain dihedral angles as a measure of the difference in angles preferences. Comparisons are discussed for each CAA below and shown in figure 2.

Results

Energy Function Modifications

Explicit Unfolded State Energy

We have calculated unfolded state energies for the CAAs and the NCAAs we have added to Rosetta. The fragment based method of calculating unfolded energies can place the central residue in a position where it experiences far fewer contacts than if it were in a folded protein or at a protein interface. The largest effect of the low contact number is that it under estimates the attractive component of the energy function for larger amino acids. This gives larger amino acids a bias when designing because they contain more atoms.

Energy Function Weighting

The weights on the energy function terms have been optimized using the procedure described above and the final weights are shown in table1. The weights on the Lennard-Jones inter residue attractive term were kept fixed during the weight fitting while all others were allowed to be free. The weights on the shared terms remain close with the exception of the Lennard-Jones inter-residue repulsive energy and solvation energy.

Energy Term	Stock Weight	Modified Weight
Inter-repulsive	0.44	0.63
Inter-attractive	0.80	0.80
Solvation	0.65	1.16
Pair	0.49	-
Bb/bb HB	0.59	0.67
Bb/sc HB	1.17	1.45
Sc/sc HB	1.10	1.19
Dunbrack	0.56	-
Omega	0.50	-
Rama	0.20	-
Reference	1.00	-

Torsion	-	0.27
Intra-repulsive	-	0.32
Intra-attractive	-	0.54
Unfolded	-	0.90

Table 2.1 Weights on the stock Rosetta energy function and on the modified energy function.

CAA Sequence and Rotamer Recovery

Sequence and rotamer recovery benchmarks were run using the stock and modified energy functions as described in the methods section. X1 rotamer recovery for the stock energy function was 84% overall, 93% in the core, and 74% on the surface. X1 and χ2 rotamer recovery for the stock energy function was 64% overall, 74% in the core, and 53% on the surface. X1 rotamer recovery for the modified energy function was 75% overall, 91% in the core, and 59% on the surface. X1 and χ2 rotamer recovery for the modified energy function was 53% overall, 71% in the core, and 37% on the surface. The overall sequence recovery was 35% for the stock energy function and 28% for the modified energy function. When the weight fitting protocol is run using the stock energy function with the reference energy term replaced with the explicit unfolded energy term the sequence recovery is 30%.

Rotamer recoveries between the two energy functions are comparable and indicate that the modified energy function can find the low energy side chain conformations of CAAs. The difference in sequence recovery between the two energy functions is larger. The modified energy function is however at a disadvantage because of it uses of the explicit unfolded state energy term instead of the reference energy term. The reference energy term adds an additional 20 fit-able parameters that can be optimized during the weight fitting protocol. This allows for finer turning of amino acid preferences and higher sequence recovery.

Rotamer Library Creation

Canonical Amino Acid Rotamer Library Creation

Rosetta currently uses the 2002 update to the Dunbrack backbone-dependent rotamer library[20]. To test the MakeRotLib protocol we have used it to create rotamer libraries for all CAAs except glycine and proline, and compared them how well they overlap. We however use the notation developed by Lovell *et al.* to describe the rotamers because of its clarity and brevity[29].

The overall RMS side chain dihedral angle distance and percent overlap of top rotamers for all amino acids is shown in the table 2.

CAA	RMS Distance (degrees)			Percent Overlap (%)		
	low	high	average	low	high	average
ARG	5.8	11.2	7.7	57	100	87
ASN	0.3	18.1	12.6	0	100	67
ASP	0.5	21.6	7.9	0	83	40
CYS	0.2	15.1	6.1	50	100	98
GLN	11.1	18.5	15	33	100	76
GLU	3.3	15	7.7	18	69	45
HIS	7.9	17.1	12.1	60	100	86
ILE	4	18.7	9.7	50	100	81
LEU	1.7	19.9	9.4	0	100	72
LYS	2.8	10	5.6	36	100	79
MET	3.3	10.4	5.9	56	100	86
PHE	0.6	17.9	4.6	33	100	50
SER	0.3	19.4	7	50	100	97
THR	0	27.8	7.8	0	100	91
TRP	5.4	14.7	9	33	100	73

TYR	0.6	18.4	4.9	33	100	53
VAL	1.5	21.9	8.6	50	100	88

Table 2.2 Comparison of the top 95% of CAA rotamers predicted by the MakeRotLib protocol to the rotamers given by the Dunbrack rotamer library. Low, high, and average values are calculated over all φ / ψ bins where the Dunbrack rotamer library reports more than 10 observations. A high percent overlap indicates that the rotamers predicted by the MakeRotLib protocol are in agreement with the rotamers predicted by the Dunbrack rotamer library. A low average RMS distance indicates that the dihedral angles for rotamer bins that overlap are in good agreement.

Arginine

Arginine has 4 χ angles with 3 χ1 rotamer wells (**mpt**), 3 χ2 rotamer wells (**mpt**), 3 χ3 rotamer wells (**mpt**), and 3 χ4 rotamer wells (**mpt**), for a total of 81 possible rotamers. At the -110/130 φ/ψ bin, the top 35 rotamers capture 97% of the Dunbrack rotamers and 95% of the MakeRotLib rotamers, while at the -60/40 φ/ψ bin, the top 35 rotamers capture 98% of the Dunbrack rotamers and 96% of the MakeRotLib rotamers. Overlap for the -110/130 φ/ψ bin is 77% while overlap for the -60/-40 φ/ψ bin is 91%. Deviations in the β-strand regions are due to the MakeRotLib protocol having a stronger preference for the χ1 **m** rotamer than the Dunbrack library. The MakeRotLib protocol favors compact side chain conformations with the side chain packing against the backbone. For example in the -110/130 φ/ψ bin the top rotamer is **mmt-85** which is the 17th most popular Dunbrack rotamer. Dunbrack favors a more extended conformation as evidenced by the top 14 rotamers having a χ2 of 180. Low overlap probabilities occur in the regions where the Dunbrack library has low counts on the extreme of the β-sheet region and the 3/10-helical region. Of the overlapping rotamers, the average RMS angle distance is 7.7 degrees for both the -110/130 and -60/-40 φ/ψ bins. χ1-3 rotamer angles cluster well around -60, 60, and 180. χ4 rotamers agree with the Dunbrack rotamers and cluster close to -85, 180, and 85. We do not see the χ4 = 105 rotamer in the -110/130 φ/ψ bin and only once in the -60/-40 φ/ψ bin. Others have shown however that when built with ideal bond angles there is a moderate clash but that in examples of that rotamer in crystal structures there are bond angle deviations that relieve the strain and permit the rotamer[29].

Asparagine

Asparagine has 2 χ angles with 3 χ1 rotamer wells (**mpt**), 6 χ2 rotamer wells (centered on -120, -60, -10, 40, 80, 140). At the -110/130 φ/ψ bin the top 9 rotamers capture 96% of the Dunbrack rotamers and 97% of the MakeRotLib rotamers, while at the -60/40 φ/ψ bin the top 10 rotamers capture 96% of the Dunbrack rotamers and 99% of the MakeRotLib rotamers. Overlap for the -110/130 φ/ψ bin is 56% while overlap for the -60/-40 φ/ψ bin is 80%. Asparagine is a difficult residue to match because of the large number of rotamers that span a full rotation about the χ2 dihedral. If we look at the distribution of the Dunbrack library for χ2 angles for rotamers that are seen more then 10% probability are we find they mostly fall near -60, -20, 20, and 60. Dunbrack uses a large number of rotamers to cover the spread of angles and the MakeRotLib protocol does not find rotamers with χ2 near 0. This significantly lowers the overlap. Additionally the χ1 preferences of the MakeRotLib protocol differ from the Dunbrack library which also lowers the overlap. The MakeRotLib protocol strongly favors rotamers with a χ1 of **m** followed by **p** and then **t** while the Dunbrack is more evenly distributed. The top rotamers predicted by the MakeRotLib protocol have a higher percent overlap for the α-helical region than the β-strand region. Of the overlapping rotamers the average RMS angle distance is 12.3 for the -110/130 φ/ψ bin and 15.2 for the -60/-40 φ/ψ bin.

Aspartic Acid

Aspartic acid has 2 χ angles with 3 χ1 rotamer wells (**mpt**), 3 χ2 rotamer wells (centered on -60, 0, 60). At the -110/130 φ/ψ bin the top 6 rotamers capture 98% of the Dunbrack rotamers and 96% of the MakeRotLib rotamers, while at the -60/40 φ/ψ bin the top 6 rotamers capture 97% of the Dunbrack rotamers and 96% of the MakeRotLib rotamers. Overlap for the -110/130 φ/ψ bin is 50% while overlap for the -60/-40 φ/ψ bin is 50%. Aspartic acid is symmetric which is not taken into account by the MakeRotLib protocol. Consequently the MakeRotLib protocol finds rotamers with χ1 of -60, 60, and 180 and χ2 of -70 and -110, as well as -55 and 125 which places the side chain in the same position. The MakeRotLib protocol is unable to match any of the rotamers near 0 which are

often high probability consequently lowering the overlap. Of the overlapping rotamers the average RMS angle distance is 8.5 for the -110/130 φ/ψ bin and 12.3 for the -60/-40 φ/ψ bin.

Cysteine

Cysteine has 1 χ angle with 3 χ1 rotamer wells (**mpt**), for a total of 3 rotamers. At the -110/130 φ/ψ bin the top 2 rotamers capture 99% of the Dunbrack rotamers and 99% of the MakeRotLib rotamers, while at the -60/40 φ/ψ bin the top 3 rotamers capture 100% of the Dunbrack rotamers and 100% of the MakeRotLib rotamers. Overlap for the -110/130 φ/ψ bin is 100% while overlap for the -60/-40 φ/ψ bin is 100%. Overall the agreement between the MakeRotLib protocol and the Dunbrack library is the highest with an average percent overlap of 98% and an average RMS distance of 6.1 degrees. There is however a distinct ψ dependence shown in the banding pattern in figure 2. In the α-helical region the second most preferred rotamer shifts from **t** to **p** around the -20 and -30 ψ bins. The shift does not occur for the MakeRotLib protocol. Lovell et al. [29] indicate that the **p** is disfavored with α-helical φ/ψ. Of the overlapping rotamers the average RMS angle distance is 5.8 for the -110/130 φ/ψ bin and 6.1 for the -60/-40 φ/ψ bin. χ1 rotamer angles cluster well around -60, 60, and 180.

Glutamine

Glutamine has 3 χ angles with 3 χ1 rotamer wells (**mpt**), 3 χ2 rotamer wells (**mpt**), 4 χ3 rotamer wells (-120/0, -80/-40, 0/45, 80,120). At the -110/130 φ/ψ bin the top 13 rotamers capture 95% of the Dunbrack rotamers and 96% of the MakeRotLib rotamers, while at the -60/40 φ/ψ bin the top 16 rotamers capture 95% of the Dunbrack rotamers and 98% of the MakeRotLib rotamers. Overlap for the -110/130 φ/ψ bin is 69% while overlap for the -60/-40 φ/ψ bin is 75%. As with asparagine the χ3 dihedral is seen adopting angles that span a full rotation. Dunbrack χ3 angles cluster differently depending on the χ2. When χ2 is **m** or **p** the χ3 angles cluster at -120, -40, 40, and 120, when the χ2 is t the χ3 angles cluster at -70, 0, 70, 180. MakeRotLib χ3 angles cluster differently depending on the χ2. When χ2 is **m** or **p** the χ3 angles cluster at -100, -60, 0, and 100, when the χ2 is t the χ3 angles cluster at -110, 20, 60, and 110. The MakeRotLib protocol does not find all the rotamers with a χ3 of

0 because they are wide and have large standard deviations, this brings the overlap down. Of the overlapping rotamers the average RMS angle distance is 15 for the -110/130 φ/ψ bin and 16.4 for the -60/-40 φ/ψ bin. The standard deviations reported by the Dunbrack library are often 20 degrees or more indicating that the MakeRotLib protocol is finding the correct rotamer wells but that the minimum of well differs between the two methods. χ1-2 rotamer angles cluster well around -60, 60, and 180.

Glutamic acid

Glutamic acid has 3 χ angles with 3 χ1 rotamer wells (**mpt**), 3 χ2 rotamer wells (**mpt**), 4 χ3 rotamer wells (centered on -60, 0, and 60), for a total of 27 rotamers. At the -110/130 φ/ψ bin the top 12 rotamers capture 95% of the Dunbrack rotamers and 97% of the MakeRotLib rotamers, while at the -60/40 φ/ψ bin the top 14 rotamers capture 95% of the Dunbrack rotamers and 98% of the MakeRotLib rotamers. Overlap for the -110/130 φ/ψ bin is 58% while overlap for the -60/-40 φ/ψ bin is 57%. Glutamic acid is symmetric which is not taken into account by the MakeRotLib protocol. Consequently the MakeRotLib protocol finds rotamers with χ3 of -120 and 60, as well as 120 and 60 which places the side chain in the same position but is not counted as matching to the Dunbrack rotamers artificially lowering the overlap. Of the overlapping rotamers the average RMS angle distance is 10.1 for the -110/130 φ/ψ bin and 6.9 for the -60/-40 φ/ψ bin. χ1-2 rotamer angles cluster well around -60, 60, and 180.

Histidine

Histidine has 2 χ angles with 3 χ1 rotamer wells (**mpt**), 2 χ2 rotamer wells (centered on -90, 90, and 180), for a total of 9 rotamers. At the -110/130 φ/ψ bin the top 7 rotamers capture 100% of the Dunbrack rotamers and 98% of the MakeRotLib rotamers, while at the -60/40 φ/ψ bin the top 7 rotamers capture 99% of the Dunbrack rotamers and 98% of the MakeRotLib rotamers. Overlap for the -110/130 φ/ψ bin is 86% while overlap for the -60/-40 φ/ψ bin is 86%. The **m180** and **t180** rotamers are significantly populated in the Dunbrack rotamer library however the MakeRotLib

protocol does not find them to be rotamers. The standard deviations for the rotamer are large with the **βmt** rotamer 32.4, the **βtt** rotamer at 29.4, the **αmt** rotamer at 21.7, and the **αtt** rotamer at 26.6. The absence of these rotamers lowers the overlap. Of the overlapping rotamers the average RMS angle distance is 13.8 for the -110/130 φ/ψ bin and 12.6 for the -60/-40 φ/ψ bin. χ1 rotamer angles all cluster well around the expected -60, 60, and 180. χ2 rotamer angles all cluster well around the expected -65, 65. The Dunbrack library clusters closer to -80, 80 and is the reason for the overall RMS angle distance of 12.1 and the lowest 7.9. The imidazol ring of histidine can occupy a wide range of angles without large clashes as evidenced by the higher than average standard deviations of the Dunbrack library and the MakeRotLib protocol.

Isoleucine

Isoleucine has 2 χ angles with 3 χ1 rotamer wells (**mpt**) and 3 χ1 rotamer wells (**mpt**). At the -110/130 φ/ψ bin the top 4 rotamers capture 98% of the Dunbrack rotamers and 95% of the MakeRotLib rotamers, while at the -60/40 φ/ψ bin the top 4 rotamers capture 98% of the Dunbrack rotamers and 97% of the MakeRotLib rotamers. Overlap for the -110/130 φ/ψ bin is 50% while overlap for the -60/-40 φ/ψ bin is 50%. For the -110/130 φ/ψ bin and for the -60/-40 φ/ψ bin both methods favor the **mm** and the **mt** rotamers more than 80%. The overlap between top rotamers for both bins is 50% because of differences in the third and forth rotamers. Dunbrack prefers **tt** and **mp**, while the MakeRotLib protocol prefers **pt** and **tp** for β-strand bin. Dunbrack prefers **tt** and **tp**, while the MakeRotLib protocol prefers **pt** and **mp** for α-helical bin. For the α-helical bin the results are unexpected as Renfrew *et al* [16] have shown that in the context of a dipeptide, the distribution of rotamer probabilities is more even and that the preferred rotamers are **tp**, **tt**, **pt**, **mt**, and **mm** . Of the overlapping rotamers the average RMS angle distance is 8.6 for the -110/130 φ/ψ bin and 15.6 for the -60/-40 φ/ψ bin. χ1-2 rotamer angles cluster well around -60, 60, and 180. However the -60/-40 φ/ψ bin the **tp** is skewed to 40, 160.

Leucine

Leucine has 2 χ angles with 3 χ1 rotamer wells (**mpt**) and 3 χ1 rotamer wells (**mpt**), for a total of 9 rotamers. At the -110/130 φ/ψ bin the top 4 rotamers capture 98% of the Dunbrack rotamers and 100% of the MakeRotLib rotamers, while at the -60/40 φ/ψ bin the top 3 rotamers capture 97% of the Dunbrack rotamers and 99% of the MakeRotLib rotamers. Overlap for the -110/130 φ/ψ bin is 75% while overlap for the -60/-40 φ/ψ bin is 100%. Both the Dunbrack rotamer library and the MakeRotLib protocol favor the **mt** and **tp** rotamers in most φ/ψ bin with probabilities >90%. Major differences in the overlap are generally the result of different preferences in the third and/or fourth most favorable rotamer. Of the overlapping rotamers the average RMS angle distance is 9.5 for the -110/130 φ/ψ bin and 11.2 for the -60/-40 φ/ψ bin. χ1-2 rotamer angles cluster well around -60, 60, and 180. Although the **tt** rotamer is often skewed, to 190, 140 by the MakeRotLib protocol. This skew can place the rotamer out of overlap range and therefore decrease the overall overlap. The skew is not consistent with the Dunbrack rotamer but is consistent with the preferred angles of Lovell et al.[29].

Lysine

Lysine has 4 χ angles with 3 χ1 rotamer wells (**mpt**), 3 χ2 rotamer wells (**mpt**), 3 χ3 rotamer wells (**mpt**) and 3 χ4 rotamer wells (**mpt**) for a total of 81 rotamers. At the -110/130 φ/ψ bin the top 24 rotamers capture 95% of the Dunbrack rotamers and 98% of the MakeRotLib rotamers, while at the At the -60/40 φ/ψ bin the top 26 rotamers capture 95% of the Dunbrack rotamers and 95% of the MakeRotLib rotamers. Overlap for the -110/130 φ/ψ bin is 92% while overlap for the -60/-40 φ/ψ bin is 85%. Low overlap probabilities occur in the regions where the Dunbrack library has low counts near the boundary of the 3/10-helical region. Of the overlapping rotamers the average RMS angle distance is 5.7 for the -110/130 φ/ψ bin and 6.5 for the -60/-40 φ/ψ bin. χ1-4 rotamer angles all cluster well around the expected -60, 60, and 180.

