Rotamer-specific statistical potentials for protein structure modeling

by

Jungkap Park

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Mechanical Engineering)
in the University of Michigan
2013

Doctoral Committee:

Professor Kazuhiro Saitou, Chair

Associate Professor Angela Violi

Assistant Professor Matthew Young

Associate Professor Yang Zhang

Acknowledgements

First and foremost, I would like to thank my advisor, Kazuhiro Saitou, for his gentle guidance, great inspiration, and financial support throughout my graduate study. Whenever I got lost in my research or frustrated with unexpected difficulties, he was always patient and understanding. He gave me much freedom to explore diverse research areas and to develop myself independently. I sincerely thank my dissertation committee members, Yang Zhang, Angela Violi and Matthew Young for their time, advice, and patience through the doctoral program.

I wish to express my thanks to Naesung Lyu, who first introduced me my advisor, Kazu and helped me to start smoothly my graduate study in the lab. I have been lucky to great lab members, Karim Hamza, Mohammed Shalaby, Jihun Kim and Jean Chu.

I have had the opportunity to work with great collaborators. Gus Rosania and Ye Li helped me to work on ChemReader project. Although it's not directly involved in my thesis, it helped me a lot to build research insight and skills.

I have been very lucky to meet great people in Michigan. I thanks to Jonghwa Yoon, Janghee Jeong, Seungjea Lee, Seunghwan Lee, Jeongseok Kim, Dongsuk Kum, Jonggirl Ok, Joosup Lim, Donghoon Song, Minjang Jin, Kyung-eun Lee, Youngki Kim, Youjin Choi, Mingoo Seok, Kyungjoon Lee, JaesunSeo, Seoungchul Yang, InhaPaick, Minjoong Kim, Jinyoung Kim,

Sanghun Lee, Soohyung Park, Nayoung Park, Myoungdo Chung and Hyoncheol Kim. I will remember all the great memories made during my time in Michigan.

Last, but definitely not least, I deeply thanks to my family (parents, wife and sisters, sister's husband) for their patience and long-term support in my pursuit of this degree. The biggest thank-you goes to my wife Hyunjoo, who has always encouraged me and showed her faith in my potential. Without their supports and loves, I would not have been able to go all along my winding road through the program.

Table of Contents

List of Tables

List of Figures

Abstract

An accurate potential energy functions is a key component for successful protein structure modeling because they are used for scoring structures and potentially deriving the optimization. Over the years, the accumulation of high-resolution X-ray structures in the Protein Data Bank (PDB) has allowed us to derive a variety of improved, more specific potentials. These potentials are called knowledge-based (or statistical) potentials because they are based on information extracted from sets of known protein structures. Due to their excellent balance between accuracy and computational efficiency, statistical potentials are widely used in folding simulation, protein design and protein structure prediction.

In order to derive more accurate potential energy function, one has to take into the environment-dependence of the specificity of interactions. In other words, since the surrounding circumstances are inhomogeneous and anisotropic on the same scale as the interacting atoms or residues, multibody contributions are important for accurate account of cooperative effects of molecular interactions. On the other hand, protein residues have great flexibility because their single covalent bonds allow rotation of the atoms they join. It is energetically favorable for residues to adopt only a limited number of staggered conformations, known as rotamers. Thus, depending on the rotameric state, the residue conformation and intra-residue interaction vary significantly within protein structures, resulting in different solvent accessibility and different electric polarization effect as well as different steric effect on residue elements. However,

existing energy potentials only reflect in some average sense the energy dependence of the residue flexibility.

The major goal of this thesis is the design and development of statistical potentials that take into account the rotamer-dependence of interactions. We hypothesized that the rotameric state of residues is related to the specificity of interactions within protein structures. We first investigated how amino acid residues in PDB structures show different interaction patterns with the environment depending on their rotameric states. Observed rotamer-specific environmental features were incorporated to a scoring function, ProtGrid for protein designs. Inter-residue or residue-solvent interactions are approximated by using pre-computable grid-based energy terms. Thus it not only takes into account the rotameric state of residues but also computes energies faster than atomic-level energy functions. Our tests demonstrated that the ProtGrid is superior to widely used Rosetta energy function in prediction of the native amino acid types and rotameric states.

Next, we formulated a rotamer-specific atomic statistical potential, named ROTAS that can be used in protein structure prediction. The ROTAS potential extends an existing orientation-dependent atomic potential (GOAP) by including the influence of rotameric states of residues on the specificity of atomic interactions. We tested its performance using various decoy sets for native structure recognition. The results showed that ROTAS performs better than other competing potentials not only in native structure recognition, but also in best model selection and correlation coefficients between energy and model quality. Finally, we applied the ROTAS potential to the problem of side-chain prediction. In our benchmark testing, compared with the existing popular side-chain modeling programs, ROTAS achieved comparable or even better prediction accuracy.

We expect that the effectiveness of our energy functions would provide insightful information for the development of many applications which require accurate side-chain modeling such as homology modeling, protein design, mutation analysis, protein-protein docking and flexible ligand docking.

Chapter 1: Introduction

1.1 Background

Proteins are essential components of organisms and participate in nearly all of the structural, catalytic, sensory, and regulatory functions within living systems. In general, the function of a protein is directly dependent on its three dimensional folded structure, which is implicitly specified by the sequence of amino acids in the protein polymer. While the advent of whole-genome sequencing has greatly increased the number of protein sequences, a vast number of protein structures are not determined yet because experimental methods such as X-ray crystallography or NMR spectroscopy methods require considerable time and cost (Figure 1-1). Thus structural bioinformatics research has devoted much effort into developing computational methods for the accurate and predictive association of protein sequence, structure and function[1].

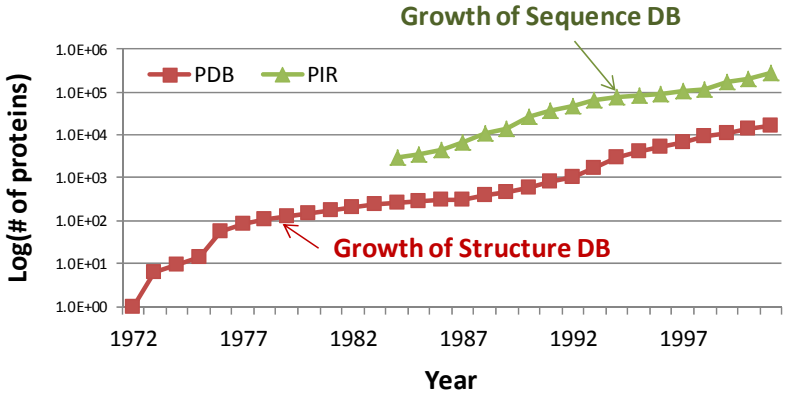

Figure 1-1. Growth of the sequence and 3D structure databases (log-scaled). Protein Data Bank (PDB) and Protein Information Resources (PIR) are a representative structure database and sequence database, respectively.

An accurate potential energy functions is a key component for successful protein structure prediction because they are used for scoring structures and potentially deriving the optimization. It is expected that a good potential energy function has a global energy minimum at the native state with a middle-range funnel biased toward the native state[2]. In principle, such an energy function can be obtained from quantum mechanics[3]. However, this has not been computationally feasible for macromolecules such as proteins or nucleic acids. As an alternative approach, molecular mechanics potentials apply a series of approximations relative to quantum level treatment of all electrons. Typically they model proteins as a collection of particles (atoms or residues) connected by springs (covalent bonds)[4]. These potential energy functions include AMBER[5], CHARMM[6], ECEPP[7], GROMOS[8], OPLS[9], and so forth. They vary in the functional forms used for each potential energy term as well as in the numerical values for the associated parameters. Molecular mechanics potentials are typically used in molecular dynamic simulations, which require the calculation of the forces at every time step. So their energy terms are usually pairwise potentials (i.e., additive two-body forces) and thus ignore the contribution of multibody effect.

Another alternative approach for the design and construction of potential energy functions is to make use of the information embedded in the known protein structures[10–14]. Such energy functions, called statistical potentials or knowledge-based potentials are derived by converting the observed frequencies of residue or atomic interactions in a database of protein structures into the free energies of corresponding interactions. Any aspect of structural features which characterize important interactions in the folded structures can be incorporated into the derivation of statistical potentials (Figure 1-2). The conversion is typically done employing the Boltzmann law[15–17]. Although their physical interpretations are still debated[18–20], due to their

simplicity, accuracy and computational efficiency, various statistical potentials have been developed and used with considerable success in many areas such as fold recognition and threading[12,21–24], protein structure prediction[25,26], refinement[27], protein design[28,29], mutation-induced stability prediction[30] , binding[31,32] and aggregation[33].

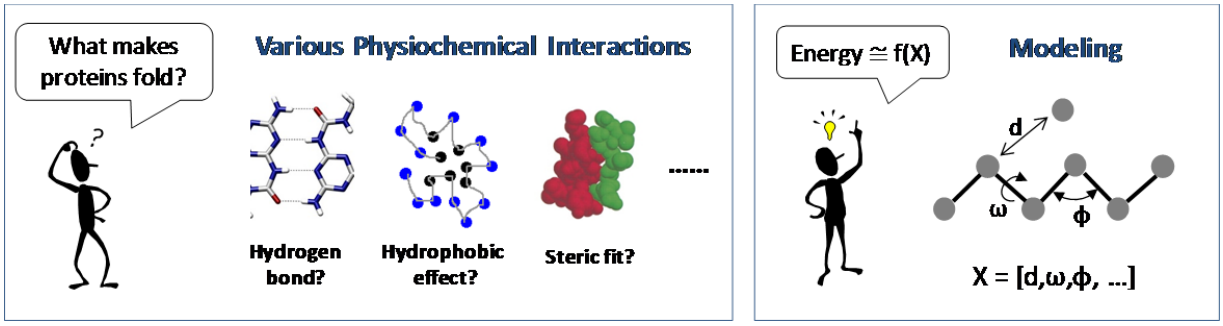

Figure 1-2. There are various factors contributing to the stability and folding of proteins. Understanding of protein energetics allows us to approximate the interaction energy in terms of structural features.

The key idea in the development of statistical potentials is how to decompose the 3-D network of interactions in protein structures. Typical pairwise potentials cannot accurately describe non-bonded interactions in protein structures. As the folded protein structures are tightly packed and surrounded by solvent molecules, the surrounding circumstances of interacting atoms are inhomogeneous and anisotropic. Also, due to the bond connectivity, there are always correlated interactions from nearby bonded atoms. Thus, more detailed and complex structural features involving multibody effects have been incorporated into the formulation of statistical potentials[34–40]. Such multibody potentials are not only able to describe the 3-D interactions more completely but also able to account for cooperative effects of molecular interactions more accurately than typical pairwise potentials.

By virtue of the dramatic increase in the number of structures in the PDB database, various statistical potentials which take into account multibody effects have been developed. For example, sequential segments of various lengths have proved useful for prediction of secondary structure[41–44]. Four body potentials were used to improve cooperativity of main-chain hydrogen-bonds[45,46]. A variety of structural motifs (i.e., residue clusters) has been identified to better characterize tightly packed protein structures[47–50]. Delaunay tessellation technique also has been employed as a means of defining multibody interactions[51,52]. Local environment templates which could account for maximum 17 residues have been introduced to more accurately capture cooperative effects in protein structures[38]. A secondary structure specific implementation of pairwise potentials has demonstrated its superiority to typical residue pairwise potentials[53,54]. In atomic-level statistical potentials, the introduction of orientation dependencies of interactions into typical distance-dependent atomic pairwise potentials has achieved substantial improvements in high-resolution modeling and refinement[37,40,55–60]. With advanced computational processing power, these multibody potentials are widely used in many practical applications for more accurate results.

1.2 Motivation

The basic building blocks of protein structures are amino acid residues, which have distinguishing physiochemical properties depending on the side-chain conformation. Most single covalent bonds in residues allow rotation of the atoms they join, so that the residues have great flexibility. Due to local steric interactions (e.g. overlapped electron orbitals), it is energetically favorable for residues to adopt only a limited number of staggered conformations, known as rotamers[61–64]. Depending on the rotameric state, the residue conformation and intra-residue

interaction vary significantly. In the perspective of quantum mechanics, the electron density distribution around each nucleus can vary depending on the molecular conformation[65–67]. The varied electron density distributions may result in different dipole moments and induce charge reorientations, which are reflected in dispersion forces and electrostatic forces. Thus, residue atoms may experience different solvent accessibility and different electric polarization effect as well as different steric effect depending on the rotameric state.

Existing statistical potentials described above do not model the flexibility of residues explicitly. Residue-level potentials which have only one interaction site per residue simply ignore the flexibility of residue conformation. In case of atomic potentials, the orientation dependent energy terms may be able to account for the anisotropic environment around each atom. However, the orientation dependence of atomic interactions is also based on rigid blocks[68]or rigid atom fragments (three atoms that are consecutively bonded)[55–57]. Thus they cannot reflect the influence of rotameric states on the specificity of atomic interactions no matter how complete a description of the relative orientation and position between interacting atoms may be.

1.3 Thesis Goal

The major goal of this thesis is the design and development of statistical potentials that take into account the rotamer-dependence of interactions. We hypothesized that the rotameric state of residues has a significant relationship with the specificity of non-bonded interactions within protein structures. In particular we are concerned with the following problems:

- Determine how amino acid residues in PDB structures show different interaction patterns with surrounding residues depending on the rotameric state of the residues.
- Development and validation of rotamer-specific statistical potentials

- Application of derived statistical potentials to side-chain modeling

We expect that the incorporation of accurate modeling of residue flexibility in potential energy functions would provide insightful information for the development of many applications which require accurate side-chain modeling such as homology modeling[69], protein design[70], mutation analysis[71], protein-protein docking[72] and flexible ligand docking[73].

1.4 Thesis Outline

In Chapter 2, we first investigated how amino acid residues in PDB structures show different interaction patterns with the environment depending on their rotameric states. We performed a statistical analysis of the local steric environment of residues in high-resolution protein structures to identify the rotamer-specific environmental features. A grid-based representation of rotameric states and their environments allows us to analyze the steric effect of the local environment on the rotameric states of residues. Based on the observation that different rotameric states have distinguishing interaction patterns with surrounding residues, we devised a scoring function, named ProtGrid which takes into account the obtained rotamer-specific environmental features. Inter-residue or residue-solvent interactions are approximated by pre-computable grid-based energy terms. Thus it not only takes into account the rotameric state of residues but also computes energies faster than atomic-level energy functions. We applied ProtGrid to protein design problems which aim to find an optimal combination of amino acid types and their rotameric states for a desired protein backbone structure. The effectiveness of rotamer-specific information for protein design problems was demonstrated.

The promising results from Chapter 2 lead us to formulate a rotamer-dependent atomic statistical potential that can be used in protein structure prediction. In Chapter 3, the derivation and

implementation of the energy function, named ROTAS are described. The ROTAS potential extends orientation-dependent atomic potentials by including the influence of rotameric states of residues on the specificity of atomic interactions. It was clearly found that the rotameric state significantly influence on the specificity of atomic interactions. Furthermore, such rotamer-dependencies are not limited to specific type or certain range of interactions. We tested its performance using various decoy sets for native structure recognition and compared to those of several existing atomic-level statistical potentials which incorporate orientation-dependent energy terms. The results showed that ROTAS performs better than other competing potentials not only in native structure recognition, but also in best model selection and correlation coefficients between energy and model quality.

In Chapter 4, we applied the ROTAS potential to side-chain prediction problem which requires an accurate modeling of side-chain conformations. In order to maximize the prediction ability, we introduced weights for energy terms and optimized them against a separate training set. Its prediction accuracy was evaluated with another separate testing set. In our benchmark testing with the existing popular side-chain modeling programs such as SCWRL4[74], OPUS-Rota[75], and OSCAR-star[76], ROTAS achieved comparable or even better prediction accuracy.

Finally, we summarize and discuss the results of the thesis and outline some important future directions in Chapter 5.

Chapter 2: Rotamer-specific Environmental Features for Computational Protein Design

2.1 Abstract

Computational protein designs demand an efficient energy function which takes into account the rotameric state of residues. Existing residue-level energy functions do not account for the rotameric state, and atomic-level ones take much time for large and complex design problems. Here, we present a new energy function, ProtGrid which not only takes into account the rotameric state of residues but also approximates interactions by using pre-computable grid-based energy terms. Based on a statistical analysis relating the rotameric states and the local steric environment of residues, rotamer-specific environmental features are obtained and incorporated to the ProtGrid scoring function. It was found that the rotamer-specific environment features can characterize not only steric features but also hydrophobic features of residues, which implies that the mutability and mutational directions of residues are strongly influenced by the local steric environment. Our tests demonstrate the ProtGrid is superior to widely used Rosetta energy function in prediction of the native amino acid types and rotameric states. We expect that ProtGrid together with the rotamer-specific environmental features can contribute to low-level protein design stage by restricting search algorithms to the proper region of search space with much less computational complexity.

2.2 Introduction

Computational protein design methods require an efficient search strategy with an accurate energy function[77,78]. In general, the objective of protein design problems is to find an optimal combination of amino acid types and their rotameric states for a desired protein backbone structure with minimum free energy (Figure 2-1)[79,80]. As amino acid side chains prefer to adopt only a limited number of conformations, known as rotamers[61–63], computational protein design methods usually rely on a sample space that depends on a rotamer library, which is a set of statistically significant discrete side-chain conformations[81,82]. The free energy of amino acid sequence for a target structure is estimated by an energy function based on models of protein energetics[28,83]. Since each residue can be changed into a different rotamer of the same type of amino acid or mutated into a different type during the energy minimization, it is required to explore a large combinatorial search space for assigning rotamers simultaneously, which results in an NP hard optimization problem[84].

Figure 2-1. Computational protein design aims to find an optimal combination of amino acid types and their rotameric states for a desired protein backbone structure with minimum free energy

Common limitations in current protein design methods include the size of search space, (i.e. the size of sequence/rotamer space) and the choice of potential energy function. Due to the discrete representation of rotameric states, large-scale protein designs suffer from the problem of combinatorial explosion. Although various stochastic and heuristic search algorithms are employed to attempt larger and more complex designs[85–96], they can never guarantee that an optimal solution has been found[77]. Several continuous modeling approaches based on the mean-field description of the energy landscape have been developed for computational protein design[97–101]. However, as the mean-field approach alters the original energy landscape, its solutions are erroneous[102]. In addition, it is highly likely to fail to converge to a feasible solution when the number of allowable rotamers is increased at each residue.

For the potential energy function, current methods usually combine common energy terms in molecular mechanics with a few empirical terms accounting for solvation effect and entropic contribution[28,29]. Due to computational efficiency, the energy terms are generally pairwise additive. However, pairwise energy functions have a limitation to capture important cooperative interaction patterns determining amino acid types and their rotameric states. For example, the solvation energy which is one of the critical factors in protein design is not inherently pairwise decomposable. Furthermore, atomic-level pairwise energy functions whose computational complexity is $O({n_{atom}}^2)$ would not be desirable to address practical protein design problems such as large-scale design and/or multi-state design problems[78,103,104].

For the design of large and complex protein structures, multiple stages with different levels of approximation can be useful in reducing the search space gradually[105]. Such multi-level optimization allows different levels of detail of the model and levels of sophistication of the

objective function for the multiple stages. Indeed, many drug design applications adopt similar strategies such that the computational complexity of scoring functions and/or the resolution of molecular model gradually increase as the simulation proceeds[105,106]. It is also noticed that restricting search algorithms to the proper region of rotamer space could improve side-chain prediction accuracy dramatically[107]. Although various coarse-grained energy functions for protein structure prediction and folding have been developed[22,33,108,109], they are not applicable to protein design problems because the rotameric state of residues are ignored in those energy functions. Mostly each residue has only one interaction site. Thus, in order to address protein design problems in advance of atomic-level modeling, it is necessary to develop an energy function which not only takes into account the rotameric state of residues but also computes energies faster than atomic-level energy functions

The energy functions for reducing the search space should effectively capture important interaction patterns contributing to the stability and folding of proteins. The most significant factor would be steric complementarity. Since the close-packing limits the number of possible combinations of the side-chains[110], energy functions that model only steric packing interactions are often used for protein core designs[89,111–113] and side-chain predictions[114–116]. In fact, initial energetic screening for reducing the rotameric search space often relies on steric complementarity[117,118]. Hydrophobic interactions are also believed to be important determinants of side-chain conformational preference in protein structures[119,120]. Non-polar atoms of residues prefer to be buried in protein cores and polar atoms prefer residing on the surface, exposed to water. Due to the expensive computation of an explicit solvent model, protein design energy functions generally employ a continuum-solvent model[121]. In order to design stable, well packed

soluble proteins, the steric energy term and the solvation energy term must be well balanced in protein design energy functions.

Since side-chain packing is a short-range phenomenon which is mainly driven by local steric and hydrophobic interactions, the local environment of residues can be related to the preferences of their amino acid types and rotameric states. Many statistical studies have revealed that amino acids have conformational preferences in terms of secondary structure[122–126] or local backbone conformation[127–131]. Thus the local environment is exploited in various applications for side-chain modeling. For example, in the context of comparative modeling, local structural information from homologous proteins has been used in side-chain modelling[132–135]. These methods take advantage of the relationship between the local environment and the conservation of side-chain dihedral angles. The successful applications of environment-specific substitution matrixes also suggest that the local structural environment within proteins significantly influences the preference of amino acid types[136–138]. Above all, nearly all side-chain prediction and protein design methods are based on using backbone-dependent rotamer libraries.

In this chapter, we present an energy function, ProtGrid, which ranks different amino acid types and their rotameric states based on the local environment information. First of all, a statistical analysis was performed on the local environment of residues in a large sample of high resolution structures. We extracted rotamer-specific environmental features for each rotameric state, which represent preferred relative position and orientation of surrounding residues. It is found that they can characterize not only the relevant steric interaction patterns explicitly but also the hydrophobic features of amino acid side-chains implicitly. In addition to common energy terms for protein designs, the obtained rotamer-specific environmental features are incorporated into

ProtGrid. We approximated inter-residue or residue-solvent interactions by grid-based energy terms such that the energy calculation can be much less computationally demanding than conventional pairwise-atomic energy functions. In ProtGrid, typical 20 amino acid types are expanded to 154 different residue types such that the specificity of inter-residue or residue-solvent interactions depends on the rotameric state.

Using 50 proteins, we tested ProtGrid and compared it to widely used Rosetta energy function in prediction of the native amino acid types and rotameric states. We also analyzed the contribution of different energy terms (steric complementarity, electrostatic interactions, solvation, rotamer-specific environmental features) to the overall prediction accuracy for each amino acid type. We report that the local environment of residues involves dominant factors determining amino acid types and their rotameric states of residues in folded protein, and can be efficiently and effectively modeled by grid-based approaches.

2.3 Materials and Methods

In this section, we describe how the relationship between rotameric states and the local steric environment of residues was analyzed, and then how a side-chain prediction test was performed with a devised scoring function, ProtGrid, which exploits the local environment information. The description consists of six parts: preparation of PDB structures, grid-based description model for the local steric environment of residues, analysis of the relationship between the rotameric state and the local steric environment of residues, extraction of rotamer-specific environmental features, derivation of scoring function, ProtGrid for protein designs, and test of side-chain prediction. The details are as follows.

Table 2-1.The composition by amino acid type of the 950,079 residue sample

Amino Acid	Number of residues		Number of rotamers	
Ala	86605	9.1%	86605	9.6%
Cys	12844	1.4%	12711	1.4%
Asp	52991	5.6%	50138	5.6%
Glu	49729	5.2%	46497	5.1%
Phe	42520	4.5%	41455	4.6%
Gly	74462	7.8%	74462	8.2%
His	22634	2.4%	21203	2.3%
Ile	58901	6.2%	57886	6.4%
Lys	40453	4.3%	34093	3.8%
Leu	95387	10.0%	90429	10.0%
Met	15638	1.6%	13367	1.5%
Asn	39928	4.2%	37299	4.1%
Pro	46627	4.9%	41954	4.6%
Gln	31766	3.3%	26579	2.9%
Arg	40841	4.3%	33288	3.7%
Ser	58017	6.1%	57423	6.4%
Thr	55493	5.8%	55093	6.1%
Val	73552	7.7%	73120	8.1%
Trp	15338	1.6%	14087	1.6%
Tyr	36353	3.8%	35366	3.9%
Total	950079	100.0%	903055	100.0%

2.3.1 Preparation of PDB structures

We obtained a set of protein X-ray structures longer than 40 residues with a maximum R-factor of 0.25 and a resolution better than 2 Å from the protein sequence culling server, PISCES[139]. Also, protein chains were filtered out with a 25% sequence identity cutoff in order to have a set of non-homologous protein structures. A total 9087 protein structures were selected and downloaded from the Protein Data Bank (PDB)[140]. The program REDUCE[141] was used to add hydrogen atoms and optimize side-chain orientations in all proteins. Residues with multiple side-chain conformations were modified such that only the side-chain conformations with atoms

having the highest occupancy and/or lowest temperature factors were used. It is possible to consider all the multiple conformations with weights calculated from the occupancy rate. However, since the portion of such residues is very small in the training set, they would not significantly affect on the result. To reduce the influence of the uncertainty in the PDB coordinates on our analysis, residues missing heavy atoms or having any backbone atom with temperature factor B > 40 were excluded. The side-chain conformation of residues was classified into rotamers which are defined in the Penultimate Rotamer Library[142]. Finally, among 950,079 collected residues, 903,055 residues were assigned to one of rotameric states (Table 2-1).

