# Exploration of the Disambiguation of Amino Acid Types to Chi-1 Rotamer Types in Protein Structure Prediction and Design

by

Jarrett Sean Johnson

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Chemical Biology)
in the University of Michigan
2018

Doctoral Committee:

Professor Yang Zhang, Chair
Professor Charles Brooks III
Assistant Professor Aaron Frank
Professor Anna Mapp
Professor Zhaohui Xu

Jarrett Sean Johnson

jarrettj@umich.edu

ORCID iD: 0000-0003-2252-2814

For my family, friends, and fellows who have travelled the path of life with me.

# ACKNOWLEDGMENTS

First, I want to thank Professor Yang Zhang for allowing me the opportunity to perform research in his lab. Professor Zhang has shown a considerable amount of passion in his and his students' work. Whether a student is a beginning programmer or an experienced bioinformatician, Professor Zhang provides an appropriate challenge for him or her. Throughout the past four years, I have had the unique experience of conducting both *in vitro* and *in silico* experiments and honing a skill set that is extremely valuable in structural biology. I deeply appreciate my advisor's initial patience in allowing me to improve my competency in both areas simultaneously. Most importantly, I am grateful for Professor Zhang granting me the academic freedom to construct the ideas in this dissertation; this was the primary reason for my decision to join this lab. I believe Professor Zhang's expertise and imagination can solve endless scientific problems of today and tomorrow.

I would also like to extend my gratitude to all members of my committee. Professor Anna Mapp has shown outstanding leadership in the Program of Chemical Biology and has continuously provided me valuable support throughout my graduate career. Professor Charles L. Brooks III provided me a solid foundation and perspective on the relationship of molecular ensembles and its constituents' behaviors on the microscopic level. I would also like to thank Professor Zhaohui Xu for allowing me to rotate in his lab and his advice on structural biology and experimental designing.

Finally, I appreciate Professor Aaron Frank for his assistance in my prediction approaches and providing a relaxed environment to discuss science.

At the University of California, Los Angeles (UCLA), Professor Catherine Clarke and Dr. Christopher Allan introduced me to scientific research as an undergraduate student. Much of my successes and opportunities have been realized by their incredible mentorship.

I want to also give special thanks to the undergraduate students who I have mentored and the Undergraduate Research Opportunity Program (UROP) for providing these students with a solid introduction to and foundation for undergraduate research. I am very proud to have mentored Logan Hansen throughout his entire undergraduate career and experience his growth as an undergraduate researcher. I wholeheartedly appreciate Logan's assistance in performing experiments in the laboratory and his mentorship to several of other undergraduate students. I have also had the pleasure to mentor Thomas Stewart and Mariela Hartudo Garcia who have spent a considerable time in the lab and also carried out their independent research projects. I would also like to personally thank my undergraduate software developers Humza Hemani, Peter Tolsma and Zachary Orban for their programming contributions in the developments of knowledge-based energy functions. There are many other undergraduate students whom I have mentored, and although much of my time was devoted toward their scientific and academic success, I am very humbled to have met every single one of them. Some notable findings of my undergraduates' work are also featured in this dissertation.

TABLE OF CONTENTS

## LIST OF FIGURES

# LIST OF TABLES

# LIST OF APPENDICES

# ABSTRACT

A protein's global fold provide insight into function; however, function specificity is often detailed in sidechain orientation. Thus, determining the rotamer conformations is often crucial in the contexts of protein structure/function prediction and design. For all non-glycine and non-alanine types, chi-1 rotamers occupy a small number of discrete number of states. Herein, we explore the possibility of describing evolution from the perspective of the sidechains' structure versus the traditional twenty amino acid types. To validate our hypothesis that this perspective is more crucial to our understanding of evolutionary relationships, we investigate its uses as evolutionary, substitution matrices for sequence alignments for fold recognition purposes and computational protein design with specific focus in designing beta sheet environments, where previous studies have been done on amino acid-types alone. Throughout this study, we also propose the concept of the "chi-1 rotamer sequence" that describes the chi-1 rotamer composition of a protein. We also present attempts to predict these sequences and real-value torsion angles from amino acid sequence information.

First, we describe our developments of log-odds scoring matrices for sequence alignments. Log-odds substitution matrices are widely used in sequence alignments for their ability to determine evolutionary relationship between proteins. Traditionally, databases of sequence information guide the construction of these matrices which illustrates its power in discovering distant or weak

homologs. Weak homologs, typically those that share low sequence identity (<30%), are often difficult to identify when only using basic amino acid sequence alignment. While protein threading approaches have addressed this issue, many of these approaches include sequenced-based information or profiles guided by amino acid-based substitution matrices, namely BLOSUM62. Here, we generated a structural-based substitution matrix born by TM-align structural alignments that captures both the sequence mutation rate within same protein family folds and the chi-1 rotamer that represents each amino acid. These rotamer substitution matrices (ROTSUMs) discover new homologs and improved alignments in the PDB that traditional substitution matrices, based solely on sequence information, cannot identify.

Certain tools and algorithms to estimate rotamer torsions angles have been developed but typically require either knowledge of backbone coordinates and/or experimental data to help guide the prediction. Herein, we developed a fragment-based algorithm, Rot1Pred, to determine the chi-1 states in each position of a given amino acid sequence, yielding a chi-1 rotamer sequence. This approach employs fragment matching of the query sequence to sequence-structure fragment pairs in the PDB to predict the query's sidechain structure information. Real-value torsion angles were also predicted and compared against SCWRL4. Results show that overall and for most amino-acid types, Rot1Pred can calculate chi-1 torsion angles significantly closer to native angles compared to SCWRL4 when evaluated on I-TASSER generated model backbones.

Finally, we've developed and explored chi-1-rotamer based statistical potentials and evolutionary profiles constructed for *de novo* computational protein design. Previous analyses which aim to energetically describe the preference of amino acid types in beta sheet environments (parallel vs antiparallel packing or n- and c-terminal beta strand capping) have been performed with amino acid types although no explicit rotamer representation is given in their scoring functions. In our

study, we construct statistical functions which describes chi-1 rotamer preferences in these environments and illustrate their improvement over previous methods. These specialized knowledge-based energy functions have generated sequences whose I-TASSER predicted models are structurally-alike to their input structures yet consist of low sequence identity.

# Chapter 1: Introduction to Protein Informatics

## 1.1 Protein Folding

Proteins are one of the major four biomolecule classes which assist in a plethora of biological processes within cells including signaling, metabolism, providing structure, molecular transport, etc... Compositionally, proteins are a polymer of unique, covalently linked organic compounds known as amino acids. The identity of the amino-acid is granted by one of at least twenty functional groups (R) which confer specific physiochemical properties. Hence, the overall physiochemical property of a protein is the product of its amino acid composition which influences the protein's folding and/or function. During the polymerization process, covalent linkages of two amino acids result in the formation of an amide bond also described as peptide bonds in the context of protein synthesis. Peptide bond formation is accomplished by the nucleophilic addition-elimination reaction involving the nucleophilic amino group of an amino acid and the carboxyl group of another amino acid or the nascent peptide chain removing a water molecule in the process. Figure 1.1 shows an example tripeptide after two polymerization events involving three total single amino acids.

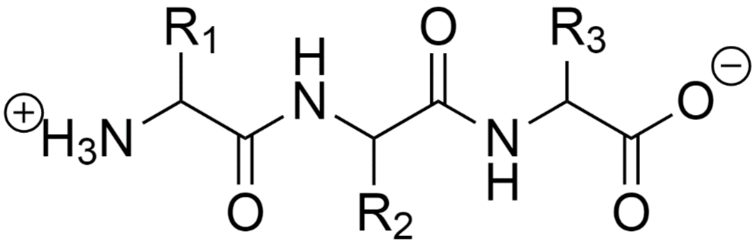


**Figure 1.1.** Zwitterion representation of a generic tripeptide.

The resulting covalent bonds give rise to the formation of the protein's backbone and are the basis to support higher-order protein structures. The folding process can be simply described by four protein structure levels:

Four levels of protein structure:

1) Primary Structure: the linear sequence of amino-acids held together covalently by peptide bonds.
2) Secondary Structure: regular sub-structural elements of a protein bound primarily by hydrogen bonds. Major classes of secondary structure include alpha-helix, beta-strand and coil. Alpha-helix and beta-strands have deeper classifications based on specific hydrogen bonding patterns. The protein backbone assumes a set of $\phi$ and $\psi$ torsion angles within defined ranges to accommodate these secondary structures [2].
3) Tertiary Structure: the resulting three-dimensional structure of a protein monomer proceeding after collapse of the hydrophobic core. Interactions on the sidechain-level greatly influence the tertiary structure.
4) Quaternary Structure: the three-dimensional assembly containing multiple monomers that forming a biological unit. Monomers typically associate by non-covalent interactions.

While quaternary structures are optional for some proteins, many biological functions require the formation of peptide complexes to carry out a particular function.

Figure 1.2 illustrates the different stages of the folding process.

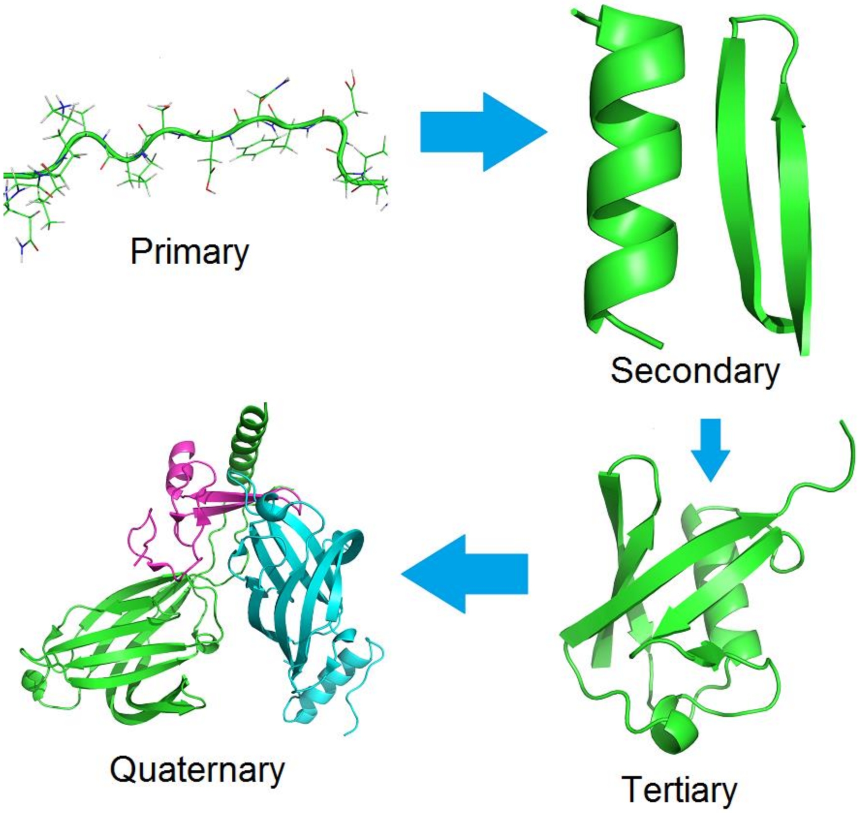


**Figure 1.2.** The four levels of protein structure.

## 1.2 Protein Databases

### 1.2.1 The UniProt Database

Managed by the UniProt Consortium, the UniProt database is a meta-repository that combines manually (UniProtKB/Swiss-Prot) and automatically (UniProbKB/TrEMBL) annotated sequences

[3, 4]. Currently, there are currently over 114 million sequence entries submitted in the repository comprising of over 38 billion total amino acids. Submission to the UniProt database requires some form of evidence of the expressed protein via Edman degradation [5] or tandem mass spectrometry (MS/MS) [6]. Annotated sequences here include a robust set of information including the protein's reported function and potential cellular location. Although users are free to use the entire database for analysis, the UniProt Consortium also offers several clustered sets of sequences, UniRef [7] and UniParc [8], to remove sequence-level redundancy.