Methionine

Methionine has 3 χ angles with 3 χ1 rotamer wells (**mpt**), 3 χ2 rotamer wells (**mpt**), and 3 χ3 rotamer wells (**mpt**). At the -110/130 φ/ψ bin the top 12 rotamers capture 96% of the Dunbrack rotamers and 96% of the MakeRotLib rotamers, while at the -60/40 φ/ψ bin the top 11 rotamers capture 95% of the Dunbrack rotamers and 96% of the MakeRotLib rotamers. Overlap for the -110/130 φ/ψ bin is 92% while overlap for the -60/-40 φ/ψ bin is 91%. Of the overlapping rotamers the average RMS angle distance is 5.5 for the -110/130 φ/ψ bin and 5.5 for the -60/-40 φ/ψ bin. χ1-2 rotamer angles all cluster well around the expected -60, 60, and 180. χ3 angles cluster around -70, 70, and 180 with the exception of the **mmp** rotamer in the α-helical φ/ψ where the χ3 is 105. This value is consistent with the Dunbrack library for that rotamer.

Phenylalanine

Phenylalanine has 2 χ angles with 3 χ1 rotamer wells (**mpt**), 2 χ2 rotamer wells (centered on 90 and 0), for a total of 6 rotamers. At the -110/130 φ/ψ bin the top 4 rotamers capture 100% of the Dunbrack rotamers and 99% of the MakeRotLib rotamers, while at the -60/40 φ/ψ bin the top 4 rotamers capture 99% of the Dunbrack rotamers and 98% of the MakeRotLib rotamers. Overlap for the -110/130 φ/ψ bin is 50% while overlap for the -60/-40 φ/ψ bin is 50%. As with aspartic acid, glutamic acid, and tyrosine, phenylalanine is symmetric which is not taken into account by the MakeRotLib protocol. Consequently the MakeRotLib protocol finds rotamers with χ1 of -60, 60, and 180 and χ2 of -90 and 90 which places the side chain in the same position. The Dunbrack rotamers with χ2 near 0 are not scene. That rotamer well is wide as evidenced by the large standard deviations. Lovell et al. note that phenylalanine rotamers with a χ2 near 0 often have bond angle deviations that would not be captured by the MakeRotLib protocol[29]. Of the overlapping rotamers the average RMS angle distance is 5.6 for the -110/130 φ/ψ bin and 5.7 for the -60/-40 φ/ψ bin.

Serine

Serine has 1 χ angle with 3 χ1 rotamer wells (**mpt**). At the -110/130 φ/ψ bin the top 3 rotamers capture 100% of the Dunbrack rotamers and 100% of the MakeRotLib rotamers, while at the -60/40

φ/ψ bin the top 3 rotamers capture 100% of the Dunbrack rotamers and 100% of the MakeRotLib rotamers. Overlap for the -110/130 φ/ψ bin is 100% while overlap for the -60/-40 φ/ψ bin is 100%. For the -110/130 φ/ψ bin the Dunbrack and MakeRotLib protocol differ significantly in their preferred rotamers. Dunbrack orders rotamers **t,m**, and **p** at 46%, 45%, and 9% respectively. While MakeRotLib orders the rotamers **m,p** and **t** at 79%, 11%, and 10% respectively. For the -60/40 φ/ψ bin the MakeRotlib protocol only favors the **m** rotamer, at 98%. The Dunbrack is much more evenly distributed which allows for the comparison of all rotamers and is the reason for the 100% overlap. The narrow distribution is the result of improper ideal coordinates that place the hydroxyl hydrogen in a position to clash with the backbone. Of the overlapping rotamers the average RMS angle distance is 5.5 for the -110/130 φ/ψ bin and 8.4 for the -60/-40 φ/ψ bin.

Threonine

Threonine has 1 χ angle with 3 χ1 rotamer wells (**mpt**). At the -110/130 φ/ψ bin the top 2 rotamers capture 98% of the Dunbrack rotamers and 100% of the MakeRotLib rotamers, while at the -60/40 φ/ψ bin the top 2 rotamers capture 99% of the Dunbrack rotamers and 100% of the MakeRotLib rotamers. Overlap for the -110/130 φ/ψ bin is 100% while overlap for the -60/-40 φ/ψ bin is 100%. Three φ/ψ bins have zero overlap probabilities: 80/-10, 70/170, 70/-10. Both Dunbrack and the MakeRotLib protocol have very strongly prefer (>95%) a different rotamers and the overlap is therefore 0%. The bins are on the borders of the 3/10 helical region and have very low counts. As with serine the **p** rotamer is significantly populated in the α-helical φ/ψ bin. Of the overlapping rotamers the average RMS angle distance is 8.4 for the -110/130 φ/ψ bin and 9.4 for the -60/-40 φ/ψ bin. χ1 rotamer angles all cluster well around the expected -60, 60, and 180.

Tryptophan

Tryptophan has 2 χ angles with 3 χ1 rotamer wells (**mpt**), and 3 χ2 rotamer wells (centered on -120, 0, 120). At the -110/130 φ/ψ bin the top 5 rotamers capture 99% of the Dunbrack rotamers and 98% of the MakeRotLib rotamers, while at the -60/40 φ/ψ bin the top 6 rotamers capture 98% of the

Dunbrack rotamers and 100% of the MakeRotLib rotamers. Overlap for the -110/130 φ/ψ bin is 80% while overlap for the -60/-40 φ/ψ bin is 67%. At both α-helical and β-strand φ and ψ the MakeRotlib protocol does not find the **m0** rotamer. The standard deviation in the Dunbrack library is given as 23.4 degrees for -110/130 φ/ψ bin and 22.2 degrees for the -60/-40 φ/ψ bin suggesting that it is quite wide. The absence of this rotamer lowers the overlap in these regions. Additionally for the -60/-40 φ/ψ bin the Dunbrack rotamer library gives the **m-90** rotamer a probability of 1.6% and the rotamer library of Lovell *et al.* gives the same rotamer a 0% while it is the most favorable rotamer for the MakeRotLib protocol at 49%. Constructing this rotamer in the context of a dipeptide with helical φ and ψ shows no major clashes. However, if it is constructed in the context of a α-helix, there is a large clash with the neighbor side chain in the helix. Indeed the **m-90** rotamer in the Dunbrack library has a significantly shifted angles χ1 = -90, χ2 = -120. The large RMS angle distance is 30.99 for this rotamer placing it out of the cutoff range and decreasing the overlap in the -60/-40 φ/ψ bin. Of the overlapping rotamers the average RMS angle distance is 9.9 for the -110/130 φ/ψ bin and 8.4 for the -60/-40 φ/ψ bin. χ1 rotamer angles all cluster well around the expected -60, 60, and 180. χ2 rotamer angles all cluster well around the expected -90 and 90 but are missing the 0 rotamer.

Tyrosine

Tyrosine has 2 χ angles with 3 χ1 rotamer wells (**mpt**), 2 χ2 rotamer wells (centered on 90 and 0). At the -110/130 φ/ψ bin the top 4 rotamers capture 99% of the Dunbrack rotamers and 99% of the MakeRotLib rotamers, while at the -60/40 φ/ψ bin the top 4 rotamers capture 98% of the Dunbrack rotamers and 97% of the MakeRotLib rotamers. Overlap for the -110/130 φ/ψ bin is 50% while overlap for the -60/-40 φ/ψ bin is 50%. Of the overlapping rotamers the average RMS angle distance is 6.2 for the -110/130 φ/ψ bin and 4.9 for the -60/-40 φ/ψ bin.

Valine

Valine has 1 χ angles with 3 χ1 rotamer wells (**mpt**). At the -110/130 φ/ψ bin the top 2 rotamers capture 97% of the Dunbrack rotamers and 100% of the MakeRotLib rotamers, while at the -60/40

φ/ψ bin the top 2 rotamers capture 97% of the Dunbrack rotamers and 100% of the MakeRotLib rotamers. Overlap for the -110/130 φ/ψ bin is 50% while overlap for the -60/-40 φ/ψ bin is 50%. Looking at the over overlap performance the results are good with the MakeRotLib protocol finding 88% of the Dunbrack rotamers. There is a distinct ψ dependence shown in the banding pattern in figure 2. In the β-strand region the preferred rotamer shifts from m to t between the -150 and-160 ψ bins. The shift occurs between the -140 and -150 ψ bins for the MakeRotLib protocol. For the -60/-40 φ/ψ bin the Dunbrack rotamer library both strongly prefer (>90%) the **t** rotamer. This result was unexpected since we have previously shown that when using a dipeptide model system and using a full molecular-mechanics force field or high-level quantum mechanics to calculate the energy of a dipeptide free from long-range interactions that could bias the rotamer probabilities, the rotamer distribution for each rotamer well was approximately even for the MM and the QM [16]. Of the overlapping rotamers the average RMS angle distance is 10.5 for the -110/130 φ/ψ bin and 16.5 for the -60/-40 φ/ψ bin. χ1 rotamer angles cluster well around -60, 60, and 180. Both the MakeRotLib protocol and the Dunbrack rotamer library prefer the **t** rotamer for the α-helical region. The angle preferred by the MakeRotlib protocol for the **t** rotamer differs from the Dunbrack by 14 degrees which is higher than the average for all seen positions. This trend was also seen Renfrew *et al.* with the full CHARMM potential preferring a **t** rotamer with a value of ~190 degrees.

The assumption of ideal bond lengths and bond angles speed up protein design calculations. If the same assumption is made during rotamer creation rotamers, amino acids that show slight bond angle deviations in certain conformations can be obscured as was seen for phenylalanine and tyrosine. The ideal assumption can also skew rotamer wells since the only degrees of freedom are torsions. Our modified energy function additionally doesn't take into account internal electrostatic interactions which are important for the small polar amino acids like aspartic acid and asparagine. The MakeRotLib protocol does not take into account increases in rotamer stability that are induced through interactions with neighboring side chains such as hydrogen bonding. These interactions can

bias rotamer libraries and are dangerous if using the probabilities to compute energies but it may also prevent strained rotamers (which would be compensated by other beneficial interactions) from being sampled because they would not be included in the database[29]. Additionally for amino acids the size of arginine or larger, the dipeptide model system used in the protocol allows rotamers that place the amino acid side chain in a position that would clash with the backbone of neighboring side chains if it were present. This could however lead to more accurate sampling of rotamers at protein termini which would most likely be under represented in a knowledge based rotamer library. Additionally it does not take into account symmetry in amino acid side chains which is captured by knowledge-based rotamer libraries using predefined bins. Symmetry is difficult to handle in a general case for all NCAAs one would wish to create rotamer libraries for and care must be taken to capture the appropriate rotamers with symmetric amino acids. The symmetry problem can be worked around by doubling the number of rotamers to sample both symmetric pairs but at the expense of having to perform essentially duplicate energy calculations during a design simulation.

Directly comparing the results of our protocol to those of knowledge-based rotamer libraries is currently the best test of its performance. Our method of creating rotamers unfortunately suffers because it does not take into account electronic effects that have not been adequately captured by the molecular mechanics terms and our energy function which are captured by rotamer libraries. However, the knowledge-based rotamer libraries can be biased because of the long range side chain / side chain interactions[16]. In this study we have identified that tryptophan rotamers with α-helical φ and ψ, like valine and leucine rotamers, are biased because of long range effects in an α-helix. Wide rotamer wells are not captured well by our protocol.

Noncanonical Amino Acid Rotamer Library Creation

The list of NCAAs that were added to Rosetta and for which rotamer libraries have been created is listed in the materials and methods section. Here we present a few examples in detail.

2-Indanyl-Glycine

2-indanyl-glycine is a hydrophobic amino acid that has been added to Rosetta. The residue was initially designed as a constrained phenylalanine with particular χ1 torsional preferences, to investigate the importance of certain residues in the substance P peptide and how that effects its binding to the tachykinin NK-1 receptor [30]. 2-indanyl-glycine exists in 2 conformers due to the pucker of the 5-membered ring. The structures of both conformers are shown in figure 3. The "down" conformer is 1.45 kcals/mol higher in energy than the "up" conformer as determined by QM when both structures were minimized in preparation for rotamer creation. No structures containing 2-indanyl-glycine have been deposited in the protein databank. The amino acid has 1 χ angle about the Cα-Cβ bond. The 5-membered ring mimics the β-branched structure of valine and the rotamers are similar as shown in table3. χ1 distribution of the "down" conformation has less spread than the "up" because of the side chain backbone clashes that occur at rotamers other than **t**.

Name	Φ	Ψ	Probability (%)	X1 (degrees)	Std. Dev. (degrees)
2IG "down"	-110	130	0.9963	178.3	10
			0.0036	-76.3	7.6
			0.0001	73.7	10.7
	-60	-40	0.9990	177.8	9.6
			0.0009	-81.5	7.2
			0.0001	68.8	10.3
2IG "up"	-110	130	0.9112	-179.9	10.6
			0.0834	-69.8	8.9
			0.0054	47.3	6.5
	-60	-40	0.9577	179.1	11.6
			0.0411	-72	9.1
			0.0011	43.8	6.7
VAL	-110	130	0.9408	178	6.1
			0.0338	57.8	9.5
			0.0254	-62.5	12.7
	-60	-40	0.9181	171.9	5.2
			0.0515	68	10.1
			0.0304	-61	11.2

Table 2.3 The rotamers of 2-indanyl-glycine predicted by the MakeRotLib protocol with the rotamer for valine from the Dunbrack rotamer library for β-strand and α-helical φ and ψ.

α-Methyl-Tryptophan

α-Methyl-tryptophan is a tryptophan derivative that has been added to Rosetta. α-methyl-tryptophan is taken up and retained by the brain because of it resemblance to serotonin. Labeled α-methyl-tryptophan is commonly used as a brain imaging tool [31]. It is identical to the canonical tryptophan amino acid with the addition of a methyl group replacing the Hα as seen in figure 4A. The addition of the methyl group restricts the rotamers that the side chain can adopt as shown table 4. The tryptophan χ2 rotamers near 0 are wide with large standard deviations. The addition of the methyl group in α-methyl-tryptophan causes a clash with the χ2 = 0 rotamer and limits the rotamers that the amino acid can have to 6.The χ1 of α-methyl-tryptophan cluster around **m, p**, and **t** and the χ2 cluster around -90 and 90. Additionally the methyl group also restricts the φ and ψ backbone dihedrals the residue can occupy, as shown in figure 4B. No structures have been deposited in the protein databank containing α-methyl-tryptophan.

Name	Φ	Ψ	Prob (%)	X1	X2	Std. Dev. 1	Std. Dev. 2
				Degrees			
AMT	-110	130	0.5772	-70.9	-91.7	7	9.8
			0.2789	-173.9	81.4	4.3	4
			0.1065	-79	76.8	4.6	20.2
			0.0258	44.1	104	4.8	2.7
			0.0109	175.8	-91.3	6.3	5.8
			0.0007	39.1	-80.6	7.6	3.6
	-60	-40	0.5034	-66.4	-94.2	10.6	9.7
			0.3157	-68	88.1	10.3	10.2
			0.1017	177.5	87.1	8.6	7
			0.0620	179.2	-87.6	9.4	8.7
			0.0100	40.3	-77.6	8.1	5.1
			0.0073	35.9	102.8	9.1	6.6
TRP	-110	130	0.5385	-69	90.5	6.3	11.8
			0.1645	-67	3.4	9.2	23.4

			0.1212	-69.7	-92.5	10.7	10.2
			0.0984	179.3	-100.5	15.7	11.7
			0.0660	178.9	88.2	5.3	11
			0.0091	-177.6	18	10.6	26.6
			0.0014	60.9	-89.8	9.3	8.8
			0.0008	61.5	87.7	10	10
			0.0001	66	-6.3	8.2	42.3
	-60	-40	0.2687	-179.3	85.5	7.7	8.6
			0.2511	179.7	-107.7	11.7	14.4
			0.2030	-73.6	109.2	12.1	14.5
			0.1242	-70.5	-11.5	10.4	22.2
			0.0794	68.8	-89.6	7.4	6.8
			0.0516	-173.7	16.7	11.1	36.1
			0.0162	-89.8	-119.8	14.8	22.4
			0.0054	73	91.3	17.8	12
			0.0004	67.4	-6.8	7.8	37.8

Table 2.4 The rotamers of α-methyl-tryptophan predicted by the MakeRotLib protocol with the rotamer for tryptophan from the Dunbrack rotamer library for β-strand and α-helical φ and ψ.

Homoserine

Homoserine is a medium sized, unbranched, polar residue that has been added to Rosetta. Homoserine differs from the canonical serine due to the addition of a methylene group in the side chain, essentially making a longer serine residue. Homoserine is a precursor in the biosynthesis of several amino acids. It is small and flexible and could be advantageous in designing hydrogen bonds at protein interfaces as seen in figure 5. χ1-2 cluster around the **m, p**, and **t** rotamers. There are no structures in the protein databank that contain an unmodified homoserine.

Name	Φ	Ψ	Prob (%)	X1	X2	Std. Dev. 1	Std. Dev. 2
				(degrees)			
HSE	-110	130	0.7381	-58.9	-62.8	10.9	12.8
			0.0790	-177.1	56.7	21.2	22.8
			0.0649	-176.6	176.6	23.2	26.6
			0.0621	-60.9	177.8	25.5	26.6
			0.0368	-176.9	-67.4	4.3	3.4
			0.0104	52.7	178.6	21.5	24.7
			0.0075	-68.7	69.9	20.6	20.9
			0.0010	51.4	-77.3	19.6	14.6

			0.0001	54.3	87.6	18.4	13.9
	-60	-40	0.6652	-178.3	56.8	0.1	0.1
			0.1178	-59.1	-62.1	11.7	12.2
			0.0939	-175.7	174.7	10.2	11.1
			0.0850	-58.9	-178.4	11.1	10.8
			0.0210	-169.4	-74.3	10.9	10.2
			0.0089	-68.2	69.6	10.6	12.1
			0.0079	48.9	178.8	10.4	11.3
			0.0004	47.6	-75.5	0.1	5.9
			0.0000	52.8	88	8.9	7.5

Table 2.5 The rotamers of homoserine predicted by the MakeRotLib protocol β-strand and α-helical φ and ψ.