The resulting list of proteins was separated into training and testing sets of 4,902 and 50 proteins respectively. The training set of proteins was used for extracting rotamer-specific environmental features and for optimizing parameters in ProtGrid devised for the side-chain prediction test. The testing set of 50 proteins was used only for the side-chain prediction test. As a result, the training and testing sets contain 892,636 and 10,419 residues in rotameric states, respectively.

2.3.2 Grid-based description of the local steric environment of residues

Each residue and its spatially neighbor residues are aligned within a reference coordinate frame and then the local steric environment of the residue is described by grid-based steric field in a box. Spatially neighbor residues of a residue within protein structures are defined using a distance cutoff criterion. That is, any residue pair whose C_α atom distance is less than 15Å is considered as spatially neighbor residues. It is assumed that such spatially neighbor residues determine the local steric environment of the considered residue. For each residue, their backbone atoms including N, C_α, C and O were superposed to fixed reference positions and then

its spatially neighbor residues are aligned in the reference frame accordingly such that the local steric descriptor is invariant to translation and/or orientation of the residue (Figure 2-2).

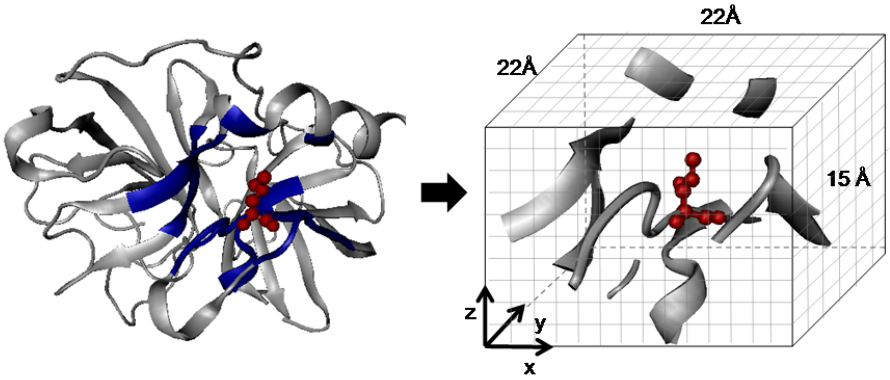

Figure 2-2. A schematic representation for computing a steric descriptor for the local structural environment. Neighboring residues are aligned into the reference coordinate system. The steric fields are computed at 1Å spaced grids in the 22x22x15Å box. In the reference coordinate, the Z-axis and X-axis are parallel to the vector from CB to CA atoms and the vector from CA to N atoms, respectively.

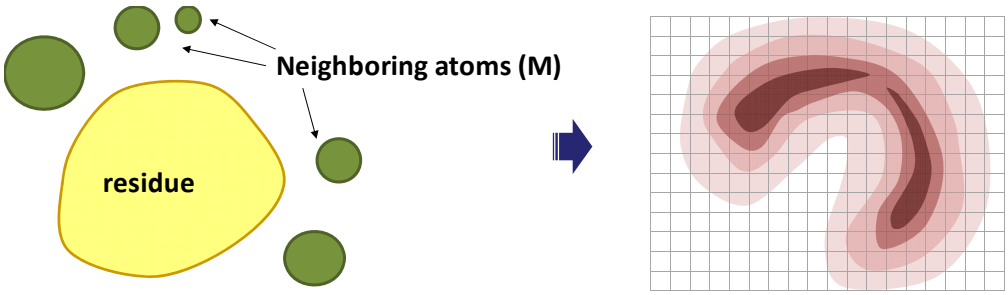

Figure 2-3. Grid-based description of the local steric environment of residues. Steric field potentials by neighboring atoms are computed at grid points.

Given the aligned set of neighbor atoms, steric fields are computed at the vertices of a predefined regularly spaced grid (Figure 2-3). The size of a bounding box enclosing grid points is

22x22x15Å box, such that side-chain atoms of the considered residue can fit into the box with a margin of 5.0 Å at any rotameric state. Grid points are generated with a 1.0 Å spacing (with a total of 7,260 grid points) and probe atoms are placed at the grid points. The steric interaction between a probe atom and a neighbor atom is calculated by the Lennard-Jones (LJ) 6-12 potential. Thus the steric field potential v_i at grid point i is computed by summing the interactions between the probe atom and each one in the neighbor atoms, M:

$$v_i(M) = \begin{cases} 0 & E_{vdw}(M,i) \leq 0 \\ 50 & E_{vdw}(M,i) \geq 50 \\ E_{vdw}(M,i) & \text{otherwise} \end{cases}$$
$$E_{vdW}(M,i) = \sum_{j \in M} \sqrt{A_{probe} \times A_j} \left\{ \left(\frac{R_{probe} + R_j}{d_{ij}} \right)^{12} - 2 \left(\frac{R_{probe} + R_j}{d_{ij}} \right)^{6} \right\} \quad (1)$$

where A_j and A_{probe} indicate the Van der Waals constant, R_j and R_{probe} are the Van der Waals radius, and d_{ij} is the distance between i and j.

The van der Waals constants and raddi are those used in the Rosetta Energy function[95] and the probe atom is the hydrogen atom in this study. The LJ potential is steep close to the van der Waals surfaces of the atom. In order to avoid infinity at the atomic centers, cut-off values are applied to grid points where the steric field potential is extremely high. In this study, a cut-off value of 50 kcal/mol was used. Also, negative steric field potentials were equivalent to zero steric field potential in this study. Once the steric field is generated by a set of neighbor atoms M, the local steric environment descriptor is given by

$$D_M = (v_i(M) | \, i \in G) \quad (2)$$

where G is a set of grid points considered in the analysis, and i indicates the grid index. In this study, we considered grid points that are within 12 Å from C_α atom position, which results in total 5692 grid points in G.

2.3.3 Analysis of the relationship between the rotameric state and the local steric environment of residues

In order to see how rotameric states are related to the local steric environment of residues, we attempted to compare rotamers in terms of their conformations, physiochemical characteristics, and local steric environment. We first defined the steric descriptors for rotamer conformation and rotamer environment as follows.

- Rotamer conformation descriptor

  Analogous to the local steric environment of residues, the rotamer conformation can be described by grid-based steric field in a box. Let M^a denotes a set of side-chain atoms of rotamer a. Then the rotamer conformation descriptor, D_{rot}^a is simply $D_{M^a} = (v_i(M^a) | \, i \in G)$. For the representative rotamer conformations of each rotameric state, we used common-atom χ values reported in the rotamer library[142].

- Rotamer steric environment descriptor

  Given a set of local steric environment descriptors of residues in the rotameric state a, we computed a histogram of steric potential at every grid point. Here, the steric potential observed ranges from 0 to 50 and is divided into 50 bins of width 1. Let the total number of observed residues in the rotameric state a be N_{res}^a and the frequency of the steric field potential at grid i, v_i falling in bin k be $f_i^a(k)$. Then the observed probability of the steric energy is

$$p_i^a(k) = f_i^a(k) \Big/ \sum_k f_i^a(k), \qquad k \in B = \{1 \dots 50\} \tag{3}$$

where B is a set of indices of the bins. The probability distributions of the steric energy can be regarded as features of the rotamer environment. Accordingly, the rotamer environment steric descriptor can be defined as

$$D_{env}^a = (p_i^a(k) |\, i \in G, k \in B) \tag{4}$$

For the similarity measure, we simply calculated the cosine of the angle between two descriptor vectors. That is, the similarity coefficient, S, between two descriptors, $D^a$and $D^b$was calculated as

$$S(D^a, D^b) = \frac{D^a \cdot D^b}{||D^a|| \cdot ||D^b||} \tag{5}$$

Thus, for a pair of rotamer a and b, rotamer conformation similarity, $S_{rot}(a, b)$ and rotamer steric environment similarity, $S_{env}(a, b)$ correspond to $S(D_{rot}^a, D_{rot}^b)$ and $S(D_{env}^a, D_{env}^b)$, respectively.

While geometric features such as the rotamer conformation or the local steric environment of rotamers can be described based on steric field potentials in grid space, it is not straightforward to compare various physiochemical characteristics of different amino acid types in a quantitative way. So we measured amino acid similarity based on correlations in the amino acid substitution matrix. In fact, it is believed that correlation coefficients derived from BLOSUM substitution matrix elements can reflect the physiochemical similarity between different amino acid types [143]. In this work we calculated the correlation coefficients from the BLOSUM64 matrix and used them as a measure for amino acid similarity.

2.3.4 Rotamer-specific environmental feature

As the same rotamer can be observed in many different environments, many grid points have large variations in the steric field value. In order to identify the rotamer-specific environmental feature, it would be necessary to focus on the grid points which are highly likely to have either high or low steric field potentials. That is, using certain threshold values, we identified two sets of grid points being relevant to each rotamer, high steric potential grids (G_H) and low steric potential grids (G_L) as follows.

$$G_H = \{i \mid P(v_i \geq t) > 0.8, i \in G\} \text{ and} G_L = \{i \mid P(v_i < t) > 0.8, i \in G\} \quad (6)$$

where $P(v_i > t)$ is the probability that the steric field potential at grid *i* is higher than or equal to a cutoff threshold, t and $P(v_i \leq t)$ is the probability of v_i being lower than t. Here the cutoff threshold, t was assigned to 2 kcal/mol. It can be seen that the grid points in G_H represent the positions which are frequently occupied by neighbor atoms and the grid points in G_L indicate the positions which are frequently voided in the protein structures or exposed to water.

There are always errors in the estimated probability due to a limited number of structures in the sample. Some of the grid points accounted above might not be relevant to the rotamer-specific environmental feature. In general, collected grid points in G_H or G_L form a few clusters in which grids are connected to each other. As a non-hydrogen atom covers at least 4 grid points by its van der Waals radius, small clusters would not be physically meaningful. In fact, those grids in small clusters could be accidently included by the error in the estimated probability. Thus we excluded grids from G_H and G_L if they form clusters whose size is less than a certain threshold. Since atoms are closely bonded or packed together in protein structures, the threshold was set to 10.

2.3.5 Scoring function

The effective energy of the rotamer is computed as a linear combination of the following nine energy terms:

$$ProtGrid = W_{atr}E_{atr} + W_{rep}E_{rep} + W_{elec}E_{elec} + W_{solv}E_{solv} + W_{rot}E_{rot} + W_{aa}E_{aa} + W_{env}^{H}E_{env}^{H} + W_{env}^{L}E_{env}^{L} - E_{ref} \quad (7)$$

where E_{atr} and E_{rep} are terms for the steric complementarity, E_{elec} is the electrostatic interaction term, E_{solv} is the solvation energy term, E_{rot} is the rotamer intrinsic energy term, E_{aa} is the amino acid preference term, and E_{ref} is an empirical reference energy assigned to each amino acid for approximating the energy of the unfolded state. E_{env}^{H} and E_{env}^{L} are terms taking the rotamer-specific environmental features into account.

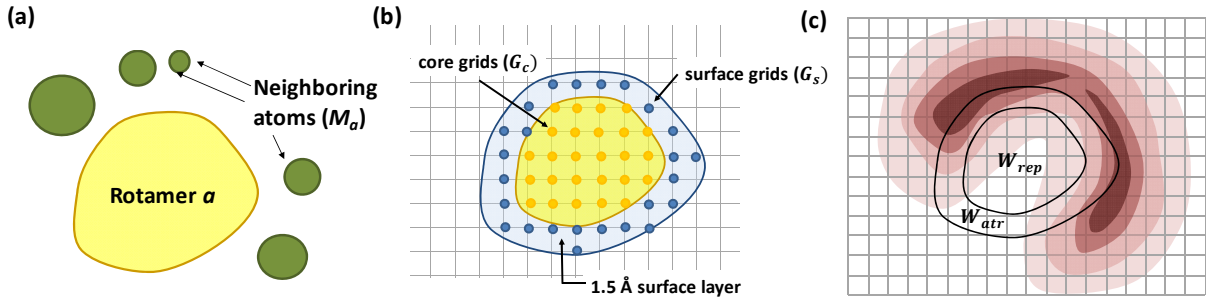

Figure 2-4. Grid-based modeling of steric complementarity. (a) Rotamer *a* is surrounded by a set of neighboring atoms (M_a), (b) Core grids and surface grids of the rotamer a are identified, (c) Steric field potentials are computed from neighboring atoms (M_a). Repulsive energy and attractive energy are the sum of steric field potentials at core grids and surface grids, respectively.

2.3.5.1 Steric complementarity

The steric complementarity has two terms for repulsive interactions (E_{rep}) and attractive interactions (E_{atr}), which can be estimated by the degree of steric clash and contact between considering residue's atoms and its neighbor atoms, respectively. Figure 2-4 illustrates how the steric complementarity between rotamer a and its neighboring atoms, M_a is computed by a grid-based approach. We first identified surface grids (G_S) and core grids (G_C) for each rotamer according to the steric field values of the rotamer conformation descriptor. Any grid point with a steric field value larger than the cutoff threshold, t (= 2 kcal/mol) is considered to be the core of rotamers. The surface grids are those located in a 1.5 Å surface layer of the core grids. Thus, given surface grids G_S^a and core grids G_C^a for rotamer a, the steric complementarity between rotamer a and the environment $D_M = (v_i)$ is calculated by the following two terms.

$$E_{atr} = -\sum_{i \in G_S^a} v_i \tag{8}$$

$$E_{rep} = \sum_{i \in G_C^a} v_i \tag{9}$$

2.3.5.2 Electrostatic interaction

The electrostatic interaction energy between the rotamer and the local environment is calculated based on a simple Coulombic model as follows:

$$E_{elec} = \sum_{i \in G} q_i \cdot e_i \cdot H(-v_i) \tag{10}$$

where$e_i(M)$ is the electric field potential of the local environment, $q_i(M^a)$ is the charge strength of rotamer a at grid point i. It is assumed that if v_j doesn't have positive steric potential

values,then the grid point j is not occupied by any neighbor atom. Thus, the unit step function, $H(-v_i)$ allows summing the electrostatic interactions only outside of neighbor atoms. $e_i(M)$ and $q_i(M^a)$ are calculated by following equations:

$$e_i(M) = \sum_{k \in M} \frac{q_k}{{d_{ik}}^2} \tag{11}$$

$$q_i(M^a) = \sum_{j \in M^a} q_j \cdot \exp(-0.5 \cdot {d_{ij}}^2) \tag{12}$$

where q_k and q_j are partial charges, and d_{ik} is the distance between i and k. The partial charges of the atoms of the rotamer are discretized in grid points by a bell-shaped Gaussian-type function with an attenuation factor, 0.5. This provides a "local smearing" effect which place more weight on interactions close to the atom, with a smooth transition to more distal points. Here, the atomic charges are taken from CHARMM27 parameter set.

2.3.5.3 Solvation energy

Solvation energy is evaluated using the Lazaridis and Karplus (LK) solvation model. The solvation free energy density of the rotamer is discretized in grid points, assuming that the density value is constant within each grid cell. Thus the solvation energy is simply evaluated by summing the density values over grid points which are occupied by neighbor atoms:

$$E_{solv} = -\sum_{i \in G} s_i \cdot \Delta_i \cdot H(v_i) \tag{13}$$

whereΔ_i is the volume of grid cell i, and s_i is the solvation free energy density of the rotamer at grid point i, which is calculated by

$$s_i(M^a) = \sum_{j \in M^a} \frac{\alpha_j}{4\pi d_{ij}^2} \cdot \exp\left\{-\left(\frac{d_{ij} - R_j}{\lambda_j}\right)^2\right\} \tag{14}$$

where R_j, λ_j and α_j are the van der Waals radius, correlation length, proportionality coefficient given by the LK model, respectively, and d_{ij} is the distance between *i* and j.

2.3.5.4 Rotamer intrinsic energy and amino acid preferences

As in the Rosetta energy function, the intrinsic energy of a rotamer was statistically derived from PDB structures in the training set. The probability of a particular rotamer and amino acid for a given backbone conformation (ϕ, ψ) was estimated by observing each of the 154 rotamers within $10° \times 10°$ bins in ϕ, ψ dihedral angle space. Thus, the energy was calculated by taking the negative log of the probabilities:

$$E_{rot}(a) = -\ln(P(rot(a)|\, \phi, \psi)) \tag{15}$$

Also, the energy term representing amino acid preferences was also calculated in the same way.

2.3.5.5 Rotamer-specific environmental feature

In the scoring function, a rotamer is more favored when relevant environment features of the rotamer are matched to the local steric environment being considered better than those of other rotamers. The degree of matching is calculated by examining steric field values at grid points in $G_H$or G_L:

$$E_{env}^H = -\frac{1}{|G_H^a|} \sum_{i \in G_H^a} v_i \tag{16}$$

$$E_{env}^L = \frac{1}{|G_L^a|} \sum_{i \in G_L^a} v_i \tag{17}$$

The derivation of these energy terms are different from that of typical statistical potentials employing the Boltzmann law to convert the observed frequencies of interactions into the free energies of corresponding interactions. This is because there is no sufficient PDB data for estimation of the observed probability of the steric energy, $p_i^a(k)$. Instead, we exploited the obtained high steric potential grids (G_H) and low steric potential grids (G_L) for each rotamer to devise energy terms.

2.3.5.6 Amino acid reference energies and weights of energy terms

The 20 amino acid reference energies and the weights for the energy terms were determined by maximizing the sum of the following objective function over 30,000 residues which were randomly selected from the training set:

$$log\left(\frac{\exp(-E_{eff}(r_j))}{\sum_{i=1}^{N_{rot}}\exp(-E_{eff}(r_i))}\right) \quad (18)$$

where*j* is the index of native rotamer for considering residues, N_{rot} is the number of rotamers in the rotamer library, and $E_{eff}(r_i)$ is effective energy of the rotamer r_i. The rotamer library used in this study contains all 152 rotameric states defined in the penultimate rotamer library plus Aly and Gly amino acids as two independent rotameric states. In this optimization procedure, only one residue was changed at a time and all other residues were kept in their native conformation. For the optimization algorithms, we employed a simulated annealing (SA)[144] followed by a sequential quadratic programming (SQP). In SA, the initial temperature is set to 100, and the temperature is updated by an annealing schedule factor of 0.95 with the re-annealing interval, 100. Starting from random reference energies and weights, at maximum, 100,000 iterations were

tried, and the termination tolerance on the objective function value for both methods was set to 10^{-6}.

2.3.6 Test of side-chain prediction

We tested the devised scoring function with 10,419 residues in the testing set. For each residue in proteins, and with all neighbor side-chains fixed in their native conformation, all amino acids and rotamers in the rotamer library (total 154 rotameric states) were tested. The prediction is regarded as success when the native rotamer is within the top five predictions. This testing method has been used by other studies to validate developed scoring functions[107,114,116,145].

2.4 Results and Discussion

2.4.1 Grid-based description of the local steric environment of residues

The grid-based steric field for capturing spatial features of the residue environments enables us to analyze and compare the local steric environment of residues effectively. Usually, the residue environment has been described in terms of local structure motifs depending on the complexity ranging from large super-secondary structures to very short segments (e.g. five or more continuous residues). However, such fragment-based approaches have difficulties when comparing different residue environments in a quantitative manner because of the heterogeneity of amino acid types and different length of motifs. In addition, continuous residues on the same secondary structure segment could result in very different local environment descriptors if side-chain atoms are facing surrounding residues in different directions (see Figure 2-5 for examples). As grid-based representation of steric fields decreases the spatial resolution of molecular conformation, it could lessen the effect of changes in the steric descriptor associated with minor variations in molecular conformations. Therefore, as long as steric fields of neighbor residues

have common spatial features, the local environment descriptors would be highly similar even though residues are on different secondary structure classes or the chemical composition of neighbor residues is different.

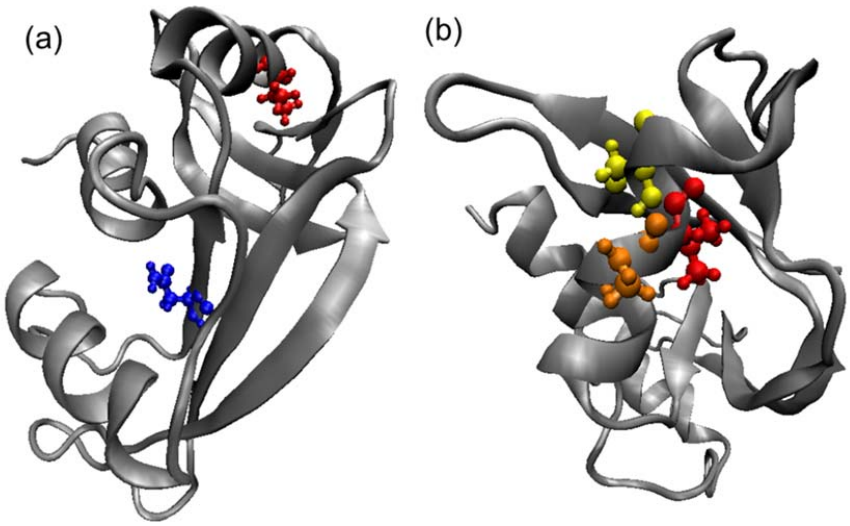

Figure 2-5. (a) Two residues, Asn44 (blue) and Val57 (red) are depicted as ball-and-stick models on the cartoon structure of ribonuclease (PDB id 7rsa). The environment similarity coefficient between Asn44 and Val57 is very high (0.88) although they are on different secondary structure classes. (b) In contrast, the environment similarity coefficients between Val57 and the very next residues on each side (yellow and orange) are only 0.69 and 0.72, even though they are continuously positioned on the same secondary structure. This is because side-chain atoms of those residues are toward different directions, which result in different steric descriptors

We computed the rotamer conformation descriptor for every rotameric state and compared them to each other. Since the rotamer conformation descriptor omits information on the chemical composition of rotamers, many rotamer pairs have high similarity coefficients. In fact, most of amino acids can be grouped by their molecular structures: {Asp, Asn, Leu}, {Ser, Cys}, {Glu, Glu}, {Lys, Arg}, {Phe, Tyr} and {Val, Ile, Thr}. Even rotamers of His, Trp or Pro which have

unique side-chain geometries could show high rotamer similarity coefficients with other rotamer conformations as shown in Figure 2-6. In the perspective of steric complementarity, such highly similar rotamers might be substitutable with each other maintaining close-packing in protein structures.