### 1.2.2 Structure Databases

#### 1.2.2.1 The Protein Databank

Curated and managed by members of the Research Collaboratory for Structural Bioinformatics (RCSB), the Protein Databank [9] offers an archive to store experimentally resolved structures of large biomolecules including proteins and nucleic acids. To date, there are 140,109 submitted biological structures submitted to the repository, and notably, the number of PDB entries is three orders of magnitude less than that of the UniProt database [10]. Structures that populate this database are often discovered by x-ray crystallography [11] and nuclear magnetic resonance (NMR) [12] techniques; however, recent progress in cryo-electron microscopy (cryo-EM) [13] offers a third mechanism to probe biomolecular structure although structure resolution is typically lower than the former two methods. On the other hand, cryo-EM can offer great value in elucidating the overall structure large molecular complexes [14].

#### 1.2.2.2 SCOPe

Often, structural biologists are curious about the structural-evolutionary relationship between proteins. The Structural Classification of Protein—Extended (SCOPe) database [15] provides domain-level structural fold classification to entries submitted in the PDB. The classification is hierarchy where the top (most general) classes discriminate proteins by their main secondary structure composition or by special features (e.g. all α, all β, α+ β, α/ β, membrane, coiled-coil, etc…). A more detailed lineage of the protein are then described by closely related structures and typically follow the following hierarchy: Class → Fold → Superfamily → Family → Protein → Species → PDB entry ID.

### 1.3 Substitution Matrices and Sequence Alignments

As mentioned previously, primary structures of the protein highlight the unidirectional sequence of amino-acids of a protein. Collapsing amino-acids into their respective one-letter representation, a protein sequence is formed. Generally, similar sequences will often fold to similar three-dimensional structures [16]; however, attempts to determine similarity without computer assistance can often be non-trivial especially if they are evolutionarily distant or do not share a similar sequence length.

Accurate scoring functions are crucial to determine the optimal alignment a sequence pair. At the heart of these sequence alignment functions lie scoring matrices which quantify the propensity of amino acids types to mutate into other types throughout evolution [17]. The BLOSUM [18] and PAM [19] substitution matrix families are the most well-known; however, BLOSUM62 is the most ubiquitously used as it is typically the default scoring metric in sequence algorithms including BLAST. The construction of these matrices depend on sequence information from evolutionary related proteins. The elements that compose the matrix usually take the form of a log-odds ratio:

$$Sij = \log\left(\frac{p_{obs}}{p_{exp}}\right)$$

where $p_{obs}$ and $p_{exp}$ are the observed and expected mutation pair probability, respectively. Further details regarding how these probabilities are obtained are discussed in Chapter 2. Moreover, the probabilities inherit in the BLOSUM matrices, specifically, are sampled from the Blocks database [20], a repository that stores multiple sequence alignments of conserved regions across protein families.

The Needleman-Wunsch approach [21] is a deterministic algorithm to determine the best global alignment between a protein sequence pair. Adopting dynamic programming approaches, Needleman-Wunsch simplifies the problem of global alignment into scoring of individual residue pairs. For each residue pair, three candidate scores are determined from the possibilities if the residue pair aligns, if a gap is should be introduced instead, or if a gap should be extended [22]. The total score for that pair is the maximum of those three possibilities plus the alignments that precede it.

The time complexity for this approach is O(MN), where M and N are the lengths of either protein sequence since every protein pair is considered. Space complexity is also O(MN) due to the construction of matrices required to store the scores. However, recently, more efficient algorithms to reduce time and space requirements have been proposed [23].

### 1.4 Structural Alignments

Structural alignments are the preferred method to determine protein pair relationships. Identifying proper and robust ways to quantify structural similarity are still being investigated; however, several methods including RMSD [24] determination, DALI [25], and TM-score [26], are three of the most widely used criteria to describe structural relationship between protein pairs. Alignment methods that consider spatial coordinates of a molecule are typically performed in two stages:

superposition of one molecule onto its partner and a value in the form of spatial deviation between the molecules' components.

RMSD (root-mean squared deviation) calculations are typically determined as a two-stage approach: calculation of a rotation matrix [27] to perform a superposition of one molecule onto another and RMSD calculation of the superimposed pair of molecules. The Kabsch algorithm is widely used to determine the optimal rotation matrix which simultaneous searches for the RMSD minimum. The universal form of RMSD is shown as:

$$RMSD = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$$

where *N* atoms of structures *x* and *y* are spatially compared.

Using this criterion to quantify structural similarity of protein pairs presents some major limitations. For example, some localized area which are poorly aligned and are very spatially distant will significantly impact the global RMSD. Also, larger protein pairs will typically favor higher RMSD values compared to smaller pairs since the deviations are compounded. These concerns along with the absence of a theoretical limit on this equation makes it difficult to provide a standard to determine similarity and protein fold relationships.

### 1.5 TM-score

These drawbacks are the primary motivation for the invention of new methods to quantify structural similarity. From any arbitrary protein pair, there should be some quantifiable measurement and cutoff to determine evolutionary relationship. Approaches including MaxSub [28] finds the optimal substructure between a pair of proteins, but information not within 3.5

Ångstroms of the superposition is not considered in the MaxSub scoring function. Alignments with near-complete template alignment coverage and 50% template alignment coverage can yield the same MaxSub score; thus, Zhang et al. developed the Template-Modeling score (TM-score) to quantify the deviation of all aligned residues. [26].

The equation for TM-score assumes the following form:

$$TMscore = max\left[\frac{1}{L_{Target}}\sum_{i=1}^{L_{ali}}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right]$$

where $L_{target}$ is the length of the target protein and $L_{ali}$ is the length of the aligned residues. The symbol $d_0$ is a normalization constant to minimize the effect on protein length on the TM-score. Moreover, the theoretical minimum and maximum scores for any protein pair are 0 and 1, respectively. Zhang et al. performed statistical studies to determine the correlation of the TM-score and protein fold relationship. From the study, the authors note that a TM-score of at least 0.5, based on extreme value distribution model, confidently suggests the structure pair belongs to the same fold family [29].

While adaptation of the TM-score is increasing, there is still some minor disadvantages. While this approach is, for the most part, length-independent, extremely small proteins still exhibit TM-score anomalies.

### 1.6 Rotamers

Each amino acid type contains a unique functional (also known as R-) group that provides it a physio-chemical identity. While backbone torsion angles' degrees of freedom offer a variety of different local conformations, sidechains also display rotational entropy. For most amino acids, a set of chi ($\chi$) torsion angles (e.g. $\{\chi^1, \chi^2, \ldots\}$) are adaptable around single bonds. The size of the $\chi$ angle set is dependent on the length, or number of rotatable bonds, of the sidechain. Rotamers are conformational poses or instance of the varying angles of the $\chi$ torsion set. The minimum requirements for the first order of $\chi$ angles ($\chi^1$) are atoms at the sidechain $\beta$ and $\gamma$ position. Therefore, alanine and glycine do not have sufficient atoms to adopt a rotamer. Throughout the following studies, however, we confer one rotamer state upon alanine and glycine although no realistic, physical representation exists. Another exception to this rule is the proline sidechain. Due to proline's cyclical nature, its Cβ-γ is not directly rotatable; however, the torsion angle is still considered for proline across this bond. The dihedral angle here, instead, describes its "puckering" conformation as one of two states: *syn*- or *anti*-proline.

### 1.6.1 Chi-1 Rotamers

Our study focuses primarily on the first-order of the sidechain torsion angle set, or $\chi^1$. Newman projections of the atoms involved in this set give insight into the rotational entropy along the Cβ-γ bond. Due to the heterogeneity of these atoms, conformation preferences are observed. Alleviation of steric strain is a predominate force in the placement of involved atoms giving rise to three stereochemical conformations. These confirmations are described as *gauche-, trans, or gauche+* depending on the interaction between the backbone nitrogen atom and the sidechain Cβ atom. Unexpectedly, the *trans* is not the most observed confirmations here as the interaction between the backbone carbonyl group and the sidechain beta atom is the more profound interaction. Therefore, another consequence of this preference is the disproportionate observed frequencies of the $\chi^1$ rotamer.

As shown in Figure 1.2, the $\chi^1$ rotational isomers are well-defined and discrete. Orders of $\chi$ angles that exhibit this property are typically descried as “rotameric”, and, generally, as the angle order increases, the rotameric characteristics diminish.

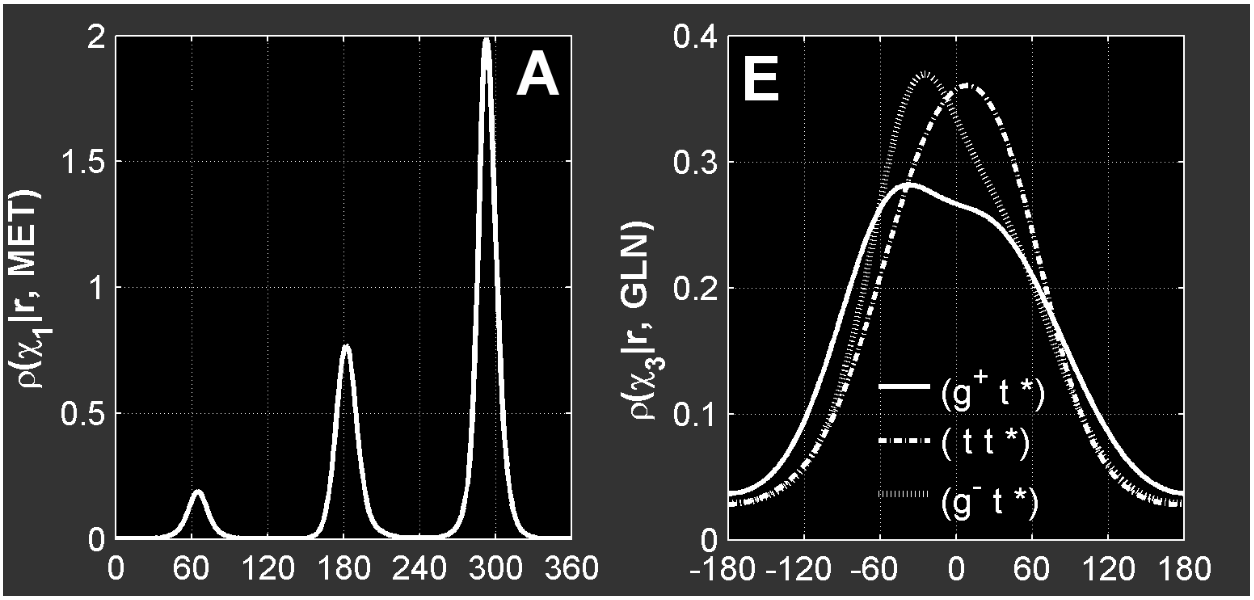


**Figure 1.3. Sidechain dihedral distributions of MET chi-1 and GLN chi-3.** Relative observed probability distributions of methonine’s chi-1 angles and Glutamine’s chi-3 angles. Curves on the right panel indicate various glutamine’s chi-1 and chi-2 rotamer states. [1]

Fortunately, we can classify each $\chi^1$ angle as an independent state and investigate whether we can replace amino-acid type designations with $\chi^1$ rotamer states. Our exact definition of rotamer states are adopted from Dunbrack’s library [30] . Each amino acid has three discrete $\chi^1$ state except for proline (two states possible), alanine, and glycine (where the latter two only has one rotamer state defined). Since these independent states inherently consists of some structural information of the sidechain, the overall hypothesis is that there is an improvement of quality in fold recognition for protein structure prediction and energy function accuracy for evolutionary protein design.

## 1.7 Computational Protein Design

There exists a complimentary problem to protein structure prediction. Instead of elucidating the native structure for a given sequence, consider attempting to discover the best sequence available from a given backbone structure or optional desired function. This inverse approach to protein structure prediction is known as protein design. Protein design can take many forms. Human intuition can often provide information regarding specific mutations to enhance or diminish a protein's function. Moreover, we can experimentally mutate a protein's sequence randomly and subject the mutants to a procedure which screens them based on a specific property (e.g. foldability, binding affinity to a ligand, etc…). We can achieve similar goals *in silico* via use of human-developed mathematical models, or force fields, which simulate the selection process. While the aforementioned design forms are often labelled as "protein engineering", *de novo* protein design studies exist in attempts to build a protein sequence "from scratch" given either backbone coordinates or some sort of desired target protein function.