Conclusions

The ability to use NCAAs in computational protein design programs could lead to new therapeutics and biological tools and is a first step to general molecular design. We have developed a modified version of the Rosetta energy function that is comparable to the standard energy function in both rotamer recovery and sequence recovery and most importantly can be used to score both CAAs and NCAAs.

Rotamer libraries are an essential part of protein modeling. We have also developed methods to create rotamer libraries that are compatible with NCAAs, and we have shown that they are able to find the majority of CAA side chain rotamers. Additional uses of the rotamer creation protocol could be the creation of context dependent rotamer libraries for situations that may be under- represented in protein structures and therefore difficult to model using knowledge-based potentials. Examples of such context dependent situations are pre/post proline positions, terminal positions, common terminal modifications, and rotamers that involve hydrogens [32]. The assumption that amino acid side chains are rotameric has been discussed in the past, with the majority of research showing they in fact are[18, 20, 29]. We have found that low energy conformations are seen the most frequently; the average cluster-member/cluster-centroid distance is low for the lowest energy rotamers, and that the shape indicates it would fit well to a Gaussian or modal distribution. However, some of the higher

energy rotamers (lower probability structures) do not fit well to a Gaussian distribution and do not appear to be rotameric (figure 1C).

The modification to Rosetta presented here now allows for the design of peptides and proteins with NCAAs. The NCAAs added to this point have α-amino acid backbones. NCAAs do not however have to be simple side chain substitutions. Extensions of the tools created here could be applied to scaffolds other than just α-peptide backbone, such as peptoids[33] or other foldamers.

Figure 2.1 Rotamer library creation protocol. The steps of the MakeRotLib protocol shown for leucine with φ = -60 and ψ = -40. For a given φ and ψ a set of leucine dipeptides is created with side chain angles initially set to all chi1 and chi2 values in 5 degree intervals (A). Each dipeptide is minimized keeping the φ and ψ fixed (B). Side chain dihedral values are clustered (C). The lowest energy set of side chain dihedrals in each cluster is used as a rotamer (D). See text for more description.

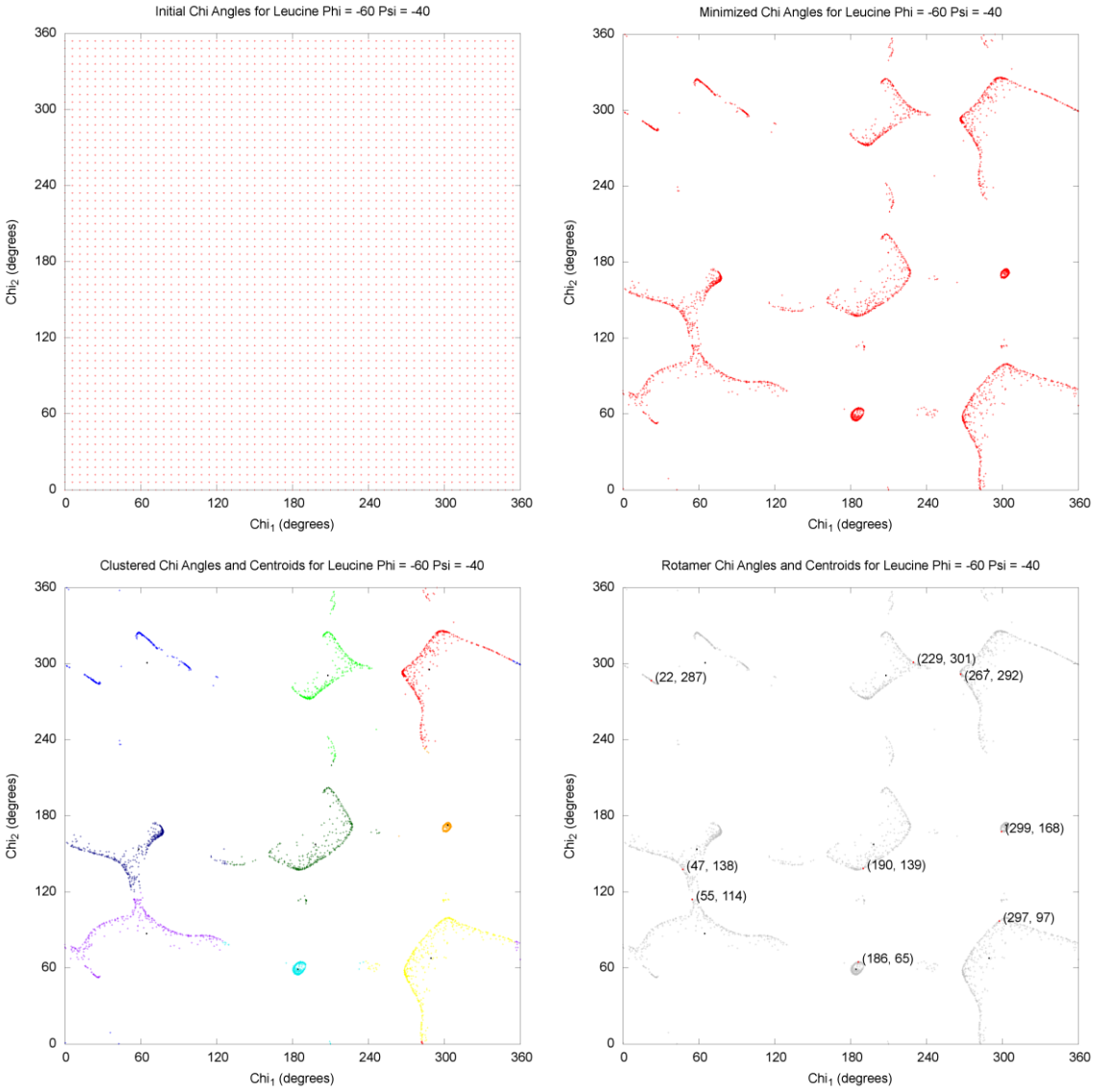
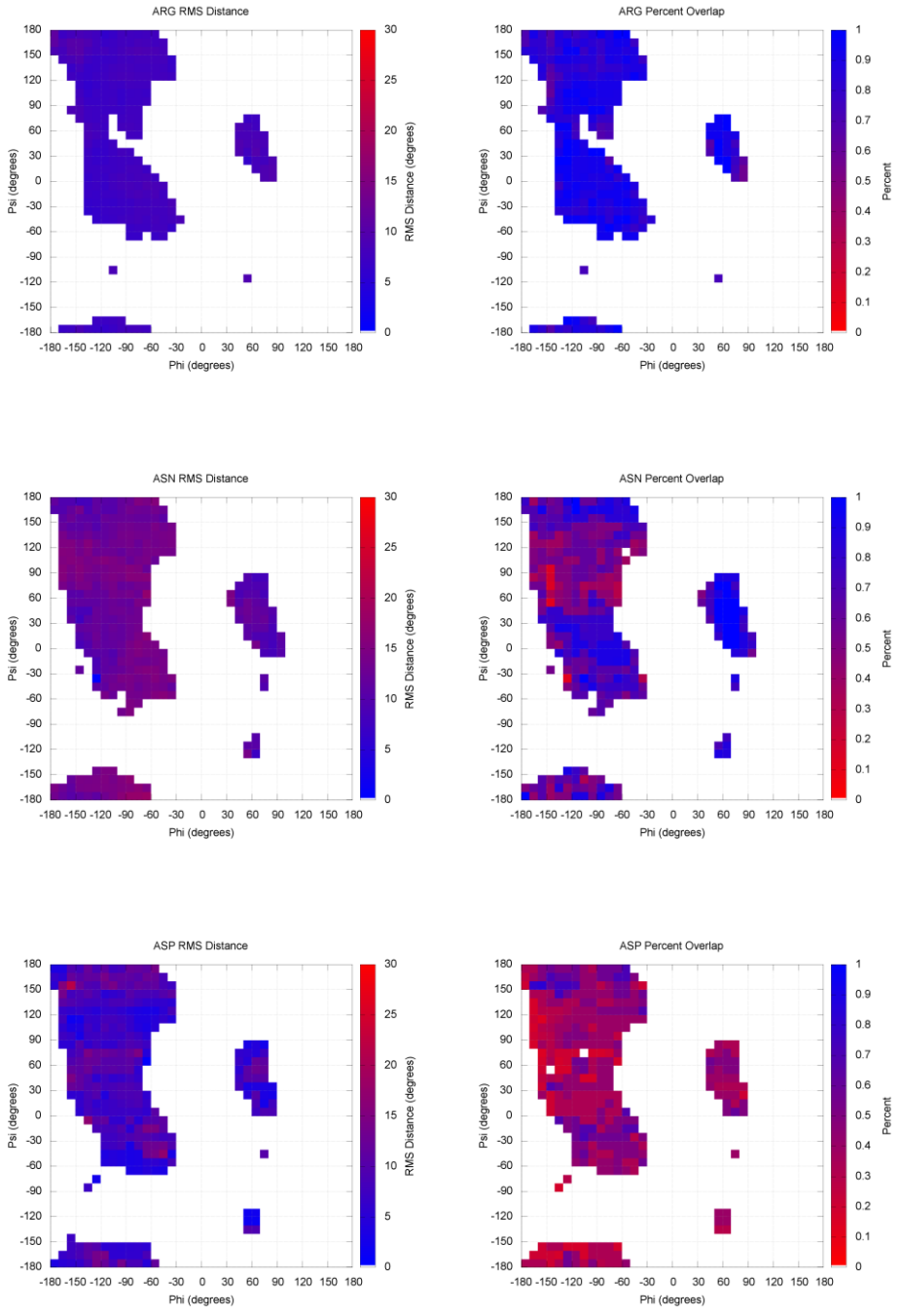
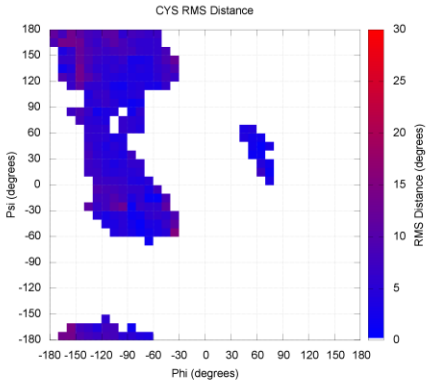
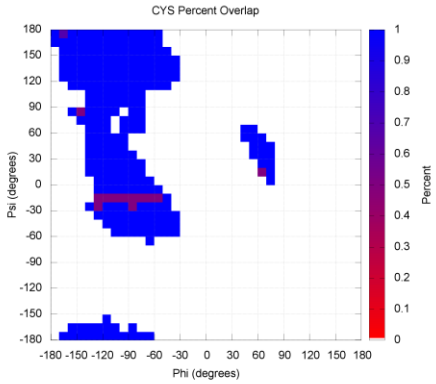
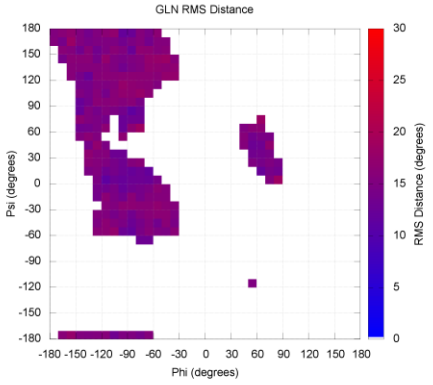


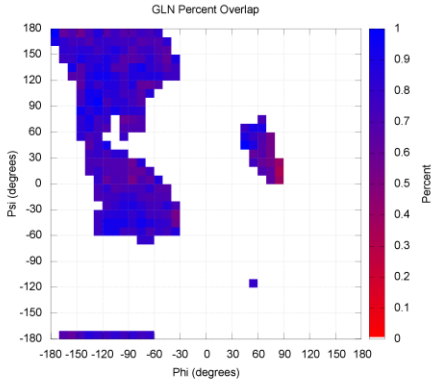
Figure 2.2 Percent overlap and RMS distance for the top 95% of rotamers between the Dunbrack rotamer library and the rotamer predicted by the MakeRotlib protocol. For each φ /ψ bin with more than 10 observations in the Dunbrack rotamer library, the percent overlap between the rotamer bins that comprise the top 95% of rotamer bins is calculated. For each pair of rotamer bins that overlap the root mean square distance in degrees is calculated. See methods for additional details on creation and results for details on analysis.

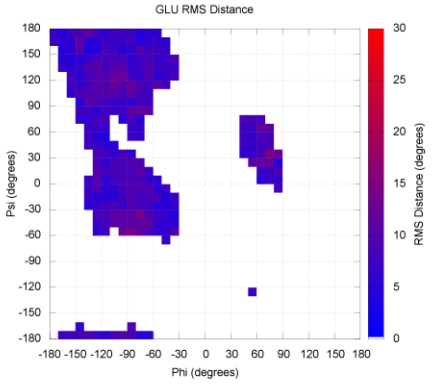


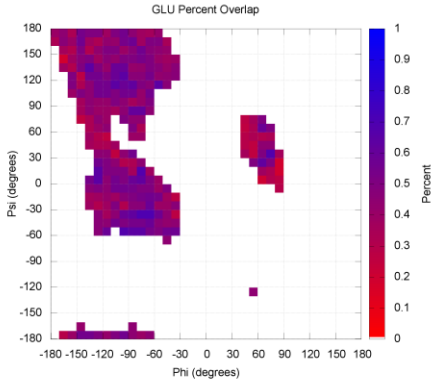

CYS RMS Distance
Psi (degrees)
Phi (degrees)
RMS Distance (degrees)

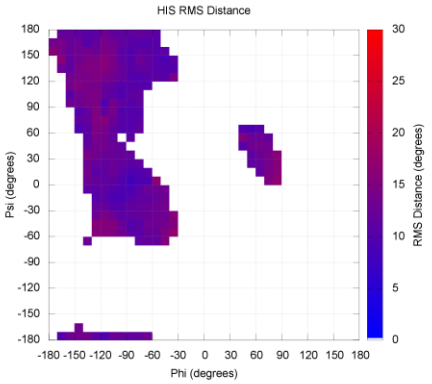

CYS Percent Overlap
Psi (degrees)
Phi (degrees)
Percent

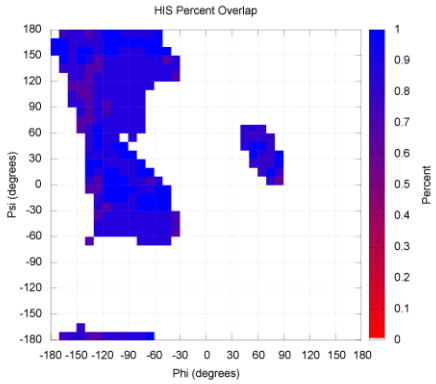

GLN RMS Distance
Psi (degrees)
Phi (degrees)
RMS Distance (degrees)

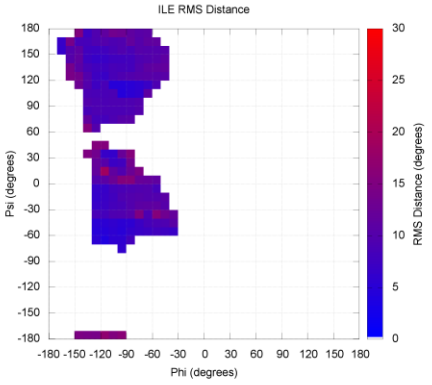

GLN Percent Overlap
Psi (degrees)
Phi (degrees)
Percent

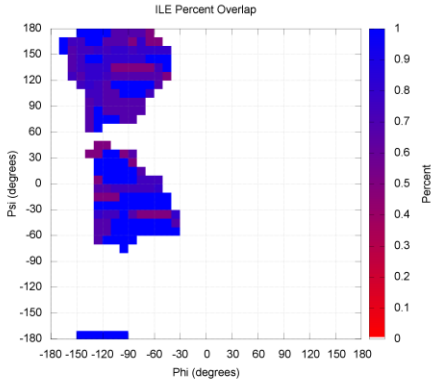

GLU RMS Distance
Psi (degrees)
Phi (degrees)
RMS Distance (degrees)

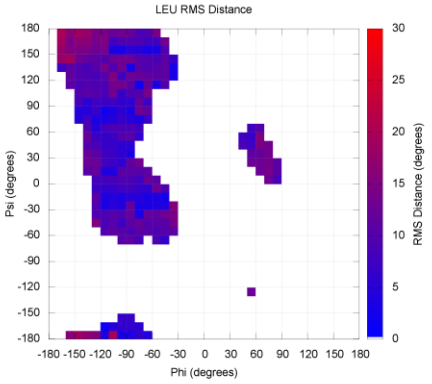

GLU Percent Overlap
Psi (degrees)
Phi (degrees)
Percent

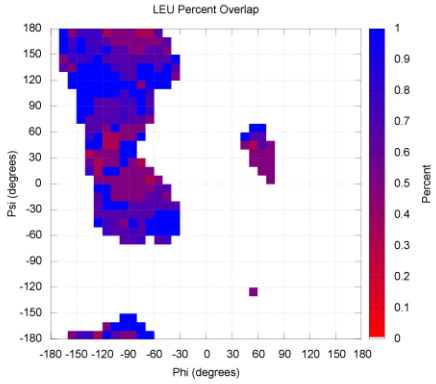

HIS RMS Distance
Psi (degrees)
Phi (degrees)
RMS Distance (degrees)

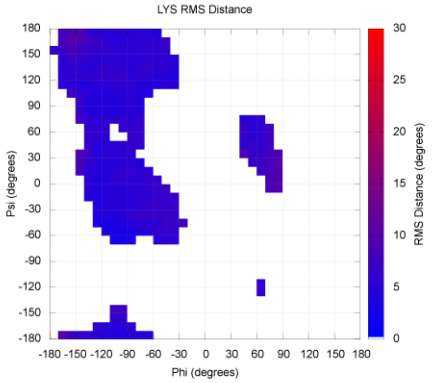

HIS Percent Overlap
Psi (degrees)
Phi (degrees)
Percent

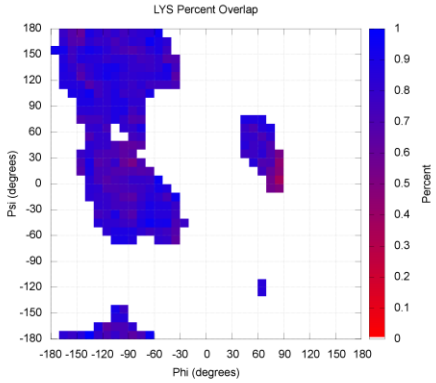

ILE RMS Distance
Psi (degrees)
Phi (degrees)
RMS Distance (degrees)