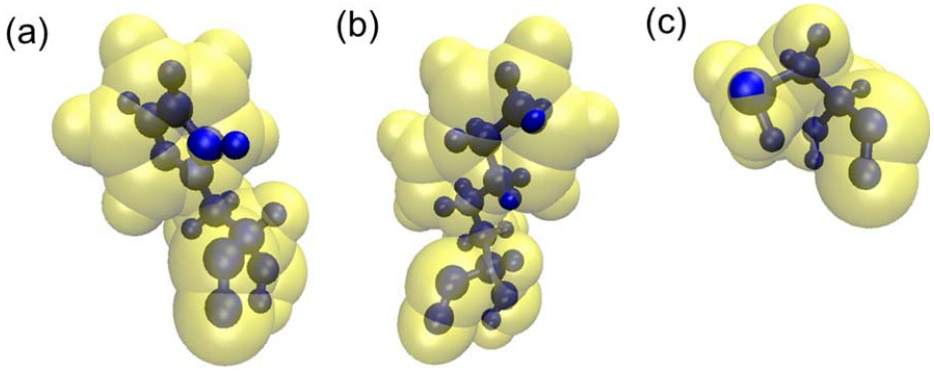

Figure 2-6. Comparisons of different rotamer conformations. Rotamer similarity scores of (a) 0.9943 for Hist t-160° (blue) vsPhe t80° (yellow), (b) 0.9826 for Lys tptt (blue) vsTrp t-105° (yellow) and (c) 0.9481 for Cys p (blue) vs Pro $C_\gamma$endo (yellow)

2.4.2 Relationship between the rotameric state and the local steric environment of residues

Next, we looked at how conformation or amino acid type of rotamers is related to the local steric environment of residues in protein structures. For this, the rotamer conformation similarity, rotamer steric environment similarity and amino acid similarity were computed for every possible rotamer pair as described in Materials and Method section. Figure 2-7 shows their relationships. In order to include these three features in the scatter plot, each data point (i.e. rotamer pair) was colored according to its amino acid similarity scores. Cleary, rotamer pairs with high amino acid similarity scores (e.g. > 0.5) tend to be in similar local environments. However, even though rotamer pairs show high rotamer similarity scores (e.g. > 0.9), they are

widely distributed in the rotamer steric environment descriptor space. Also, rotamer pairs with high rotamer steric environment similarity scores (e.g. > 0.9) are also scattered widely over the rotamer similarity axis. Correlation coefficients in Table 2-2 also confirm the same observations in a quantitative way.

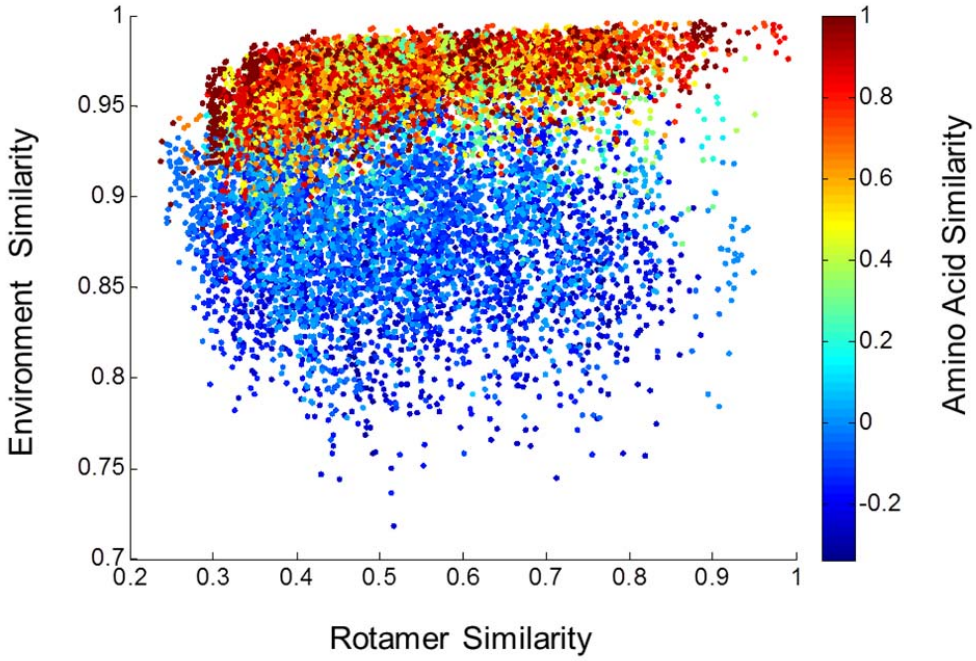

Figure 2-7.Relationship between rotamer similarity, environment similarity and amino acid similarity of all rotamer pairs.

Table 2-2 Relationship between rotamer similarity, environment similarity and amino acid similarity

Rotamer Similarity vs. Environment Similarity	0.109
Rotamer Similarity vs. Amino Acid Similarity	-0.0021
Amino Acid Similarity vs. Environment Similarity	0.7518

The high correlation between amino acid similarity scores and rotamer steric environment similarity scores reflects the underlying physical principles of amino acid substitutions well. In general, hydrophobic residues and hydrophilic residues tend to be conserved as a class and small residues are not replaced by large ones in the interior of proteins. For example, the similarity score between Asp and Glu is 0.8165 while the similarity score between Asp and Leu is -0.2422, which means that substitutions between Asp and Glu are more frequently occurred than those between Asp and Leu in the process of protein evolution. Interestingly, Asp-Glu rotamer pairs show higher rotamer steric environment similarity scores ($0.94 \leq S \leq 0.99$) than any Asp-Leu rotamer pair ($0.75 \leq S \leq 0.8$). In addition, we found that the correlation coefficient between rotamer steric environment similarity scores and amino acid similarity scores increases as BLOSUM cluster percentage increases (Figure 2-8). In other words, the substitution rates observed in more closely related protein sequences are more strongly correlated with rotamer environment similarity scores than those observed in divergent proteins. It can be seen that small changes that occur over short time scales are more likely constrained by the local steric environment than large changes that occur over long evolutionary time scales. These observations support the idea that the local steric environment is a significant factor affecting on the mutability and mutational direction of residues in protein structures.

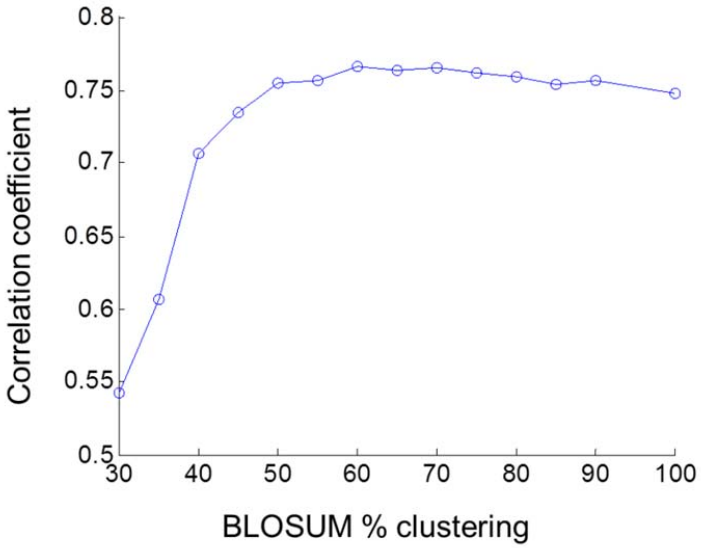

Figure 2-8.Correlation between rotamer environment similarity scores and amino acid similarity scores with different BLOSUM % clustering.

2.4.3 Rotamer-specific environmental features

The rotamer-specific environmental features for each rotamer consist of two sets of grid points: high occupancy grids (G_H) and low occupancy grids (G_L). It can be seen that the grid points in G_H represent the positions which are frequently occupied by neighbor atoms and the grid points in $G_L$indicate the positions which are frequently voided in the protein structures or exposed to water. In the perspective of the steric complementarity, it was expected that rotamer core-grids (inside rotamer atoms) and rotamer surface-grids (outside surface of rotamer atoms) correspond to low occupancy grids and high occupancy grids, respectively. However, as shown in Figure 2-9, while most of core-grids are involved in G_L, many of surface-grids do not belong in G_H. Rotamer core-grids usually correspond to the blue segments at the center of panes. Different rotamers show different spatial patterns of $G_H$or G_L. It is noted that both Glu (Figure 2-9. (a)) and Asp (Figure 2-9. (b)) rotamers have a larger cluster of low occupancy grids outside core-grids while

the Leu (Figure 2-9. (c)) rotamer has only one cluster of low occupancy grids for the rotamer core-grids, which are surrounded by high occupancy grids. The large cluster of low steric potential grids observed in the rotamer-specific environmental features of Glu and Asp rotamers would indicate a space for water molecules which favorably form hydrogen bonds to the carboxylic acid of Asp and Glu. In contrast, the Leu rotamer tends to be buried inside proteins due to its hydrophobicity.

Furthermore, we could observe that the location of low occupancy grids is strongly related to the orientation of hydrophilic atoms in side-chains. Figure 2-10illustrates the rotamer-specific environmental features for Lys's rotamers. While those four rotamers have the same χ_1 and χ_2 angles with the same chemical composition, it is clearly found that the location of low occupancy grids tends to strongly depend on χ_3 or χ_4 angles which influence on the orientation of hydrophilic atoms in Lys. Together, these findings imply that that the rotamer-specific environment features can characterize not only steric features but also hydrophobic features of amino acid side-chains.

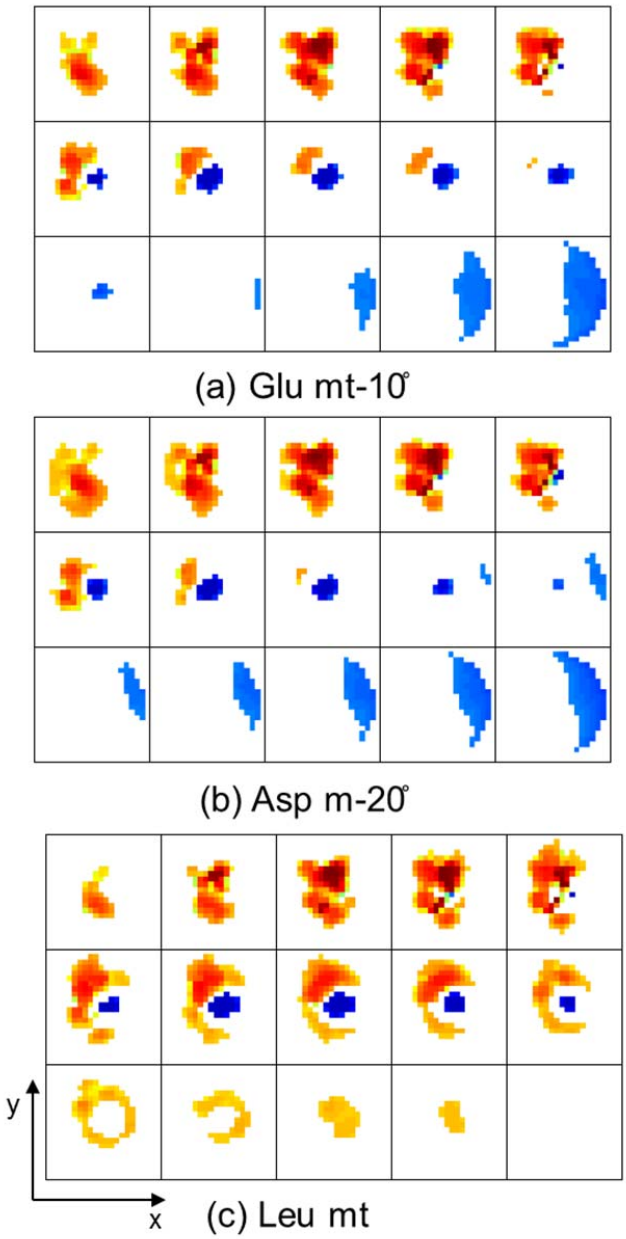

Figure 2-9. Illustrative examples of rotamer-specific steric environment features: (a) Glu mt-1, (b) Asp m-2 and (c) Leumt. Each pane represent surface grids parallel to the X-Y plane (see Figure 2-3 for the reference axis); Panes are arranged in the row-wise direction such that the bottom surface of the 3D box is present at top-left pane. The colors of grids are mapped to average steric field potentials, G_H(red)-G_L(blue).

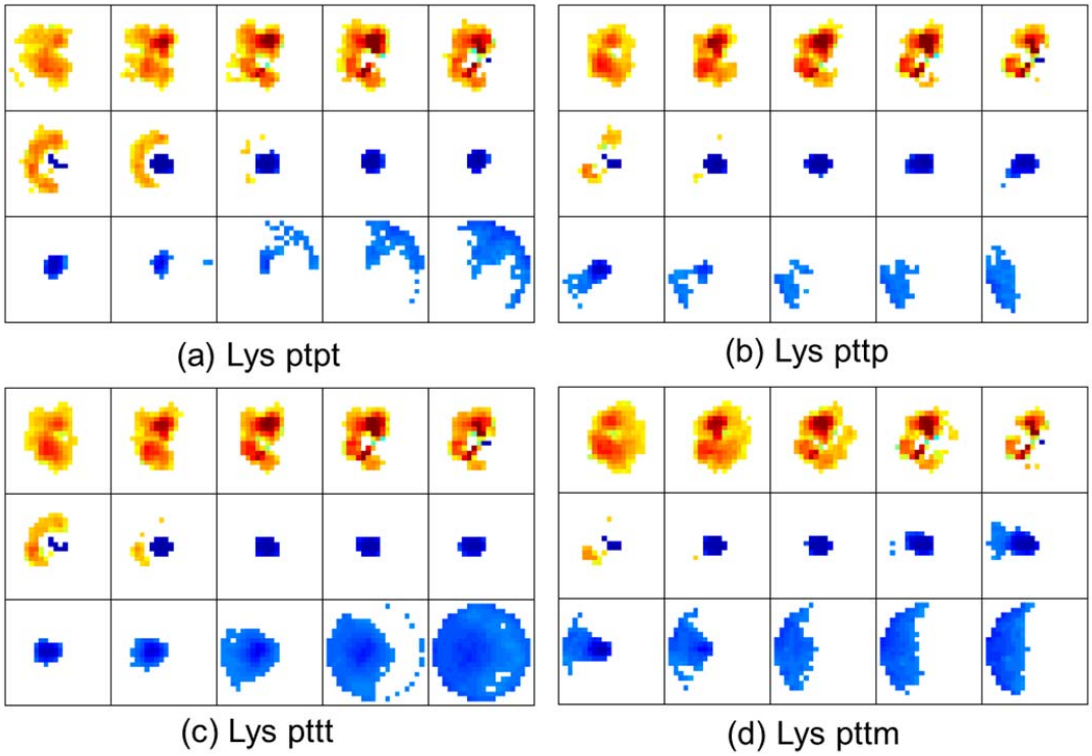

Figure 2-10. Rotamer-specific features for four Lys's rotamers: (a) ptpt, (b) pttp, (c) pttt, and (d) pttm. Panes are arranged in the same way as in Figure 2-9.

2.4.4 Derived scoring function

In order to demonstrate the usefulness of the rotamer-specific environmental features for side-chain modeling, we devised a scoring function, ProtGrid, for ranking different amino acid types and their rotameric states of residues. ProtGrid is a linear combination of the following energy terms: steric complementarity, electrostatic interaction, solvation energy, rotamer intrinsic energy, amino acid preference, reference energy for each amino acid, and terms that are specifically designed to account for the rotamer-specific environmental features. In principle, the last terms measure the similarity between the local steric environment of residues and the rotamer-specific environmental features such that a rotamer is more favored when its rotamer-

specific environmental features are matched to the local steric environment being considered better than those of other rotamers.

ProtGrid approximates inter-residue and solvent-residue interactions using pre-computable grid-based energy terms such that the computational cost for the energy calculation could be reduced significantly. In protein designs, it is desirable to balance between thermodynamic accuracy and computational tractability. In fact, on the grid, the local environment of residues is represented by the steric and electric field potentials, and the rotamer is also characterized by the core/surface grids, electric charge and solvent free energy density. The grid-based terms measure how well the physiochemical properties of the rotamer fit into the steric and electric field of the local environment of residues. The grid-based approximation allowed us to remove the computational cost, $O({n_{atom}}^2)$ from the energy calculation, which is required by typical pairwise-atomic energy functions. Furthermore, since every term in ProtGrid is linear function of grid values of the local environment and the rotamer, there is no problem to re-formulate ProtGrid as a pairwise-residue function such that existing search algorithms for protein designs can be readily incorporated.

2.4.5 Test of side-chain prediction

ProtGrid was tested on a set of 50 known protein structures, and compared to Rosetta energy function. Prior to the test, the weights of energy terms and amino acid reference energies of both scoring functions were determined using the same optimization procedure and training structures for a fair comparison (see Materials and Methods). We assumed that the native amino acid type and rotameric state is the global free energy minimum of the scoring functions given the native local environment. The testing structures were used neither for optimizing parameters in the

scoring functions nor extracting the rotamer-specific environmental features. In order to see how each energy term contributes to performance, the accuracy of seven versions of ProtGrid and three version of Rosetta were tested, each of which consisted of a different combination of energy terms. The weights and reference energies were kept the same as for the full versions.

Table 2-3 shows the summarized results of the side-chain prediction test. The success rate is defined by the ratio of testing residues of which the native amino acid types and rotameric states are within the top five predictions. First of all, the success rates of the full versions of ProtGrid and Rosetta, (i.e. ProtGrid-7 and Rosetta-3) were 74.3% and 72.6%, respectively. The objective function values also clearly confirmed that ProtGrid-7 outperformed Rosetta-3. Any version of ProtGrid could even score lower objective function value than any version of Rosetta. This may be because ProtGrid describe the steric complementarity of side-chain conformations more effectively than the Lennard Jones potential implemented in Rosetta energy function.

Table 2-3. Side-chain prediction test results for 50 testing structures. Different combinations of energy terms are tested. Included energy terms are marked by circles.

	Steric	Rot. Specific env.		Solv.	Elec.	H-Bond	Stat. + Ref.	objective function value	Success rate
		High	Low						
ProtGrid-1	O						O	24434.66	70.8%
ProtGrid-2	O	O					O	24025.23	72.0%
ProtGrid-3	O		O				O	24155.45	72.1%
ProtGrid-4	O	O	O				O	23764.69	73.0%
ProtGrid-5	O	O	O		O		O	23003.85	75.1%
ProtGrid-6	O			O	O		O	23723.66	72.3%
ProtGrid-7	O	O	O	O	O		O	23134.28	74.3%
Rosetta-1	O						O	30490.04	66.5%
Rosetta-2	O			O			O	26202.17	71.7%
Rosetta-3	O			O		O	O	25876.01	72.6%

When including terms concerning the rotamer-specific environmental features, the success rate increased and found to be higher than any version of Rosetta. Compared to ProtGrid-1, the success rates of ProtGrid-2 and ProtGrid-3 were improved by 1.2% and 1.3%, respectively. This indicates that both high and low occupancy grids are effective to improve the side-chain prediction accuracy. Also, while the solvation energy and electrostatic interaction terms in ProtGrid-6 increased the success rate by 1.5% than ProtGrid-1, ProtGrid-5 including terms on the rotamer-specific environmental features and electrostatic interactions increased the success rate by 4.3% and thus yielded the highest success rate and the lowest object function value in the test. Cleary, these results illustrate that the rotamer-specific environmental features could assist in finding the native amino acid types and rotameric states in protein structures.

Table 2-4. Comparison of the success rates of the three scoring functions, ProtGrid-5, ProtGrid-7 and Rosetta-3 for residue burial or secondary structure

	residue burial		secondary structure					
	core	surface	α-helix	β-strand	turn	bridge	3_{10} helix	coil
ProtGrid-5	86.6%	57.5%	73.0%	79.0%	74.0%	73.8%	75.2%	74.2%
ProtGrid-7	85.8%	56.7%	72.0%	78.0%	73.2%	75.4%	74.2%	73.9%
Rosetta-3	85.4%	53.1%	69.7%	78.6%	71.0%	72.1%	68.9%	71.1%

We also compared the success rates of ProtGrid-5, ProtGrid-7 and Rosetta-3 for contact and surface residues, and for six different secondary structures (Table 2-4). In order to classify core or surface residues in protein structures, we assumed that residues are on protein surface if the percentage of the residue exposed surface area is larger than 17%. And the secondary structure types of the residues were assigned by STRIDE software[146]. In all three functions, core residues show better success rates than surface residues. Regardless of the residue burial or the secondary structure, ProtGrid scoring functions show better performance than the full version of Rosetta

energy function. On the other hand, it is noted again that ProtGrid-5 which omits the solvation energy term based on the Lazridis-Karplus (LK) model shows the best accuracy for all categories except bridge secondary structure. This may demonstrate that the rotamer-specific environment feature models the hydrophobic effect more effectively than the LK model in the side-chain prediction.

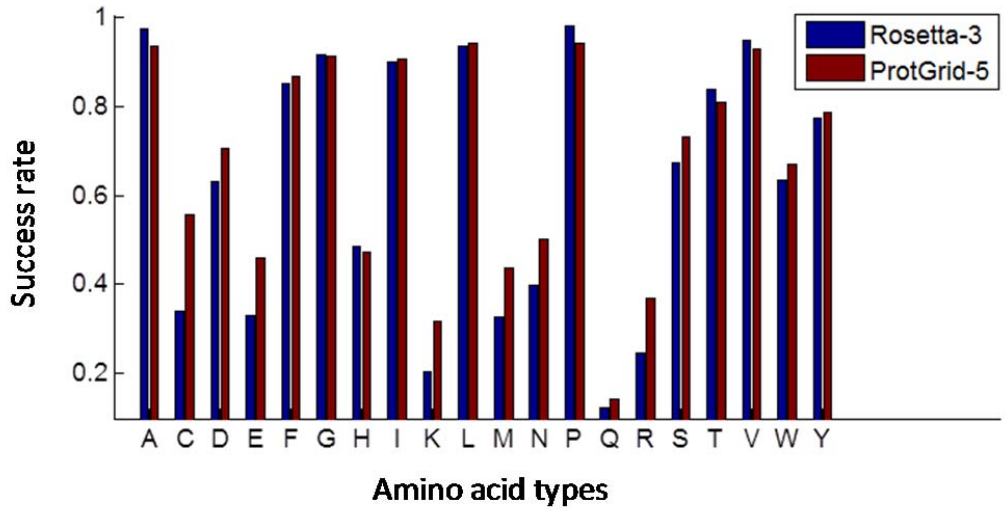

Figure 2-11. Comparison of success rates for 20 amino acids types.

Figure 2-11 shows a comparison of success rates over different amino acid types for ProtGrid-5 and Rosetta-3, each of which seems to be the most accurate version of the scoring functions. Overall, ProtGrid-5 shows comparable or higher success rates than Rosetta-3. In particular, the success rates of ProtGrid-5 for charged polar amino acids such as Asp, Glu, Lys or Arg, are significantly higher than those of Rosetta-3. Table 2-5 shows how the success rates for each amino acid vary depending on different versions of ProtGrid. ProtGrid-1 was used as a reference such that the success rates of other versions were subtracted by those of ProtGrid-1. In ProtGrid-2, the success rates of most amino acids increased except a few nonpolar amino acids such as

Phe, Ile, Leu or Val. We found that their high occupancy grids are not distinguishing. On the other hand, ProtGrid-3 improved the success rate of Cys significantly, which has unique low occupancy grids. Above all, ProtGrid-5 shows a dramatic improvement in the success rates for polar or charged amino acids such as Cys, Asp, Asn, Glu, Gln, Ser, Lys or Arg. In Figure 2-11, we can also found that the success rates of ProtGrid-5 for these polar residues are much higher than those of Rosetta-3.