Major milestones in the *de novo* protein design have been accomplished within the last few decades [31]. Historical achievements in the field include the first design of a protein domain by Mayo et al. [32] and the design of Top7, which consists of a fold reportedly never observed in nature by Kuhlman et al. [33]. Modern protein design algorithms typically consist of two major components: a search algorithm to sample candidate sequences and a force field to apply quantitative "selection" upon these sequences.

## 1.8 Protein Design Decoy Search by Monte Carlo Simulations

Monte Carlo simulations [34] provide a stochastic, non-biased method to generate candidate sequences, usually performed by single (or a relatively smaller number) of mutations iteratively.

Coupled with a force field, the Monte Carlo procedure can guide the algorithm to select candidate mutations based on the predicted energy or score provided by the energy function. For typical energy minimization processes, candidate sequences that are determined to have lower energy than its predecessor are accepted for further mutations while higher energy sequences are rejected. The major limitation to this approach is this will quickly reach a dead-end and the algorithm will be "trapped" in a local minima [35]. There are several ways to address this problem. First, one can apply a certain principle to occasionally accept sequences with higher energy. The Metropolis-Hastings criterion [36, 37] periodically accepts higher energy sequences depending on two factors: a randomly generated value $p$ between 0 and 1 and the magnitude of the energy gain. The Metropolis criterion is dictates that a sequence with increased energy $\Delta E$ will be accepted if:

$$p > \exp(-\frac{\Delta E}{kT}) \tag{1.1}$$

where $k$ and $T$ are the Boltzmann constant and temperature, respectively. Other approaches to escaping local minima and discovering the global minima include Simulated Annealing [38] which initiates the simulation at a high temperature and slowly decreases the temperature (e.g. decreases chance of accepting high energy sequences) throughout the process. Replica-Exchange Monte Carlo [39] has also shown promise, especially in protein structure prediction [40], to generate low energy models. Here, multiple trajectories occupy a unique simulation temperature (selected temperatures often span across a large range) and independently undergoes the simulation but occasionally swaps its decoy with a neighbor of an adjacent temperature index.

One of the major challenges of protein design, also shared by the protein structure prediction community, is the development of accurate force fields which can identify optimum sequence and structure pairs [41, 42]. Most algorithms, therefore, claim uniqueness from their algorithm and implemented force fields to guide their optimization procedures.

## 1.9 Force Fields

Force fields are often represented as a sum of smaller energy functions and are usually chosen to serve a specific purpose. Some classical physics-based force fields include CHARMM [43], AMBER [44], and OPLS [45]. These functions assume a functional form very similar to each other, but contains slight variations in the energy components in parameterization. These functions are also typically derived from first principals from physics, chemistry, and/or quantum mechanics and perform accurately on relatively smaller molecules. Molecular dynamic simulations using these models have recently achieved the ability fold proteins slightly above 92 amino acids; however will spend days to complete [46].

$$U(\vec{R}) = \sum_{bonds} k_i^{bond}(r_i - r_0)^2 + \sum_{angles} k_i^{angle}(\theta_i - \theta_0)^2 +$$

$$\sum_{dihedrals} k_i^{dihed}\,[1 + \cos(n_i\phi_i + \delta_i)] + \sum_i \sum_{j \neq i} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \tag{1.2}$$

$$\sum_i \sum_{j \neq i} \frac{q_i q_j}{\varepsilon r_{ij}}$$

For problems that involve a larger scale of protein folding and design, deriving fast, yet accurate, energy functions is crucial. One proposal to address this challenge is to use real protein data and empirics to inform mathematical functions. From this, statistical effective energy functions [47] were invented. These energy functions are based on energetic relationships with probabilities found in nature and are often derived from Boltzmann's equations. One basic form of the SEEF is typically exhibited as the following equation:

$$E_i = -kT\ln\left(\frac{p(i)}{p(ref)}\right)$$

In this equation, we can assign an energy *E* to state *i* if we have knowledge of its probability and the probability of the system's reference state *ref*. While the calculation of a target state's probability is often apparent, the selection of an appropriate reference state is usually non-trivial. In fact, many published atomic statistical potentials differ mainly by the chosen reference state [48-50] . One theoretical assumption that is typically made when considering these equations is that the population where probabilities are sampled from represents a Boltzmann distribution. In cases where SEEFs are applied to protein problems, statistical are often extracted from the PDB although the database itself doesn't accurately represent this special distribution. Nevertheless, these SEEFs are still extremely accurate and are used widely across state-of-the-art force fields [51, 52].

Throughout this dissertation, the target states' probabilities and the selection of the reference state used in the following energy functions will be described for each potential. Detailed derivations for our approach in designing SEEFs are outlined in Chapter 4.

### 1.10 Questions Explored by the Dissertation

A protein's ability to fold is primarily dictated by the sidechain's physiochemical properties, granted by its amino acid type, and their relative positions within the sequence. The detailed function, however, is often determined by the specific orientation of several rotamers. $X^1$ rotamers provide sidechain structural information as well as inherit knowledge of the amino acid type. We hypothesize that the using rotameric representations over simple amino acid ones yields better performance in fold recognition for protein structure prediction and in generating native-like sequences in computational protein design. To support this statement we propose several experiments and share results based on the following questions addressed in the following chapters:

Chapter Two: Do $\chi^1$ based substitution matrices perform better than amino acid-based ones in sequence alignment and fold recognition?

Chapter Three: Can we reliably predict the $\chi^1$ rotamer class and real-value dihedral angle from sequence information alone?

Chapter Four: Are statistically energy functions that draw on the statistics of $\chi^1$ rotamers from the PDB useful for protein design applications? Can we improve designs of beta sheets using these potentials?

## 1.11 References

1. Shapovalov, M.V. and R.L. Dunbrack, Jr., *A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions.* Structure, 2011. **19**(6): p. 844-58.
2. Ramachandran, G.N., C. Ramakrishnan, and V. Sasisekharan, *Stereochemistry of polypeptide chain configurations.* J Mol Biol, 1963. **7**: p. 95-9.
3. Consortium, T.U., *UniProt: the universal protein knowledgebase.* Nucleic Acids Res, 2017. **45**(D1): p. D158-D169.
4. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.* Nucleic Acids Res, 2003. **31**(1): p. 365-70.
5. Edman, P., *A method for the determination of amino acid sequence in peptides.* Arch Biochem, 1949. **22**(3): p. 475.
6. Mcluckey, S.A., G.L. Glish, and K.G. Asano, *The Coupling of an Atmospheric Sampling Ion-Source with an Ion Trap Mass-Spectrometer.* Abstracts of Papers of the American Chemical Society, 1988. **196**: p. 81-Anyl.
7. Suzek, B.E., et al., *UniRef: comprehensive and non-redundant UniProt reference clusters.* Bioinformatics, 2007. **23**(10): p. 1282-8.
8. Leinonen, R., et al., *UniProt archive.* Bioinformatics, 2004. **20**(17): p. 3236-7.
9. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
10. Zhang, Y., *Progress and challenges in protein structure prediction.* Curr Opin Struct Biol, 2008. **18**(3): p. 342-8.
11. Kendrew, J.C., et al., *3-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis.* Nature, 1958. **181**(4610): p. 662-666.
12. Wuthrich, K., *Protein structure determination in solution by NMR spectroscopy.* J Biol Chem, 1990. **265**(36): p. 22059-62.
13. Jiang, W., et al., *Semi-automated icosahedral particle reconstruction at sub-nanometer resolution.* Journal of Structural Biology, 2001. **136**(3): p. 214-225.
14. Costa, T.R.D., A. Ignatiou, and E.V. Orlova, *Structural Analysis of Protein Complexes by Cryo Electron Microscopy.* Methods Mol Biol, 2017. **1615**: p. 377-413.
15. Fox, N.K., S.E. Brenner, and J.M. Chandonia, *SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures.* Nucleic Acids Res, 2014. **42**(Database issue): p. D304-9.
16. Bowie, J.U., R. Luthy, and D. Eisenberg, *A method to identify protein sequences that fold into a known three-dimensional structure.* Science, 1991. **253**(5016): p. 164-70.
17. Schwartz, R.M.D., M.O, *Matrices for detecting distant relationships.* Atlas of protein sequence and structure, 1978. **5**: p. 353-58.
18. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.
19. M.O. Dayhoff, R.M.S., B.C. Orcutt, *A Model of Evolutionary Change in Proteins.* Atlas of protein sequence and structure, 1978. **5**.
20. Pietrokovski, S., J.G. Henikoff, and S. Henikoff, *The Blocks database--a system for protein classification.* Nucleic Acids Res, 1996. **24**(1): p. 197-200.
21. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J Mol Biol, 1970. **48**(3): p. 443-53.
22. Gotoh, O., *An improved algorithm for matching biological sequences.* J Mol Biol, 1982. **162**(3): p. 705-8.
23. Chakraborty, A. and S. Bandyopadhyay, *FOGSAA: Fast Optimal Global Sequence Alignment Algorithm.* Sci Rep, 2013. **3**: p. 1746.
24. Kabsch, W., *Solution for Best Rotation to Relate 2 Sets of Vectors.* Acta Crystallographica Section A, 1976. **32**(Sep1): p. 922-923.

25. Holm, L. and P. Rosenstrom, *Dali server: conservation mapping in 3D.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W545-9.
26. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins, 2004. **57**(4): p. 702-10.
27. Kabsch, W., *Discussion of Solution for Best Rotation to Relate 2 Sets of Vectors.* Acta Crystallographica Section A, 1978. **34**(Sep): p. 827-828.
28. Siew, N., et al., *MaxSub: an automated measure for the assessment of protein structure prediction quality.* Bioinformatics, 2000. **16**(9): p. 776-85.
29. Xu, J. and Y. Zhang, *How significant is a protein structure similarity with TM-score = 0.5?* Bioinformatics, 2010. **26**(7): p. 889-95.
30. Dunbrack, R.L., Jr. and F.E. Cohen, *Bayesian statistical analysis of protein side-chain rotamer preferences.* Protein Sci, 1997. **6**(8): p. 1661-81.
31. Zanghellini, A., et al., *New algorithms and an in silico benchmark for computational enzyme design.* Protein Sci, 2006. **15**(12): p. 2785-94.
32. Dahiyat, B.I., C.A. Sarisky, and S.L. Mayo, *De novo protein design: towards fully automated sequence selection.* J Mol Biol, 1997. **273**(4): p. 789-96.
33. Kuhlman, B., et al., *Design of a novel globular protein fold with atomic-level accuracy.* Science, 2003. **302**(5649): p. 1364-8.
34. Zhang, C.T. and K.C. Chou, *Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition.* Biophys J, 1992. **63**(6): p. 1523-9.
35. Dill, K.A., et al., *The protein folding problem.* Annu Rev Biophys, 2008. **37**: p. 289-316.
36. Hastings, W.K., *Monte-Carlo Sampling Methods Using Markov Chains and Their Applications.* Biometrika, 1970. **57**(1): p. 97-&.
37. Metropolis, N., et al., *Equation of State Calculations by Fast Computing Machines.* Journal of Chemical Physics, 1953. **21**(6): p. 1087-1092.
38. Khachaturyan, A.G., S.V. Semenovskaya, and B.K. Vainshtein, *Statistical Thermodynamical Approach to the Problem of Determination of Phases of Structure Amplitudes.* Kristallografiya, 1978. **24**(5): p. 905-916.
39. Sugita, Y. and Y. Okamoto, *Replica-exchange molecular dynamics method for protein folding.* Chemical Physics Letters, 1999. **314**(1-2): p. 141-151.
40. Zhang, Y., *I-TASSER server for protein 3D structure prediction.* BMC Bioinformatics, 2008. **9**: p. 40.
41. Baker, D., *An exciting but challenging road ahead for computational enzyme design.* Protein Sci, 2010. **19**(10): p. 1817-9.
42. Gordon, D.B., S.A. Marshall, and S.L. Mayo, *Energy functions for protein design.* Curr Opin Struct Biol, 1999. **9**(4): p. 509-13.
43. Brooks, B.R., et al., *Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations.* Journal of Computational Chemistry, 1983. **4**(2): p. 187-217.
44. Cornell, W.D., et al., *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995).* Journal of the American Chemical Society, 1996. **118**(9): p. 2309-2309.
45. Jorgensen, W.L., D.S. Maxwell, and J. TiradoRives, *Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids.* Journal of the American Chemical Society, 1996. **118**(45): p. 11225-11236.
46. Nguyen, H., et al., *Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent.* J Am Chem Soc, 2014. **136**(40): p. 13959-62.
47. Sippl, M.J., *Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures.* J Comput Aided Mol Des, 1993. **7**(4): p. 473-501.
48. Zhang, C., S. Liu, and Y. Zhou, *Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential.* Protein Sci, 2004. **13**(2): p. 391-9.