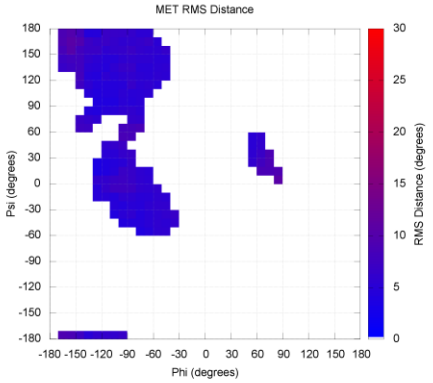

ILE Percent Overlap
Psi (degrees)
Phi (degrees)
Percent

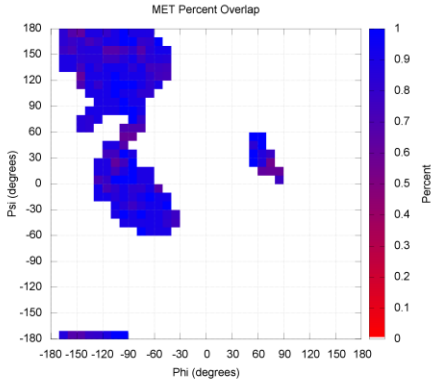

LEU RMS Distance
Psi (degrees)
Phi (degrees)
RMS Distance (degrees)

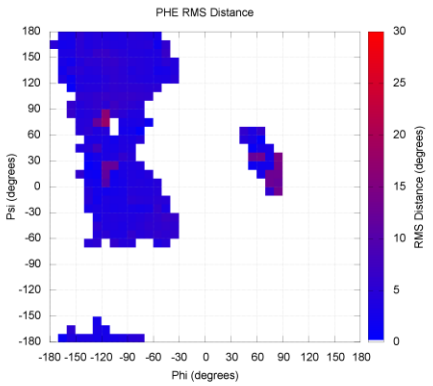

LEU Percent Overlap
Psi (degrees)
Phi (degrees)
Percent

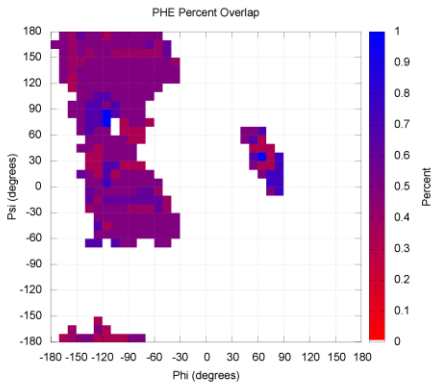

LYS RMS Distance
Psi (degrees)
Phi (degrees)
RMS Distance (degrees)

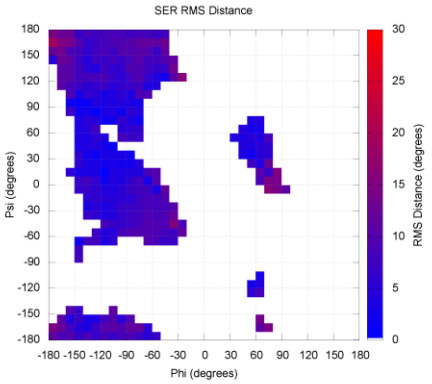

LYS Percent Overlap
Psi (degrees)
Phi (degrees)
Percent

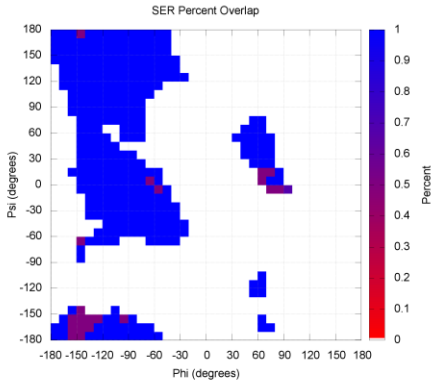

MET RMS Distance
Psi (degrees)
Phi (degrees)
RMS Distance (degrees)

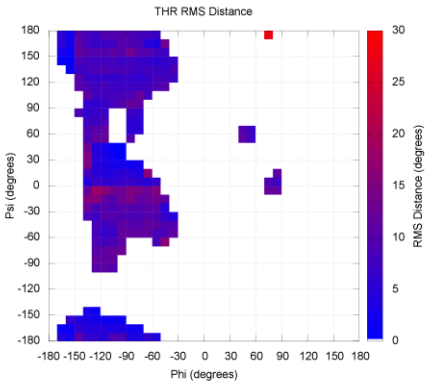

MET Percent Overlap
Psi (degrees)
Phi (degrees)
Percent

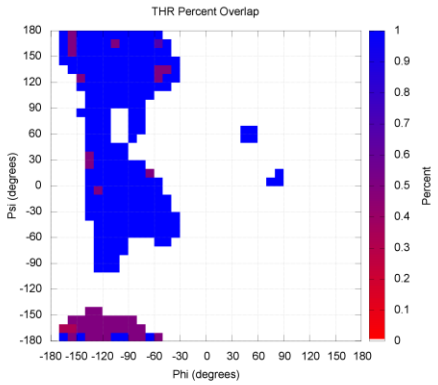

PHE RMS Distance
Psi (degrees)
Phi (degrees)
RMS Distance (degrees)

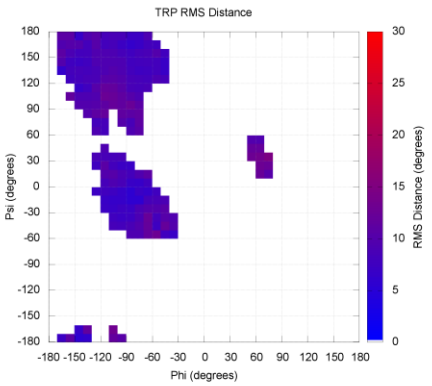

PHE Percent Overlap
Psi (degrees)
Phi (degrees)
Percent

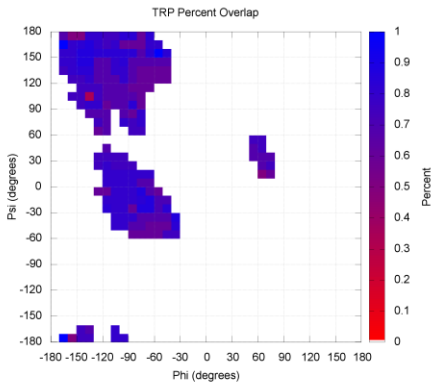

SER RMS Distance
Psi (degrees)
Phi (degrees)
RMS Distance (degrees)

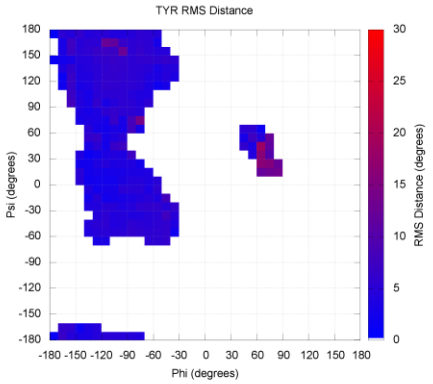

SER Percent Overlap
Psi (degrees)
Phi (degrees)
Percent

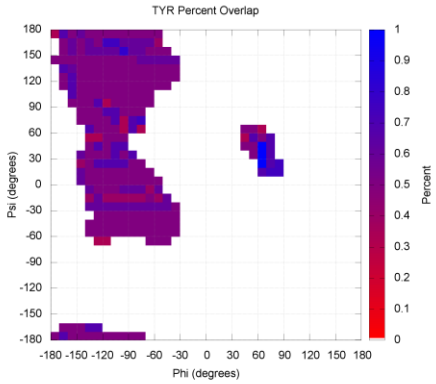

THR RMS Distance
Psi (degrees)
Phi (degrees)
RMS Distance (degrees)

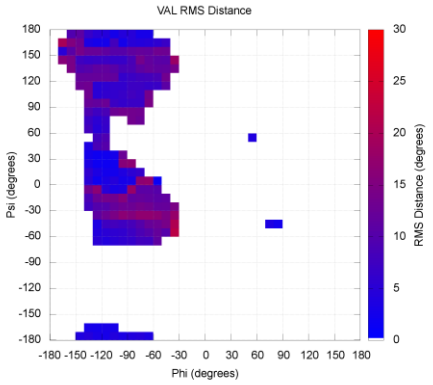

THR Percent Overlap
Psi (degrees)
Phi (degrees)
Percent


TRP RMS Distance
Psi (degrees)
Phi (degrees)
RMS Distance (degrees)


TRP Percent Overlap
Psi (degrees)
Phi (degrees)
Percent


TYR RMS Distance
Psi (degrees)
Phi (degrees)
RMS Distance (degrees)


TYR Percent Overlap
Psi (degrees)
Phi (degrees)
Percent


VAL RMS Distance
Psi (degrees)
Phi (degrees)
RMS Distance (degrees)

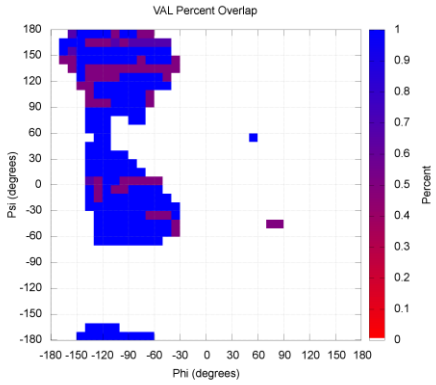

VAL Percent Overlap
Psi (degrees)
Phi (degrees)
Percent

Figure 2.3 The structure of 2-indynal-glycine. The structure of 2-indynal-glycine is shown in a dipeptide context with φ = -150 and ψ = 150.The different pucker state of the five member ring of 2-indynal glycine are modeled as separate amino acid type by Rosetta because of the difficulty in using rotamer libraries to capture coordinated movements that involved rotation about multiple dihedral angles. There is a 1.45 kcal/mol energy difference between the “down” conformer (left) and the “up” conformer (right) with the “up” conformer being lower in energy.

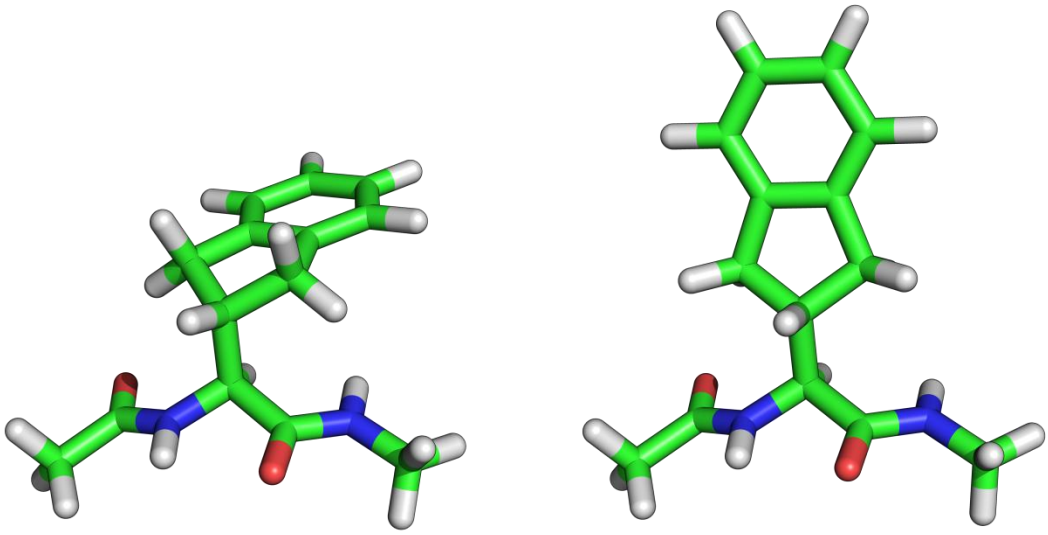
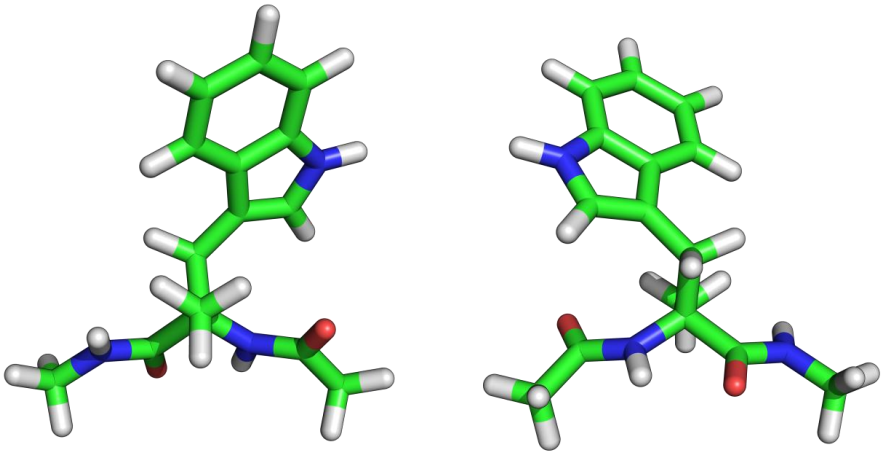
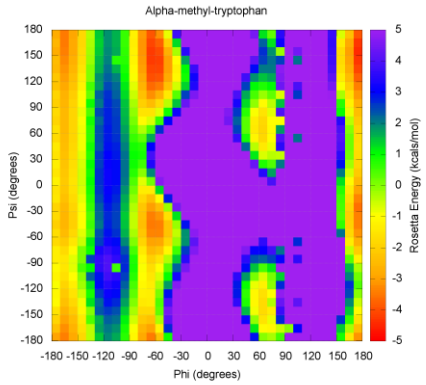


Figure 2.4 Structure of α-methyl-tryptophan.The structure of α-methyl-tryptophan is shown in a dipeptide context with φ = -150 and ψ = 150, images are rotated 180 degrees with respect to each other (top). Plots of backbone the energy landscape of α-methyl-tryptophan and tryptophan (bottm left) and canonical tryptophan (bottom right) as calculated by Rosetta. Calculations were done in a didpeptide context where the backbone φ and ψ were fixed, the side chain was repacked and minimized for each phi and psi bin in 5 degree intervals. Colors represent energy of the didpeptide in kcals/mol with red being the lowest energy and most preferred backbone conformation.





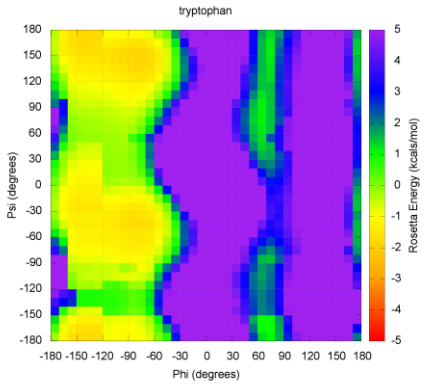
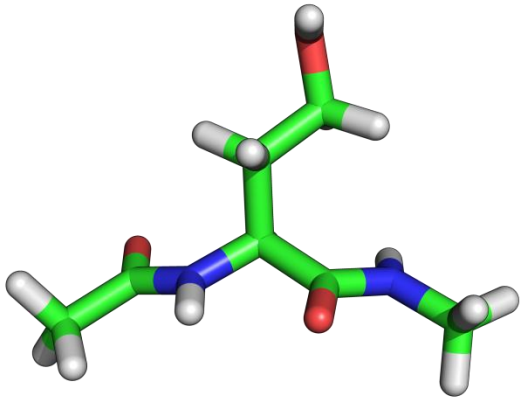


Figure 2.5 The structure of homoserine in a didpeptide context with φ = -150 and ψ = 150.



Biblography

1. Dahiyat, B.I. and S.L. Mayo, *De Novo Protein Design: Fully Automated Sequence Selection.* Science, 1997. **278**(5335): p. 82-87.

2. Jain, T., *Configurational-bias sampling technique for predicting side-chain conformations in proteins.* Protein Science, 2006. **15**(9): p. 2029-2039.

3. Rothlisberger, D., et al., *Kemp elimination catalysts by computational enzyme design.* Nature, 2008. **453**(7192): p. 190-5.

4. Milton, R.C., S.C. Milton, and S.B. Kent, *Total chemical synthesis of a D-enzyme: the enantiomers of HIV-1 protease show reciprocal chiral substrate specificity [corrected].* Science, 1992. **256**(5062): p. 1445-8.

5. Fung, H.B. and Y. Guo, *Enfuvirtide: a fusion inhibitor for the treatment of HIV infection.* Clin Ther, 2004. **26**(3): p. 352-78.

6. Horng, J.C. and D.P. Raleigh, *phi-Values beyond the ribosomally encoded amino acids: kinetic and thermodynamic consequences of incorporating trifluoromethyl amino acids in a globular protein.* J Am Chem Soc, 2003. **125**(31): p. 9286-7.

7. Hendrickson, W.A., J.R. Horton, and D.M. LeMaster, *Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure.* EMBO J, 1990. **9**(5): p. 1665-72.

8. Banerjee, R., S. Chattopadhyay, and G. Basu, *Conformational preferences of a short Aib/Ala-based water-soluble peptide as a function of temperature.* Proteins, 2009. **76**(1): p. 184-200.

9. Bradley, P., K.M.S. Misura, and D. Baker, *Toward High-Resolution de Novo Structure Prediction for Small Proteins.* Science, 2005. **309**(5742): p. 1868-1871.

10. Butterfoss, G.L. and B. Kuhlman, *COMPUTER-BASED DESIGN OF NOVEL PROTEIN STRUCTURES.* Annual Review of Biophysics and Biomolecular Structure, 2006. **35**(1): p. 49-65.

11. Creasy, D.M. and J.S. Cottrell, *Unimod: Protein modifications for mass spectrometry.* PROTEOMICS, 2004. **4**(6): p. 1534-1536.