Table 2-5. Comparison of the success rates for each amino acid for different versions of ProtGrid

Amino Acid	ProtGrid-1	ProtGrid-2	ProtGrid-3	ProtGrid-4	ProtGrid-5	ProtGrid-6	ProtGrid-7
ALA	92.8%	-0.1%	1.1%	0.9%	0.6%	1.2%	2.2%
CYS	36.9%	1.0%	18.4%	16.5%	18.4%	4.9%	17.5%
ASP	63.8%	5.6%	-5.1%	0.0%	6.7%	1.1%	-0.2%
GLU	24.6%	4.6%	2.0%	6.4%	21.3%	7.4%	11.8%
PHE	87.9%	-1.3%	1.3%	-0.6%	-1.3%	1.9%	0.4%
GLY	86.9%	0.5%	4.1%	4.2%	4.1%	1.4%	4.9%
HIS	42.1%	2.3%	4.6%	5.4%	5.0%	-4.2%	2.7%
ILE	91.4%	-1.2%	0.9%	-0.8%	-0.8%	1.5%	0.8%
LYS	18.7%	2.3%	-6.0%	-3.6%	13.0%	11.7%	9.1%
LEU	94.8%	-1.2%	0.3%	-1.3%	-0.7%	1.0%	0.4%
MET	36.1%	1.4%	7.5%	7.5%	7.5%	3.4%	8.8%
ASN	38.8%	5.0%	2.4%	7.8%	11.3%	-4.3%	4.5%
PRO	88.2%	5.7%	0.6%	5.7%	5.7%	3.8%	8.4%
GLN	9.5%	0.6%	1.8%	3.3%	4.7%	-0.9%	0.9%
ARG	24.5%	2.2%	1.1%	3.5%	12.4%	3.0%	7.0%
SER	67.4%	2.6%	4.1%	6.9%	5.8%	-1.3%	6.7%
THR	78.8%	1.2%	1.8%	2.9%	2.2%	-0.6%	2.6%
VAL	94.2%	-1.6%	-0.2%	-1.1%	-1.6%	1.4%	-0.2%
TRP	66.2%	0.7%	1.4%	0.7%	0.7%	0.7%	1.4%
TYR	78.9%	0.2%	1.6%	0.2%	-0.2%	1.4%	0.2%

ProtGrid-1 is used as a reference such that the success rates of other versions were subtracted by those of ProtGrid-1.
The color of cells varies according to changes in the success rates, increase(red)-decrease(blue)

In case a pair of rotamers has high scores for both rotamer conformation similarity and rotamer steric environment similarity, ProtGrid would have a difficulty to decide which rotamer is more probable under a given environment. Let rotamer a and b be "mutually substitutable", when $S(D_{env}^{a}, D_{env}^{b}) > 0.7$ and $S(D_{rot}^{a}, D_{rot}^{b}) > 0.9$. We counted the number of mutually substitutable rotamers for each rotamer. Figure 2-12shows how the success rates for amino acid types vary with the number of mutually substitutable rotamers. Overall, the success rate is inversely proportional to the number of mutually substitutable rotamers. In fact, it was found that, for 67.7% of testing residues, at least one mutually substitutable rotamer was ranked in the top five scoring rotamers. This percentage increased up to 90.5% when we considered only residues whose native rotamers had more than 10 mutually substitutable rotamers. This would imply that, only been given a local steric environment, it is difficult to distinguish native rotamers and their mutually substitutable rotamers. Interestingly, amino acid types with low success rates (e.g. Gln, Lys or Arg) are mostly able to form non-steric interactions such as hydrogen bonds or salt bridges. Therefore, we can expect that the success rate can be increased by a subsequent stage of refinement, which can consider non-steric interactions for top-ranked rotamers together with their mutually substitutable rotamers.

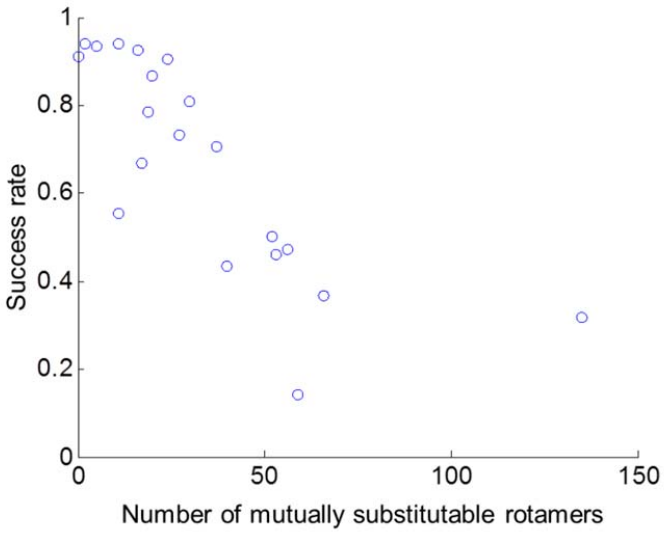

Figure 2-12. Relationship between the success rate and the number of similar rotamer pairs for amino acid types.

2.5 Conclusion

We have investigated an effective way to exploit local steric environment information for rapidly ranking possible amino acid types and their rotameric states of protein residues. The local environment of residues in a large sample of high-resolution structures was modeled by a grid-based steric field. We found that rotamer pairs with high amino acid similarity scores are appeared in similar local environments, which implies that the mutability and mutational direction of residues are strongly influenced by the local steric environment. Moreover, as the rotamer-specific environmental features account for the relative position and orientation of distinct steric effects, hydrophobic features of amino acid side-chains also could be included implicitly.

We have devised a scoring function, ProtGrid, which takes commonly used interaction energies for protein designs but also the obtained rotamer-specific environmental features in to account.

Since ProtGrid approximates commonly used inter-residue or residue-solvent interactions with pre-computable grid-based energy terms, its computational complexity is dramatically reduced than typical pairwise-atomic energy functions. We tested the derived scoring function and compared its accuracy with Rosetta energy function. For every testing residue, all possible amino acid types and rotamers were scored whereas all neighbor residues were fixed in their native conformations. ProtGrid could rank the native amino acids and rotameric states within the top five predictions for 75.1% of testing residues at best, which performs better than Rosetta energy function. In particular, the test results demonstrate that the rotamer-specific environmental features could effectively improve the prediction accuracy. We expect that ProtGrid together with the rotamer-specific environmental features can contribute to low-level protein design stage by restricting search algorithms to the proper region of search space with much less computational complexity, and thus will pave the way for solving large-scale and/or multi-state protein design problems.

Chapter 3: A Rotamer-dependent, Atomic Statistical Potential for Assessment and Prediction of Protein Structures

3.1 Abstracts

Statistical potentials are widely used as essential tools in protein structure modeling and quality assessment. They are derived from experimentally determined protein structures aiming to extract relevant structural features that characterize the tightly folded structures. In this study, we have developed a new statistical potential energy function, named "rotamer-dependent atomic statistical potential" (ROTAS) that extends an orientation-dependent atomic potential (GOAP) by including the influence of local structural environments on the specificity of atomic interactions. Here, the local structural environment is defined in terms of the rotameric state of residues taking into account that protein residues prefer to adopt only a few rotamer conformations. In ROTAS, the interaction between two atoms is specified by not only the distance and five angle parameters but also two state parameters which concern the rotameric state of the residues to which the interacting atoms belong. It has clearly found that the rotameric state is significantly related to the specificity of atomic interactions. Compared to GOAP, such rotamer-dependencies are not limited to specific type or certain range of interactions. The performance of ROTAS was tested using 13 sets of decoys and compared to those of several existing atomic-level statistical potentials which incorporate orientation-dependent energy terms. The results show that ROTAS performs better than other competing potentials not only in native structure recognition, but also in best model selection and correlation coefficients between

energy and model quality. In particular, the relative improvement of ROTAS over GOAP implies that the local structural environment can be incorporated for fine-tuning of atomic-level statistical potentials. Furthermore, the rotameric state of residues may be used to describe the local environment of residue atoms. The ROTAS potential is freely available in https://sites.google.com/a/umich.edu/rotas/.

3.2 Introduction

Statistical potentials are energy functions derived from a database of experimentally determined protein structures[10–12]. In general, they employ the Boltzmann law to convert the observed frequencies of interactions in protein structures into the free energies of corresponding interactions[15,16]. Although their physical interpretations are still debated[18–20], due to their simplicity, accuracy and computational efficiency, various statistical potentials have been developed and frequently used with considerable success in many areas such as fold recognition and threading[21–24], protein structure prediction[25,26], refinement[27], protein design[28,29], mutation-induced stability prediction[30] , binding[31,32] and aggregation[33].

Despite the progress in methodology and theory, and the dramatic increase in the number of structures in the PDB database[140], the accuracy of statistical potentials still requires improvement for accurate protein structure prediction[147–149]. The simplest statistical potential would be residue-level pairwise contact (or distance-dependent) potential. However, it has turn out that residue-level pairwise potentials are inadequate for ab initio folding or difficult fold recognition[150–152]. Thus, more detailed representation (*e.g.*, semi-residue or atom-level) and/or more complex structural features involving multibody interactions or orientation-dependences have been incorporated into the formulation of statistical potentials[34–40]. In fact, any aspect of

structural features (*e.g.*, contact area, torsion angle, solvent accessibility, secondary structural states) that can characterize the tightly packed interior of native structures can be employed to enhance the ability of statistical potentials[53,147,153–157].

Over the years, the introduction of orientation dependencies of interactions into typical distance-dependent atomic pairwise potentials has achieved substantial improvements in high-resolution modeling and refinement. For example, dDFIRE combine the DFIRE distance-dependent potential and the orientation-dependent interactions between polar and non-polar atoms and between polar atoms by treating each polar atom as a dipole[56]. Similarly, it has shown that incorporating side-chain orientation-dependent energy term into a distance-dependent potential using an ideal random-walk chain as reference state (RW) could improve the overall performance over the case with only the RW pairwise potential[57]. OPUS-PSP is another orientation-dependent atomic statistical potential, which describes orientation dependence on side-chain packing interactions based on 19 rigid-body blocks decomposed from 20 amino acid residues[68]. A recently introduced potential, GOAP that accounts for orientation-dependences of all heavy atom types in proteins, was successfully validated using various decoy sets[55]. The inclusion of higher order multibody interaction is a possible means of improving the specificity of statistical potentials[59].

In order to derive more accurate potential energy function, one has to take into account the influence of various structural environments on the specificity of atomic interactions as well as the relative position and orientation. The surrounding circumstances are inhomogeneous and anisotropic on the same scale as the interacting atoms. In the perspective of quantum mechanics, the electron density distribution around each nucleus can significantly vary depending on the

local environment of atoms[65–67]. The change in electron density distributions may result in varied dipole moments and induce charge reorientations, which are reflected in dispersion forces and electrostatic forces. On the other hand, most single covalent bonds in residues allow rotation of the atoms they join, so that the residues have great flexibility. Due to the local steric interactions (*e.g.,* overlapped electron orbitals), residues prefer to adopt only a limited number of staggered conformations, known as rotamers[61–63]. Depending on the rotameric state, the residue conformation and intra-residue interaction vary significantly, resulting in different solvent accessibility and different electric polarization effect as well as different steric effect on residue atoms.

In this study, we define the local structural environment in the context of rotameric state of residues and derive a new potential energy function, named "rotamer-dependent atomic statistical potential" (ROTAS). Describing the local environment around atoms in terms of the rotameric state enables us to classify the environmental state into a few statistically significant discrete states[142,158]. The interaction between two atoms is specified by not only the distance and five angle parameters but also two state parameters which concern the rotameric state of the residues to which the interacting atoms belong. In a comparison between ROTAS and GOAP, it has clearly found that the rotameric state is significantly related to the specificity of atomic interactions. Furthermore, such rotamer-dependencies are not limited to specific type or certain range of interactions.

We tested ROTAS on various sets of decoys generated from different methods and compared its performance to those of several existing all-atom statistical potentials which incorporate orientation-dependent energy terms. The results show that ROTAS performs better than other

competing potentials not only in the native structure recognition, but also in the best model selection and the correlation coefficients between energy and model quality. In other words, ROTAS successfully extends orientation-dependent atomic statistical potentials by including the influence of local structural environments on the specificity of atomic interactions. The rotameric state of residues turned out to be an effective structural feature for fine-tuning of atomic-level statistical potentials.

3.3 Materials and Methods

3.3.1 Derivation of ROTAS

In the ROTAS potential, the interaction between two atoms is described by the spatial distance, relative orientation and local structural environments as illustrated in Figure 3-1(a). Basically, it extends the description of inter-atomic interaction in GOAP by including the local environment (i.e. rotameric states). The detailed description for how the local structural environment is defined in terms of the rotameric state is explained in the next section. Here we focus on the formulation of the ROTAS potential.

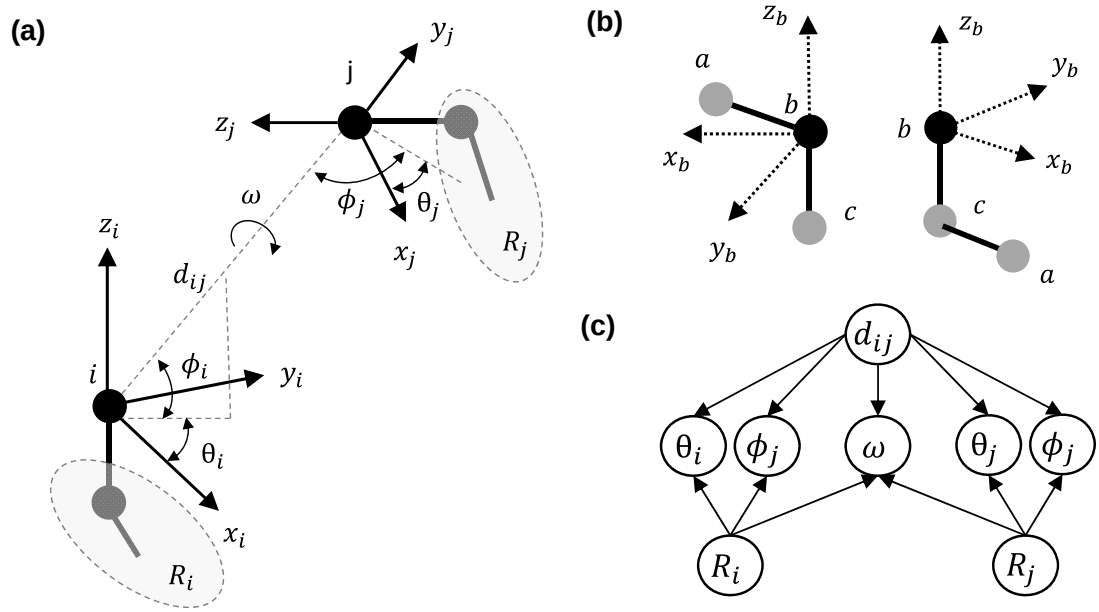

Figure 3-1. (a) Description of the distance and relative orientation of two local coordinate frames for interacting atom types *i* and *j*. (b) Local coordinate frames for atom *b* in two cases: atom *b* has two bonded atoms *a* and *c* (left), and atom *b* has only one bonded atom *c* with next bonded atom *a* (right). (c) Bayesian network structure representing conditional independence of parameters defined in the ROTAS potential

First, we attach local coordinate frames to all the types of atoms considered. In this study, we only consider the interaction between heavy atoms and distinguish 167 residue-specific heavy atom types. There are two cases for attaching the local coordinate frame to heavy atoms (see Figure 3-1 (b)): atom *b* has two bonded atoms *a* and *c* (left figure), and atom *b* has only one bonded atom *c* with next bonded atom *a* (right figure). In both cases, unit direction vectors for the local coordinate frame can be defined as follows:

$$z_b = \frac{r_b - r_c}{|r_b - r_c|};\ y_b = \frac{z_b \times (r_a - r_b)}{|z_b \times (r_a - r_b)|};\ x_b = y_b \times z_b \tag{1}$$

Once the local coordinate frames are defined, the interaction between atom i and j is then specified by eight parameters: $d_{ij}, \theta_i, \phi_i, \theta_j, \phi_j, \omega, R_i \text{and} R_j$ (see Figure 3-1 (a)). Here, d_{ij}, θ_i, ϕ_i are the spherical coordinates of atom j with respect to the local frame of atom i, and ω is a torsional angle around d_{ij}, and $R_i \text{and} R_j$ represent the local structural environments of atoms i and j, respectively. The equation of the ROTAS potential can be obtained using the inverse Boltzmann law:

$$E(d_{ij},\ \theta_i,\ \varphi_i,\ \theta_j,\ \varphi_j,\ \omega,\ R_i, R_j) = -k_B T \ln \frac{P^{obs}(d_{ij},\ \theta_i,\ \varphi_i,\ \theta_j,\ \varphi_j,\ \omega,\ R_i, R_j)}{P^{exp}(d_{ij},\ \theta_i,\ \varphi_i,\ \theta_j,\ \varphi_j,\ \omega,\ R_i, R_j)} \quad (2)$$

where k_B is the Boltzmann constant and T is the absolute temperature. P^{obs} is the probability of a particular state ($d_{ij},\ \theta_i,\ \varphi_i,\ \theta_j,\ \varphi_j,\ \omega,\ R_i, R_j$) observed in a sample of known protein structuresand P^{exp} the expected probability of the same state in a reference state where the interaction is zero. Considering that there are a finite number of known protein structures, we assume conditional dependencies of parameters as shown in Figure 3-1 (c) to obtain sufficient statistics. Namely, the angular parameters are assumed as independent of each other at the given distance and local structural environments, which has also been similarly assumed in other studies[55,56,159]. Considering the independence assumption, the joint probability can be written as

$$P(d_{ij},\ \theta_i,\ \varphi_i,\ \theta_j,\ \varphi_j,\ \omega,\ R_i, R_j)$$
$$= P(d_{ij})P(R_i)P(\theta_i|d_{ij}, R_i)P(\phi_i|d_{ij}, R_i)P(R_j)P(\theta_j|d_{ij}, R_j)P(\phi_j|d_{ij}, R_j)P(\omega|d_{ij}, R_i, R_j) \quad (3)$$

Applying Bayes' rule, the conditional probability, $P(\omega|d_{ij}, R_i, R_j)$ can be rewritten as

$$P(\omega|d_{ij}, R_i, R_j) = \frac{P(\omega|d_{ij}, R_i)P(\omega|d_{ij}, R_j)P(d_{ij})}{P(\omega|d_{ij})P(d_{ij})} \quad (4)$$

Integrating equation (2)~(4) gives the final equation for the ROTAS potential energy function:

$$\begin{aligned} &E(d_{ij}, \theta_i, \varphi_i, \theta_j, \varphi_j, \omega, R_i, R_j) \\ &= E(d_{ij}) + E(\theta_j|d_{ij}, R_j) + E(\phi_j|d_{ij}, R_j) + E(\omega|d_{ij}, R_j) + E(\theta_i|d_{ij}, R_i) \\ &+ E(\phi_i|d_{ij}, R_i) + E(\omega|d_{ij}, R_i) - E(\omega|d_{ij}) + E(R_i) + \mathrm{E}(R_j) \end{aligned} \quad (5)$$

Here, $E(R_i)$and $E(R_j)$ can be seen as intra-energy terms associated to the local structural environments (*e.g.*, rotamer intrinsic energy). Assuming that the stability of overall folded structure is mainly determined by non-bonded interactions, we ignore these terms in this study.

3.3.2 Defining the local structural environment

The local structural environment of each atom type is defined by the rotameric state of the residue to which the atom belongs. The observed side-chain dihedral angles cluster around ideal values, such as +60°, -60°, and 180° dihedral angles expected between two sp3 hybridized atoms (Figure 3-2). A rotameric state is a combination of these ideal dihedral angles that describes the residue conformation, assuming the bond lengths and angles are fixed. Since long residues such as Met, Lys or Arg have too many rotameric states to obtain sufficient statistics for each rotamer, we associate up to two side-chain dihedral angles whose rotating bonds are within 3 bond lengths from the considered atom to its local structural environment. For example, the local structural environment of CB, CG and CD atoms in Lys is defined by a combination of {X1, X2}, {X2, X3} and {X3, X4} dihedral angles, respectively. One exception is the backbone oxygen atom, which is related to {X1, X2} angles because it frequently interacts with side-chain atoms depending on

backbone ψ angle. Also, every atom in Pro is associated to only X1 angle because X2 dihedral angle is strongly correlated with X1 dihedral angle.

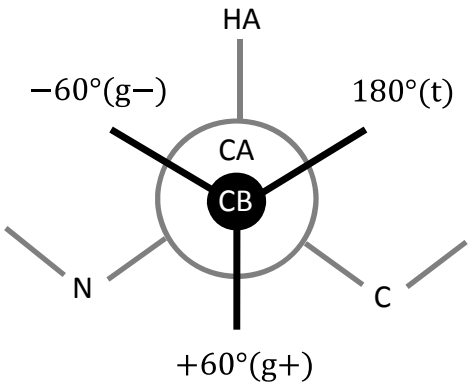

Figure 3-2. Newman diagram of three favored X1 angles in proteins. The -60, +60, and 180 angles are often referred to as gauche minus(g-), gauche plus(g+), and trans(t), respectively.

For each side-chain dihedral angle, we divide the dihedral angle space into three or two regions. The dihedral angle between two sp3 hybridized atoms is classified into three distinct rotameric states: 0°~120° (g+), -120°~0° (g-), and 120°~240° (t). Last dihedral angles of Asn, Asp, Gln, Glu, His, Trp, Phe and Tyr are non-rotameric[158]. For those non-rotameric dihedral angles, we divide the dihedral angle space into two regions, {(0~π), (-π~0)}. X1 dihedral angle of Pro is also divided into two regions, positive or negative. All 167 heavy atom types and their associated dihedral angles for defining the local structural environments are listed in Table 3-1.

Table 3-1. All 167 residue-specific heavy atom types and associated side-chain dihedral angles for defining their local structural environments.

Amino Acids	Dihedrals	Associated Atoms	Number of rotamric states
GLY	-	C, O, N, CA	1
ALA	-	C, O, N, CA, CB	1
CYS	X_1	C, O, N, CA, CB, SG	3

SER	X_1	C, O, N, CA, CB, OG	3
THR	X_1	C, O, N, CA, CB, OG1,CG2	3
PRO	X_1	C, O, N, CA, CB, CG, CD	3
VAL	X_1	C, O, N, CA, CB, CG1, CG2	3
ILE	X_1, X_2	C, O, N, CA, CB, CD1, CG1, CG2	9
LEU	X_1, X_2	C, O, N, CA, CB, CG, CD1, CD2	9
ASP	X_1, X_2	C, O, N, CA, CB, CG, OD1, OD2	6
ASN	X_1, X_2	C, O, N, CA, CB, CG, OD1, ND2	6
GLU	X_1, X_2	C, O, N, CA, CB, CG	9
	X_2, X_3	CD, OE1, OE2	6
GLN	X_1, X_2	C, O, N, CA, CB	9
	X_2, X_3	CG, CD, OE1, NE2	6
MET	X_1, X_2	C, O, N, CA, CB	9
	X_2, X_3	CG, SD, CE	9
ARG	X_1, X_2	C, O, N, CA, CB	9
	X_2, X_3	CG	9
	X_3, X_4	CD, NE, CZ	9
	X_4	NH1, NH2	3
LYS	X_1, X_2	C, O, N, CA, CB	9
	X_2, X_3	CG	9
	X_3, X_4	CD, CE, NZ	9
HIS	X_1, X_2	C, O, N, CA, CB, CG, ND1, CD2	6
	X_2	CE1, NE2	2
PHE	X_1	C, O, N, CA, CB	3
	X_1, X_2	CG, CD1,CD2	6
	X_2	CE1, CE2, CZ	2
TRP	X_1, X_2	C, O, N, CA, CB, CG, CD1, CD2	6
	X_2	NE1, CE2, CE3, CZ2, CZ3, CH2	2
TYR	X_1	C, O, N, CA, CB	3
	X_1, X_2	CG, CD1, CD2	6
	X_2	CE1, CE2, CZ, OH	2

3.3.3 Construction of distance-dependent pairwise potential

In ROTAS, the distance-dependent pairwise energy term does not involve the rotamer-dependence. While the observed distance-dependent pairwise probability $P^{obs}(d_{ij})$ can be calculated straightforwardly, a reference state needs to be defined to compute the expected

probability $P^{exp}(d_{ij})$. Because the focus of this work is the effect of rotamer-dependence on the performance of potential energy function, we simply employed the DFIRE[35] reference state. The DFIRE reference state is an ideal gas system in which atoms are uniformly distributed, and has been successfully applied in other studies[55,56,160]. The DFIRE-based distance-dependent potential energy can be calculated by

$$E(d_{ij}) = -RT \log\left[\left(\frac{d^{cut}}{d_{ij}}\right)^{\alpha} \cdot \frac{N^{obs}(d_{ij})}{N^{obs}(d_{ij}^{cut})}\right] \quad (1)$$

where $N^{obs}(d_{ij})$ is the number of observed atom pair i and j at distance d, and α is a scaling factor such that $N^{exp}(d)$ increases in d^{α}. Beyond a distance cutoff d_{ij}^{cut}, it is assumed that both observed and expected pairwise distributions are equal. Here we set d_{ij}^{cut} = 15 Å and α = 1.61 as suggested by the original work[35]. To obtain the distribution, the bin width is set to 0.5 Å from 0 to 15 Å. When estimating the observed probability and evaluating the distance-dependent pairwise potential, atom pairs that are in the same residue are excluded.

In addition to DFIRE, we constructed other widely used distance-dependent potentials such as RAPDF[153], KBP[161], DOPE[36] and RW[57] and tested each of them in ROTAS in order to examine the influence of different reference states on the performance of ROTAS. The same structural database, distance cutoff and bin width were applied.