49. Zhang, J. and Y. Zhang, *A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction.* PLoS One, 2010. **5**(10): p. e15386.

50. Zhou, H. and J. Skolnick, *GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction.* Biophys J, 2011. **101**(8): p. 2043-52.

51. Thomas, P.D. and K.A. Dill, *Statistical potentials extracted from protein structures: How accurate are they?* Journal of Molecular Biology, 1996. **257**(2): p. 457-469.

52. BenNaim, A., *Statistical potentials extracted from protein structures: Are these meaningful potentials?* Journal of Chemical Physics, 1997. **107**(9): p. 3698-3706.

# Chapter 2: Chi-1 Substitution Matrix from Structural Alignments by TM-align

## 2.1 Introduction

Substitution matrices describe the evolutionary relationship, usually in the form of conservation, between two amino acids in a protein family [17]. Traditional substitution matrices describe the ease of these changes among amino acids, and are typically useful in scoring relationships between two, and even multiple, sequences by alignment [17]. Notably, major substitution matrix families, BLOSUM [18] and PAM [19], serve as the scoring metric in alignment procedures that involve dynamic programming, including (PSI-) BLAST [53, 54] and protein threading algorithms. BLOSUM62 is perhaps the most widely used substitution matrix across most alignment platforms due to its ability to identify strong and weak homologs of protein sequences. While, this method is suitable to identify related sequences, it falls short in highlighting related proteins with low sequence identity but significant structural similarity.

Other forms of log-odds substitution matrices are represented here which use structural information to generate the elements in the table. Upon measuring structure similarity between proteins to construct these matrices, we utilize the TM-score [26] by TM-align [55] between protein backbones found in the Protein Databank (PDB) [9]. This structural score similarity is a widely adopted measurement to describe structural similarity.

For this study, we explored the impact of reclassifying amino-acid types into their structural $\chi^1$ rotamer types. Ideally, all possible rotamer types across all possible $\chi$ angles could be realized as its own type. Unfortunately, due to the relatively low number of structural information in the PDB that could sufficiently describe the distribution of rotamers across all proteins plus the flexibility of higher order chi angles, we decided to only consider the first $\chi$ dihedral. We define these states similar to those previously described [30]. Three $\chi^1$ angle states exist for most amino acid types, except for alanine and glycine (which are considered to consist of only one rotamer type). Furthermore, proline's $\chi^1$ rotamer types assume either *syn* or *anti*.

We record the performance of our substitution matrix by its ability to identify homologs compared to the BLOSUM62 matrix. We identify one major potential of improvement to previous substitution matrices is the extraction of structural data from the PDB. Often, comparison of sequence information between two proteins offers sufficient information to determine a potential relationship. However, a structurally related pair may lack sufficient similarity on the sequence level [10] to be identified by sequence alignment methods. Purely sequence-based substitution matrices thus cannot adequately describe some structural underlying factors. Current methods to address this, usually found in protein threading methods to fit or "thread" the sequence onto a structure template [16]. These often involve a more thorough approach, including information such as predicted information regarding the protein's secondary structure, backbone torsion angles, and solvent accessibility [56]. Protein family profiles (or position-specific scoring matrices) are exhaustively used as well and can also be applied outside of protein fold-recognition [57]. Previously, we designed a fixed-backbone protein algorithm, EvoDesign [58], whose scoring system is greatly influenced through a structural-based sequence profile augmented by information gained in the BLOSUM62 matrix. The substitution matrices described here can also be applied to evolutionary approaches to protein design. The construction of a rotamer substitution matrix (ROTSUM) and the rotamer profile are described and assessed here.

## 2.2 Material and Methods

### 2.2.1 Rotamer Sequence

For the purposes of our studies, we discretize $\chi^1$ rotamers based on both amino-acid type and dihedral angle. The angles are binned into one of three categories for non-proline residues: *gauche+* [0°, 120°), *trans* [120°, 240), or *gauche-* [-120°, 0°). For example, phenylalanine $\chi1$ rotamer angles that are *gauche+*, *trans*, or *gauche-* are designated as F1, F2, or F3, respectively. The first character depicts the amino acid single character representation and is followed by the rotameric class represented as an integer. Furthermore, a single byte, alphanumeric representation of these rotamer types, suitable for *in silico* sequence alignments, are mapped shown in Appendix A and an example rotamer sequence is illustrated in Figure 2.1. Alanine and glycine do not have rotameric possibilities, thus its designation is always A1 or G1. Proline has *syn-* [0°, 240°) or *anti-* [-120°, 0°) confirmations and are designated as either P1 or P2. Therefore, a total number of 55 $\chi1$ rotameric types are available for the twenty common amino acids. Herein, we define a rotamer sequence as an array of these rotameric designations. We can, thus, express protein sequences as either a canonical amino-acid sequence or rotamer sequence. An automated, auxiliary tool to determine a rotamer sequence from protein structure, Rot1Calc, is also available on the Rot1Pred site https://zhanglab.ccmb.med.umich.edu/Rot1Suite/.

```
MQIFV...RLRGG
M1Q3I1F3V2......R2L3R2G1G1
ajRMv...mZmNN
```

**Figure 2.1. Rotamer Sequence.** Top: Canonical amino acid sequence represented as a sum of one-byte character. (20 possibilities). Middle: Readable sum of double byte representation of a chi-1 rotamer states. First byte is one of twenty amino acid types. The second byte is the rotamer class. (55 possibilities). Bottom: Sum of single byte representation of the rotamer character.

### 2.2.2 Construction of the Rotamer Substitution Matrix

To consider the evolutionary relationship between $\chi$1 rotamers, a substitution matrix was created to describe the propensity of a $\chi$1 rotamer to mutate into another across evolution. In contrast to the Henikoff method for generation of a substitution matrix, we cannot simply use a sequence database to accomplish this task since detailed structural information of rotamer dihedrals are required. Instead, we considered closely related protein structures found in the PDB to determine rotamer evolution. The first step in generating the matrix involves an all-against-all protein structure alignment of a non-redundant (by 70% sequence identity) subset of the PDB. Here, we generated TM-score cutoffs of 0.5 to 0.9 in 0.05 increments. From all pairwise alignments where the TM-score is above a certain threshold, we calculated the total $\chi$1 rotamer substitution frequency for each aligned positions. To prevent inaccuracies that may arise from alignment boundaries, alignment pairs that consists of an unaligned neighbor are excluded. From the observed substitution frequencies, we also calculated the expected substitution frequencies $e_{ij}$ from the individual $\chi$1 rotamer probabilities (Eqn. 1, 2). Our final substitution matrix $S$ is calculated as a log-odds ratio in bit units (Eqn. 3).

$$p_i = qii + \sum_{i,i \neq \mathrm{j}}^{55} \frac{q_{ij}}{2} \tag{2.1}$$

$$e_{ij} = p_i * p_j \tag{2.2}$$

$$S_{ij} = log_2(\frac{q_{ij}}{e_{ij}}) \tag{2.3}$$

Naturally, substitution matrices' distributions with more stringent structural similarity cutoffs increase in distance from the expected (or background) distribution of rotamer pairs. An effective and robust substitution matrix should contain enough information of the structure information as possible without too much bias toward very closely related proteins. This will provide enough information to identify both close and distant protein homologies. Quantification of the distance between two distributions are often described by relative entropy which defines how much information is gained from divergence of a second distribution [59]. Although the derivation of the ROTSUM matrices are fundamentally different from previous sequence-derived sequence-based substitution matrices, the different matrix types can still be compared. We define the information gain for a distribution from the background distribution (H):

$$H = \sum_{i=1}^{20} \sum_{j=1}^{i} q_{ij} \times s_{ij} \tag{2.4}$$

ROTSUM matrices uniquely take the form of a 55×55 substitution matrix; thus, our relative entropy calculations were performed on a "collapsed", 20×20 version of the ROTSUM matrices (named here as COLLMX). We obtained these matrices by adding all rotamer probabilities belonging to the same amino-acid type. Relative entropy values of 0 indicate that the rotamer (or amino-acid) pair distributions are non-distinguishable from the expected distribution.

### 2.2.3 Alignment Quality Generated by ROTSUM and Rotamer Sequences

Here, we implement a conventional Needleman-Wunsh [21] approach to compare the quality of the ROTSUM vs BLOSUM62. For our benchmark, we employ the MUSTER500 [56] dataset which consists of 500 proteins. Each component of the dataset was aligned against all proteins in the non-redundant PDB. All hits equal to and above 30% sequence identity (determined by the BLOSUM62 matrix) are removed from consideration. The significance of the alignment was determined by the Z-score of the alignment, where the Z-score is equal to the number of standard deviations from the mean alignment score of the template library to the specific query:

$$Z_{ali,i} = \frac{S_{ali,i} - < S_{ali} >}{\sigma_{ali}} \tag{2.5}$$

$S_{ali,i}$ is the score of alignment *i* calculated for a given pair of sequences containing the query and a template sequence. $<S_{ali}>$ and $\sigma_{ali}$ are the average score of all template alignments against the query and the standard deviation of the score distribution, respectively. TM-scores for the top alignments (highest Z-score) for each query when subjected to either the ROTSUM or BLOSUM62 matrix are then determined.

### 2.2.4 Sequence alignment

The first scoring function involves only the ROTSUM72 substitution matrix to assess its individual impact against BLOSUM62. For a given query position $i$ and template position $j$, the alignment score is defined as

$$S_{ij}(aa_i, aa_j) = M_{BL}(aa_i, aa_j) \tag{2.5}$$

where *aa* represents the amino acid type. Next, we utilize another scoring function specifically for $\chi^1$ rotamer sequences of the query sequence rather than amino acid types. The scoring function is defined as

$$S_{ij}(X_i^1, X_j^1) = M_{ROT}(X_i^1, X_j^1) \tag{2.6}$$

where $X^1$ represents the $\chi^1$ rotamer type. Finally, we derived a combined scoring function which considers both relationships between the $\chi^1$ rotamer type and the amino acid type. The scoring function is a weighted, linear combination of the two previous equations:

$$S_{ij}(aa_i, aa_j, X_i^1, X_j^1) = w_{BL}M_{BL}(aa_i, aa_j) \ + w_{ROT}M_{ROT}(X_i^1, X_j^1) \tag{2.7}$$

where $w_{BL}$ and $w_{ROT}$ are the respective weighting factors for the amino acid and $\chi^1$ rotamer substitutions. In this study, the ROTSUM weighting factor $w_{ROT}$ was evaluated from 0.0 to 1.0 in increments in 0.1 while satisfying the constraint: $w_{ROT} + w_{BL} = 1.0$.

### 2.2.5 Sequence- and Structure-derived profiles

For the template library, profile matrices and be generated from structure-derived multiple sequence alignments. The details of the profile construction procedure is outlined in Chapter 4. Our dataset assumes we do not have prior information regarding its three-dimensional structure. To construct a profile for our each protein in this set, the source of sequences used to create the

MSA comes from alignments generated by Section 2.2. We select query-anchored sequence alignments whose Z-scores values are above 1.5 and subsequently combine them into an MSA and construct the profile matrix.