12. Garavelli, J.S., *The RESID Database of Protein Modifications as a resource and annotation tool.* PROTEOMICS, 2004. **4**(6): p. 1527-1533.

13. Hornbeck, P.V., et al., *PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation.* PROTEOMICS, 2004. **4**(6): p. 1551-1561.

14. Bock, A., et al., *Selenocysteine: the 21st amino acid.* Mol Microbiol, 1991. **5**(3): p. 515-20.

15. Srinivasan, G., C.M. James, and J.A. Krzycki, *Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA.* Science, 2002. **296**(5572): p. 1459-62.

16. Renfrew, P.D., G.L. Butterfoss, and B. Kuhlman, *Using quantum mechanics to improve estimates of amino acid side chain rotamer energies.* Proteins, 2008. **71**(4): p. 1637-1646.

17. Rohl, C.A., et al., *Protein Structure Prediction Using Rosetta*, in *Numerical Computer Methods, Part D.* 2004, Academic Press. p. 66 - 93-66 - 93.

18. Dunbrack, R.L., *Rotamer Libraries in the 21st Century.* Current Opinion in Structural Biology, 2002. **12**(4): p. 431-440.

19. Engh, R.A. and R. Huber, *Accurate bond and angle parameters for X-ray protein structure refinement.* Acta Crystallographica Section A, 1991. **47**(4): p. 392-400.

20. Jr, R.L.D. and F.E. Cohen, *Bayesian statistical analysis of protein side-chain rotamer preferences.* Protein Science : A Publication of the Protein Society, 1997. **6**(8): p. 1661–1681-1661–1681.

21. Neria, E., S. Fischer, and M. Karplus, *Simulation of activation free energies in molecular systems.* The Journal of Chemical Physics, 1996. **105**(5): p. 1902-1921.

22. Lazaridis, T. and M. Karplus, *Effective energy function for proteins in solution.* Proteins: Structure, Function, and Genetics, 1999. **35**(2): p. 133-152.

23. Kortemme, T., A.V. Morozov, and D. Baker, *An Orientation-dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein-Protein Complexes.* Journal of Molecular Biology, 2003. **326**(4): p. 1239-1259.

24. Kuhlman, B., et al., *Design of a Novel Globular Protein Fold with Atomic-Level Accuracy.* Science, 2003. **302**(5649): p. 1364-1368.

25. Gordon, D.B., S.A. Marshall, and S.L. Mayo, *Energy functions for protein design.* Curr Opin Struct Biol, 1999. **9**(4): p. 509-13.

26. Petrella, R.J., T. Lazaridis, and M. Karplus, *Protein sidechain conformer prediction: a test of the energy function.* Folding and Design, 1998. **3**(5): p. 353-377.

27. Brooks, B.R., et al., *CHARMM: A program for macromolecular energy, minimization, and dynamics calculations.* Journal of Computational Chemistry, 1983. **4**(2): p. 187-217.

28. Wang, G. and R.L. Dunbrack, Jr., *PISCES: a protein sequence culling server.* Bioinformatics, 2003. **19**(12): p. 1589-91.

29. Lovell, S.C., et al., *The penultimate rotamer library.* Proteins: Structure, Function, and Bioinformatics, 2000. **40**(3): p. 389-408.

30. Josien, H., et al., *Design and Synthesis of Side-Chain Conformationally Restricted Phenylalanines and Their Use for Structure-Activity Studies on Tachykinin NK-1 Receptor.* Journal of Medicinal Chemistry, 1994. **37**(11): p. 1586-1601.

31. Diksic, M., *Labelled alpha-methyl-L-tryptophan as a tracer for the study of the brain serotonergic system.* J Psychiatry Neurosci, 2001. **26**(4): p. 293-303.

32. Ho, B. and D. Agard, *Identification of new, well-populated amino-acid sidechain rotamers involving hydroxyl-hydrogen atoms and sulfhydryl-hydrogen atoms.* BMC Structural Biology, 2008. **8**(1): p. 41-41.

33. Butterfoss, G.L., et al., *A preliminary survey of the peptoid folding landscape.* J Am Chem Soc, 2009. **131**(46): p. 16798-807.

Using NCAAs in Peptide/Protein Interface Design

Introduction

Computational protein design is becoming an increasingly powerful tool that has enabled us to create new protein folds, increase the binding affinity of protein/protein interfaces, and modify or predict the specificity of protein/protein binding pairs [1-5].

Computational protein design programs have traditionally only used the 20 CAAs during the design process. This is in part due to the difficulties of incorporating NCAAs in to proteins in vivo and the limitations on size during chemical synthesis [6]. Furthermore since computational protein design programs often make use of knowledge-based potentials which are incompatible with NCAAs because there is not enough structural information to derive meaningful statistics [7]. In the previous chapter we described methods we have developed to incorporate NCAAs into the computational protein design program Rosetta, developed in our lab. These changes allow us to design proteins using NCAAs. In this chapter we describe the use of these methods to increase the binding affinity of a subdomain of the calpastatin peptide for a domain of the protein calpain.

The calcium dependant cysteine protease, calpain, is involved in many important pathways[8]. There are many isoforms of calpain some of which are ubiquitous and some of which are tissue specific. Two ubiquitous isoforms of calpain, caplain-1 and calpain-2 also know as mu-calpain and m-calpain respectively, are the most well studied of the various isoforms. These two isoforms of calpain function as a heterodimers. An 80 kD subunit comprises the first 4 domains (DI – DIV) of calpain1 or calpain2 and a 30 kD subunit also known as calpain4 comprises the last 2 domains (DV and DVI) and the other partner in the heterodimer[9]. The 80 kD subunits of the isoforms share approximately 60% sequence identity and are structurally similar. The heterodimer interface is between DIV and DVI.

DIV and DVI are structurally similar and both contain five EF-hand domains. Four of the EF-hand domains in each DIV and DVI bind calcium. The fifth EF-hand domains of each interact to form the heterodimer interface as shown in figure 1. Upon calcium binding by the EF-hand motifs in DIV and DVI and in DII, the heterodimer undergoes a conformational change that brings the residues that comprise the protease active site into alignment, activating the enzyme [10, 11]. This conformational change also allows calpastatin to bind, which inhibits calpain. Calpastatin is an unstructured protein that contains an N-terminal domain that is involved in membrane localization, known as the L domain. The L-domain is followed by four repeats of three subdomains: A, B, and C. Subdomains A and C interact as amphipathic α-helices binding to hydrophobic patches on DIV and DVI, respectively, of either calpain-1 or calpain-2. Subdomain B interacts with the active site of the catalytic DII, forming a loop around the active site cysteine to avoid cleavage. Calpastatin is able to inhibit 4 calpain heterodimer complexes.

The number of proteins targeted for proteolysis by calpain implicates it in a variety of diseases [12, 13]. Inhibitors of calpain could be of potential therapeutic use to treat the symptoms of these diseases. We have computationally redesigned positions on the interface between subdomain C of calpastatin and DVI of calpain-1 by allowing NCAAs at the calpastatin positions.

Materials and Methods

Peptide/Protein Interface Design Protocol

The program Rosetta 3.0 was used to perform the interface redesigns. The design protocol iterates between two phases, a backbone perturbation phase that searches through different backbone conformations, and a design phase that searches for low energy sequences to fit the current backbone. The perturbation phase has 2 parts: a backbone perturbation and a round of "rotamer trials." First one of the following backbone perturbations is performed on the peptide/protein complex: a "small" move, where φ or ψ of a randomly chosen residue on the peptide is rotated by up to 3 degrees, a

"shear" move, where the φ of a random residue on the peptide is rotated up to 3 degrees and the ψ of the preceding residue is rotated by an equal amount in the opposite direction, a ridged-body translation of the peptide in the binding pocket, or a ridged-body rotation of the peptide in the binding pocket. Each one of these perturbations is followed by a round of rotamer trials for each residue whose energy increased as a result of the perturbation. Rotamer trials is a fast side chain optimization routine where for each residue in a set of residues, the best rotamer is chosen given the current context; it proceeds over the set of residues in a random order. The design phase of the protocol also consists of two parts: a round of the "pack rotamers" routine followed by gradient minimization. The pack rotamers routine is a rotamer optimization routine that tries to find the best combination of rotamers given a set of residues using a simulated annealing Monte Carlo/Metropolis search. The routine randomly chooses a single rotamer to replace (rotamers can be from different amino acid types than the current amino acid at the position) and determines the energy of the complex if the change is made. If the energy of the complex decreases the change is accepted, if the change increases the energy, the change is accepted based on the Metropolis criterion. Following the pack rotamers routine, gradient based minimization of the complex is performed. Both backbone and side chain dihedrals of the peptide and side chain dihedrals of the protein as well as the distance between the peptide and the protein are allowed as degrees of freedom. All residue side chains on the peptide were allowed to repack but only residues within 6 angstroms of any residue in the peptide were allowed to be repacked. To generate a single design we perform 50 iterations of 100 cycles of the perturbation phase followed by 1 cycle of the design phase. The protocol is not designed to find a new binding mode but to allow enough flexibility in the interface to allow the possible incorporation of NCAAs. All designs were created using the 2.0 angstrom resolution crystal structure of a calpain-4 domain DVI bound to a 19mer peptide of calpastatin comprising subdomain C of the first inhibitory repeat (protein databank code 1NX1). Only 11 residues of the peptide were resolved in the crystal structure (positions 601–611). The structure contains a homodimer of DVI in the asymmetric unit with a calpastatin bound to each monomer. The CA RMSD between the calpain chains is 0.28 angstroms.

Calpain chain A and calpastatin chain C were used for the design because the b-factors of residues at the calpastatin binding site were lower than in the other interface. Before designing, the entire protein was repacked and minimized using standard Rosetta routines.

Allowing all NCAAs at all positions on the peptide is too computationally intensive given the current resources. To pick out point mutations for the initial rounds of experimental testing we did 500 independent runs allowing all NCAAs at a single position while keeping the sequence of the other positions fixed for each position in the peptide. Additionally we performed 256 independent runs where we individually tried each NCAA at each position in the peptide. Results were evaluated based on the total energy of the structure and the predicted binding energy of the structure calculated as the difference in energy of the complex and the unbound chains.

Purification of Calpain, Calpastatin, and Calpastatin Mutants

Calpain was expressed as a GST-fusion protein in *E. coli* that had been transfected with a pet41b vector that contained the gene encoding porcine calpain-1.The cells were grown in Luria-Bertani broth with 50ug/ml kanamycin at 37 degrees Celsius to an OD_{600} of 0.6 at which point 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to induce expression. Cells were grown for an additional 4 hours and harvested by centrifugation. Cells were resuspended in a buffer containing 50mM $NaPO_4$, 150mM NaCl, 5mM BME at pH 8.0 and lysed by sonication. Lysate was centrifuged at 12500g for 30 minutes and the supernatant was run over a GSTrap-FF column that had been equilibrated with the lysis buffer. After loading, the column was washed with 10 column volumes of the lysis buffer before the GST-calpain was eluted with 50mM Tris and 10mM reduced glutithione, pH 8.0. The eluent was monitored by absorbance at 280 nm. 5 mL fractions were collected. Those fractions which absorbed at 280 nm were pooled and divided in half. Thrombin was added to the first half to separate the calpain from the GST and the cleavage reaction was allowed to cleave overnight at 4 Celsius. Both halves were further purified with a Sephacryl S-200 gel filtration column using a buffer containing 50mM $NaPO_4$, 50mM NaCl, 5mM BME, 1mM $CaCl_2$, 1mM EDTA at pH 8.0.

Proteins were concentrated through centrifugation and found to be pure and ran true to predicted size on SDS-PAGE gels.

Wild type and 4-methyl-phenylalanine mutant peptides were synthesized by the Tuffs University Core Facility. The sequence used was PDDAIDALSDDXT-amide, where for the wild type peptide the X is a phenylalanine and for the mutant it is a 4-methyl-phenylalanine. The sequence was label with a fluocine dye through an N-terminal β-alanine linker.

Fluorescence Polarization Binding Assays

Purified calpain was manually titrated in to a solution containing 500nM calpastatin with 50mM NaP, 50mM NaCl, 5mM BME, 1mM $CaCl_2$, 1mM EDTA at pH 8.0, till the change in fluorescence polarization reached a plateau. Binding assays were performed at room temperature, 3 polarization readings were averaged for each concentration. Disassociation constants were calculated by fitting the data to a single state binding model using Sigma plot software.

Results

Computational Positional Results

Calpastatin binds as an amphipathic α-helix in a hydrophobic pocket between EF hands 1 and 2 of calpain DVI. Todd *et al.* identified calpastatin positions leu606 and phe610 as being the main residues involved in binding based on the crystal structure[9]. We have preformed design using the protocols described above. The results of the design runs were screened based on the predicted total energy and predicted change in binding energy. At positions 601, 603, 604, 605, and 608, Rosetta was unable to identify any mutations that scored better than the wild type residue. The modifications made to the energy function tend to favor mutations to large amino acids. The average attractive forces calculated for each amino acid are under estimated in the fragment based approach that was used to calculate the explicit unfolded energy. Predictions have been screened to remove designs that are a result of this. The predictions were analyzed by looking at structures for each mutation that had the best total

energy and the best binding energy. When the same protocol is run on the wild type sequence the total energy of the complex is -57.9 kcals/mol and the predicted binding energy is -13.4 kcals/mol. The results for all positions are summarized in table 1.

Position	Amino Acid	Total Energy	Binding Energy
602	2-amino-heptanoic acid	-62.3	-14.2
	allo-isoleucine	-59.3	-13.9
	alpha-aminoadipic acid	-59.9	-13.2
	fluoro-alanine	-58.2	-13.5
	trifluoro-alanine	-58.4	-13.5
	norvaline	-60.2	-13.9
606	homophenylalanine	-59.4	-15.9
	fluoro-leucine	-60.6	-13.6
	trifluoro-leucine	-59.7	-13.8
607	amino-butyiric acid	-61.9	-15.1
	2-allyl-glycine	-62.0	-14.4
	beta-chloro-alanine	-61.0	-15.2
	fluoro-alanine	-60.8	-14.9
	trifluoro-alanine	-61.4	-14.7
609	1-methyl-histidine	-61.8	-15.2
	homoserine	-59.5	-14.2
610	4-methyl-phenylalanine	-60.0	-14.2
611	3-methyl-histidine	-61.3	-14.9
	4-hydroxy-phenylglycine	-59.6	-14.7

Table 3.6 Summary of the Rosetta predictions for the redesign of the calpain/calpastatin interface that could improve the binding affinity.

Position 602

The wild type alanine packs against Leu106, Leu102, and the hydrophobic moiety of Gln105 (figure 2A). Fluoro-alanine and trifluoro-alanine (figure2E and 2F) are both able to fit in the space occupied by the alanine and have comparable energies to the wild type design. It has previously been shown that single mutations of hydrophobic residue to a fluorinated analog have been shown to stabilize proteins[14]. The rotamers of the surrounding residues are not disturbed by either mutation. Allo-isoleucine and norvaline (figure 2C and 2G) are able to interact with the same residue as the wild type alanine but interact with more hydrophobic surface area. The overall structure of the calpastatin helix

is slightly perturbed as a result of the mutation and the N-terminal end of the peptide moves out of the binding pocket by approximately 1 angstrom to accommodate the slightly larger side chains. The CB2 of the allo-isoleucine is in the same place as the CB of the alanine. 2-amino-heptanoic acid and α-aminoadipic acid (figure 2B and 2D) are also predicted to increase the binding affinity if placed in this position. The additional length of these residues allows them to interact with more of the hydrophobic surface area. Furthermore, α-aminoadipic acid can additionally form a weak salt bridge with Lys124, further stabilizing the interaction.

Position 606

The leucine at position 606 is one of the primary residues involved in binding[9]. It is buried in a large hydrophobic pocket that forms upon calcium binding and interacts with Leu102, Leu106, Ile121, and Trp166 (figure 3A). The leucine residue is unable to fill the entire pocket; however, mutation of this residue to a homophenylalanine is able to occupy more of the space in the cavity (figure 3B). The epsilon carbons of the homophenylalanine are placed in the same position on the leucine CD1 and CD2 leaving the rest of the side chain to penetrate deeper in to the pocket. The pocket is also able to accommodate fluoro-leucine and trifluoro-leucine (figure 3B and 3C). The larger fluorine atoms allow this modified leucine to occupy more of the pocket. Rosetta considers the mutations more favorable than the wild type leucine.

Position 607

The wild type serine forms a weak hydrogen bond with His129 (figure 4A). Rosetta predicts that small hydrophobic amino acids placed in the hydrophobic pocket made by residues Val125, Ile603, and Arg128 will increase the binding affinity. Amino butyric acid and β-chloro-alanine (figure 4B and 4C) are predicted to increase the binding affinity by approximately 2 kcals/mol while 2-allyl-glycine, fluoro-alanine, and trifluoro-alanine (figure 4D-F) by approximately 1 kcal/mol. None of these mutations affect the position of the peptide in the binding pocket.

Position 609

The wild type aspartic acid makes a hydrogen bond with Trp166, one of three hydrogen bonds between the peptide and the protein (figure 5A). The residues preferred by Rosetta both keep this hydrogen bond intact. 1-Methyl-histidine (figure 5B) forms an ideal hydrogen bond with Trp166. The hydrogen to acceptor distance is 1.9 angstroms. The aliphatic part of the methyl-histidine packs against phe99, leu102, lys170, and ala605. Homoserine (figure 5C) is also able to make the hydrogen bond to trp166. The hydrogen to acceptor distance is 2.1 angstroms. The difference in functional groups between the asp and the homoserine allows the homoserine to form more ideal geometry.

Position 610

The wild type phenylalanine is buried in a large hydrophobic pocket and along with Leu606 forms the main hydrophobic interface. The phenylalanine interacts with Trp166, His129, Leu132, Val125, Ile169, Phe224, and the hydrophobic portion of Gln173 (figure 6A). The crystal structure shows that the pocket is not entirely filled by the phenylalanine. Rosetta predicts that a 4-methyl-phenylalanine (figure 6B) can fill more of the cavity and creates more hydrophobic contacts without disrupting the overall binding, and would therefore have an increased binding affinity. This design has been tested and shown to increase the binding affinity (discussed bellow).