3.3.4 Construction of orientation-dependent pairwise potential

In order to obtain smooth and continuous estimates of the observed probability distribution of angular parameters $\{\theta_i, \varphi_i, \omega\}$ for a particular distance and rotameric state (d_{ij}, R_i) from a finite sample data, we employed kernel density estimation. Suppose that $\{\theta_s\}_{s=1\ldots N}$ is a set of angles

θ_i collected at a given distance d_{ij} and rotameric state R_i. Then the probability density $p(\theta_i|d_{ij}, R_i)$ can be calculated using von Mises distribution as the kernel:

$$p(\theta_i|d_{ij}, R_i) = \frac{1}{N}\sum_{s=1}^{N} K_{VM}(\theta_i; \theta_s, \kappa)$$

$$K_{VM}(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp[\kappa \cdot \cos(\theta - \mu)] \quad (1)$$

where K_{VM} denotes the von Mises kernel function, κ is the kernel bandwidth controlling the smoothness of the kernel and $I_0$is the Bessel function of the first kind of order 0. Here, we set κ = 8.21 which is equivalent to $\sigma = \pi/9$ in the normal distribution. The distances d_{ij} were discretized into 0.5 Å bins which span from 2to 15Å. The kernel density estimator is computed at π/9 grid points that are ranged from -π to π (in case of ϕ, from -π/2 to π/2).

The relative orientation between atoms is significantly affected by chain connectivity constrains when the atoms are positioned in residues that are close in the sequence. In order to reduce the chain (or bond) connectivity effect on the estimates of orientation-dependent probability, we applied a sequence separation as done in other studies[22,37,55]. In this study, only atom pairs that are separated by at least 6 residues along the protein chain are considered.

Despite the use of kernel density estimation, in the case of rarely observed rotameric states in protein structures, there is still a problem of insufficient sample data. For example, the number of Ile rotamers in (+60°,+60°) dihedral pair is less than 1,000 in our database. In such case, rather than using poorly estimated probability density $p^{obs}(\theta_i|d_{ij}, R_i)$, we calculated the corrected probability density $p_{corr}^{obs}(\theta_i|d_{ij}, R_i)$ as a linear combination of $p^{obs}(\theta_i|d_{ij}, R_i)$ and $p^{obs}(\theta_i|d_{ij})$:

$$p_{corr}^{obs}(\theta_i|d_{ij},R_i) = \frac{1}{1+\sigma N(d_{ij},R_i)} P^{obs}(\theta_i|d_{ij}) + \frac{\sigma N(d_{ij},R_i)}{1+\sigma N(d_{ij},R_i)} P^{obs}(\theta_i|d_{ij},R_i) \quad (2)$$

where $N(d_{ij},R_i)$ is the number of observations used to estimate $P^{obs}(\theta_i|d_{ij},R_i)$ and σ is a parameter that controls how many observations must be sampled such that both $P^{obs}(\theta_i|d_{ij},R_i)$ and $P^{obs}(\theta_i|d_{ij})$ would have equal weights. Here we set σ = 1/100.

The expected probability distribution of angles can be calculated from a reference state in which the relative orientation of atom pair is determined randomly. Thus the expected probability is calculated by:

$$P^{exp}(\theta) = P^{exp}(\omega) = const = \frac{1}{2\pi}$$

$$P^{exp}(\phi) = \frac{1}{M}\int_{-\pi/2}^{\pi/2} \cos(\psi)\cdot K_{VM}(\phi;\psi,\kappa)d\psi \quad (3)$$

where M is a normalization factor such that the integration of $P^{exp}(\phi)$ from -π/2 to π/2 becomes one. $P^{exp}(\phi)$ is calculated numerically because there is no analytical way for integrating above equation.

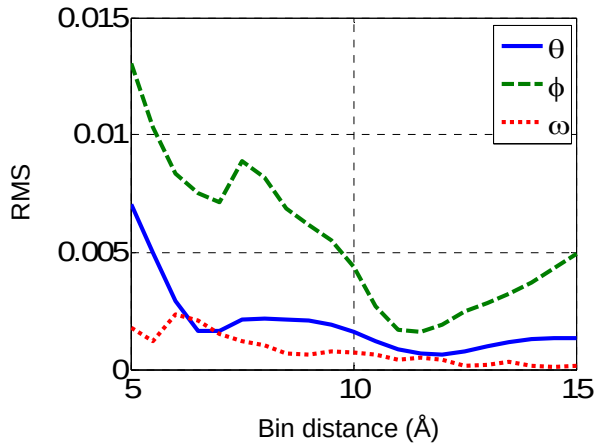

Figure 3-3. The distance dependence of root mean square of ($P^{obs} - P^{exp}$) for angular parameters. The observed probability distribution is calculated over all pairs of atom types. The thin, dashed and dotted curves corresponds to θ, ϕ and ω, respectively.

3.3.5 Interaction cutoff for the ROTAS potential

Although the distance bin between 14.5 and 15Å was used as the cutoff in the construction of distance-dependent pairwise potential, we calculate the energy score within 10 Å and ignore the long-range tail of potentials beyond 10 Å. In fact, most physical interactions between atoms rapidly converge to zero beyond 8~10 Å. However, statistically derived potentials are likely to have fluctuations in the long-range, which inherently resulted from the statistical uncertainties. For example, Figure 3-3 reveals that the deviations of the observed probability from the expected probability for angular parameters do not consistently decrease as the atom-pair distance increases. It is noted that the root mean square of $(P^{obs}(\phi|d) - P^{exp}(\phi|d))$ increase after 12 Å. In addition, it was reported that distance-dependent pairwise potentials between hydrophobic atom pairs have either repulsive or attractive tail in the long range, even if no electrostatic interaction exists[18]. Thus it's not always beneficial to include the long-range interactions in

statistical potentials. We simply set the interaction cutoff to 10 Å without fine-tuning against a specific training dataset.

3.3.6 Preparation of PDB structures

We obtained a set of protein X-ray structures with a maximum R-factor of 0.25 and a resolution better than 2 Å from the protein sequence culling server, PISCES[139]. Also, protein chains were filtered out with a 40% sequence identity cutoff in order to have a set of non-homologous protein structures. A total 9321 protein structures were selected and downloaded from the Protein Data Bank (PDB)[140]. The program REDUCE[141] was used to optimize the flip states of Asn, Gln, and His in all protein structures. Residues with multiple side-chain conformations were modified such that only the side-chain conformations with atoms having the highest occupancy and/or lowest temperature factors were used.

3.3.7 Performance evaluation using decoy sets

We tested the ROTAS potential on various sets of decoys generated by different methods. A total of 13 decoy sets, including 4state_reduced[162], fisa[25], fisa_casp3[25], lmds[163], hg_structal, ig_structal, ig_structal_hires, lattice_ssfit[164], moulder[165], Rosetta[166], I-TASSER[57], AMBER99[167] and CASP5-8[147], were used The first 8 decoy sets were downloaded from the Decoys 'R' Us database[168] (http://dd.compbio.washington.edu/). The moulder decoy set produced by iterative target-template alignment and comparative-modeling methods was download from the Sali lab (http://salilab.org/decoys/). Three ab-initio simulation based decoy sets, Rosetta, I-TASSWER, Amber99 were obtained from http://depts.washington.edu/bakerpg/decoys/, http://zhanglab.ccmb.med.umich.edu/decoys/, and http://cssb.biology.gatech.edu/amberff99/, respectively. The CASP5-8 decoy set collected from the CASP5-CASP8 experiments was

downloaded from http://zhanglab.ccmb.med.umich.edu/RW/ (cleaned version). The decoy models in this set were generated by a large variety of groups and methods participated in the CASP experiments.

The performance of ROTAS potential was compared to those of four other existing atomic potentials which take into account the orientation-dependencies on the interactions between atoms, blocks or side-chains: dDFIRE[56], OPUS_PSP[68], RWplus[57], and GOAP[55]. The binary programs for these potentials were downloaded from the corresponding authors' websites. Because ROTAS can be seen as an extended version of GOAP, we constructed our own GOAP potential energy function using the same structure database and techniques that were used for the construction of ROTAS. In this manner we reduced the possibility that estimation of probability distribution, specific computational implementation, or other technical aspects could affect the results, so that the improvements of ROTAS compared to GOAP can be fairly demonstrated.

The performance of statistical potentials is evaluated by three aspects: (1) the recognition of native structure from decoys, (2) the selection of the best (most native-like) decoy model and (3) the correlation between the energy score and model quality. The quality of decoy models was assessed by TM-score which measures the similarity between two protein structures by a score between (0, 1] [169].

3.4 Results and Discussion

3.4.1 The influence of rotameric states on atomic interactions

We constructed both ROTAS and GOAP potentials using the same structure database and techniques as described in Methods section. Figure 3-4 shows the energy profiles of ROTAS

and GOAP for four different atom pairs. First of all, all examples clearly show that the energy profiles of ROTAS significantly vary depending on the rotameric state. While GOAP only reflects in some average sense the preferred orientation between interacting atoms, ROTAS adjusts the preferred orientation accurately depending on the rotameric state. The first example shows the disulfide interaction between Cys SG atoms (Figure 3-4 (a)). The torsional angular term $E(\omega|d_{ij}, R_i)$ has two distinct favored positions regardless of the rotameric state. However, $E(\theta_i|d_{ij}, R_i)$ shows slightly different curves. The most favored positions for θ_i are 90 Å, -72 Å and 72° for three rotameric states of Cys, g+, g-, and t, respectively. This might be due to close steric interactions between the backbone atoms and Cys SG. The second example is a typical hydrogen bond interaction between Ser O and Gly N at a distance of 3 Å (Figure 3-4 (b)). It is observed that different relative position of Ser OG atom significantly affects on the hydrogen bond interaction between backbone atoms. Figure 3-4 (c) shows an example of a non-polar interaction between Ile CG2 and Val CG1 at a distance of 5 Å. In this example, the GOAP potential shows very similar energy profiles with a particular rotameric state, (X1 = g- and X2 = t), which is the most populated rotamer for Ile (59% of Ile residues observed in this rotamer). The last example shows a polar interaction between Lys Nz and Asp OD2 at a distance of 7 Å. It is noted that, although the pair distance is relatively longer rather than previous examples, the energy profiles of different rotameric states significantly differ. This suggests that the rotamer-dependency is not limited to short range interactions resulting from strong steric effects.

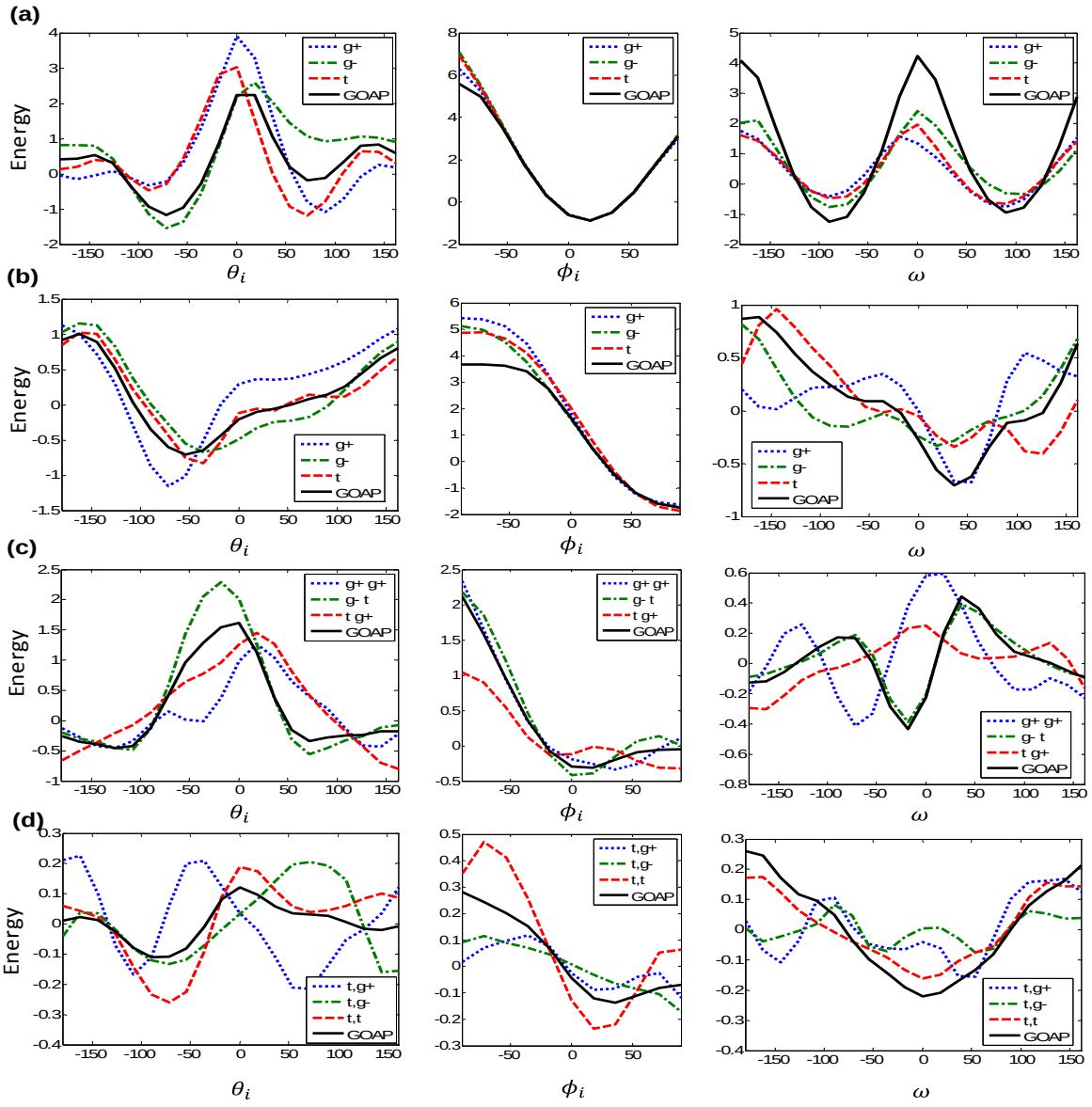

Figure 3-4. Examples of the rotamer dependence of the energy terms, $E(\theta_i|d_{ij},R_i)$, $E(\phi_i|d_{ij},R_i)$, and $E(\omega|d_{ij},R_i)$ in ROTAS potential. (a) Disulfide bond interaction for *i*and*j* = Cys SG at $d_{ij}=2$ Å, (b) hydrogen bond interaction for *i* = Ser O and *j* = Gly N at $d_{ij}=3$ Å, (c) nonpolar interaction for *i* = Ile CG2 and *j* = Val CG1 at $d_{ij}=5$ Å, and (d) polar interaction for *i* = Lys NZ and *j* = Asp OD2 at $d_{ij}=7$ Å.

3.4.2 Native structure recognition

We assessed the performance of ROTAS in terms of its ability for recognizing the native structures from decoy models and compared it with those of four other statistical potentials. In this test, the performance was assessed by two measures: the number of targets having the native structure ranked as the lowest energy score and Z-score of the native structure. The Z-score represents the energy gap between the energy of native structure (E_{native}) and the averaged energy of all decoys ($< E_{decoy} >$) in units of the energy standard deviation of all decoys (σ_{decoy}), which is defined as:

$$Z = \frac{E_{native} - < E_{decoy} >}{\sigma_{decoy}} \quad (1)$$

The lower the Z-score, the better the potential is for recognizing the native structures. The results of the native structure recognition are summarized in Table 3-2. ROTAS could recognize total 409 native structures correctly out of 469 targets, which is the best success rate (87.2%) in the comparison. Although RWplus and GOAP record the highest success rate on I-TASSER and Amber 99, respectively, for the remaining 11 decoy sets, ROTAS recognized native structures more or equal than other potentials. GOAP recognized 399 native structures (85.1% success rate) with the average Z-score of -3.35. These results are consistent with those in the GOAP article which reported that the success rate and the average Z-score of GOAP are 81.3% (226 out of 278) and -3.57, respectively.

Table 3-2. Performance on native structure recognition

Decoy set	Targets	dDFIRE	OPUS_PSP	RWplus	GOAP	ROTAS
4state_reduced	7	7 (-4.15)	7 (-4.49)	6 (-3.50)	7 (-4.67)	7 (**-5.07**)
fisa	4	3 (-3.80)	3 (-4.24)	3 (-4.78)	3 (-3.98)	3 (**-4.83**)
lmds	10	6 (-2.44)	8 (**-5.63**)	7 (-1.03)	8 (-4.34)	8 (-5.47)
fisa_casp3	5	4 (-4.73)	5 (-6.33)	4 (-5.17)	4 (-6.65)	4 (**-7.48**)
hg_structal	29	15 (-1.25)	18 (-2.28)	12 (-1.70)	20 (-2.46)	**22** (**-2.51**)
ig_structal	61	26 (-0.82)	22 (-1.13)	0 (1.11)	44 (-1.91)	**46** (**-2.25**)
ig_structal_hires	20	16 (-2.00)	15 (-1.79)	0 (0.31)	18 (-2.68)	18 (**-3.11**)
lattice_ssfit	8	8 (**-10.08**)	8 (-6.56)	8 (-8.77)	8 (-7.94)	8 (-8.90)
moulder	20	18 (-2.74)	19 (**-4.83**)	19 (-2.84)	19 (-3.53)	19 (-3.76)
rosetta	59	12 (-0.43)	40 (-3.62)	20 (-1.21)	43 (-3.66)	**48** (**-4.18**)
I-TASSER	56	48 (-5.03)	49 (-5.40)	**56** (-5.77)	48 (-5.81)	49 (**-7.31**)
Amber99	47	27 (-3.42)	20 (-2.58)	16 (-2.38)	**38** (-4.38)	37 (**-4.48**)
CASP5-8	143	98 (-1.34)	134 (**-2.45**)	106 (-1.67)	139 (-2.26)	**140** (-2.43)
Total	469	288 (-2.16)	348 (-3.08)	257 (-1.98)	399 (-3.35)	**409** (**-3.80**)

Numbers outside the parentheses are the numbers of correctly recognized native structures and the ones in the parentheses are the average Z-scores of the native structures. The best scores are highlighted in bold type.

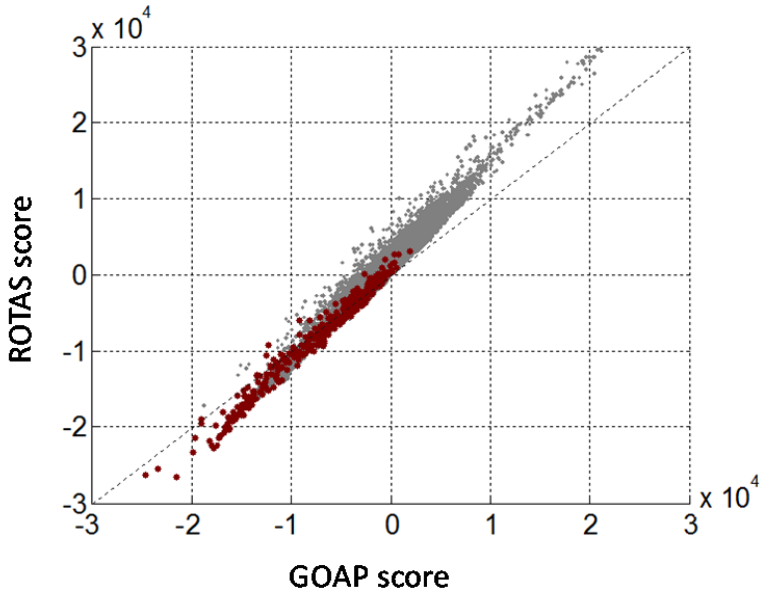

Figure 3-5. Relationship between the energy scores of ROTAS and GOAP for all native and decoy structures.

The relative improvement of ROTAS over GOAP can be clearly seen in the average Z-scores. While GOAP correctly recognized the native structures comparable to ROTAS, it is noticed that ROTAS shows consistently improved Z-scores over all decoy sets tested here. Figure 3-5 shows the relationship between the energy scores of ROTAS and GOAP for all native (red) and decoy (gray) structures used in the test. It can be easily confirmed that ROTAS scores native structures with lower energies and decoy models with higher energies, compared to GOAP.

We found that the performance of ROTAS in native structure recognition is largely affected by experimental methods used to determine the native structures. The success rate of ROTAS is 89% for targets whose native structures were determined by X-ray crystallography, whereas the success rate significantly decreases to 60% when the native structures were determined by NMR spectroscopy (Table 3-3). Furthermore, both the average success rate and Z-score decrease for

low-resolution native structures. This might be because the ROTAS potential was constructed based on high-resolution X-ray structures. The large margin of error in the location of atoms in low-resolution structures (*e.g.*, > 2.2 Å) would decrease the confidence of computed energy score. This trend is also observed for other potential energy functions except RWplus which performs very well on NMR native structures. In fact, the RWplus potential can correctly recognize all 18 native NMR structures in the I-TASSER decoy set with low Z-scores.

Table 3-3. The ability of ROTAS on native structure recognition as a function of native structure resolution

Exp. method	Resolution	Targets	Rank1	Z
NMR	-	25	15 (60%)	-3.32
X-ray	all	444	394 (89%)	-3.82
	R <= 1.8	152	143 (94%)	-4.91
	1.8 <= R < 2.2	171	153 (89%)	-3.71
	2.2<= R < 2.8	102	86 (84%)	-2.78
	2.8 < R	19	12 (63%)	-1.79

Numbers in parentheses are the ratio of Rank1 structures

3.4.3 Best model selection

We also assessed the ability of ROTAS in selecting the best models without native structures. This is more difficult and realistic task than the native structure recognition because, in practice, potential energy functions are used to find more and more native-like conformations in an iterative way when the native structure is not known. Thus, good potential energy function should be able to score the most native-like decoy model in the lowest energy. In this study, we use TM-score[169] to assess the quality of decoy models quantitatively. The TM-score measures the similarity between two protein structures by a score between (0, 1]. It is reported that TM-

score is more accurate than other measures such as RMSD or GDT_TS because TM-score is sensitive to overall topology rather than local substructures[170].

Table 3-4 summarizes the result of the best model selection by dDFIRE, OPUS_PSP, RWplus, GOAP and ROTAS for 13 decoy sets. Measures $\log P_{B1}$ and $\log P_{B10}$ are the log probability of selecting the best (highest TM-score) model as the lowest energy model or among the top 10 lowest energy models, respectively. Suppose the top i^{th} scoring conformation x_i has the TM-score rank of R_i in n decoy models, then the log probability can be calculated as

$$\log P_{BN} = \log(\frac{\min_{i=1...N} R_i}{n}) \quad (1)$$

In both measures, GOAP and ROTAS shows better performance than other three potentials, dDFIRE, OPUS_PSP and RWplus. The average $\log P_{B1}$ by GOAP is slightly better than that by ROTAS, whereas the average $\log P_{B10}$ by ROTAS is better than that by GOAP. This indicates that the lowest energy model by GOAP is likely to be better in TM-score than that by ROTAS. However, when we consider the top 10 lowest energy models, ROTAS tend to include better TM-score decoy models in the top 10 than GOAP.