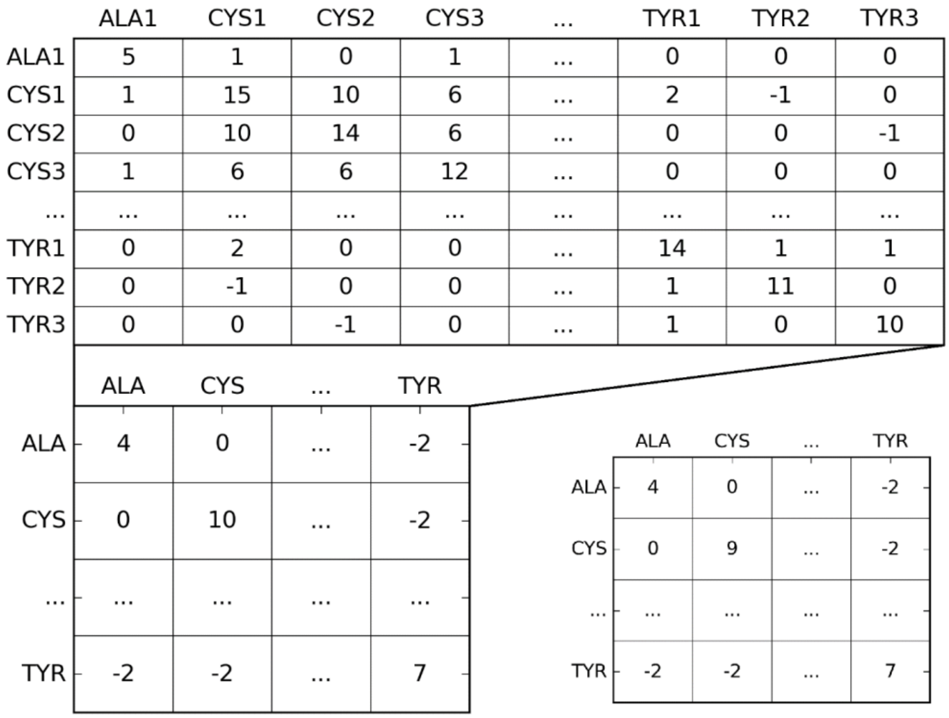
For a given query position *i* and template position *j*, the alignment score is defined as:

$$S_{ij} = \sum_{k=1}^{55} \left( Q(i,k) \times T(j,k) \right) + SS(i,j) + C \tag{2.8}$$

where *Q* and *T* are the sequence-derived query and structure-derived template profiles, respectively. Symbol *k* represents the 55 $\chi^1$ rotamer types (where each column in the profile is assigned to a unique $\chi^1$ rotamer type). The second term in Eqn. 2.8 returns a value that represents a match ($SS_{match}$ = 0.65) or mismatch ($SS_{mismatch}$ = -0.65) between the predicted secondary structure element at position *i* of the query sequence and the secondary structure element of position *j* of the template sequence. Finally, the last term, *C*, is simply a constant previously tuned.

## 2.3 Results

### 2.3.1 ROTSUM Matrices



|  | ALA1 | CYS1 | CYS2 | CYS3 | ... | TYR1 | TYR2 | TYR3 |
|---|---|---|---|---|---|---|---|---|
| ALA1 | 5 | 1 | 0 | 1 | ... | 0 | 0 | 0 |
| CYS1 | 1 | 15 | 10 | 6 | ... | 2 | -1 | 0 |
| CYS2 | 0 | 10 | 14 | 6 | ... | 0 | 0 | -1 |
| CYS3 | 1 | 6 | 6 | 12 | ... | 0 | 0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| TYR1 | 0 | 2 | 0 | 0 | ... | 14 | 1 | 1 |
| TYR2 | 0 | -1 | 0 | 0 | ... | 1 | 11 | 0 |
| TYR3 | 0 | 0 | -1 | 0 | ... | 1 | 0 | 10 |

|  | ALA | CYS | ... | TYR |
|---|---|---|---|---|
| ALA | 4 | 0 | ... | -2 |
| CYS | 0 | 10 | ... | -2 |
| ... | ... | ... | ... | ... |
| TYR | -2 | -2 | ... | 7 |

|  | ALA | CYS | ... | TYR |
|---|---|---|---|---|
| ALA | 4 | 0 | ... | -2 |
| CYS | 0 | 9 | ... | -2 |
| ... | ... | ... | ... | ... |
| TYR | -2 | -2 | ... | 7 |

**Figure 2.2. ROTSUM72 Matrix.** (Top) Preview of the 55×55 chi-1 rotamer substitution matrix. (Bottom-left) Truncated substitution matrix of the rotamer substitution matrix collapsed into their respective amino acids. (Bottom-right) Truncated BLOSUM62 matrix as comparison to the matrix previewed in (b).

We employ TM-score as measurement to quantify structural evolutionary relationship since the correlation of the score and evolutionary relationship is greater than RMSD [29]. TM-align also provides an alignment between structure pairs without knowledge of either protein's sequence. We circumvent the need of a motif database e.g. Blocks database [20] or phylogenetic information. In doing so, we avoid any use of including explicit amino acid sequence information. We clustered structures into groups whose backbones share a TM-score above a given cutoff. Figure 2.2 exhibits the truncated and rounded form of the symmetric ROTSUM72 matrix. The full real-value form of each ROTSUM matrix is available on the Rot1Align website (see Section 2.4). The integral character proceeding “ROTSUM” represents the decimal TM-score cutoff (0.72 in this case). We illustrate ROTSUM72 due to its similar relative entropy compared to BLOSUM62’s (see Section 3.2). The bottom-left matrix shows the collapsed form of ROTSUM72 matrix and shown adjacent to the BLOSUM62 matrix for comparison. One hallmark of the ROTSUM matrices is their ability to quantify rotamer conservation throughout evolution. For a given amino acid type, a particular rotamer state will tend to maintain that rotamer state. Interestingly, even across amino acid types, rotamer classes tend to also be conserved. For example, the log odds score for the ASP1 to GLU1 mutation is greater than either intra-amino acid rotamer mutation score (ASP1 to ASP2 or ASP1 to ASP3). Although the derivation of the ROTSUM72 matrix contains absolutely no *a priori* knowledge of the sequences used to compose the rotamer substitutions (TM-align requires only Cα trace of either protein), many substitution scores between the two matrices nearly are identical.

### 2.3.2 Relative Entropy

Figure 2.3 illustrates the relative entropies for the ROTSUM matrices and their respective 20×20 collapsed forms. As expected, substitution matrices derived from pairwise structure alignments with greater TM-score cutoffs increase in information gain from background distributions. The relative entropy for BLOSUM62 is shown as a horizontal line for comparison. Drastic relative entropy loss of the rotameric representation collapsing to the amino acid types indicates support for our hypothesis: knowledge of a position’s rotamer state provides more information for the alignment procedure to match two positions than just knowing the amino-acid type alone.

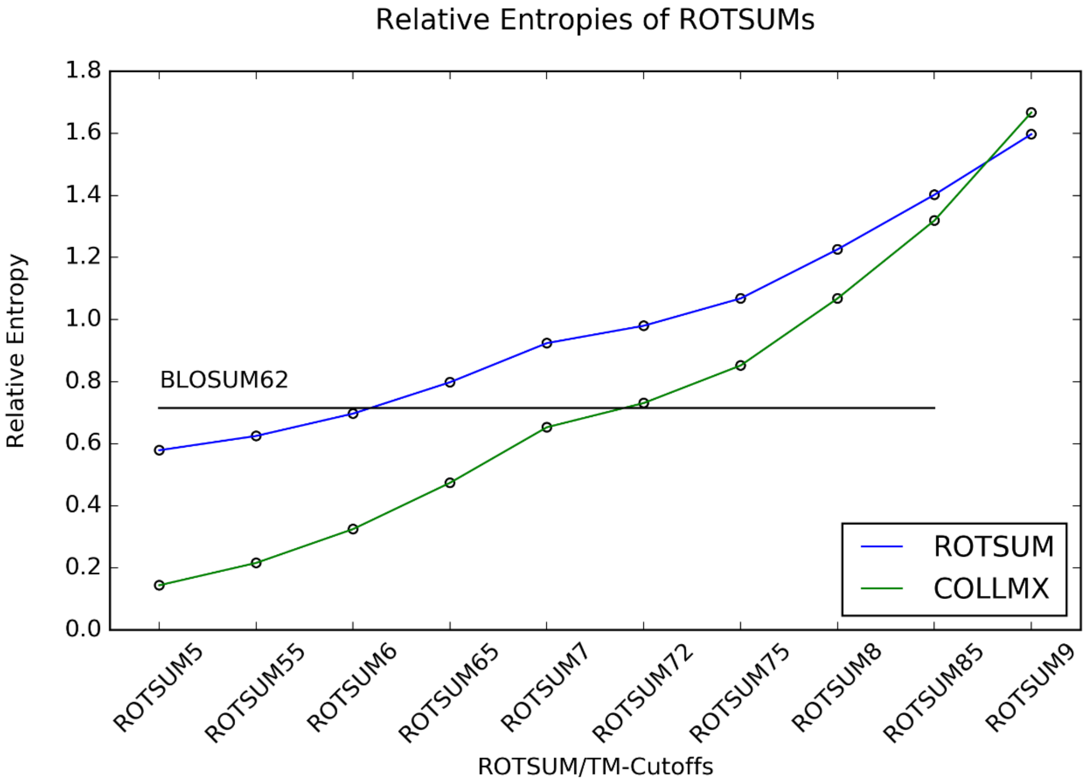


**Figure 2.3. Relative Entropies of ROTSUM Matrices**. Calculated relative entropies (in bit units) are shown for all generated ROTSUM matrices under different TM-score cutoffs. For each cutoff, both relative entropies for the 55×55 matrix (blue) and its corresponding 20x20 collapsed (green).

### 2.3.3 Fold Recognition with ROTSUM

#### 2.3.3.1 Rotamer vs Amino Acid Sequence Alignment by Dynamic Programming

To assess the individual effects of ROTSUM on recognizing protein fold against BLOSUM62, we aligned each protein sequence in the dataset against a non-redundant (by 70% sequence identity) PDB library. We identify the optimal alignments guided by a substitution matrix with the Needleman-Wunsch dynamic programming approach. For both cases, we employ both gap opening (11) and gap extension (1) penalty [22]. Scores of the alignments were calculated only via ROTSUM8 or BLOSUM62. Upon completion of all possible pairwise alignments between a query and all protein templates in a nonredundant protein database, we determine and sort the Z-score for each alignment. The resulting alignment with the highest Z-score was used as input for structural alignment in determining the TM-score between the query and best template structure. Figure 2.4 shows the best TM-score in the top N alignments for ROTSUM8 and BLOSUM62 (N=1, 2, or 5). The average TM-scores for the best discovered alignments were 0.392 and 0.376 for ROTSUM8 and BLOSUM62, respectively (p-value = 3.6E-4). Although there is a considerable increase in the TM-score using ROTSUM, there are still cases where BLOSUM62 can identify folds readily while ROTSUM cannot and vice-versa. Illustrative examples of contrasting performance are shown in Figure 2.5.

Additionally, we attempted to improve the sequence alignment quality by implementing a scoring function which combines ROTSUM8 and BLOSUM62. We repeated this combination several times while varying the weights of the respective matrices $w_{R8}$ (for ROTSUM8) and $w_{B62}$ (for BLOSUM62) under the restraint $w_{R8} + w_{B62} = 1.0$. Table 2.1 summarizes the average TM-score, RMSD, and fraction alignment coverage amongst all top 1 hits for all protein queries. The most optimal weight combination ($w_{R8} = 0.8$ and $w_{B62} = 0.2$) yields an average TM-score of 0.393 (p =

3.7E-5). This supports the notion that ROTSUM matrices provides a relatively stronger impact to the quality of the alignment than BLOSUM62.