Position 611

The backbone of the wild type threonine makes hydrogen bonds to the side chain of gln173 while the hydroxyl group of the side chain is solvent exposed (figure 7A). Rosetta predicts that 3-Methyl-histidine (figure 7B) is able to pack its hydrophobic side in to the hydrophobic part of the Lys177 side chain and Ala174 while the nitrogen is solvent exposed. Additionally 4-hydroxy-phenylglycine (figure 7C) is able to pack against Lys177 and Ala174. It does expose one face of the phenyl ring to solvent which could affect the solubility of the peptide.

Fluorescence Polarization Binding Assays

Fluorescence polarization binding assays were conducted using the wild type peptide and a designed peptide where phe610 was mutated to a 4-methyl-phenylalanine. Assays were conducted as described as above using the two peptides with calpain and a GST-calpain fusion protein. The disassociation constant for the wild type calpastatin peptide with the cleaved-calpain and GST-calpain is 1.60μM and 1.50μM respectively. The disassociation constant for the 4-methyl-phenylalanine mutant calpastatin peptide with the cleaved-calpain and GST-calpain is 0.73μM and 0.81μM respectively (figure 9). There is excellent agreement between the data and the fit with all correlation coefficient above 0.98.

Conclusions

We have shown that the use of NCAAs in computational protein design is a potentially powerful that can be used to increase the binding affinity of a peptide-protein complex. The addition of only 3 additional atoms to a key residue in the binding interface is enough to lower the disassociation constant almost two fold. The design of inhibitors of calpain is an area of active research and the design of inhibitors that bind to area other than the active site is thought to increase the specificity of the inhibitors for calpain[12]. The small molecule inhibitor 3-(4-iodophenyl)-2-mercapto-(Z)-2-propenoic acid (also known as PD150606) discovered by Wang *et al.* [15] binds to calpain in the same hydrophobic pocket as position 610 and resembles the 4-methyl-phenylalanine predicted by Rosetta[9] and found to lower the disassociation constant. The structure of the inhibitor bound to the calpain has been solved (protein databank code 1NX3) and is shown superimposed with our design in figure 8. The high degree of structural similarity between the inhibitor and 4-methyl-phenylalanine and the similarity between the predicted binding mode and the structure of the bound inhibitor gives us confidence that our peptide is binding in a similar fashion [9, 15]. It is clear that additional experimental screening needs to be done as well as testing in additional model systems, but we are encouraged by the results we have gotten to this point.

Figure 3.1 The structure of calpain and calpastatin. A) the calpain-1 DI-DVI (green) with calpain-4 DVI (cyan) with a calpastatin subdomains A,B, and C (magenta). Dashed lines are where there was no density in the crystal structure for calpastatin. B) Close up view of the interaction between subdomain C of calpastatin and DVI of calpain-4 (colors as in A).

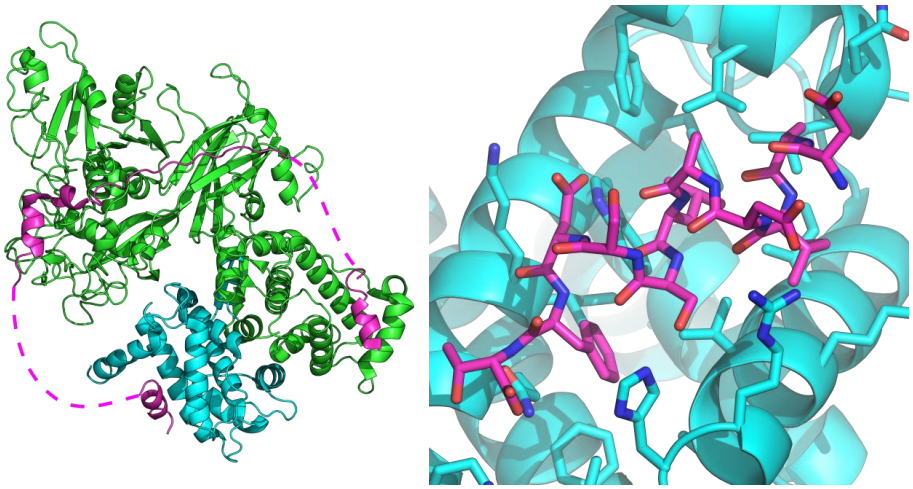
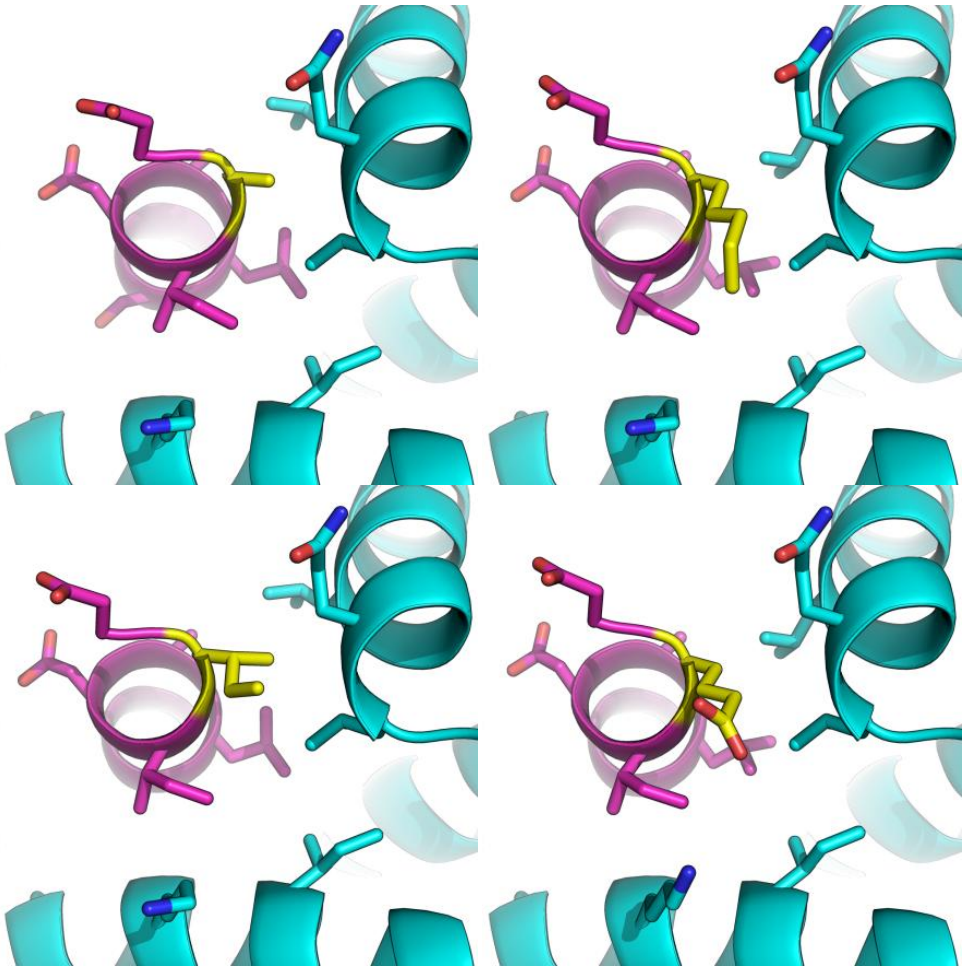
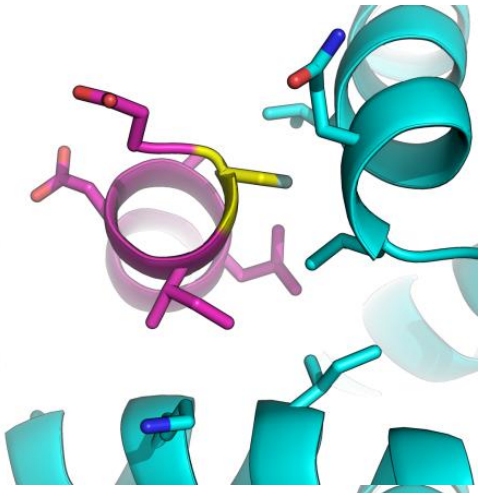
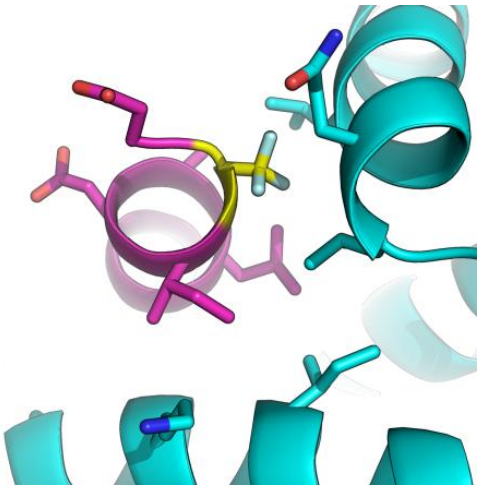


Figure 3.2 Rosetta predictions for calpastatin position 602, colors as in figure 1 with 602 in yellow.Wild type alanine (A), 2-amino-heptanoic acid (B), allo-isoleucine (C), α-aminoadipic acid (D), fluoro-alanine (E), trifluoro-alanine (F), and norvaline (G).







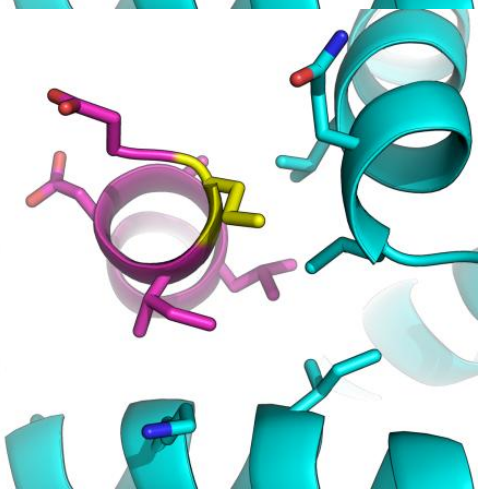


Figure 3.3 Rosetta predictions for calpastatin position 606, colors as in figure 1 with 606 in yellow. Wild type leucine (A), homophenylalanine (B), trifluoro-leucine (C), and fluoro-leucine (D).

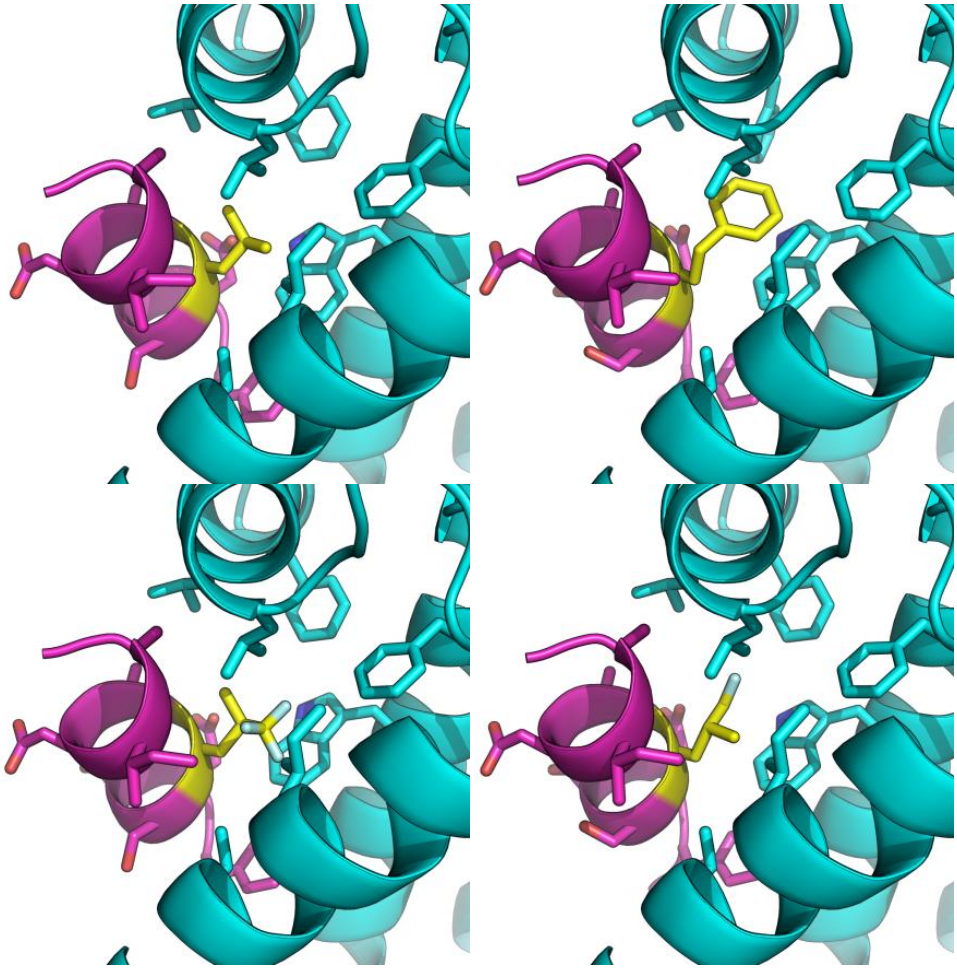


Figure 3.4 Rosetta predictions for calpastatin position 607, colors as in figure 1 with 607 in yellow. Wild type serine (A), 2-allyl-glycine (B), amino-butyeiric acid (C), β-chloro-alanine (D), fluoro-alanine (E), and trifluoro-alanine (F).

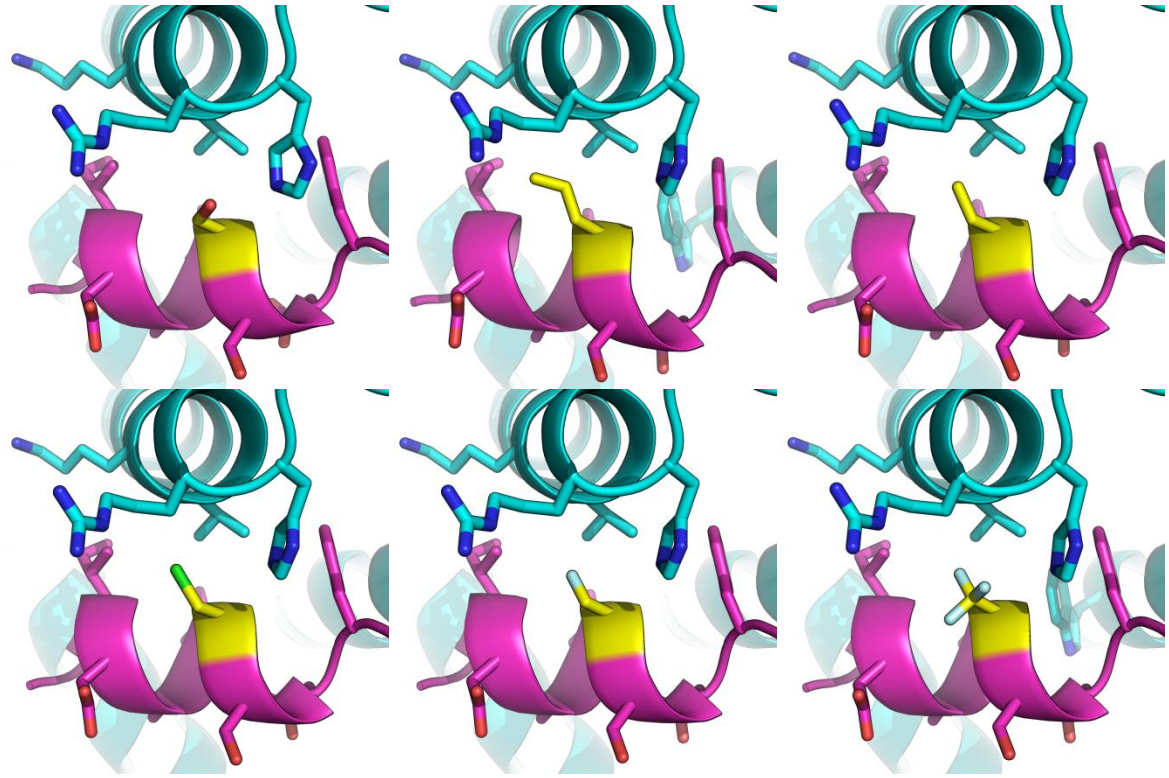


Figure 3.5 Rosetta predictions for calpastatin position 609, colors as in figure 1 with 609 in yellow. Wild type aspartic acid (A), 1-methyl-histidine (B), and homoserine (C).

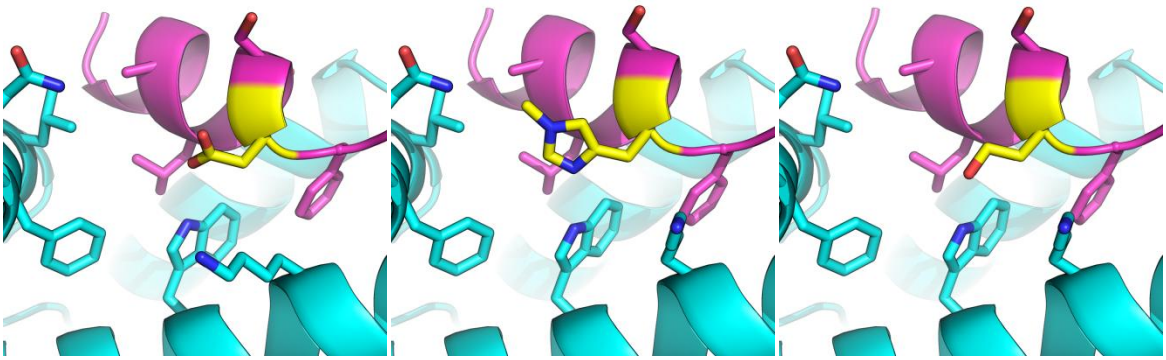


Figure 3.6 Rosetta predictions for calpastatin position 610, colors as in figure 1 with 610 in yellow. Wild type phenylalanine (A), and 4-methyl-phenyl-alanine (B).