Table 3-4.Performance on best model selection

	dDFIRE		OPUS_PSP		RWplus		GOAP		ROTAS	
	$logP_{B1}$	$logP_{B10}$	$logP_{B1}$	$logP_{B10}$	$logP_{B1}$	$logP_{B10}$	$logP_{B1}$	$logP_{B10}$	$logP_{B1}$	$logP_{B10}$
4state_reduced	-3.604	-5.844	-4.031	**-6.142**	-2.796	-5.704	-4.675	-6.042	**-4.997**	-6.100
fisa	-2.685	-4.029	-1.568	-3.607	-2.184	-4.059	**-3.112**	-4.337	-2.226	**-5.191**
lmds	-1.513	-3.394	-1.084	-3.362	-1.041	-3.449	**-1.918**	**-3.573**	-1.825	-3.570
fisa_casp3	-1.420	-3.242	-0.813	-3.133	-1.189	**-4.226**	**-1.561**	-3.328	-1.295	-3.781
hg_structal	-2.444	**-3.329**	-2.549	-3.174	-2.500	-3.329	-2.419	-3.292	**-2.551**	-3.306
ig_structal	-2.061	-3.583	**-2.595**	**-3.764**	-2.137	-3.558	-2.172	-3.693	-1.956	-3.672
ig_structal_hires	-1.843	-2.661	-1.929	**-2.820**	**-1.949**	-2.806	-1.905	-2.710	-1.827	-2.765
lattice_ssfit	-1.603	-3.679	-1.031	-3.532	-1.131	**-4.098**	-1.238	-2.719	**-1.651**	-3.009
moulder	-3.175	-4.787	-2.706	-4.619	-3.059	-4.905	**-3.835**	-5.083	-3.717	**-5.118**
rosetta	-1.303	-3.448	**-1.761**	-3.179	-1.719	**-3.660**	-1.652	-3.563	-1.512	-3.590
I-TASSER	-1.834	**-3.871**	-1.259	-3.599	-1.782	-3.727	-1.773	-3.612	**-1.865**	-3.686
Amber99	-3.644	-5.434	-3.027	-4.719	-3.478	-4.935	-4.092	-5.643	**-4.247**	**-5.891**
CASP5-8	-1.890	-2.800	-1.358	-2.766	-1.877	**-2.808**	**-1.906**	-2.803	-1.873	-2.798
Total	-2.109	-3.575	-1.905	-3.437	-2.111	-3.562	**-2.263**	-3.601	-2.233	**-3.656**

3.4.4 Correlation between the energy score and decoy model quality

Next, we examined the correlation of the energy score and the quality of decoy models in order to assess the ability of ROTAS in guiding conformation sampling to near-native states. In an energy landscape perspective, a good potential energy function should not only be able to make a deep energy minimum with steep wall at the native state but also be able to form a middle-range funnel biased toward the native state. In Table 3-5, we compare the performance of potentials as assessed by both their Pearson correlation coefficient(r) and the Kendall's rank correlation coefficient(τ) between the energy score and TM-score. Overall, the performance of potentials does not show significant difference depending on the correlation measures. We find that

ROTAS shows the best performance in both measures. GOAP yields the second best performance in the average correlation coefficients. dDFIRE and RWplus have comparable performance although the average correlation coefficients of RWplus is slightly better than those of dDFIRE. OPUS_PSP performs significantly worse than the other potentials tested although its performance comes in third in the native structure recognition. Figure 3-6 shows some examples of the correlation between ROTAS energy and TM-score from different decoy sets.

Table 3-5. Performance on correlation coefficients between energy score and model quality

Decoy set	**dDFIRE**		**OPUS_PSP**		**RWplus**		**GOAP**		**ROTAS**	
	r	τ	r	τ	r	τ	r	τ	r	τ
4state_reduced	-0.693	-0.483	-0.590	-0.399	-0.605	-0.417	-0.766	-0.550	**-0.783**	**-0.562**
fisa	-0.461	-0.321	-0.282	-0.189	-0.462	-0.315	**-0.476**	**-0.327**	-0.442	-0.297
lmds	**-0.248**	**-0.168**	-0.091	-0.054	-0.147	-0.095	-0.228	-0.149	-0.227	-0.149
fisa_casp3	**-0.251**	**-0.168**	-0.090	-0.063	-0.236	-0.152	-0.161	-0.102	-0.182	-0.117
hg_structal	-0.796	-0.618	-0.752	-0.553	-0.806	**-0.630**	-0.808	-0.609	**-0.811**	-0.602
ig_structal	-0.766	-0.308	-0.779	-0.340	-0.782	-0.277	**-0.851**	**-0.377**	-0.836	-0.372
ig_structal_hires	-0.844	-0.373	-0.832	-0.403	-0.879	-0.411	**-0.890**	**-0.436**	-0.860	-0.401
lattice_ssfit	-0.068	-0.047	-0.050	-0.033	**-0.096**	**-0.059**	-0.034	-0.025	-0.043	-0.029
moulder	-0.832	**-0.670**	-0.755	-0.600	-0.792	-0.642	-0.823	-0.660	**-0.833**	-0.665
rosetta	-0.265	-0.176	-0.192	-0.113	-0.350	**-0.237**	-0.330	-0.212	**-0.351**	-0.221
I-TASSER	**-0.522**	**-0.303**	-0.281	-0.195	-0.485	-0.290	-0.465	-0.276	-0.456	-0.271
Amber99	-0.609	-0.339	-0.421	-0.201	-0.526	-0.313	-0.692	-0.355	**-0.721**	**-0.357**
CASP5-8	-0.594	-0.488	-0.440	-0.354	-0.611	-0.501	-0.593	-0.490	**-0.613**	**-0.502**
Total	-0.581	-0.380	-0.465	-0.297	-0.584	-0.382	-0.603	-0.394	**-0.612**	**-0.396**

r : Pearson's correlation coefficient
τ : Kendall's rank correlation coefficient

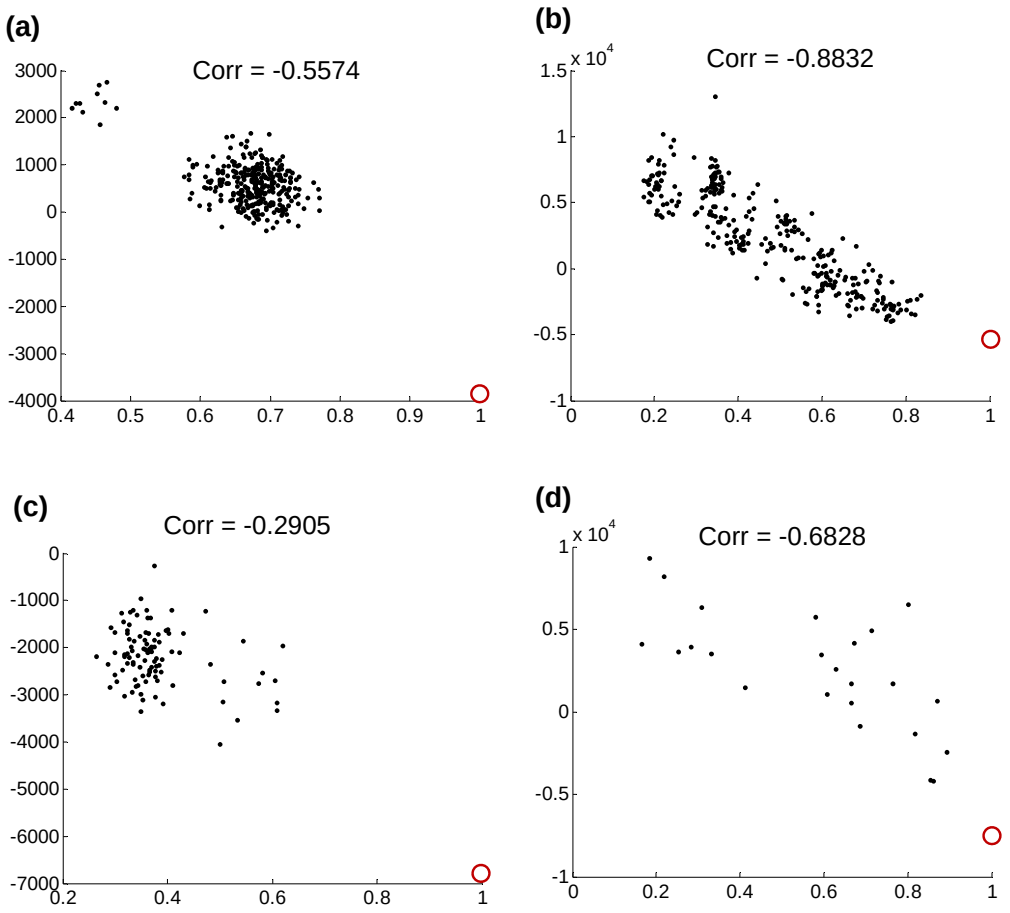

Figure 3-6. Examples of Pearson correlation between ROTAS energy and TM-score: (a) 1SCP_ in I-TASSER, (b) 1CAU in Moulder, (c) 1LOU in Rosetta and (d) T0324 in CASP7. The native structures are included and represented as empty circle at TM-score = 1.

3.4.5 Interaction cutoff effect on the performance

The interaction cutoff effect on the performance of ROTAS and GOAP was examined. The performances of ROTAS and GOAP are significantly affected by the interaction cutoff (Figure 3-7). Interaction cutoffs between 7 and 10 Å maximize the number of correctly recognized native structures and minimize the average Z-score for both potentials. Increasing or decreasing the cutoff outside of this range makes the performance for native structure recognition worse

dramatically. The performance of ROTAS and GOAP for recognizing the best models is maximized around 11~13 Å. On the other hand, as the interaction cutoff increases, the average correlation coefficient decreases. But the slopes around 13~15 Å are almost zero. Although the optimal interaction cutoff varies depending on the evaluation criteria, we confirm that the long-range interactions in statistical potentials could reduce the performance of potentials and an interaction cutoff of 10 Å for ROTAS gives a moderate performance on various evaluation criteria. It should be noticed that even though optimal interaction cutoffs are applied to individual potentials, ROTAS performs better than GOAP.

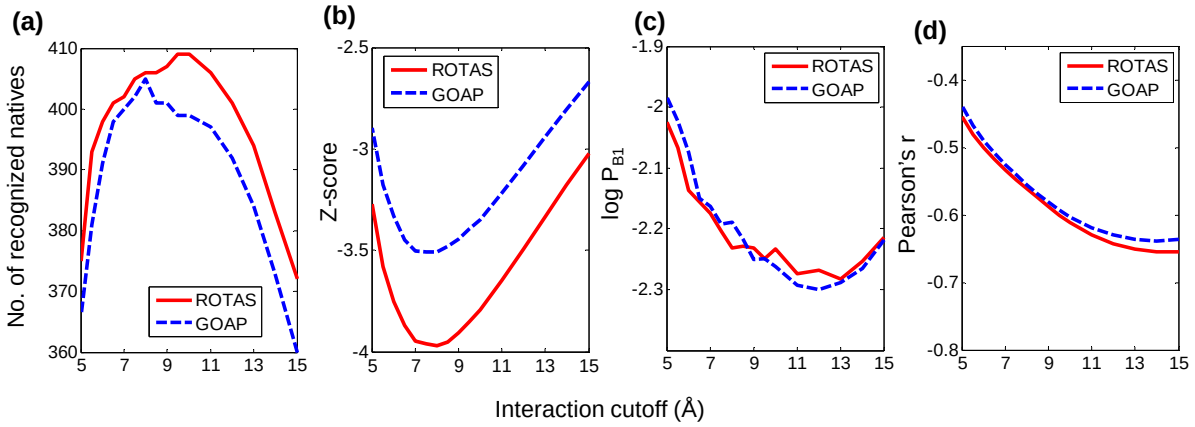

Figure 3-7. Relation between the cutoff distance and the performance of ROTAS and GOAP: (a) Number of correctly recognized native structures (b) Average Z-score, (c) Average log P_{B1} and (d) Average Pearson's correlation coefficient.

It is noticed that the highest average correlation coefficient is obtained when we consider all the long-range interactions available in the potentials. However, in this case, the native structures are poorly recognized. A similar observation that a scoring function producing a good linear correlation is normally less capable of recognizing the native state has been reported in a previous study[171]. A theoretical study argue that the potential energy of near-native

conformations might not be linearly related to their distances from the native state[172]. Also, since a shorter interaction cutoff would increase ruggedness of the energy landscape[173], the energy score of decoy models might be affected by small structural differences sensitively.

Table 3-6. Performance of different distance-dependent pairwise potentials in ROTAS

Ref. state	Rank1	Z-score	$\log P_{B1}$	$\log P_{B10}$	Pearson's r	Kendall's τ
DFIRE	409	-3.795	-2.233	**-3.656**	-0.612	-0.396
DOPE	409	-3.810	-2.172	-3.576	-0.566	-0.358
RW	408	-3.818	**-2.258**	-3.645	**-0.617**	**-0.401**
RAPDF	409	**-3.867**	-2.185	-3.592	-0.578	-0.367
KBP	409	-3.638	-2.276	-3.630	-0.609	-0.393

3.4.6 Different reference states for distance-dependent pairwise potential

We applied five widely-used reference states including DFIRE, DOPE, RW, RAPDF and KBP for the distance-dependent pairwise potential in ROTAS and compared their performances. To rigorously compare the influence of the reference state on the performance, we constructed all five distance-dependent pairwise potentials using the same structure database, the same cutoff distance, and the same bin width. Table 3-6 summarizes the performance results on the 13 decoy sets. It is not clear to find the best reference state outperforming other reference states. In terms of Rank1, there is little difference on the performance. Each reference state shows strength on difference evaluation criteria as incorporated into ROTAS. The RAPDF reference state gives the best average Z-score whereas the DFIRE reference state shows the best average $\log P_{B10}$. The RW reference state shows the best performance on $\log P_{B1}$ and both correlation measures. Overall, the DFIRE and RW reference states are found to show better performance than other three reference states in ROTAS.

3.5 Conclusion

We have developed a new statistical potential energy function, named "rotamer-dependent atomic statistical potential" (ROTAS) that extends orientation-dependent atomic statistical potential (GOAP) by including the influence of local structural environments on the specificity of atomic interactions. The local structural environment is defined based on the rotameric state of residues, taking into account that different rotamer conformations may result in different solvent accessibility, electric polarization, and steric effects on residue atoms. The interaction between two atoms is specified by not only the distance and five angle parameters but also two state parameters which concern the rotameric state of the residues to which the interacting atoms belong. It has clearly found that the rotameric state is significantly related to the specificity of atomic interactions. Furthermore, such rotamer-dependencies are not limited to specific type or certain range of interactions.

The performance of ROTAS has been tested using various sets of decoys and compared to those of several existing all-atom statistical potentials which incorporate orientation-dependent energy terms. For a fair comparison, we implemented our own GOAP potential using the same structure database and techniques used for the construction of ROTAS. The results show that ROTAS performs better than other competing potentials not only in native structure recognition, but also in best model selection and correlation coefficients between energy and model quality. In particular, the relative improvement of ROTAS over GOAP indicates that the local structural environment can be incorporated for a fine-tuning of atomic-level statistical potentials. Furthermore, the rotameric state of residues may be used to describe the local environment of protein atoms. The effectiveness of ROTAS would provide insightful information for the

development of other applications which require accurate side-chain modeling such as homology modeling[69], protein design[70], mutation analysis[71], protein-protein docking[72] and flexible ligand docking[73].

Chapter 4: Side-chain modeling with ROTAS

4.1 Abstracts

Accurate modeling of side-chain conformations is an essential task in high-resolution refinement of protein structures. It aims to find the optimal rotamer combination from all possible combinations of side-chain rotamers for a given backbone structure. In this study, we have applied the ROTAS potential to side-chain modeling. A composite energy function which includes a modified Lennard Jones potential, rotamer-intrinsic energy terms and the ROTAS potential was devised. The weights of these energy terms were optimized to achieve the maximum number of correctly predicted rotamers on a training set of 50 protein structures. Our scoring function was combined with a Monte Carlo search algorithm to predict all the side-chains onto a backbone structure. In our benchmark testing, compared with the existing popular side-chain modeling programs such as SCWRL4, OPUS-Rota and OSCAR-star, ROTAS achieved comparable or even better prediction accuracy. In particular, our scoring function showed advantages in predicting the conformation of large amino acids including Glu, Gln, Met, Lys, Arg, His, Trp and Tyr. Considering that only a few rigid rotamers are included in this work without flexible rotamers, we still expect that there is still a room for improvement, which can be achieved by incorporating a more detailed rotamer library, (iterative) relaxation technique or flexible DEE.

4.2 Introduction

Prediction of side-chain conformations is an important step in protein structure prediction and protein design. In general, protein structure prediction can be decomposed into two levels of tasks with different resolution: the generation of native-like backbone structures and the prediction of side-chains on these backbones[174]. The functional annotation of protein structure often requires detailed information about the binding site of the target protein[175]. In particular, protein side-chains make dominant contribution to molecular recognition between proteins or protein and ligand molecules[72,73]. Thus accurate side-chain modeling is essential in high-resolution refinement of predicted structure models.

In side-chain modeling, there are three main components: scoring function, side-chain rotamer library and search algorithm, First, the scoring function is used to predict the free energy of side-chain conformations for a target backbone structure[116]. Second, for the efficient description of flexibility of amino acid residues, most side-chain modeling methods use a discrete set of statistically significant side-chain conformation, namely rotamer library[81]. Third, search algorithms are employed to find the optimal combination of rotameric states with minimum value of scoring function. Based on these components, side-chain modeling can be transformed into a combinatorial search problem which aims to find the optimal rotamer combination from all possible combinations of side-chain rotamers[127,176,177]. To develop a fast and accurate method for side-chain modeling, many efforts have been made by improving scoring functions[75,76,107,116,178,179], rotamer libraries[74,94,114,158] and search strategies[98,180–189].

We previously described a rotamer-dependent atomic-level statistical potential,which shows significant improvements in decoy discrimination tests compared to other existing energy

functions. In this work, we attempt to apply the ROTAS potential to side-chain modeling. In order to maximize the prediction accuracy of ROTAS for side-chain modeling, the ROTAS potentials was combined with other energy terms including a modified Lennard Jones potential and rotamer-intrinsic energy. The weights of these energy terms were optimized to achieve the maximum number of correctly predicted rotamers on a training set of 50 protein structures. Our scoring function was combined with a Monte Carlo search algorithm to predict all the side-chains onto a backbone structure. In our benchmark testing, compared with the existing popular side-chain modeling programs such as SCWRL4[190], OPUS-Rota[75] and OSCAR-star[76,191], ROTAS achieved comparable or even better prediction accuracy.

4.3 Materials and Methods

4.3.1 Rotamer library

In this work, we use Dunbrack backbone-dependent rotamer library which was constructed by employing adaptive kernel density estimation[158]. Standard bond lengths and angles was adopted from CHARMM27 parameters[6] to build rotamer conformations on backbone structures. Rotamers having a probability < 0.5% were excluded in this work.

4.3.2 Scoring function

The effective energy of the rotamer is computed as a linear combination of the following four energy terms:

$$E_{total} = E_{rotas} + W_{atr}E_{atr} + W_{rep}E_{rep} + E_{rot} \quad (1)$$

where E_{rotas} is the ROTAS potential described in Chapter 3. E_{atr} and E_{rep} are attractive term and repulsive term of van der Waals potential. It is well known that statistical distance-dependent

pairwise potentials lack an effective short-distance repulsive component[192]. It is therefore necessary to incorporate a second van der Waals energy function to avoid steric clash[193]. Here, we used a modified a modified 6-12 Lennard-Jones (LJ) potential:

$$E_{vdW} = min\left(10,\ e_{ij}\left\{\left(\frac{d_{ij}}{a_{ij}}\right)^{-12} - 2\left(\frac{d_{ij}}{a_{ij}}\right)^{-6}\right\}\right) \quad (2)$$

where$a_{ij} = a_i + a_j$, $e_{ij} = \sqrt{e_i e_j}$, a_i,$a_j$are atomic radii and e_i, e_j are well depths. Parameters for atomic radii and well depths are adopted from CHARMM27 parameter set[6,194]. Maximum repulsive energy is limited to 10 Kcal/mol considering that we use rigid-rotamer model. The LJ potential is calculated only within a distance of 9 Å. The repulsive energy term is sum of all positive van der Waals potential energies and the attractive energy term is sum of all negative van der Waals potentials for every possible atom pairs.

$E_{rot}$is the rotamer intrinsic energy tern, which measures the preference of the rotamers. For a rotamer R_i of residue i,

$$E_{rot} = -\sum_{i=1}^{N} \gamma_{AA} \log \frac{P(R_i|\phi_i, \psi_i, A_i)}{P(\max|\phi_i, \psi_i, A_i)} \quad (1)$$

where N is the number of residues and $P(R_i|\phi_i, \psi_i, A_i)$ is the probability of a rotamer R_i whose backbone dihedral angles and residue type are ϕ_i,$\psi_i$andA_i, respectively. It is normalized to the maximum probability $P(\max|\phi_i, \psi_i, A_i)$ of residue R_i in the same backbone dihedral angles. γ_{AA}isa residue-specific scaling factor.

The residue-specific scaling factor $\gamma_{AA}$and the weights for the energy terms (W_{atr}, W_{rep}) were determined by maximizing the sum of the following objective function against a training set of 50 protein structures:

$$\sum_{k=1}^{M} \frac{\sum_{i=1}^{N} \exp(-E(i))}{M \times \{N \cdot \exp(-E(k)) + \sum_{i=1}^{N} \exp(-E(i)\}} \quad (1)$$

where M is the total number of calculated residues, N is the number of rotamers in the rotamer library for residue k, E(k) is the energy of native side-chain conformation and E(i) is the energy of rotamer i. In this optimization procedure, only one residue was changed at a time and all other residues were kept in their native conformation. For the optimization algorithms, we employed a simulated annealing (SA)[144] followed by a sequential quadratic programming (SQP). In SA, the initial temperature is set to 100, and the temperature is updated by an annealing schedule factor of 0.95 with the re-annealing interval, 100. Starting from random reference energies and weights, at maximum, 100,000 iterations were tried, and the termination tolerance on the objective function value for both methods was set to 10^{-6}.

4.3.3 Search method

We adopted a simulated annealing (SA) method for searching rotamer conformation space. First, for given a backbone structure and sequence, initial rotamer conformations are randomly selected. Then, a rotamer substitution is made at a selected position. The probability of selecting a position is proportional to the number of rotamers for the residue position. A rotamer is randomly selected and evaluated by the derived scoring function. If the new energy score is lower than the previous energy, the move is accepted. Otherwise the move is accepted with the probability $\exp[(E_{old} - E_{new})/T]$. The initial temperature T is set to 100 and is scaled by

0.95after each cycle. A Total of 20 cycles are repeated. We hold the temperature constant at each cycle for 10,000 substitutions or 1,000 successful substitutions, whichever comes first.

4.3.4 Evaluation

The prediction accuracy is defined as the ratio of correctly predicted side-chain dihedral angles within a threshold of 40° compared with their native values. The accuracy of X_1 is defined as the ratio of residues whose predicted X_1dihedral is within 40° from the native value, the accuracy $X_{1+\ldots+N}$ is defined as the ratio of residue for which all $X_1, \ldots, X_N$ dihedrals are within the 40° of the native value. All residue types except Gly and Ala have at least one dihedral angle. In the evaluation, the symmetric terminal groups of Asp, Glu, Phe and Tyr were flipped to yield optimal atom matching.

We compared the performance of our method with several popular side-chain modeling programs: SCWRL4[190], OPUS-Rota[75] and OSCAR-star[76]. The binary programs for these methods were downloaded from the corresponding authors' websites.

4.3.5 Training and testing protein sets

We obtained a set of protein X-ray structures with a maximum R-factor of 0.25 and a resolution better than 2 Å from the protein sequence culling server, PISCES[139]. Also, protein chains were filtered out with a 40% sequence identity cutoff in order to have a set of non-homologous protein structures. A total 9321 protein structures were selected and downloaded from the Protein Data Bank (PDB)[140]. The program REDUCE[141] was used to optimize the flip states of Asn, Gln, and His in all protein structures. Residues with multiple side-chain conformations were modified such that only the side-chain conformations with atoms having the highest occupancy and/or lowest temperature factors were used.

The resulting list of proteins was separated into 3 sets: training set1, training set2 and testing set. The training set1 consisting of 9,221 structures was used for constructing the ROTAS potential, and the training set 2 was used for optimizing weight parameters. The testing set of 50 proteins was used only for side-chain prediction test. The separation was done in a random manner.

4.4 Results and Discussion

Table 4-1 summarizes the results of our benchmark test. In our comparison with other three side-chain modeling programs, ROTAS showed comparable or even better prediction accuracies than other programs. The overall X_1 accuracy of ROTAS is comparable to those of OPUS_Rota and SCWRL4. OSCAR-star shows the highest X_1 accuracy with a ratio of 0.872. However, it was found that, in case of theX_{1+2}, X_{1+2+3} and $X_{1+2+3+4}$ accuracies, ROTAS outperformed other methods. In terms of computational time, SCWRL4 shows the best performance because they employ a sophisticated graph-decomposition algorithm for conformational search. Other three methods commonly use the simulated annealing algorithm. It is also found that OSCAR_star minimizes atomic clashes. However, the differences between methods are not significant.

Table 4-1. Averaged prediction accuracies of side-chain modeling by different methods

	$\mathbf{X_1}$	$\mathbf{X_{1+2}}$	$\mathbf{X_{1+2+3}}$	$\mathbf{X_{1+2+3+4}}$	$\frac{\textbf{atomic clash}}{\textbf{residue}}$	**Total CPU time**
SCWRL4	0.851	0.741	0.436	0.334	2.671	**3 min.**
OPUS_Rota	0.856	0.742	0.432	0.327	2.688	8 min.
OSCAR_star	**0.872**	0.757	0.467	0.352	**2.663**	22 min.
ROTAS	0.850	**0.760**	**0.504**	**0.386**	2.668	10 min

Figure 4-1 shows the X_1 andX_{1+2} accuracies for different residue types. In general, the prediction accuracies for hydrophobic residues are much higher than those of hydrophilic residues. This is simply because hydrophobic residues are likely to be buried but hydrophilic residues are usually flexible surface residues. The accuracy of surface residues may be improved by considering crystal contacts[85,179] because protein-protein contacts would provide more information for selecting possible rotamers on surface residues.