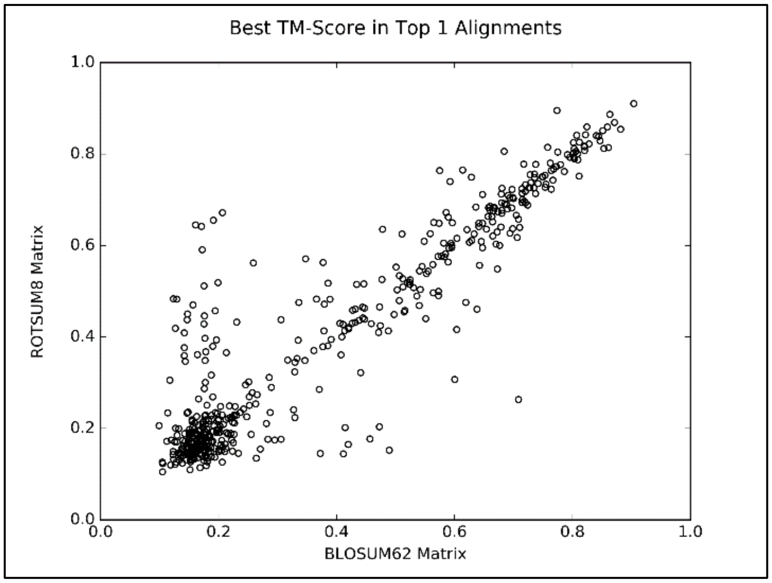


**Figure 2.4. TM-score of the best alignments provided by ROTSUM8 or BLOSUM62.** Needleman Wunsch alignments of the MUSTER test set containing 500 query proteins and their top hits given by either ROTSUM8 or BLOSUM62. Top alignments utilizing either matrix were used to determine the TM-score of the aligned residues between the query and its highest scoring alignment.

| $w_{R8}$ | $w_{B62}$ | Avg. TM-score | Avg. RMSD | Coverage |
|---|---|---|---|---|
| 0 | 1 | 0.376 | 12.799 | 0.933 |
| 0.1 | 0.9 | 0.382 (3.7E-4) | 12.668 | 0.930 |
| 0.2 | 0.8 | 0.388 (9.9E-6) | 12.546 | 0.930 |
| 0.3 | 0.7 | 0.391 (7.9E-7) | 12.322 | 0.928 |
| 0.4 | 0.6 | 0.392 (5.1E-7) | 12.299 | 0.927 |
| 0.5 | 0.5 | 0.392 (2.3E-6) | 12.374 | 0.926 |
| 0.6 | 0.4 | 0.392 (6.5E-6) | 12.420 | 0.923 |
| 0.7 | 0.3 | 0.392 (2.8E-6) | 12.420 | 0.920 |
| 0.8 | 0.2 | 0.393 (3.7E-5) | 12.358 | 0.918 |
| 0.9 | 0.1 | 0.389 (1.7E-3) | 12.400 | 0.916 |
| 1.0 | 0 | 0.391 (3.6E-4) | 12.224 | 0.914 |

**Table 2.1. NW-Alignments with combined scoring matrices.** Average TM-score, RMSD, and coverage are reported for Top 1 alignments of the weighted scoring function. Wilcoxin rank-signed test was used to determine the significance (in the form of p-value—shown in parentheses) in mean difference between the weighted scoring function and alignments generated purely by BLOSUM62.

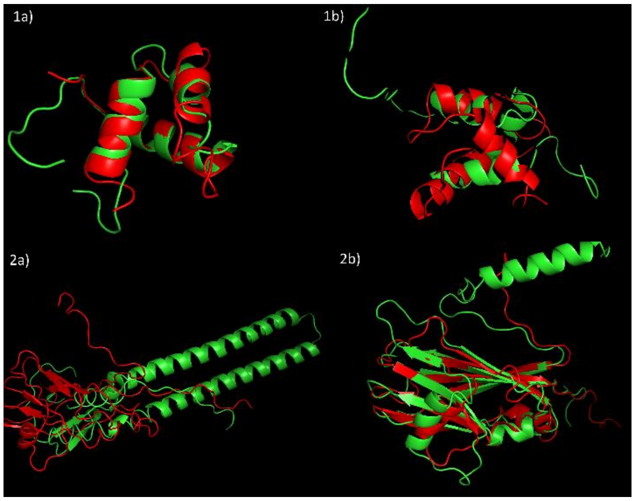


**Figure 2.5: Alignment comparisons of best alignments determined by ROTSUM8 and BLOSUM62.** Needleman-Wunsch dynamic programming was used to find the top template alignment using either ROTSUM8 or BLOSUM62. The query is shown in red cartoon while the top template found is shown in greed cartoon. Assessment values are reported here as TM-score/sequence identity in parenthesis. 1. Top template and alignment hit of the 1fexA query. 1a) Best alignment found by ROTSUM8. 0.582/0.22 1b) Best alignment found by BLOSUM62 (0.177/0.22). 2. Top template and alignment hit of the 1f15c query. 2a) Best alignment found by ROTSUM8 (0.123/0.24). 1b) Best alignment found by BLOSUM62 (0.617/0.25).

### 2.3.3.2 Correlation between Z-score and TM-score

The Z-score quantitatively describes the quality of the alignment between query and template sequence pairs amongst other templates in the sequence database and will usually directly correlate with its respective TM-score. Figure 2.6 presents the TM-score for each query's top sequence alignment's Z-score. To identify aligned structure pairs with TM-scores above 0.5 and using a Z-score cutoff of 1.6, we obtain a false positive rate and false negative rate of 14.3% and 17.9%, respectively. State-of-the-art fold recognition programs and other more robust approaches typically have false rates an order of magnitude lower than this approach. However, we simply highlight the individual performance of the matrix and its ability to discern good and poor alignments.

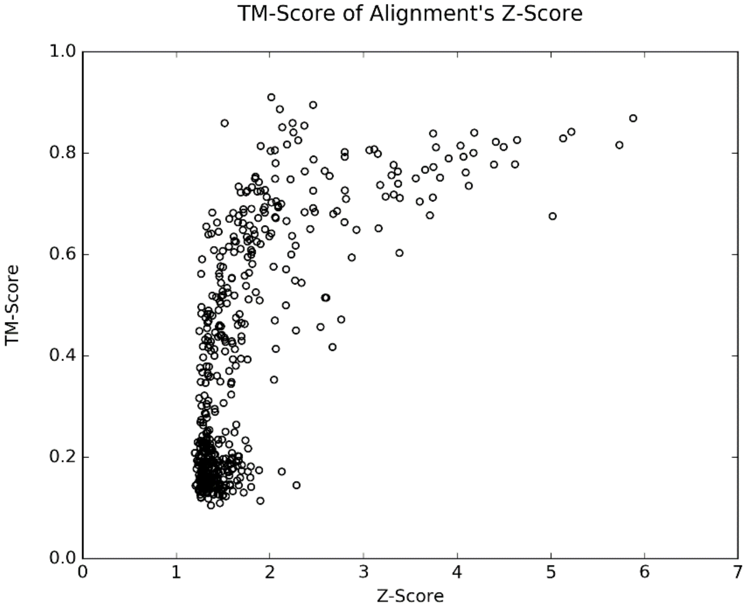


**Figure 2.6: Alignment comparisons of Top-1 hit by ROTSUM8 and BLOSUM62.** The Z-score of 500 target protein's best TM-score in top 1 alignment and their corresponding TM-scores are plotted. Horizontal line indicates a TM-score of 0.5. The vertical line represents the z-score threshold of 1.6 where the lowest false positive rate and false negative rate for TM-score > 0.5 classification are located.

#### 2.3.3.3 Rotamer-Profile Assisted Fold Recognition

Profile-profile alignments (PPAs) have shown promise in threading algorithms and are more sensitive in identifying templates for hard targets than simple sequence alignments [60]. Here, we attempted a proof-of-concept to determine if similar principles can be applied with rotamer-sequence profiles. Figure 2.7 shows the best TM-score for the top N alignments using the scoring function defined by equation 3.7. For the best TM-score in the top 1 alignment, the average TM-scores for a dataset of 100 proteins subjected to both rotamer profile-rotamer profile alignments (RPRPAs) and PPAs were 0.593 and 0.615, respectively. Also, the average best TM-score were 0.615 (RPRPAs) and 0.633 (PPAs) for top 2 alignments and 0.627 (RPRPAs) and 0.646 (PPAs) for top 2 alignments. Overall, although there was a tremendous increase in RPRPA performance compared to basic rotamer sequence alignments, RPRPAs, using this approach, did not outperform PPAs. There are several factors that affect its performance. First, all template and queries must undergo a TM-align search to identify similar structures. In cases where a fold is relatively rare compared to others, the profile may not be well-informed due to a low number of structures that occupy the multiple sequence alignment. PSI-BLAST profiles have the luxury of using sequence databases which are extremely rich in entries and deeply covers most protein family compared to the PDB. Moreover, variables tuned for this scoring function, including the constant shift and secondary structure match, have been well tuned for PPA. Future studies should also consider tuning these variables but only when well-informed profiles have been constructed.

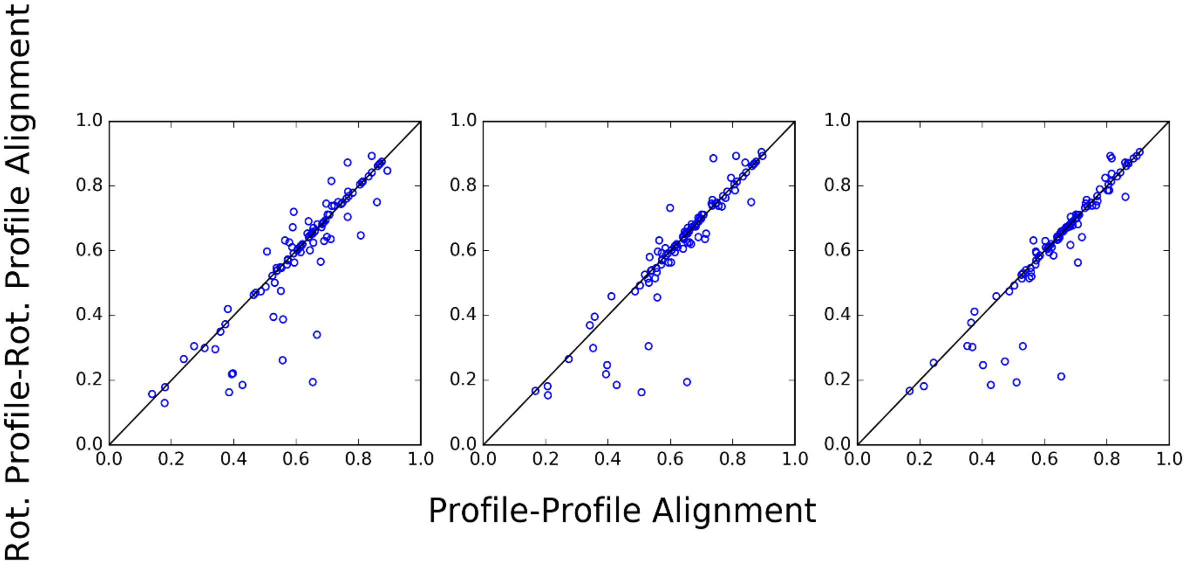


**Figure 2.7: Best TM-scores of Top N alignments by sequence-based versus rotamer-sequence based profiles**. Best TM-score in the top N alignments calculated by a rotamer-profile/rotamer-profile alignments and profile-profile alignments were determined for a dataset of 100 proteins. Left: N=1; middle: N = 2; right: N = 5.

### 2.4 Discussion

In this study, we describe the construction and performance of structurally-derived $\chi^1$ rotamer sequence-based substitution matrices (also known as ROTSUM). While the ROTSUM matrices are symmetric and its elements are expressed as log-odds, there are major intrinsic differences between them and previously used substitution matrices. First, the obtained frequencies originate from structure alignments by TM-align instead of a sequence database. Interestingly, the non-redundant database used for this experiment is significantly limited compared to that of sequence database in both size and variety. Though the structure database covers most, if not all, folds created by nature, we may not have a representative set of proteins in this database especially for those proteins that are difficult to experimentally validate. The accuracy of the structural database is perhaps also something to consider. Low-resolution crystal structures or NMR structure may not capture the intended $\chi^1$ rotamer relevant for its function. However, even structures with low assignment accuracy are also identified by use of ROTSUM.

Moreover, alignments generated by this structural alignment approach do not require the sequence of either structure to be known beforehand. Finally, the 55x55 matrix-form of ROTSUM matrices serves a unique purpose for $\chi^1$ rotamers although a 20x20 form can be utilized for canonical, amino-acid sequence alignments.