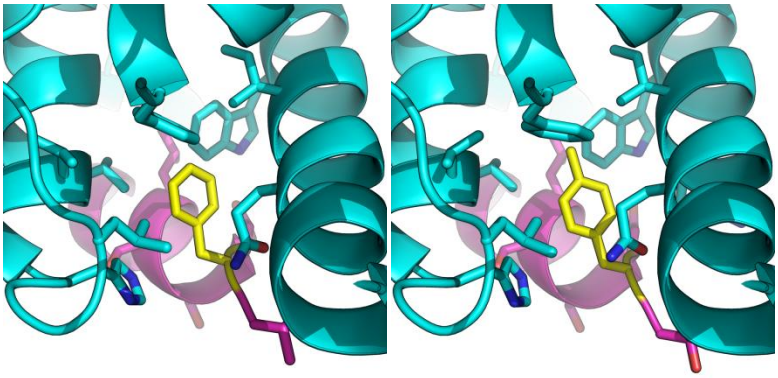


Figure 3.7 Rosetta predictions for calpastatin position 611, colors as in figure 1 with 611 in yellow. Wild type threoninie (A), 3-methyl-histidine (B), and 4-hydroxy-phenylglycine (C).

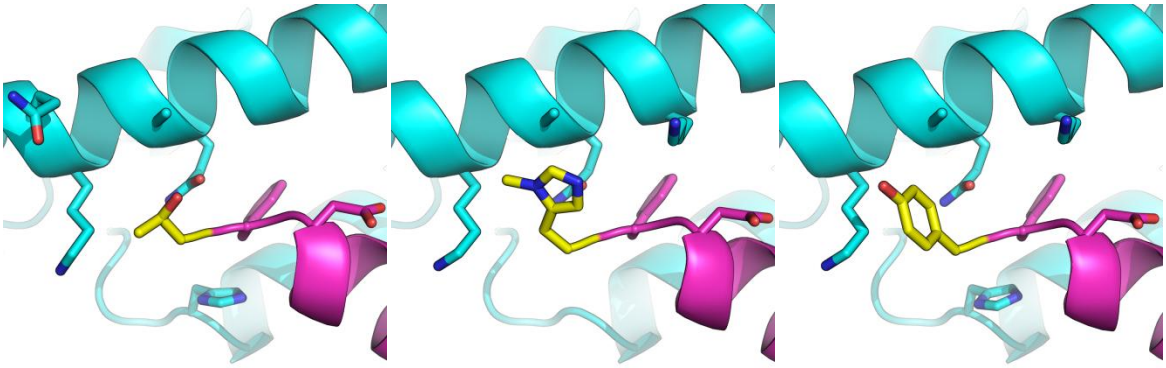


Figure 3.8 Comparison of the PD150560 inhibitor and the designed mutant. The structure of 4-methyl-phenylalanine closely resembles that of the inhibitor (A). The orientation of the PD150560 is identical to the predicted binding mode of the 4-methyl-phenylalanine (B).

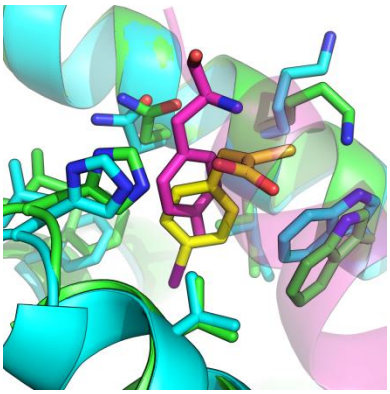
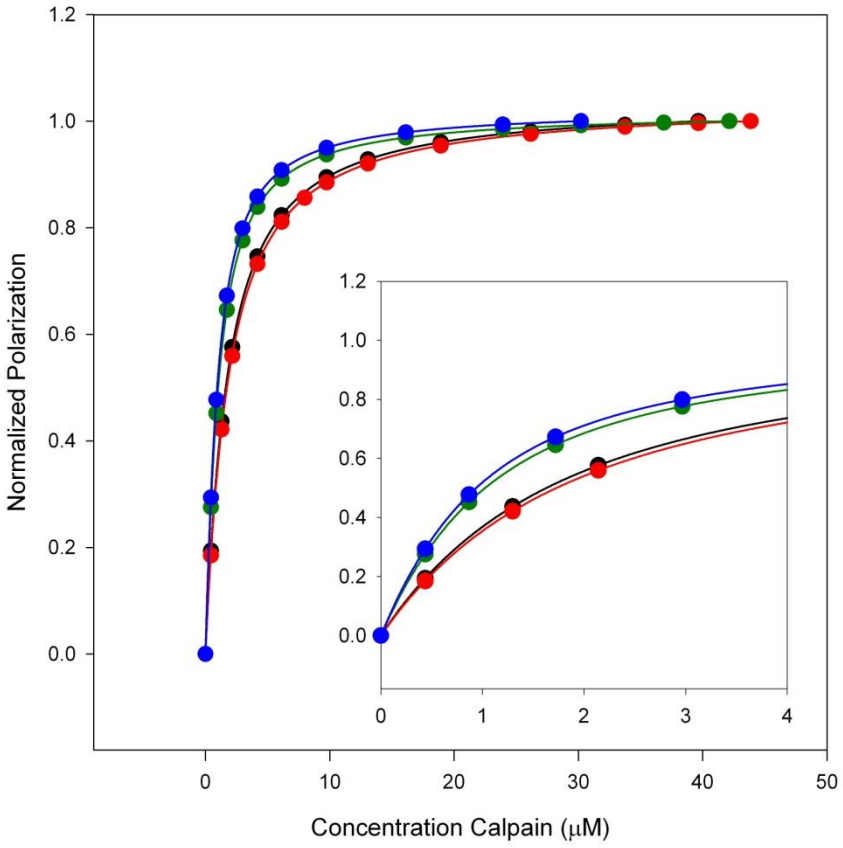


Figure 3.9 Calpain/calpastatin fluorescence polarization binging assays. Binding curves for calpain with wild type calpastatin (black,), GST-calpain fusion with wild type calpastatin (red), calpain with 4MF-calpastatin design (green), and GST-calpain with 4MF-calpastatin (blue). Inset shows detail for calpan concentrations between 0 and 4 μM. 4-methylphenylalnine mutants bind almost twofold tighter than wild type.



Bibliography

1. Kuhlman, B., et al., *Design of a Novel Globular Protein Fold with Atomic-Level Accuracy.* Science, 2003. **302**(5649): p. 1364-1368.

2. Sammond, D.W., et al., *Structure-based Protocol for Identifying Mutations that Enhance Protein-Protein Binding Affinities.* Journal of Molecular Biology, 2007. **371(**5): p. 1392-1404.

3. Kortemme, T., et al., *Computational redesign of protein-protein interaction specificity*. Nat Struct Mol Biol, 2004. **11**(4): p. 371-9.

4. Humphris, E.L. and T. Kortemme, *Design of Multi-Specificity in Protein Interfaces.* PLoS Comput Biol, 2007. **3**(8): p. e164-e164.

5. Humphris, E.L. and T. Kortemme, *Prediction of Protein-Protein Interface Sequence Diversity Using Flexible Backbone Computational Protein Design.* Structure, 2008. **16**(12): p. 1777-1788.

6. Hendrickson, T.L., V. de Crecy-Lagard, and P. Schimmel, *Incorporation of nonnatural amino acids into proteins.* Annu Rev Biochem, 2004. **73**: p. 147-76.

7. Gordon, D.B., S.A. Marshall, and S.L. Mayo, *Energy functions for protein design.* Curr Opin Struct Biol, 1999. **9**(4): p. 509-13.

8. Goll, D.E., et al., *The calpain system.* Physiol Rev, 2003. **83**(3): p. 731-801.

9. Todd, B., et al., *A Structural Model for the Inhibition of Calpain by Calpastatin: Crystal Structures of the Native Domain VI of Calpain and its Complexes with Calpastatin Peptide and a Small Molecule Inhibitor*. Journal of Molecular Biology, 2003. **328**(1): p. 131-146.

10. Hanna, R.A., R.L. Campbell, and P.L. Davies, *Calcium-bound structure of calpain and its mechanism of inhibition by calpastatin.* Nature, 2008. **456**(7220): p. 409-12.

11. Moldoveanu, T., K. Gehring, and D.R. Green, *Concerted multi-pronged attack by calpastatin to occlude the catalytic cleft of heterodimeric calpains.* Nature, 2008. **456**(7220): p. 404-8.

12. Carragher, N.O., *Calpain inhibition: a therapeutic strategy targeting multiple disease states*. Curr Pharm Des, 2006. **12**(5): p. 615-38.

13. Saez, M.E., et al., *The therapeutic potential of the calpain family: new aspects.* Drug Discov Today, 2006. **11**(19-20): p. 917-23.

14. Horng, J.C. and D.P. Raleigh, *phi-Values beyond the ribosomally encoded amino acids: kinetic and thermodynamic consequences of incorporating trifluoromethyl amino acids in a globular protein.* J Am Chem Soc, 2003. **125**(31): p. 9286-7.

15. Wang, K.K., et al., *Alpha-mercaptoacrylic acid derivatives as novel selective calpain inhibitors.* Adv Exp Med Biol, 1996. **389**: p. 95-101.

Using Noncanonical Amino Acids to Computationally Redesign an HIV Entry Inhibitor

Introduction

According to the Joint United Nations Programme on HIV/AIDS 2008 Report on the Global AIDS Epidemic (http://www.unaids.org), 25 million people have died of AIDS related causes throughout the course of the global epidemic. In 2007 there were approximately 33 million people living with AIDS/HIV, 2.7 million people became infected, and 2 million people died of AIDS related causes. In the part of the world most affected by the epidemic, sub-Saharan Africa, upwards of 5% of the population of several nations is infected and death rates due to AIDS have lowered the national life expectancy. Globally, the number of people infected each year is decreasing; however, new anti-retroviral therapeutics are still needed for prevention and treatment. New therapeutics are being developed that are designed to interfere with the mechanism by which the viral membrane and the host cellular membrane come to fuse together causing the virus to infect the cell[1]. Therapeutics that target this process are known as HIV entry or integration inhibitors.

The Kay group at the University of Utah designs HIV integration inhibitors constructed of the D-enantiomers of amino acids[2]. We have been collaborating with the Kay lab to redesign their D-peptide inhibitors to incorporate NCAAs and increase the binding affinity of the peptide for its target.

The HIV *Env* gene encodes an envelope glycoprotein called gp160. This protein is cleaved into gp120 and gp41 which are non-covalently bound to each other on the surface of the virus. Gp41 contains three main regions: a transmembrane region, shown to penetrate the host cell membrane, and two helical regions which, when expressed separately, are called the N and C peptide regions. It has been shown that expressed by themselves, the N and C peptides form a hexamer with three N peptides making a three-helix coiled-coil while the three C peptides bind anti-paralell in the grooves formed by

the coiled-coil[3]. Eckert and Kim reviewed the model of the integration process[4] described below and shown in figure 1. The process of infection begins when gp120 binds to the CD4 receptor on the surface of a host cell. The binding event triggers conformational changes in gp120 that allow it to additionally bind a coreceptor in the chemokine family. The conformational changes in gp120 induce conformational changes in gp41 that cause it to become extended. The transmembrane region of gp41 becomes buried in the host cell membrane. The regions that make up the N and C peptides are exposed; three N peptide regions form an exposed N trimer, and gp41 is acts like a bridge between the virus and host cell membranes. This complex of three gp41 molecules and three gp120 molecules is long lived (on the order of minutes) and is called the prehairpin complex. Eventually, gp41 undergoes a conformational change that brings the N and C peptides together in to a structure similar to the structure of the hexamer where the N and C peptide regions form a trimer of hairpins. This change brings the two membranes in close proximity and the virus membrane fuses with the membrane of the host cell.

The pocket where the C peptides bind to the N peptides is highly conserved[3]. Integration can be inhibited by finding molecules that target the N peptide region of gp41 in the prehairpin complex and prevent the formation of the hairpin structure from forming and the subsequent membrane fusion [4]. Inhibitors have taken the form of minimized C peptides, helical mimics, antibodies, and small molecules [1, 2].

Recently, the Kay group designed HIV integration inhibitors that target the N peptide region of gp41 using an 8mer peptide constructed of the D-enantiomers of amino acids[2]. The peptide was designed using mirror-image phage display. Mirror-image phage display works by starting with a target protein synthesized using the D-enantiomers of the amino acids. L-peptides are displayed on the phage and evolved to bind the D target. Through symmetry, the D-enantiomers of the evolved L-peptides will also bind to the L-enantiomer of the D-target[5]. Welch *et al.* used the D-enantiomer of 1QN17, a three-helix coiled-coil that mimics the N trimer region in the prehaipin complex of gp41, as a target

[6]. Three molecules of the inhibitor bind to conserved hydrophobic patches located at the interface between each pair of N peptide region. These patches are occupied by the C peptide region in the hairpin complex. The best published inhibitor, called PIE7, has an IC50 of 620 nM and a Kd of 80nM and a construct of three PIE7 peptides linked by poly-ethylene-glycol linkers has an IC50 of 250pM and a Kd of 70pM against a target that mimics the HXB2 strain of HIV. D-peptides are advantageous to use as potential therapeutics primarily because they are resistant to proteolysis[7]. Fuzeon (also known as T-20 or enfuvirtide) is the only FDA approved integration inhibitor [2, 6]. It is a 36 residue C peptide that binds to the N peptide region of the prehairpin complex. Patients must receive injections of 90 mg of fuzeon twice a day because the plasma elimination half-life is 3.8 hours as a result of proteolytic degradation[8].

The modified mirror-image phage display technique used by Welch *et al.* is only able to use the D-enantiomers of the 20 CAAs[2]. We have been collaborating with the Kay group to determine if using NCAAs in the design of their integration inhibitors can improve the binding affinity and create better therapeutics. As we have shown in the previous chapters, the use of NCAA side chains increases the number of possible sequences and in turn the number of conformations available during a design simulation. The peptide therapeutics will ultimately be created using solid state synthesis, which is compatible with NCAAs, but the diversity of sequences is currently limited to the 20 CAAs because of the constraints imposed by their design process.

Materials and Methods

Input Structure

The designs were based on unpublished structures of a new D-peptide inhibitor called PIE12. The 2 cysteine residues in the inhibitors form a disulfide bond that cyclizes the small peptides. The sequence of the peptides is the same between these cysteine residues, but the residues on either side of the cysteines have been optimized to increase the binding affinity (PIE7 sequence ACE-

KGACDYPEWQWLCAA-NH2, PIE12 sequence ACE-HPCDYPEWQWLCELGK-NH2). The N-terminal lysine and glycine linker, added to increase the solubility in PIE7, has been moved to the C-terminus, and the sequence length has increased by one residue to have two flanking residues outside of the cysteines. The Kay group solved the crystal structure of PIE12 in complex with IQN17, and a comparison of the PIE7 structure with the PIE12 structure is shown in figure 2. Three different crystal forms were obtained, and although the structures are very similar, there are slight differences between them giving us several views of the structure. The first crystal form, called G63, crystallized in space group P2(1) with a resolution of 1.55 angstroms contains three IQN17 molecules and three PIE12 molecules in the asymmetric unit. The second crystal form, called D32, crystallized in space group R3 with a resolution of 1.45 angstroms contains one molecule of IQN17 and one molecule of PIE12 in the asymmetric unit. The third crystal form, called F81, crystallized in space group P321 with a resolution of 1.45 angstroms contains one molecule of IQN17 and one molecule of PIE12 in the asymmetric unit.

The PIE7 and PIE12 structures are similar to the three IQN17 molecules in PIE 12 (G63) and PIE7 (pdb code 2R5D) have CA RMSD of 0.373 angstroms and residues between the cysteines in PIE12 (G63, chain H) and PIE7 (2R5D, chain H) have a CA RMSD of 0.181 angstroms. Differences do occur in the positions of the residues outside the cysteines. In PIE12, His3 packs against Pro4 while making a hydrogen bond to the backbone of Cys5, thus stabilizing the N-terminal region. Additionally, the additional Gly2 residue positions the backbone of the N-terminal Lys1 to form a hydrogen bond with Gln39 on IQN17 that is not possible in the PIE7 structure. On the C-terminal end, Leu16 is able to fill more of the hydrophobic patch formed by Val34 and Leu29 on IQN17 and Trp12 and Leu13 on the peptide, than the Ala15 in the same position in PIE7.

The D32 and F81 structures contain only one copy of IQN17 and the PIE12 peptide The biologically active complex was created by applying the appropriate symmetry operations in the program

Pymol[9]. The G63, D32, and F81 structures were then repacked and minimized using standard Rosetta routines.

Computational Design Protocol

The protocol used to redesign the interface of the HIV integration inhibitors is the same two phase protocol used to redesign the calpastatin/calpain interface in the previous chapter. Briefly, the protocol iterates between two phases: a backbone perturbation phase and a design and minimization phase. The perturbation phase either perturbs the backbone dihedral angles of the peptide or the orientation of the peptide in the binding pocket followed by a fast side chain packing routine. The design phase repacks the entire interface using a more rigorous packing routine and, in the process, can allow the sequence of positions on the peptide to change, followed by gradient based minimization.

PIE12 designs were initially done on the D-peptide chain H of the G63 structure. Allowing the D-enantiomers of all NCAAs at all positions in the peptide is not feasible given current computer resources. For each position in the peptide, 256 independent runs were carried out that allowed the position to mutate to any of the NCAAs. Gly2, being achiral, has backbone φ and ψ dihedral angles that are in the β sheet region of the Ramachandran plot for L-amino acids. The L-enantiomers of all the NCAAs were also tried at this position. The results of these design runs were screened based on the predicted total energy and predicted change in binding energy. Designs that were predicted to increase the binding affinity were additionally tested in the other chains of the G63 structure and on one of the chains in the D32 and F81 structures.

Results

The gp41/PIE12 interface is the most recent iteration of research conducted over the past decade to find a D-peptide integration inhibitor [2, 6]. The sequence is highly optimized and binds approximately 20 fold tighter than PIE7. Only two residue positions, Trp12 and Leu16, were found to

be amenable to mutation to NCAAs. At these two positions, Rosetta predicted that the binding affinity would increase upon mutation to a NCAA. Efforts have focused on increasing buried hydrophobic surface area of the interface.