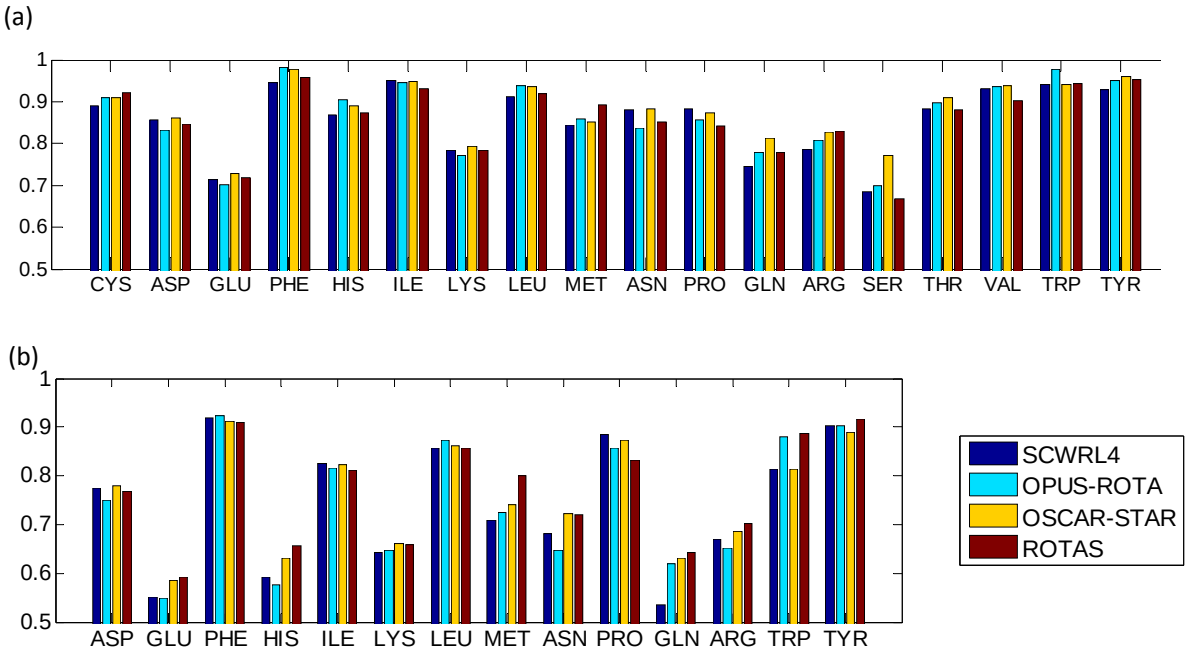

Figure 4-1. Prediction accuracy of 50 test proteins for different residue types. (a) X_1 accuracy and (b) X_{1+2} accuracy

Next, for each residue **t**ype, we considered that residues are correctly predicted only if their every side-chain dihedrals are within 40° from the native value. modeling of large amino acids.

Table 4-2 shows that ROTAS has strength in modeling large amino acids such as Glu, Gln, Met, Lys, Arg, His, Trp and Tyr. Such large amino acids have rather complex geometries than other

small amino acids such as Cys, Ser, Val or Thr. Depending on the rotameric states, their conformations are very different. This could be one reason that our ROTAS potential which takes into account the high order multibody effects shows strengths on modeling of large amino acids.

Table 4-2. Averaged prediction accuracy for different residue types when the residues are considered correctly predicted only if all side-chain dihedral angles are within 40° of their native values.

	# of dihedrals	# of residues	SCWRL4	OPUS_Rota	OSCAR_star	ROTAS
CYS	1	109	0.89	0.91	0.91	**0.92**
ASP	2	664	0.77	0.75	**0.78**	0.77
GLU	3	837	0.39	0.38	0.42	**0.46**
PHE	2	467	0.92	**0.92**	0.91	0.91
HIS	2	265	0.59	0.58	0.63	**0.66**
ILE	2	667	**0.83**	0.82	0.82	0.81
LYS	4	617	0.35	0.35	0.37	**0.39**
LEU	2	989	0.86	**0.87**	0.86	0.86
MET	3	269	0.54	0.56	0.61	**0.64**
ASN	2	488	0.68	0.65	**0.72**	0.72
PRO	2	533	**0.88**	0.86	0.87	0.83
GLN	3	465	0.37	0.39	0.40	**0.47**
ARG	4	582	0.32	0.30	0.33	**0.38**
SER	1	639	0.68	0.70	**0.77**	0.67
THR	1	606	0.88	0.90	**0.91**	0.88
VAL	1	806	0.93	0.94	**0.94**	0.90
TRP	2	167	0.81	0.88	0.81	**0.89**
TYR	2	416	0.90	0.90	0.89	**0.92**

Highlighted entries are the best ones in the respective residue type

4.5 Conclusion

We have applied the ROTAS potential to side-chain modeling. A composite energy function has been devised by including a few more energy terms and weight parameters. Our scoring function

was combined with a Monte Carlo search algorithm to predict all the side-chains onto a backbone structure. In our benchmark testing, compared with the existing popular side-chain modeling programs such as SCWRL4, OPUS-Rota and OSCAR-star, ROTAS achieved comparable or even better prediction accuracy. In particular, our scoring function shows advantages in predicting the conformation of large amino acids including Glu, Gln, Met, Lys, Arg, His, Trp and Tyr. Considering that only a few rigid rotamers are included in this work without flexible rotamers, we still expect that there is still a room for improvement which can be achieved by incorporating sophisticated search techniques extending large sub-rotamers, (iterative) relaxation technique or flexible DEE.

Chapter 5: Conclusions

This chapter concludes the dissertation. It includes a summary of the work, list of contributions and expected future extensions.

5.1 Dissertation Conclusion

The research in this dissertation focused on the design and development of statistical potentials that take into account the rotamer-dependence of interactions.First, a statistical analysis of the local environment of residues was introduced to find the rotamer-specific environmental features. Based on the observation that different rotameric states have distinguishing interaction patterns with surrounding residues, the obtained features were exploited to devise a scoring function for protein designs. The accuracy of the derived scoring function was compared to that of Rosetta energy function. It was found that the rotamer-specific environmental features could effectively improve the prediction accuracy.

The idea that the rotameric state of residues critically affects on the specificity of interactions within protein structures was applied to develop a rotamer-dependent atomic statistical potential (ROTAS) for protein structure prediction. The ROTAS potential extends orientation-dependent atomic potentials by including the influence of rotameric states of residues on the specificity of atomic interactions. Its performance was successfully demonstrated using various decoy sets. In a comparison with existing atomic potentials, ROTAS showed better performance not only in native structure recognition, but also in best model selection and correlation coefficients between

energy and model quality. In the next study, the ROTAS potential was applied to side-chain modeling. A composite scoring function with weight parameters was developed, and combined with a Monte Carlo search algorithm to predict all the side-chains onto a backbone structure. A benchmark testing showed that the developed method achieved comparable or even better prediction accuracy compared with the existing side-chain modeling programs.

5.2 Contributions

The main contribution of this dissertation is the incorporation of rotamer-dependence of interactions within protein structures into the design and development of statistical potentials. More specifically, its contribution is three-fold as follows:

- Introduction of rotamer-specific environmental features, which are exploited to devise a scoring function for protein design problem that aims to find an optimal combination of amino acid types and their rotameric states for a desired protein backbone structure
- Development of an rotamer-dependent atomic statistical potential which extends orientation-dependent atomic potentials by including the relationship between the rotameric states of residues and the specificity of atomic interactions for protein structure prediction
- Incorporation of the ROTAS potential into side-chain modeling for accurate prediction of side-chain conformations

5.3 Future Work

While there could be several extensions to the research presented in this dissertation, the following are perceived attractive:

- Developing a comprehensive descriptor for the environment of interacting residues or atoms. The environment of protein residues or atoms can be described in various ways. In addition to the rotameric state, local structural motif and/or secondary structure type may be integrated to define the local environment to describe interactions more accurately.
- Including the bond-related energies, solvation effect and entropic contribution in the present non-bonded potentials. The lack of bonding or torsional energies may lead to artificially lower scores for some distorted conformations. Also, solvation and entropy effects are not pairwise additive and thus the derived potentials do not account for these effects explicitly. Including the bond-related energies and the effects of solvation and entropy may improve the performance[195–197].
- Applying the same methodology described in this dissertation to derive scoring functions for other modeling tasks such as protein design, mutation analysis, protein-protein docking and flexible ligand docking. Since these modeling tasks usually require highly accurate side-chain modeling, the incorporation of the residue flexibility into scoring functions would improve the prediction accuracy.

Bibilography

1. Schlick T. Molecular Modeling and Simulation: An Interdisciplinary Guide (Interdisciplinary Applied Mathematics). Springer; 2002 p. 634. Available from: http://www.amazon.com/Molecular-Modeling-Simulation-Interdisciplinary-Mathematics/dp/038795404X

2. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins [Internet]. 1995 March [cited 2013 June 19];21(3):167–95. Available from: http://www.ncbi.nlm.nih.gov/pubmed/7784423

3. Senn HM, Thiel W. QM/MM methods for biomolecular systems. Angewandte Chemie - International Edition. 2009;48(7):1198–1229.

4. Weiner SJ, Kollman PA, Nguyen DT, Case DA. An all atom force field for simulations of proteins and nucleic acids. J. Comput. Chem. 1986;7:230–252.

5. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz Jr. KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. Journal of the American Chemical Society. 1995;117(19):5179–5197.

6. Brooks BR, Brooks III CL, Mackerell Jr. AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, et al. CHARMM: The biomolecular simulation program. Journal of Computational Chemistry. 2009;30(10):1545–1614.

7. Némethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, Scheraga HA. Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded

interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. Journal of Physical Chemistry. 1992;96(15):6472–6484.

8. Lindahl E, Hess B, van der Spoel D. GROMACS 3.0: A package for molecular simulation and trajectory analysis. Journal of Molecular Modeling. 2001;7(8):306–317.

9. Jorgensen WL, Tirado-Rives J. The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. Journal of the American Chemical Society [Internet]. 1988 March [cited 2013 June 14];110(6):1657–1666. Available from: http://dx.doi.org/10.1021/ja00214a001

10. Tanaka S, Scheraga HA. Medium- and Long-Range Interaction Parameters between Amino Acids for Predicting Three-Dimensional Structures of Proteins. Macromolecules [Internet]. 1976 November [cited 2013 June 2];9(6):945–950. Available from: http://dx.doi.org/10.1021/ma60054a013

11. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules [Internet]. 1985 May [cited 2013 June 2];18(3):534–552. Available from: http://dx.doi.org/10.1021/ma00145a039

12. Sippl MJ. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. Journal of molecular biology [Internet]. 1990 June 20 [cited 2013 June 2];213(4):859–83. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2359125

13. Sippl MJ. Knowledge-based potentials for proteins. Current Opinion in Structural Biology [Internet]. 1995 April [cited 2013 June 2];5(2):229–235. Available from: http://dx.doi.org/10.1016/0959-440X(95)80081-6

14. Wodak SJ, Rooman MJ. Generating and testing protein folds. Current Opinion in Structural Biology [Internet]. 1993 April [cited 2013 June 2];3(2):247–259. Available from: http://dx.doi.org/10.1016/S0959-440X(05)80160-5

15. Koppensteiner WA, Sippl MJ. Knowledge-based potentials - Back to the roots. Biochemistry (Moscow). 1998;63(3):247–252.

16. Thomas PD, Dill KA. An iterative method for extracting energy-like quantities from protein structures. Proceedings of the National Academy of Sciences of the United States of America. 1996;93(21):11628–11633.

17. McQuarrie DA. No Title. Statistical Mechanics. 1976.

18. Thomas PD, Dill KA. Statistical potentials extracted from protein structures: how accurate are they? Journal of molecular biology [Internet]. 1996 March 29 [cited 2013 May 24];257(2):457–69. Available from: http://dx.doi.org/10.1006/jmbi.1996.0175

19. Ben-Naim A. Statistical potentials extracted from protein structures: Are these meaningful potentials? The Journal of Chemical Physics [Internet]. 1997 September 1 [cited 2013 June 2];107(9):3698. Available from: http://link.aip.org/link/?JCPSA6/107/3698/1

20. Hamelryck T, Borg M, Paluszewski M, Paulsen J, Frellsen J, Andreetta C, Boomsma W, Bottaro S, Ferkinghoff-Borg J. Potentials of mean force for protein structure prediction vindicated, formalized and generalized. PloS one [Internet]. 2010 January [cited 2013 June 2];5(11):e13714. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2978081&tool=pmcentrez&rendertype=abstract

21. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature [Internet]. 1992 July 2 [cited 2013 June 2];358(6381):86–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1614539

22. Buchete N-V, Straub JE, Thirumalai D. Development of novel statistical potentials for protein fold recognition. Current opinion in structural biology [Internet]. 2004 April [cited 2013 June 1];14(2):225–32. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15093838

23. Sánchez R, Sali A. Advances in comparative protein-structure modelling. Current opinion in structural biology [Internet]. 1997 April [cited 2013 June 2];7(2):206–14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9094331

24. Miyazawa S, Jernigan RL. An empirical energy potential with a reference state for protein fold and sequence recognition. Proteins [Internet]. 1999 August 15 [cited 2013 June 2];36(3):357–69. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10409829

25. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. Journal of molecular biology [Internet]. 1997 April 25 [cited 2013 June 2];268(1):209–25. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9149153

26. Skolnick J. In quest of an empirical potential for protein structure prediction. Current opinion in structural biology [Internet]. 2006 April [cited 2013 June 1];16(2):166–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16524716

27. Li Y, Zhang Y. REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. Proteins [Internet]. 2009 August 15 [cited 2013 June 2];76(3):665–76. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2771173&tool=pmcentrez&rendertype=abstract

28. Boas FE, Harbury PB. Potential energy functions for protein design. Current Opinion in Structural Biology [Internet]. 2007;17(2):199–204. Available from: <Go to ISI>://WOS:000246330900009

29. Gordon DB, Marshall SA, Mayo SL. Energy functions for protein design. Current Opinion in Structural Biology. 1999;9(4):509–513.

30. Gilis D, Rooman M. Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. Journal of molecular biology [Internet]. 1997 September 19

[cited 2013 June 2];272(2):276–90. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9299354

31. Turjanski AG, Gutkind JS, Best RB, Hummer G. Binding-induced folding of a natively unstructured transcription factor. Shakhnovich EI, editor. PLoS computational biology [Internet]. 2008 April [cited 2013 June 2];4(4):e1000060. Available from: http://dx.plos.org/10.1371/journal.pcbi.1000060

32. Su Y, Zhou A, Xia X, Li W, Sun Z. Quantitative prediction of protein-protein binding affinity with a potential of mean force considering volume correction. Protein science : a publication of the Protein Society [Internet]. 2009 December [cited 2013 June 2];18(12):2550–8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2821273&tool=pmcentrez&rendertype=abstract

33. Bereau T, Deserno M. Generic coarse-grained model for protein folding and aggregation. The Journal of chemical physics [Internet]. 2009 June 21 [cited 2013 June 2];130(23):235106. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19548767

34. Chen WW, Shakhnovich EI. Lessons from the design of a novel atomic potential for protein folding. Protein science : a publication of the Protein Society [Internet]. 2005 July [cited 2013 June 2];14(7):1741–52. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2253347&tool=pmcentrez&rendertype=abstract

35. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein science : a publication of the Protein Society [Internet]. 2002 November [cited 2013 June 2];11(11):2714–26. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2373736&tool=pmcentrez&rendertype=abstract

36. Shen M-Y, Sali A. Statistical potential for assessment and prediction of protein structures. Protein science : a publication of the Protein Society [Internet]. 2006 November [cited 2013 May 25];15(11):2507–24. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2242414&tool=pmcentrez&rendertype=abstract

37. Buchete N-V, Straub JE, Thirumalai D. Orientational potentials extracted from protein structures improve native fold recognition. Protein science : a publication of the Protein Society [Internet]. 2004 April [cited 2013 June 2];13(4):862–74. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2280067&tool=pmcentrez&rendertype=abstract

38. Mayewski S. A multibody, whole-residue potential for protein structures, with testing by Monte Carlo simulated annealing. Proteins [Internet]. 2005 May 1 [cited 2013 June 2];59(2):152–69. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15723360

39. Munson PJ, Singh RK. Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. Protein science : a publication of the Protein Society [Internet]. 1997 July [cited 2013 June 2];6(7):1467–81. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2143734&tool=pmcentrez&rendertype=abstract

40. Wu Y, Lu M, Chen M, Li J, Ma J. OPUS-Ca: a knowledge-based potential function requiring only Calpha positions. Protein science : a publication of the Protein Society [Internet]. 2007 July [cited 2013 June 2];16(7):1449–63. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2206690&tool=pmcentrez&rendertype=abstract

41. Deane CM, Blundell TL. A novel exhaustive search algorithm for predicting the conformation of polypeptide segments in proteins. Proteins: Structure, Function and Genetics. 2000;40(1):135–144.

42. De Brevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins: Structure, Function and Genetics. 2000;41(3):271–287.

43. De Brevern AG, Valadié H, Hazout S, Etchebest C. Extension of a local backbone description using a structural alphabet: A new approach to the sequence-structure relationship. Protein Science. 2002;11(12):2871–2886.

44. Figureau A, Soto MA, Tohá J. A pentapeptide-based method for protein secondary structure prediction. Protein Engineering. 2003;16(2):103–107.

45. Fernández A, Sosnick TR, Colubri A. Dynamics of hydrogen bond desolvation in protein folding. Journal of Molecular Biology. 2002;321(4):659–675.

46. Kolinski A, Skolnick J. Discretized model of proteins. I. Monte Carlo study of cooperativity in homopolypeptides. The Journal of Chemical Physics. 1992;97(12):9412–9426.

47. Jonassen I, Eidhammer I, Conklin D, Taylor WR. Structure motif discovery and mining the PDB. Bioinformatics. 2002;18(2):362–367.

48. Karlin S, Zhu Z-Y. Characterizations of diverse residue clusters in protein three-dimensional structures. Proceedings of the National Academy of Sciences of the United States of America. 1996;93(16):8344–8349.

49. Zhu Z-Y, Karlin S. Clusters of charged residues in protein three-dimensional structures. Proceedings of the National Academy of Sciences of the United States of America. 1996;93(16):8350–8355.

50. Jonassen I, Eidhammer I, Taylor WR. Discovery of local packing motifs in protein structures. Proteins: Structure, Function and Genetics. 1999;34(2):206–219.

51. Munson PJ, Singh RK. Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence-structure alignment. Protein Science. 1997;6(7):1467–1481.

52. Singh RK. Delaunay tessellation of proteins: Four body nearest-neighbor propensities of amino acid residues. Journal of Computational Biology. 1996;3(2):213–221.

53. Zhang C, Kim SH. Environment-dependent residue contact energies for proteins. Proceedings of the National Academy of Sciences of the United States of America [Internet]. 2000 March 14 [cited 2013 June 3];97(6):2550–5. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=15966&tool=pmcentrez&rendertype=abstract

54. Benkert P, Tosatto SCE, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins: Structure, Function and Genetics. 2008;71(1):261–277.

55. Zhou H, Skolnick J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophysical journal [Internet]. 2011 October 19 [cited 2013 June 2];101(8):2043–52. Available from: http://dx.doi.org/10.1016/j.bpj.2011.09.012

56. Yang Y, Zhou Y. Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins [Internet]. 2008 August [cited 2013 June 2];72(2):793–803. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18260109

57. Zhang J, Zhang Y. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PloS one [Internet]. 2010 January 27 [cited 2013 June 2];5(10):e15386. Available from: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0015386#pone.0015386-Lu2

58. Lu M, Dousis AD, Ma J. OPUS-PSP: An Orientation-dependent Statistical All-atom Potential Derived from Side-chain Packing. Journal of Molecular Biology. 2008;376(1):288–301.

59. Bordner AJ. Orientation-dependent backbone-only residue pair scoring functions for fixed backbone protein design. BMC bioinformatics [Internet]. 2010 January [cited 2013 June 3];11(1):192. Available from: http://www.biomedcentral.com/1471-2105/11/192

60. Miyazawa S, Jernigan RL. How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins? The Journal of chemical physics [Internet]. 2005 January 8 [cited 2013 June 1];122(2):024901. Available from: http://link.aip.org.proxy.lib.umich.edu/link/?JCPSA6/122/024901/1

61. Janin J, Wodak S, Levitt M, Maigret B. CONFORMATION OF AMINO-ACID SIDE-CHAINS IN PROTEINS. Journal of Molecular Biology [Internet]. 1978;125(3):357–386. Available from: <Go to ISI>://WOS:A1978FY32300007

62. Ponder JW, Richards FM. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. Journal of molecular biology [Internet]. 1987 February 20 [cited 2011 May 16];193(4):775–91. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2441069

63. Schrauber H, Eisenhaber F, Argos P. Rotamers: to be or not to be? An analysis of amino acid side-chain conformations in globular proteins. Journal of molecular biology [Internet]. 1993 March [cited 2011 May 16];230(2):592–612. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8464066

64. Dunbrack RL, Karplus M. CONFORMATIONAL-ANALYSIS OF THE BACKBONE-DEPENDENT ROTAMER PREFERENCES OF PROTEIN SIDE-CHAINS. Nature Structural Biology [Internet]. 1994;1(5):334–340. Available from: <Go to ISI>://A1994PH94800015

65. Halgren TA, Damm W. Polarizable force fields. Current Opinion in Structural Biology [Internet]. 2001 April [cited 2013 June 20];11(2):236–242. Available from: http://dx.doi.org/10.1016/S0959-440X(00)00196-2

66. Lamoureux G, Roux B. Modeling induced polarization with classical Drude oscillators: Theory and molecular dynamics simulation algorithm. Journal of Chemical Physics. 2003;119(6):3025–3039.

67. Friesner RA. Modeling Polarization in Proteins and Protein-ligand Complexes: Methods and Preliminary Results. Advances in protein chemistry [Internet]. 2005 January [cited 2013 June 5];72(null):79–104. Available from: http://dx.doi.org/10.1016/S0065-3233(05)72003-9

68. Lu M, Dousis AD, Ma J. OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. Journal of molecular biology [Internet]. 2008 February 8 [cited 2013 June 2];376(1):288–301. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2669442&tool=pmcentrez&rendertype=abstract

69. Fiser A, Feig M, Brooks CL, Sali A. Evolution and physics in comparative protein structure modeling. Accounts of chemical research [Internet]. 2002 June [cited 2013 June 21];35(6):413–21. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12069626

70. Dahiyat BI, Sarisky CA, Mayo SL. De novo protein design: Towards fully automated sequence selection. Journal of Molecular Biology. 1997;273(4):789–796.

71. Feyfant E, Sali A, Fiser A. Modeling mutations in protein structures. Protein science : a publication of the Protein Society [Internet]. 2007 September [cited 2013 June 21];16(9):2030–41. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2206969&tool=pmcentrez&rendertype=abstract

72. Wang C, Schueler-Furman O, Baker D. Improved side-chain modeling for protein-protein docking. Protein Science [Internet]. 2005;14(5):1328–1339. Available from: <Go to ISI>://000228594900021

73. Sousa SF, Fernandes PA, Ramos MJ. Protein-ligand docking: current status and future challenges. Proteins [Internet]. 2006 October 1 [cited 2013 June 11];65(1):15–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16862531

74. Krivov GG, Shapovalov M V, Dunbrack RL. Improved prediction of protein side-chain conformations with SCWRL4. Proteins [Internet]. 2009 December [cited 2011 July 6];77(4):778–95. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2885146&tool=pmcentrez&rendertype=abstract

75. Lu M, Dousis AD, Ma J. OPUS-Rota: A fast and accurate method for side-chain modeling. Protein Science. 2008;17(9):1576–1585.