Notably, not all rotamers for a given amino-acid type are not substituted for other rotamer types at the same rate, especially bulky aromatic residues. This is perhaps expected due to their inability to move freely in their relative positions in proteins. Especially in protein design, it is fatal to consider that all three $\chi^1$ rotamer types for these aromatic residues should be given the same weight. Thus, considering discrete $\chi^1$ rotamers is more appropriate for this task. We also discover that, generally, many rotamers with similar physiochemical properties tend to mutate to other types who share similar $\chi^1$ angles. For example, the score which describes the mutation from ASP1 to GLU1 is 2.7 bits, while mutating that same rotamer to either GLU2 or GLU3 grants a score of 0.1 bits and 0.6 bits, respectively. Interestingly, the evolution of ASP1 to ASP2 or ASP3 is 3.6 bits and 2.8 bits, respectively. This suggests that the flexibility of the $\chi^1$ angle to switch to its other $\chi^1$ modes is on the same order as mutating to a glutamate residues whose rotamer shares a similar $\chi^1$ dihedral angle.

Rotamer sequence alignment provides greater performance in the substitution matrix compared to basic amino-acid types. Obviously, rotamer alignment cannot be realized unless we know the structure of the protein's sidechains. In Chapter 3, we explore methods to predict the $\chi^1$ rotamer sequence of a protein given only its amino-acid sequence information.

Online server support for $\chi^1$ rotamer alignments and the substitution matrices constructed in this study are available for academic use at https://zhanglab.ccmb.med.umich.edu/Rot1Suite/.

## 2.5 References

1. Schwartz, R.M.D., M.O, *Matrices for detecting distant relationships.* Atlas of protein sequence and structure, 1978. **5**: p. 353-58.
2. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.
3. M.O. Dayhoff, R.M.S., B.C. Orcutt, *A Model of Evolutionary Change in Proteins.* Atlas of protein sequence and structure, 1978. **5**.
4. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
5. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
6. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins, 2004. **57**(4): p. 702-10.
7. Zhang, Y. and J. Skolnick, *TM-align: a protein structure alignment algorithm based on the TM-score.* Nucleic Acids Res, 2005. **33**(7): p. 2302-9.
8. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
9. Dunbrack, R.L., Jr. and F.E. Cohen, *Bayesian statistical analysis of protein side-chain rotamer preferences.* Protein Sci, 1997. **6**(8): p. 1661-81.
10. Zhang, Y., *Progress and challenges in protein structure prediction.* Curr Opin Struct Biol, 2008. **18**(3): p. 342-8.
11. Bowie, J.U., R. Luthy, and D. Eisenberg, *A method to identify protein sequences that fold into a known three-dimensional structure.* Science, 1991. **253**(5016): p. 164-70.
12. Wu, S. and Y. Zhang, *MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information.* Proteins, 2008. **72**(2): p. 547-56.
13. Stormo, G.D., et al., *Use of the Perceptron Algorithm to Distinguish Translational Initiation Sites in Escherichia-Coli.* Nucleic Acids Research, 1982. **10**(9): p. 2997-3011.
14. Mitra, P., D. Shultis, and Y. Zhang, *EvoDesign: De novo protein design based on structural and evolutionary profiles.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W273-80.
15. S. Kullback, R.A.L., *On Information and Sufficiency.* The Annals of Mathematical Statistics, 1951. **22**(1).
16. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J Mol Biol, 1970. **48**(3): p. 443-53.
17. Xu, J. and Y. Zhang, *How significant is a protein structure similarity with TM-score = 0.5?* Bioinformatics, 2010. **26**(7): p. 889-95.
18. Pietrokovski, S., J.G. Henikoff, and S. Henikoff, *The Blocks database--a system for protein classification.* Nucleic Acids Res, 1996. **24**(1): p. 197-200.
19. Gotoh, O., *An improved algorithm for matching biological sequences.* J Mol Biol, 1982. **162**(3): p. 705-8.
20. Panchenko, A.R., *Finding weak similarities between proteins by sequence profile comparison.* Nucleic Acids Res, 2003. **31**(2): p. 683-9.

# Chapter 3: Categorical and Real-value Prediction of Chi-1 Rotamers from Sequence

## 3.1 Introduction

To date, there are approximately 24 million non-redundant entries in the UniProt database [61] yet only about 100,000 in the Protein Databank [9]. This large and growing discrepancy is making even more important to confidently match sequence and structure pairs [10]. This issue has been addressed in multiple facets including prediction of protein structural properties including phi- and psi-torsion angles in fold-recognition strategies. Understanding more probable structural properties of the protein can help identify template homologs in structure prediction. If reliably predictable, even sidechain orientation may be worthwhile to explore to identity if and any impact on fold recognition is significant compared to current methods.

Knowledge of protein backbone torsion angles can often provide more information about topological units across the structure, such as secondary structure, in its fold [2]. Additionally, side chains can adopt its own set of chi ($\chi$) dihedral angles where the size of the set is dependent on the length of the sidechain [62]. Due to steric hindrance in most cases, the sidechain's gamma atom is naturally positioned away from the backbone carbonyl group, it's been observed that the first order of chi angles ($\chi$1—defined by the four atoms N, C$\alpha$, C$\beta$, [C/S/O]b$\gamma$) is discretized and can only adopt a relatively small number of states [30] . For the same reason, these discrete rotamer types are not equally distributed. We explore the definition of rotamer types relevant for the experiments in Section 2.1.

Sequence-based torsion angle prediction is an area of structural prediction that has yet to be fully explored. Backbone phi/psi angles from sequence information have been studied previously using artificial neural networks and/or support vector machines [63, 64]. However, $\chi1$ rotamer prediction is another aspect of the problem that has not received sufficient attention. To this date and to the author's knowledge, there are no algorithms that are trained to predict rotamer configuration from sequence alone. Typically, algorithms such as SCWRL [65, 66] require exact backbone coordinates to predict the atomic positions of the residue sidechains. Software that doesn't rely on backbone coordinates do exist, however, they use additional information outside of sequence alone including chemical shifts from NMR experiments as seen in PREDITOR [67]. Here, although PREDITOR allows sequence-only submissions, the algorithm will use the most similar template's $\chi1$ rotamer information from the Protein databank. This approach doesn't yield high accuracy for many proteins, especially those with low- to no-sequence homologs.

In this study, we address the rotamer prediction problem using a sequence and secondary structure fragment-based approach to predict the $\chi1$ rotamer for each position of a sequence query. We also set out to predict its real value $\chi1$ dihedral angles at each position and compare it to torsion angles determined by a backbone coordinate-dependent method, SCWRL4 tested on I-TASSER-modelled structures [40].

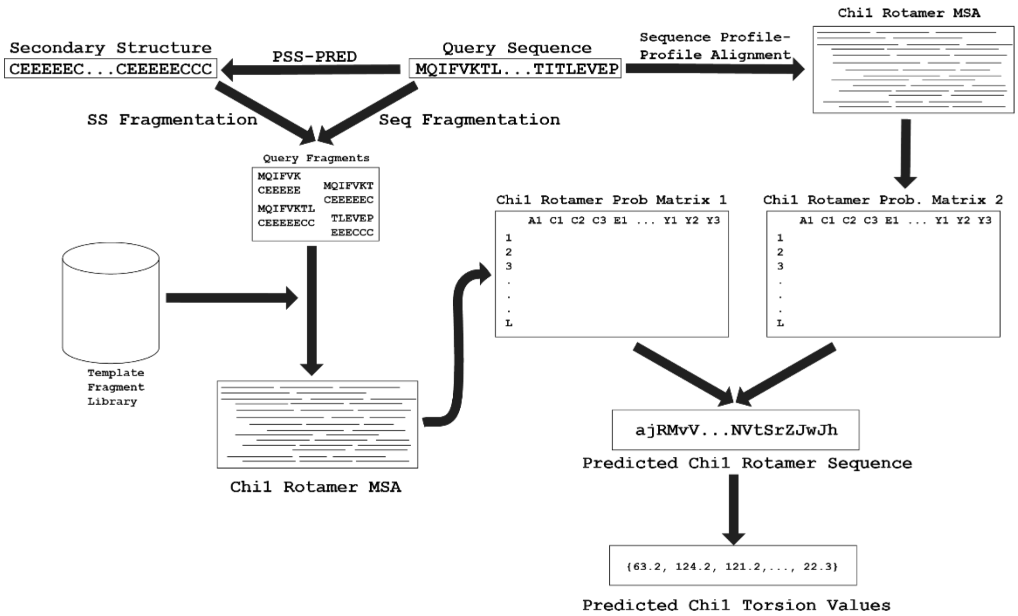


**Figure 3.1.** Rot1Pred Pipeline. Two chi-1 probability distribution matrices are generated from sequence-dependent sequence/secondary-structure alignments and profile-profile alignments.

## 3.2 Methods

### 3.2.1 Datasets

The template library used in this study is made from high resolution structures (<2.5A) that do not share sequence identity of more than 70%. PISCES [68] server was used to automatically generate the list of approximately 26,000 proteins that satisfy this criterion. Secondary structure elements and the χ1 rotamer sequence for the template structures are calculated by STRIDE [69] and Rot1Calc, respectively. Of these 26,000 proteins, 300 proteins were randomly selected as the test set. The PDBID for each dataset item is included in Appendix B. Unless specified otherwise, for each test set item, all proteins in the template library that share more than 30% sequence identity by Needleman-Wunsch sequence alignment [21] were removed for our benchmark results. Also, we generated structural models for the testing set proteins using I-TASSER, where we identified and excluded homologous templates with a sequence identity greater than 30%.

### 3.2.2 Rot1Pred: Fragment Alignment and Prediction

Starting from solely sequence information, sequence fragments are generated with varying sizes of 6 to 30 amino acids starting from every position. The secondary structure element of each position is predicted by PSS-PRED [70]. For each position across a sequence, we generate these fragments and perform ungapped sequence alignments against fragments of the template library with the same window size. A score between aligned positions are generated based on a function that is a weighted linear-combination of amino-acid type and secondary structure type match/mismatch (Eqn. 1):

$$S\left(aa_q, aa_t, ss_q, ss_t\right) = M\left(aa_q, aa_t\ \right) + 2 \times SSmatch(ss_q, ss_t) \tag{3.1}$$

Symbols *aa* and *ss* are the amino acid type and secondary structure elements of a query fragment *q* at position *i* and template fragment *t* at position *j*, respectively. *M* represents the 20x20 collapsed form of the ROTSUM95 substitution matrix. Details of this matrix (COLLMX95) is shown in Appendix C. The score of two aligned fragments is, thus, the sum of the positional pairwise scores (Eqn. 2).

$$S_{frag} = \sum_{z=1}^{N} S(aa_{q,z}, ss_{q,z}, aa_{t,z}, aa_{t,z}) \tag{3.2}$$

*N* is the length of the fragment and *z* is the relative position of the fragment length.

As the raw-score is typically biased on the length of the fragment, we determine the significance of a score against all others of the same size using Z-score (Eqn. 3.3):

$$Z_{frag} = \frac{S_{frag} - < S_{frag,N} >}{\sigma_N} \tag{3.3}$$

where $S_{frag}$ represents the score of the fragment, $<S_{frag,N}>$ represents the average score of all fragments that share the same size $N$, and $\sigma_N$ represents the standard deviation of these scores. All Z-scores from all fragment sizes are then pooled and sorted. For each position of the query, we select fragments with the 125 top Z-scores and record the frequency of each χ1 rotamer type aligned.

This process is repeated for each position in the protein. At the end of this process, we are given an Lx55 matrix that represents the distribution of χ1 rotamer types at every position. This frequency matrix is then converted to a probability matrix and adjusted by a pseudocount. The calculation of the pseudocount is derived from the equation introduced by PSIBLAST [53], but instead of using a 20x20 sequence substitution matrix, we provide a χ1 rotamer substitution matrix, named ROTSUM. The details of the construction of the ROTSUM matrix and its effects on sequence alignments will be showcased in a future study. The ROTSUM matrix used for this adjustment is downloadable from Rot1Pred website. From the adjusted probability matrix, we finally select the highest value for a specific amino-acid type for all query positions and report its predicted rotamer types.