Position	Amino Acid	Structure (Chain)	Total Energy	Δ Total Energy	Binding Energy	Δ Binding Energy
-	WT	D32	-114.6	-	-10.8	-
		F81	-148.9	-	-12.1	-
		G63(H)	-136.6	-	-11.5	-
		G63(K)	-137.1	-	-11.9	-
		G63(L)	-134.9	-	-13.7	-
12	4-methyl-tryptophan	D32	-116.2	-1.6	-11.9	-1.2
		F81	-149.4	-0.5	-13.7	-1.7
		G63(H)	-138.2	-1.6	-12.1	-0.6
		G63(K)	-137.1	0.0	-12.7	-0.8
		G63(L)	-137.8	-2.9	-14.3	-0.6
	5-methyl-tryptophan	D32	-116.2	-1.7	-11.8	-1.0
		F81	-147.5	1.4	-14.3	-2.2
		G63(H)	-139.0	-2.4	-12.6	-1.0
		G63(K)	-135.7	1.4	-12.1	-0.2
		G63(L)	-138.4	-3.5	-15.0	-1.3
	α-methyl-tryptophan	D32	-116.6	-2.0	-11.3	-0.5
		F81	-150.2	-1.3	-13.5	-1.4
		G63(H)	-140.9	-4.3	-12.0	-0.4
		G63(K)	-139.6	-2.5	-9.0	2.9
		G63(L)	-137.5	-2.6	-13.8	-0.1
16	2-amino-5-phenyl-propanoic acid	D32	-115.1	-0.5	-12.0	-1.2
		F81	-149.4	-0.5	-14.1	-2.0
		G63(H)	-138.3	-1.7	-12.4	-0.9
		G63(K)	-137.0	0.1	-12.4	-0.5
		G63(L)	-137.6	-2.6	-14.8	-1.1
	homoleucine	D32	-114.6	0.0	-11.1	-0.3
		F81	-148.7	0.2	-13.7	-1.7
		G63(H)	-137.9	-1.4	-12.8	-1.3
		G63(K)	-137.1	0.0	-12.4	-0.5
		G63(L)	-137.0	-2.0	-14.7	-1.0

Table 4.7 Summary of Rosetta predictions of point mutations to the GP41/PIE12 interface. Energies are in kcal/mol. Changes in energy are relative to the energy of the wild type chain of the same structure and chain.

Position 12

The wild type residue at position 12 is a tryptophan. Trp12 was one of the residues found by Eckert *et al*. to be important for binding in their original D-peptide design on which PIE7 and PIE12 are based [6]. Both Trp10 and Trp12 penetrate into a deep hydrophobic pocket formed at the interface between two IQN17 chains. On the protein, Trp12 packs against Val34, Lys38, and Ile37 and forms hydrogen bonds with Gln41 (figure 3A). On the peptide, Trp12 packs against Leu13, Trp10, and Leu16 and forms backbone hydrogen bonds with Leu16. Rosetta predicts three mutations at this position that may increase the binding affinity of the peptide for the protein: 4-methyl-tryptophan (4MTRP, figure 3B), 5-methyl-tryptophan (5MTRP, figure 3C), and α-methyl-tryptophan (AMTRP, figure 3D). There are no major changes to the structure of the complex upon mutation. 4MTRP and 5MTRP are tryptophan analogs that have an additional methyl group at the 4 position and 5 position, respectfully, on the six member indole ring. The additional methyl causes the 4MTRP to rotate about the chi1 angle by ~15 degrees. The rotation buries more of the six membered indol ring. The rotation also forms a hydrogen bond with Glu9 on the peptide. The hydrogen bonds between Trp12 and Gln41 on the protein and between Glu9 on the peptide (figure 3A) are not seen in all of the structures. The 5MTRP mutation functions similarly to the 4MTRP mutation, but the rotation is increased to ~25 degrees from the wild type Trp12 chi1 angle. To keep the hydrogen bond to Glu9 intact, the ψ of position 10 is modified allowing Glu9 to adjust to the 5MTRP position. The chi1 M rotamer of Trp, 4MTRP, and 5MTRP is the least favorable rotamer given the backbone φ and ψ for this position; however, the chi1 angle of 4MTRP and 5MTRP are within 1 standard deviation of the predicted mean. 4MTRP and 5MTRP are predicted to increase the binding affinity of the peptide for the protein by approximately 1 kcal/mol. The third mutation at position 12 is AMTRP. AMTRP is predicted to only increase the bind affinity of the peptide for the protein by approximately 0.5 kcal/mol. AMTRP is a tryptophan analog with the addition of a methyl group at the CA. This methyl group dramatically affects the accessible phi/ψ region that the residue can occupy (see chapter 2, figure 4). The φ and ψ angles of position 12 fall within the acceptable range for α-methylated amino acids. The restriction of dihedral space could pre-order the structure of this region of the peptide which could increase the

binding affinity of the peptide for the protein. This entropic advantage is captured to an extent in the fragment based explicit unfolded state energy discussed in chapter 2 and is the reason for the approximately 2.5 kcal/mol decrease in total energy for this mutation from the wild type sequence.

Position 16

The wild type residue at position 16 is a leucine. This leucine packs into a hydrophobic pocket against Val34, Leu29 and Leu32 on IQN17 and Leu13 and Trp12 of the peptide and additionally makes a backbone/backbone hydrogen bond to Leu13 on the peptide (figure 4A). Rosetta predicts two mutations at this position that will increase the binding affinity of the peptide for the protein: 2-amino-5-phenyl-propanoic acid (2A5PP, figure 4B) and homoleucine (HLU, figure 4B). 2A5PP is a long, hydrophobic residue that buries more hydrophobic surface area than the wild type leucine, and Rosetta predicts it to increase the binding affinity by approximately 1.1 kcal/mol. The mutation does not change the position in the binding pocket, and the CA, CB, and CG of 2A5PP are super imposable with the wild type leucine residue. The CD of 2A5PP superimposes with the CD2 of the Leu, and the phenyl group is positioned over top of Val34 and against the hydrophobic part of the Lys38 side chain. Similarly to 2A5PP, HLU superimposes on the side chain of the Leu to the CD2. The CE1 and CE2 of the HLU pack against Val34 and Leu29. Rosetta predicts that the mutation would increase the binding affinity by approximately 0.9 kcal/mol.

Conclusions

Our collaboration with the Kay lab is ongoing, and they are in the process of making and testing the above designs to see if they are able to increase binding affinity. None of the mutations is predicted to dramatically alter the structure of the peptide. While he proposed mutations are not predicted to dramatically lower the binding affinity, based on results of Weltch et al., small increases in binding affinity of the monomer translate into larger increases as a result of the avidity[2].

Figure 4.1 Proposed model of HIV virus fusion. Gp120 binds to CD4 receptor and a chemokine coreceptor. The binding cause a conformational change in gp120 which induced a conformational change in gp41 forming the prehairpin intermediate. The N-trimer region of gp41 is exposed and able to be target by inhibitors. Gp120 is present but not shown in the prehairpin intermediate. The prehairpin intermediate eventually forms a trimer of hairpins bringing the membrane together. Figure taken from Welch *et al.* [2].

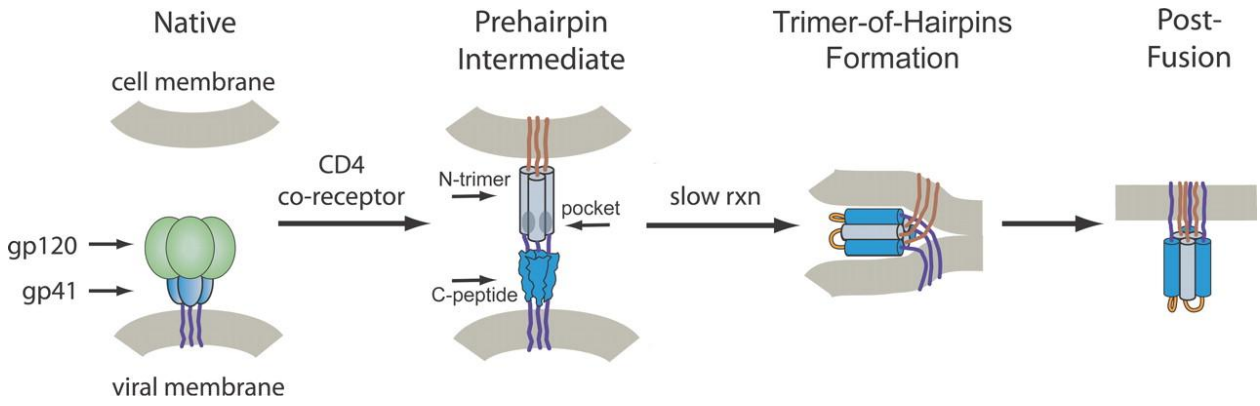


Figure 4.2 Comparison of PIE12 chain H from G63 (cyan) and PIE7 chain H from pdb code 2R5D[2]. Residues in IQN17 superimpose well with a Cα RMSD of 0.37 angstroms. Residues between the cysteines are the same sequence and also superimpose well with a Cα RMSD of 0.18 angstroms. Residues outside of the cysteines have been optimized for binding.

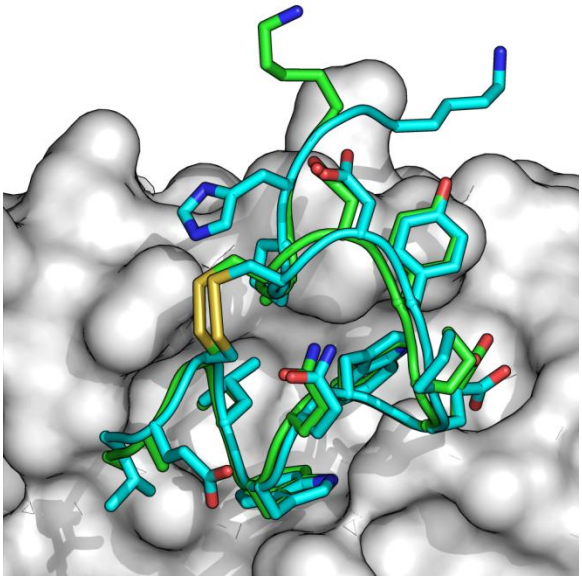


Figure 4.3 Comparison of Rosetta predicted design at peptide position 12. Wild type tryptophan (A), 4-methyl-tryptophan (B), 5-methyl-tryptophan (C), and 2-amino-5-phenyl-propanoic acid (D). Structure is of the G63 structure with the D-peptide chain H in yellow, chain A of IQN17 in green, chain C of IQN17 in magenta, and position 12 in orange. Position 12 and residues within 5 angstroms of position 12 shown as sticks, other residues shown as lines. Hydrogen bonds are shown as dashed black lines.

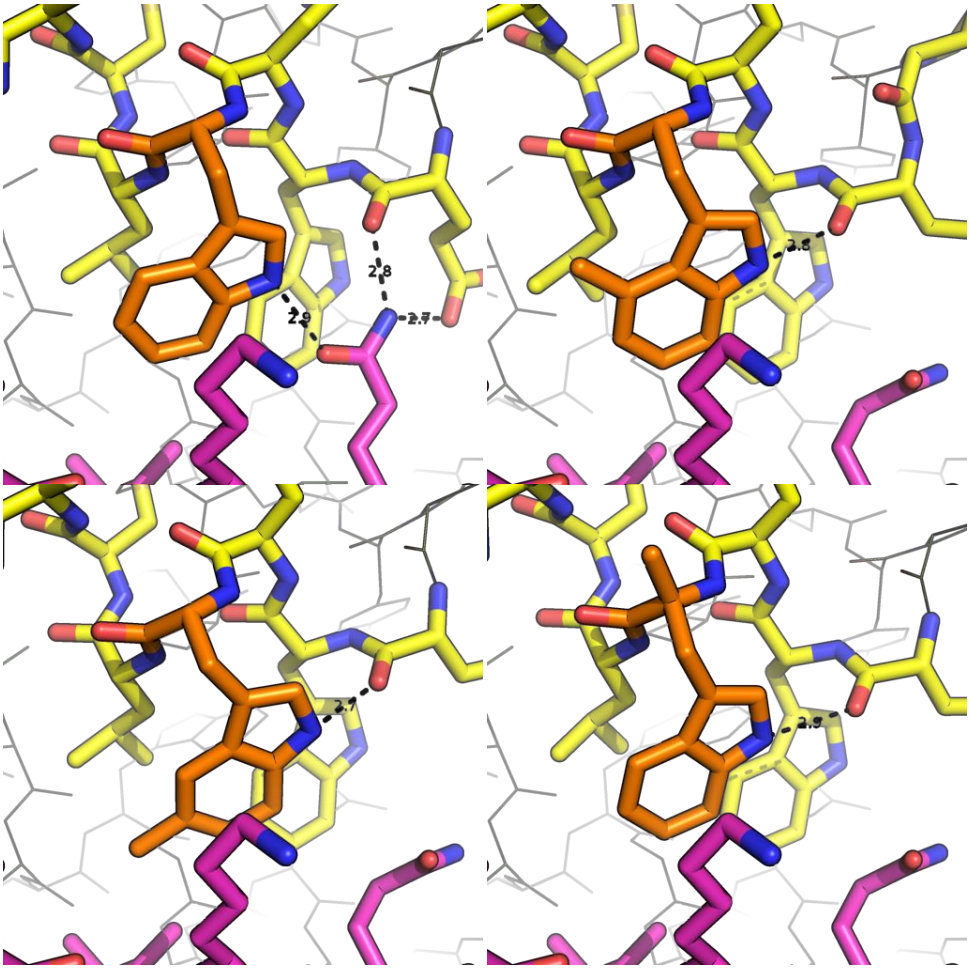
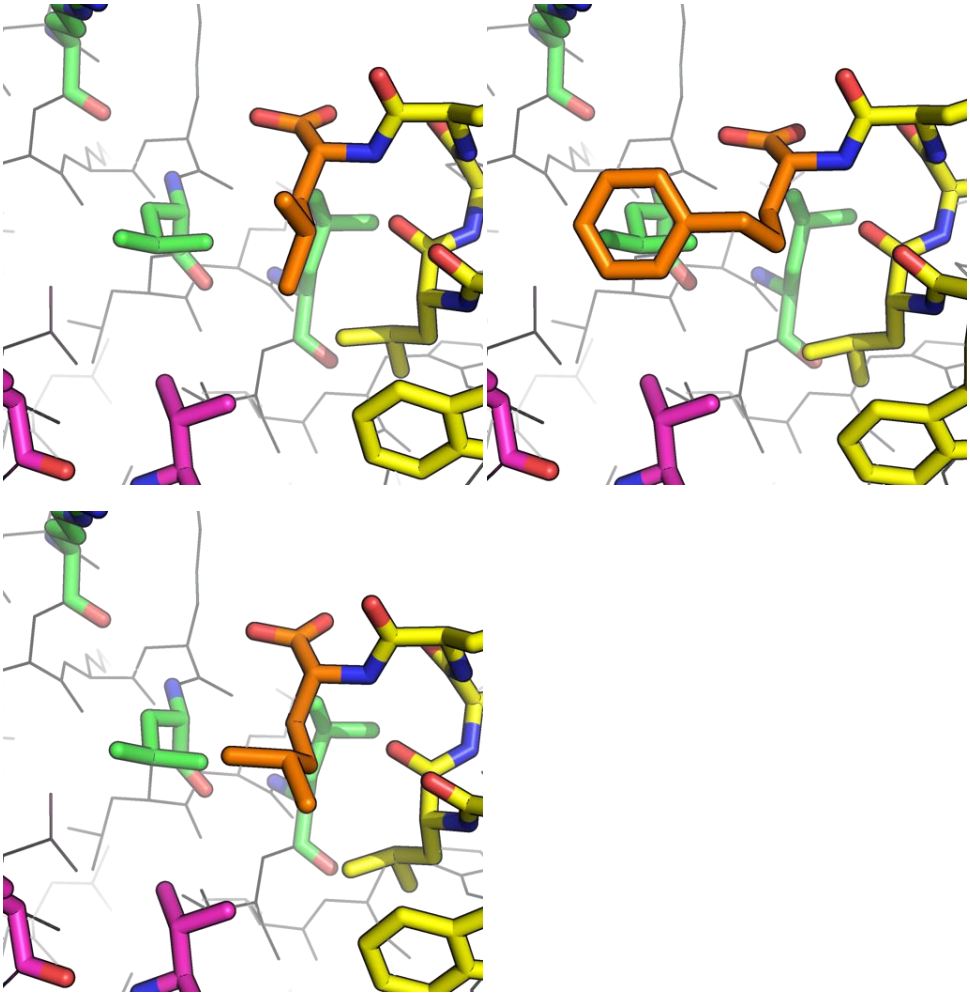


Figure 4.4 Comparison of Rosetta predicted design at peptide position 16. Wild type leucine (A), 2-amino-5-phenyl-propanoic acid (B), and homoleucine (C). Structure is of the G63 structure with the D-peptide chain H in yellow, chain A of IQN17 in green, chain C of IQN17 in magenta, and position 16 in orange. Position 16 and residues within 5 angstroms of position 16 shown as sticks, other residues shown as lines.



Bibliography

1. Leonard, J.T. and K. Roy, *The HIV entry inhibitors revisited.* Curr Med Chem, 2006. **13**(8): p. 911-34.

2. Welch, B.D., et al., *Potent D-peptide inhibitors of HIV-1 entry.* Proc Natl Acad Sci U S A, 2007. **104**(43): p. 16828-33.

3. Chan, D.C., et al., *Core structure of gp41 from the HIV envelope glycoprotein.* Cell, 1997. **89**(2): p. 263-73.

4. Eckert, D.M. and P.S. Kim, *Mechanisms of viral membrane fusion and its inhibition.* Annu Rev Biochem, 2001. **70**: p. 777-810.

5. Schumacher, T.N., et al., *Identification of D-peptide ligands through mirror-image phage display.* Science, 1996. **271**(5257): p. 1854-7.

6. Eckert, D.M., et al., *Inhibiting HIV-1 entry: discovery of D-peptide inhibitors that target the gp41 coiled-coil pocket.* Cell, 1999. **99**(1): p. 103-15.

7. Milton, R.C., S.C. Milton, and S.B. Kent, *Total chemical synthesis of a D-enzyme: the enantiomers of HIV-1 protease show reciprocal chiral substrate specificity [corrected]*. Science, 1992. **256**(5062): p. 1445-8.

8. Fung, H.B. and Y. Guo, *Enfuvirtide: a fusion inhibitor for the treatment of HIV infection.* Clin Ther, 2004. **26**(3): p. 352-78.

9. DeLano, W.L., *The Pymol Molecular Graphics System*. 2002, DeLano Scientific: Palo Alto, CA.