76. Liang S, Zheng D, Zhang C, Standley DM. Fast and accurate prediction of protein side-chain conformations. Bioinformatics [Internet]. 2011;27(20):2913–2914. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-80053986637&partnerID=40&md5=f914c52e474c2205ed6f846d80abf43c

77. Voigt CA, Gordon DB, Mayo SL. Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. Journal of Molecular Biology [Internet]. 2000;299(3):789–803. Available from: <Go to ISI>://WOS:000087540100018

78. Desjarlais JR, Clarke ND. Computer search algorithms in protein modification and design. Current Opinion in Structural Biology [Internet]. 1998;8(4):471–475. Available from: <Go to ISI>://WOS:000075505000010

79. Pokala N, Handel TM. Review: Protein Design--Where We Were, Where We Are, Where We're Going. Journal of Structural Biology [Internet]. 2001;134(2-3):269–281. Available from: http://www.sciencedirect.com/science/article/B6WM5-457V5F3-2V/2/f763480a4c0197c66a4a66af214e6a9d

80. Samish I, MacDermaid CM, Perez-Aguilar JM, Saven JG. Theoretical and computational protein design. Annual review of physical chemistry [Internet]. 2011 May 31 [cited 2011 July

19];62:129–49. Available from: http://www.annualreviews.org/doi/abs/10.1146/annurev-physchem-032210-103509?journalCode=physchem

81. Dunbrack RL. Rotamer libraries in the 21(st) century. Current Opinion in Structural Biology [Internet]. 2002;12(4):431–440. Available from: <Go to ISI>://WOS:000177571100002

82. Street AG, Mayo SL. Computational protein design. Structure. 1999;7(5).

83. Gordon DB, Marshall SA, Mayo SL. Energy functions for protein design. Current Opinion in Structural Biology. 1999;9(4):509–513.

84. Pierce NA, Winfree E. Protein Design is NP-hard. Protein Engineering Design and Selection [Internet]. 2002 October 1 [cited 2011 August 10];15(10):779–782. Available from: http://peds.oxfordjournals.org/cgi/content/abstract/15/10/779

85. Holm L, Sander C. Fast and simple Monte Carlo algorithm for side chain optimization in proteins: Application to model building by homology. Proteins: Structure, Function and Genetics. 1992;14(2):213–223.

86. Hellinga HW, Richards FM. Optimal sequence selection in proteins of known structure by simulated evolution. Proceedings of the National Academy of Sciences of the United States of America. 1994;91(13):5803–5807.

87. Godzik A. In search of the ideal protein sequence. Protein Engineering. 1995;8(5):409–416.

88. Sasai M. Conformation, energy, and folding ability of selected amino acid sequences. Proceedings of the National Academy of Sciences of the United States of America. 1995;92(18):8438–8442.

89. Dahiyat BI, Mayo SL. Protein design automation. Protein Science. 1996;5(5):895–903.

90. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. Proceedings of the National Academy of Sciences of the United States of America [Internet]. 2000;97(19):10383–10388. Available from: <Go to ISI>://WOS:000089341400019

91. Jones DT. De novo protein design using pairwise potentials and a genetic algorithm. Protein Science. 1994;3(4):567–574.

92. Desjarlais JR, Handel TM. Side-chain and backbone flexibility in protein core design. Journal of Molecular Biology [Internet]. 1999;290(1):305–318. Available from: http://www.sciencedirect.com/science/article/B6WK7-45R87DR-F3/2/cff4499c0f376fee5a90b82b86baef20

93. Allen BD, Mayo SL. Dramatic performance enhancements for the FASTER optimization algorithm. Journal of computational chemistry [Internet]. 2006 July 30 [cited 2011 November 9];27(10):1071–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16685715

94. De Maeyer M, Desmet J, Lasters I. All in one: A highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. Folding and Design. 1997;2(1):53–66.

95. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. Science (New York, N.Y.) [Internet]. 2003 November 21 [cited 2011 June 11];302(5649):1364–8. Available from: http://www.sciencemag.org.proxy.lib.umich.edu/content/302/5649/1364.full

96. Dantas G, Kuhlman B, Callender D, Wong M, Baker D. A Large Scale Test of Computational Protein Design: Folding and Stability of Nine Completely Redesigned Globular Proteins. Journal of Molecular Biology [Internet]. 2003 September 12 [cited 2011 July 6];332(2):449–460. Available from: http://dx.doi.org/10.1016/S0022-2836(03)00888-X

97. Koehl P, Delarue M. A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. Nature Structural Biology. 1995;2(2):163–170.

98. Koehl P, Delarue M. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. Journal of Molecular Biology. 1994;239(2):249–275.

99. Koehl P, Delarue M. Mean-field minimization methods for biological macromolecules. Current Opinion in Structural Biology. 1996;6(2):222–226.

100. Lee C. Predicting protein mutant energetics by self-consistent ensemble optimization. Journal of Molecular Biology. 1994;236(3):918–939.

101. Vasquez M. An evaluation of discrete and continuum search techniques for conformational analysis of side chains in proteins. Biopolymers. 1995;36(1):53–70.

102. Voigt CA, Gordon DB, Mayo SL. Trading accuracy for speed: A quantitative comparison of search algorithms in protein sequence design. Journal of Molecular Biology. 2000;299(3):789–803.

103. Lippow SM, Tidor B. Progress in computational protein design. Current Opinion in Biotechnology [Internet]. 2007;18(4):305–311. Available from: http://www.sciencedirect.com/science/article/B6VRV-4P7FHHX-1/2/9219002490e0cf85261c4f242b68915b

104. Tian P. Computational protein design, from single domain soluble proteins to membrane proteins. Chemical Society reviews [Internet]. 2010 June 20 [cited 2011 November 8];39(6):2071–82. Available from: http://pubs.rsc.org/en/content/articlehtml/2010/cs/b810924a

105. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nature reviews. Drug discovery [Internet]. 2004 November [cited 2011 July 20];3(11):935–49. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15520816

106. Lengauer T, Rarey M. Computational methods for biomolecular docking. Current Opinion in Structural Biology [Internet]. 1996 June [cited 2011 November 10];6(3):402–406. Available from: http://dx.doi.org/10.1016/S0959-440X(96)80061-3

107. Peterson RW, Dutton PL, Wand AJ. Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. Protein Science. 2004;13(3):735–751.

108. Wu Y, Lu M, Chen M, Li J, Ma J. OPUS-Ca: A knowledge-based potential function requiring only Cα positions. Protein Science. 2007;16(7):1449–1463.

109. Tozzini V. Coarse-grained models for proteins. Current opinion in structural biology [Internet]. 2005 April [cited 2013 May 23];15(2):144–50. Available from: http://dx.doi.org/10.1016/j.sbi.2005.02.005

110. Levitt M, Gerstein M, Huang E, Subbiah S, Tsai J. Protein folding: The endgame. Annual Review of Biochemistry [Internet]. 1997;66:549–579. Available from: <Go to ISI>://WOS:A1997XH20100019

111. Desjarlais JR, Handel TM. De novo design of the hydrophobic cores of proteins. Protein Science. 1995;4(10):2006–2018.

112. Dahiyat BI, Mayo SL. Probing the role of packing specificity in protein design. Proceedings of the National Academy of Sciences of the United States of America. 1997;94(19):10172–10177.

113. Lazar GA, Desjarlais JR, Handel TM. De novo design of the hydrophobic core of ubiquitin. Protein Science. 1997;6(6):1167–1178.

114. Xiang Z, Honig B. Extending the accuracy limits of prediction for side-chain conformations. Journal of Molecular Biology [Internet]. 2001;311(2):421–430. Available from: http://www.sciencedirect.com/science/article/B6WK7-457CY84-C0/2/e960222a13a120f29bf4cbe25b2ece75

115. Bower MJ, Cohen FE, Dunbrack RL. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. Journal of molecular

biology [Internet]. 1997 April 18 [cited 2011 November 9];267(5):1268–82. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9150411

116. Liang S, Grishin N V. Side-chain modeling with an optimized scoring function. Protein science : a publication of the Protein Society [Internet]. 2002 February [cited 2011 May 23];11(2):322–31. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2373451&tool=pmcentrez&rendertype=abstract

117. Goldstein RF. Efficient rotamer elimination applied to protein side-chains and related spin glasses. Biophysical Journal. 1994;66(5):1335–1340.

118. Gordon DB, Hom GK, Mayo SL, Pierce NA. Exact rotamer optimization for protein design. Journal of Computational Chemistry. 2003;24(2):232–243.

119. Dill KA. Dominant forces in protein folding. Biochemistry [Internet]. 1990 August [cited 2011 July 27];29(31):7133–7155. Available from: http://dx.doi.org/10.1021/bi00483a001

120. Oobatake M, Ooi T. Hydration and heat stability effects on protein unfolding. Progress in biophysics and molecular biology [Internet]. 1993 January [cited 2011 November 11];59(3):237–84. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8441810

121. CHOTHIA C. Hydrophobic bonding and accessible surface area in proteins. Nature [Internet]. 1974 March 22 [cited 2011 November 11];248(5446):338–339. Available from: http://dx.doi.org/10.1038/248338a0

122. Nagano K. Logical analysis of the mechanism of protein folding. I. Predictions of helices, loops and beta-structures from primary structure. Journal of molecular biology [Internet]. 1973 April 5 [cited 2011 November 11];75(2):401–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/4728695

123. Chou PY, Fasman GD. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry [Internet]. 1974 January 15

[cited 2011 November 11];13(2):211–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/4358939

124. Tanaka S, Scheraga HA. Statistical mechanical treatment of protein conformation. I. Conformational properties of amino acids in proteins. Macromolecules [Internet]. [cited 2011 November 11];9(1):142–59. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1249985

125. Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. Advances in enzymology and related areas of molecular biology [Internet]. 1978 January [cited 2011 November 11];47:45–148. Available from: http://www.ncbi.nlm.nih.gov/pubmed/364941

126. Levitt M. Conformational preferences of amino acids in globular proteins. Biochemistry [Internet]. 1978 October 3 [cited 2011 November 11];17(20):4277–85. Available from: http://www.ncbi.nlm.nih.gov/pubmed/708713

127. Dunbrack Jr RL, Karplus M. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. Journal of Molecular Biology. 1993;230(2):543–574.

128. Dunbrack Jr RL, Cohen FE. Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Science. 1997;6(8):1661–1681.

129. Dunbrack RL, Karplus M. Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains. Nature Structural Biology [Internet]. 1994 May [cited 2011 May 16];1(5):334–340. Available from: http://dx.doi.org/10.1038/nsb0594-334

130. Chakrabarti P, Pal D. The interrelationships of side-chain and main-chain conformations in proteins. Progress in biophysics and molecular biology [Internet]. 2001 January [cited 2011 November 9];76(1-2):1–102. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11389934

131. Chakrabarti P, Pal D. Main-chain conformational features at different conformations of the side-chains in proteins. Protein engineering [Internet]. 1998 August [cited 2011 November 9];11(8):631–47. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9749916

132. Ogata K, Umeyama H. The role played by environmental residues on sidechain torsional angles within homologous families of proteins: a new method of sidechain modeling. Proteins [Internet]. 1998 June 1 [cited 2011 November 9];31(4):355–69. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9626696

133. Summers NL, Carlson WD, Karplus M. Analysis of side-chain orientations in homologous proteins. Journal of molecular biology [Internet]. 1987 July 5 [cited 2011 November 9];196(1):175–98. Available from: http://www.ncbi.nlm.nih.gov/pubmed/3309349

134. Smith RE, Lovell SC, Burke DF, Montalvao RW, Blundell TL. Andante: reducing side-chain rotamer search space during comparative modeling using environment-specific substitution probabilities. Bioinformatics (Oxford, England) [Internet]. 2007 May 1 [cited 2011 August 3];23(9):1099–105. Available from: http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/9/1099

135. Sutcliffe MJ, Hayes FRF, Blundell TL. Knowledge based modelling of homologous proteins, part II: rules for the conformations of substituted sidechains. "Protein Engineering, Design and Selection" [Internet]. 1987 October [cited 2011 November 9];1(5):385–392. Available from: http://peds.oxfordjournals.org/cgi/content/abstract/1/5/385

136. Rice DW, Eisenberg D. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. Journal of molecular biology [Internet]. 1997 April 11 [cited 2011 November 9];267(4):1026–38. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9135128

137. Wako H, Blundell TL. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. II. Secondary structures. Journal of molecular biology [Internet]. 1994 May 20 [cited 2011 November 1];238(5):693–708. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8182744

138. Johnson MS, Overington JP, Blundell TL. Alignment and searching for common protein folds using a data bank of structural templates. Journal of molecular biology [Internet]. 1993 June 5 [cited 2011 November 9];231(3):735–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8515448

139. Wang G, Dunbrack RL. PISCES: a protein sequence culling server. Bioinformatics [Internet]. 2003 August 11 [cited 2011 July 15];19(12):1589–1591. Available from: http://bioinformatics.oxfordjournals.org/cgi/content/abstract/19/12/1589

140. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, et al. The Protein Data Bank. Acta crystallographica. Section D, Biological crystallography [Internet]. 2002 June [cited 2011 July 20];58(Pt 6 No 1):899–907. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12037327

141. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. Journal of molecular biology [Internet]. 1999 January 29 [cited 2012 March 21];285(4):1735–47. Available from: http://dx.doi.org/10.1006/jmbi.1998.2401

142. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins [Internet]. 2000 August 15 [cited 2011 July 20];40(3):389–408. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10861930

143. Murphy LR, Wallqvist A, Levy RM. Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein engineering [Internet]. 2000 March [cited 2011 August 10];13(3):149–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10775656

144. Ingber L. Adaptive simulated annealing (ASA): Lessons learned. CONTROL AND CYBERNETICS [Internet]. 1996 [cited 2012 March 21];25:33 – 54. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.15.2777

145. Petrella RJ, Lazaridis T, Karplus M. Protein sidechain conformer prediction: a test of the energy function. Folding and Design [Internet]. 1998 October [cited 2011 November 8];3(5):353–377. Available from: http://dx.doi.org/10.1016/S1359-0278(98)00050-9

146. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins [Internet]. 1995 December [cited 2011 July 22];23(4):566–79. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8749853

147. Rykunov D, Fiser A. New statistical potential for quality assessment of protein models and a survey of energy functions. BMC bioinformatics [Internet]. 2010 January [cited 2013 June 2];11:128. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2853469&tool=pmcentrez&rendertype=abstract

148. Bowman GR, Pande VS. Simulated tempering yields insight into the low-resolution Rosetta scoring functions. Proteins [Internet]. 2009 February 15 [cited 2013 June 2];74(3):777–88. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18767152

149. Shmygelska A, Levitt M. Generalized ensemble methods for de novo structure prediction. Proceedings of the National Academy of Sciences of the United States of America [Internet]. 2009 February 3 [cited 2013 June 2];106(5):1415–20. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2631076&tool=pmcentrez&rendertype=abstract

150. Khatun J, Khare SD, Dokholyan N V. Can contact potentials reliably predict stability of proteins? Journal of molecular biology [Internet]. 2004 March 5 [cited 2013 June 1];336(5):1223–38. Available from: http://dx.doi.org/10.1016/j.jmb.2004.01.002

151. Vendruscolo M, Najmanovich R, Domany E. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? Proteins [Internet]. 2000 February 1 [cited 2013 June 2];38(2):134–48. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10656261

152. Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein science : a publication of the Protein Society [Internet]. 1999 February [cited 2013 June 2];8(2):361–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2144252&tool=pmcentrez&rendertype=abstract

153. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. Journal of molecular biology [Internet]. 1998 February 6 [cited 2013 June 2];275(5):895–916. Available from: http://dx.doi.org/10.1006/jmbi.1997.1479

154. McConkey BJ, Sobolev V, Edelman M. Discrimination of native protein structures using atom-atom contact scoring. Proceedings of the National Academy of Sciences of the United States of America [Internet]. 2003 March 18 [cited 2013 June 2];100(6):3215–20. Available from: http://www.pnas.org/content/100/6/3215.short

155. Bagci Z, Kloczkowski A, Jernigan RL, Bahar I. The origin and extent of coarse-grained regularities in protein internal packing. Proteins [Internet]. 2003 October 1 [cited 2013 June 2];53(1):56–67. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12945049

156. POHL FM. Empirical Protein Energy Maps. Nature [Internet]. 1971 December 29 [cited 2013 June 2];234(52):277–279. Available from: http://www.nature.com/nature-newbio/journal/v234/n52/abs/newbio234277a0.html

157. Fang Q, Shortle D. A consistent set of statistical potentials for quantifying local side-chain and backbone interactions. Proteins [Internet]. 2005 July 1 [cited 2013 June 2];60(1):90–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15852305

158. Shapovalov M V, Dunbrack RL. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure (London, England : 1993) [Internet]. 2011 June 8 [cited 2013 May 28];19(6):844–58. Available from:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3118414&tool=pmcentrez&rendertype=abstract

159. Kortemme T, Morozov A V, Baker D. 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):1239–1259.

160. Yang Y, Zhou Y. Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. Protein science : a publication of the Protein Society [Internet]. 2008 July [cited 2013 June 4];17(7):1212–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2442011&tool=pmcentrez&rendertype=abstract

161. Lu H, Skolnick J. A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins [Internet]. 2001 August 15 [cited 2013 June 5];44(3):223–32. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11455595

162. Park B, Levitt M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. Journal of Molecular Biology. 1996;258(2):367–392.

163. Keasar C, Levitt M. A novel approach to decoy set generation: Designing a physical energy function having local minima with native structure characteristics. Journal of Molecular Biology. 2003;329(1):159–174.

164. Xia Y, Huang ES, Levitt M, Samudrala R. Ab initio construction of protein tertiary structures using a hierarchical approach. Journal of Molecular Biology. 2000;300(1):171–185.

165. John B, Sali A. Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Research. 2003;31(14):3982–3992.

166. Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read RJ, Baker D. High-resolution structure prediction and the crystallographic phase problem. Nature. 2007;450(7167):259–264.

167. Wroblewska L, Skolnick J. Can a physics-based, all-atom potential find a protein's native structure among misfolded structures? I. Large scale AMBER benchmarking. Journal of computational chemistry [Internet]. 2007 September [cited 2013 June 6];28(12):2059–66. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17407093

168. Samudrala R, Levitt M. Decoys "R" Us: A database of incorrect conformations to improve protein structure prediction. Protein Science. 2000;9(7):1399–1401.

169. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins [Internet]. 2004 December 1 [cited 2013 May 27];57(4):702–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15476259

170. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score = 0.5? Bioinformatics (Oxford, England) [Internet]. 2010 April 1 [cited 2013 May 27];26(7):889–95. Available from: http://bioinformatics.oxfordjournals.org/content/26/7/889.short

171. Cossio P, Granata D, Laio A, Seno F, Trovato A. A simple and efficient statistical potential for scoring ensembles of protein structures. Scientific Reports [Internet]. 2012 April 3 [cited 2013 May 21];2. Available from: http://www.nature.com/srep/2012/120403/srep00351/full/srep00351.html

172. Bogatyreva NS, Finkelstein A V. Cunning simplicity of protein folding landscapes. Protein Engineering Design and Selection [Internet]. 2001 August 1 [cited 2013 May 31];14(8):521–523. Available from: http://peds.oxfordjournals.org.proxy.lib.umich.edu/content/14/8/521.short

173. Ruvinsky AM, Vakser IA. Interaction cutoff effect on ruggedness of protein-protein energy landscape. Proteins [Internet]. 2008 March [cited 2013 May 30];70(4):1498–505. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17910068

174. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nature protocols. 2010;5(4):725–738.

175. Campbell SJ, Gold ND, Jackson RM, Westhead DR. Ligand binding: functional site location, similarity and docking. Current Opinion in Structural Biology [Internet]. 2003 June [cited 2013 May 21];13(3):389–395. Available from: http://dx.doi.org/10.1016/S0959-440X(03)00075-7

176. Eisenmenger F, Argos P, Abagyan R. A method to configure protein side-chains from the main-chain trace in homology modelling. Journal of molecular biology [Internet]. 1993 June 5 [cited 2013 June 25];231(3):849–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8515455

177. Lee C, Subbiah S. Prediction of protein side-chain conformation by packing optimization. Journal of Molecular Biology. 1991;217(2):373–388.

178. Eyal E, Najmanovich R, Mcconkey BJ, Edelman M, Sobolev V. Importance of Solvent Accessibility and Contact Surfaces in Modeling Side-Chain Conformations in Proteins. Journal of Computational Chemistry. 2004;25(5):712–724.

179. Jacobson MP, Kaminski GA, Friesner RA, Rapp CS. Force field validation using protein side chain prediction. Journal of Physical Chemistry B. 2002;106(44):11673–11680.

180. Canutescu AA, Shelenkov AA, Dunbrack RL. A graph-theory algorithm for rapid protein side-chain prediction. Protein science : a publication of the Protein Society [Internet]. 2003 September [cited 2011 July 16];12(9):2001–14. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2323997&tool=pmcentrez&rendertype=abstract

181. Desmet J, De Maeyer M, Hazes B, Lasters I. The dead-end elimination theorem and its use in protein side-chain positioning. Nature. 1992;356(6369):539–542.

182. Desmet J, Spriet J, Lasters I. Fast and Accurate Side-Chain Topology and Energy Refinement (FASTER) as a new method for protein structure optimization. Proteins: Structure, Function and Genetics. 2002;48(1):31–43.

183. Hwang J-K, Liao W-F. Side-chain prediction by neural networks and simulated annealing optimization. Protein Engineering. 1995;8(4):363–370.

184. Kingsford CL, Chazelle B, Singh M. Solving and analyzing side-chain positioning problems using linear and integer programming. Bioinformatics. 2005;21(7):1028–1036.

185. Leach AR, Lemon AP. Exploring the conformational space of protein side chains using dead- end elimination and the A algorithm. Proteins: Structure, Function and Genetics. 1998;33(2):227–239.

186. Liu Z, Jiang L, Gao Y, Liang S, Chen H, Han Y, Lai L. Beyond the rotamer library: Genetic algorithm combined with the disturbing mutation process for upbuilding protein side-chains. Proteins: Structure, Function and Genetics. 2003;50(1):49–62.

187. Xu J. Rapid protein side-chain packing via tree decomposition. Lecture Notes in Bioinformatics (Subseries of Lecture Notes in Computer Science). 2005;3500:423–439.

188. Gainza P, Roberts KE, Donald BR. Protein design using continuous rotamers. PLoS Computational Biology [Internet]. 2012;8(1). Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-84857498912&partnerID=40&md5=4951f7166322cf8364d822ccf898330a

189. Cao Y, Song L, Miao Z, Hu Y, Tian L, Jiang T. Improved side-chain modeling by coupling clash-detection guided iterative search with rotamer relaxation. Bioinformatics. 2011;27(6):785–790.

190. Krivov GG, Shapovalov M V, Dunbrack Jr. RL. Improved prediction of protein side-chain conformations with SCWRL4. Proteins: Structure, Function and Bioinformatics. 2009;77(4):778–795.

191. Liang S, Zhou Y, Grishin N, Standley DM. Protein side chain modeling with orientation-dependent atomic force fields derived by series expansions. Journal of Computational Chemistry. 2011;32(8):1680–1686.

192. Huang S-Y, Zou X. An iterative knowledge-based scoring function to predict protein-ligand interactions: I. Derivation of interaction potentials. Journal of computational chemistry [Internet]. 2006 November 30 [cited 2013 June 27];27(15):1866–75. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16983673

193. Muegge I, Martin YC, Hajduk PJ, Fesik SW. Evaluation of PMF scoring in docking weak ligands to the FK506 binding protein. Journal of Medicinal Chemistry [Internet]. 1999;42(14):2498–2503. Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-0033566211&partnerID=40&md5=5441a0ca54bc4a243ea12cbac64b6485

194. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. 1983;4:187–217.

195. Jagielska A, Wroblewska L, Skolnick J. Protein model refinement using an optimized physics-based all-atom force field. Proceedings of the National Academy of Sciences of the United States of America. 2008;105(24):8268–8273.

196. Bermejo GA, Clore GM, Schwieters CD. Smooth statistical torsion angle potential derived from a large conformational database via adaptive kernel density estimation improves the quality of NMR protein structures. Protein science : a publication of the Protein Society [Internet]. 2012 December [cited 2013 June 29];21(12):1824–36. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23011872

197. Huang S-Y, Zou X. Inclusion of solvation and entropy in the knowledge-based scoring function for protein-ligand interactions. Journal of chemical information and modeling [Internet]. 2010 March 22 [cited 2013 June 29];50(2):262–73. Available from: http://dx.doi.org/10.1021/ci9002987