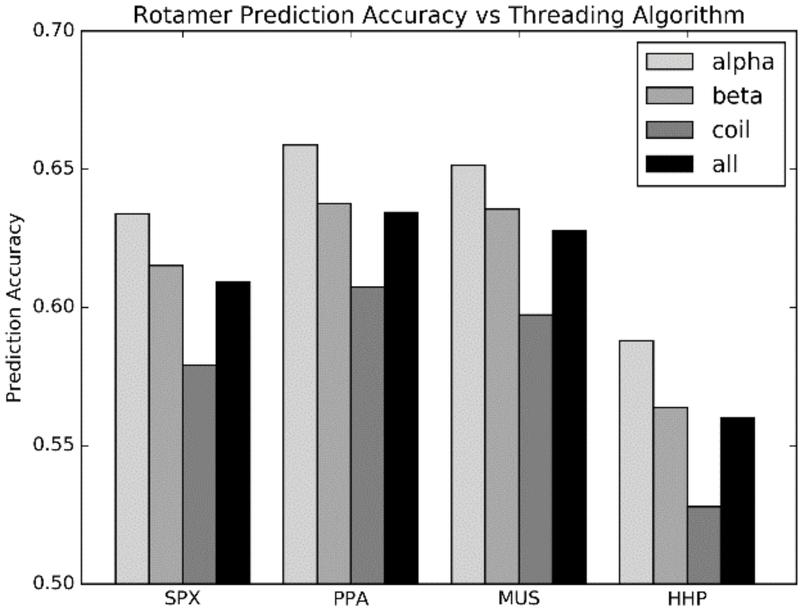


**Figure 3.2.** The effect of various fold-recognition algorithms on categorical $\chi 1$ torsion angle prediction. The prediction accuracy is determined from the most probable rotamers in a multiple rotamer-sequence alignment generated by the individual threading algorithms (Sparks-X, Profile-Profile alignment, MUSTER, and HHPred) compared to the $\chi 1$ rotamer in the native structure. Accuracies are calculated for different secondary structure elements (α helix, β sheet, and coil) and are also calculated for all positions.

### 3.2.3 Augmentation of Protein Threading with Rot1Pred

Since we select sidechain conformers from sequence alignment information, we included a new alignment source from protein sequence-structure threading programs that also require only sequence information as input. Four types of threading programs were used to generate these query-anchored sequence alignments: protein profile-protein alignment + secondary structure match (PPA), SPARKS-X (SPX) [71], MUSTER (MUS) [56], and HHPred (HHP) [72]. The generated sequence alignments are trivially converted from amino-acid to $\chi 1$ rotamer sequence alignments. New probability distributions, $Q$ are constructed by a weighted average between the adjusted probabilities in Section 2.3 and the probabilities from the threading alignments (Eqn. 4.4):

$$K_{aa,i} = \max(w_1 P_{frag_{aa,i,k}} + w_2 P_{thd_{aa,i,k}}) \tag{3.4}$$

Additionally, Rot1Pred can predict the real value of the dihedral angle. For a given query, we scale each of the angles in the top 125 fragments by their fragments' Z-score. All angles that belong to a range [0, 120), [120, 240), or (-120, 0]—except for Proline (which are separately considered) are then summed and divided by the sum of the Z-scores, yielding a "weighted average" of the χ1 dihedrals (Eqn. 5):

$$\overline{X}_{i,k}^{1} = \frac{\sum_{n=1}^{N} Z_{frag,i,n,k} \times X_{i,n,k}^{1}}{\sum_{n=1}^{N} Z_{frag,i,k}} \tag{3.5}$$

Here, *i* refers to the query position, *k* refers to the rotamer class, and *n* is the number of total rotamer characters that align to the *i*th position. The selection of *k* is determined by the most observed rotamer class seen for the query's amino acid type at position *i*. Details of the final class selection is outlined in section 2.4.

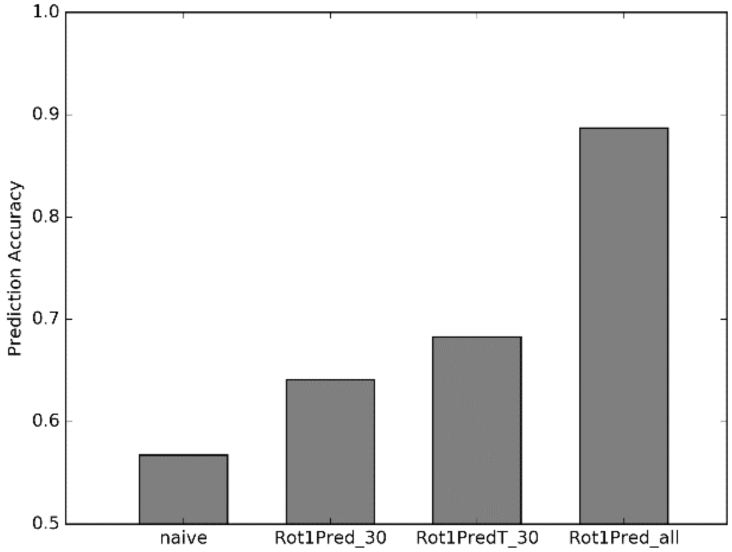


**Figure 3.3.** Summary of Rot1Pred performance on categorical chi-1 rotamer identification. Prediction accuracy was determined by this study's approach using either '30'% sequence identity cutoff for included templates in the Protein Databank or 'all' templates included.

### 3.2.4 Rot1Pred Parameterization

Several criteria were considered for tuning to optimize prediction accuracy on the training set. The following parameters were tuned and the optimum value is shown in parenthesis: minimum window size, maximum window size, relative weighting of amino-acid type and secondary structure match in fragment similarity searching, top N Z-score fragments to consider for multiple rotamer sequence alignment, and beta constant in weighting the impact of the substitution matrix in pseudocount generation. Specifically, for threading-enabled rotamer prediction, we selected specific criteria on template hits to be considered for its multiple sequence alignment: Z-score of the alignment, template structure resolution, and weighting of the adjusted probability table from fragment similarities and from threading-enabled alignments.

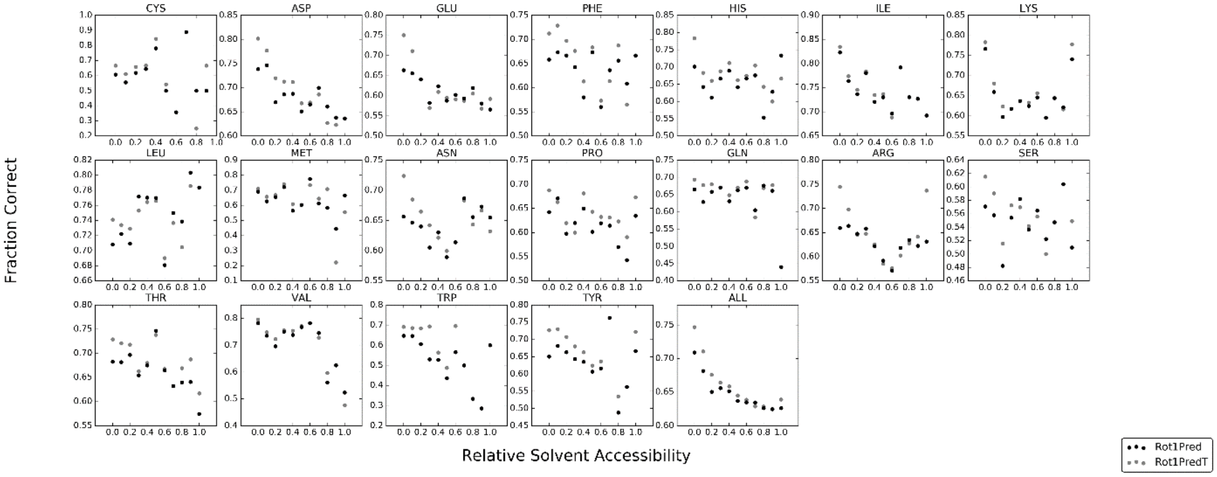


**Figure 3.4.** Effects of Sidechain solvent accessibility on prediction accuracy. Solvent accessibility for each residue type in the dataset proteins are calculated by STRIDE. Residues in the test set are binned by its amino-acid type and relative solvent accessibility in increments of 0.1. The fraction of correct rotamers for each bin are then calculated.

### 3.2.5 Prediction of Real-value $\chi$1 Dihedrals on Predicted Protein Structures

The structure of each protein in the testing set was predicted by I-TASSER. In the search of structural templates by LOMETS [73], we removed all templates whose sequence identity is over 30% of the query sequence. Following the I-TASSER simulation and clustering stage, we use the largest cluster's decoy closest to the cluster center. Since the output of this decoy is only a Cα trace, we employ Pulchra [74] to construct the other backbone atoms.

## 3.3 Results

### 3.3.1 Parameterization of Fragment Alignment Scoring Function

The choice of the range of fragment lengths used is crucial to obtain pertinent structural information from the template library. Fragments should be long enough to obtain specific data regarding the query sequence. However, fragments too long tend to compound error within the gapless alignment and will also render a temporal impact. Figure 6 illustrates an approach to optimize the range of fragment sizes used for this pipeline. While there is a clear optimum for the minima fragment size, the maximum fragment size steadily increases in accuracy. Since there are negligible diminishing returns after a size of 30, this value was chosen as the most practical considering a trade-off between accuracy and time.

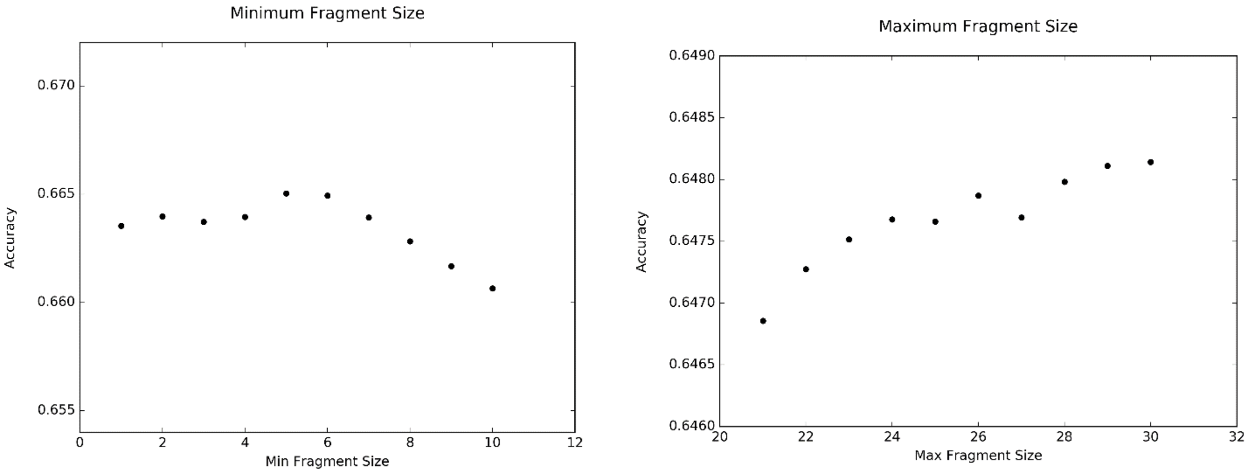


**Figure 3.5.** The effect of minimum and maximum fragment size on categorical X1 torsion angle prediction. Minimum (left) and maximum (right) fragment size.

From tens of millions of alignments generated, it’s necessary to determine which alignments qualify as valuable to help in determining structural information for the query. One approach consists of taking the top N scoring alignments and extracting their rotamer information and determining which N provides the greatest prediction accuracy. Figure 7 illustrates the results from this proposal. As the number of chosen alignments increase, so does the accuracy. At N = 250, there is no apparent drop-off in accuracy. This is likely due to the weighting of the alignments. Those with very low Z-scores will contribute little toward the rotamer prediction while the converse will have a much greater prediction impact. As there is virtually no improvement from

N = 125 to N = 250, we consider the top 125 scoring fragments in favor of saving computational memory.

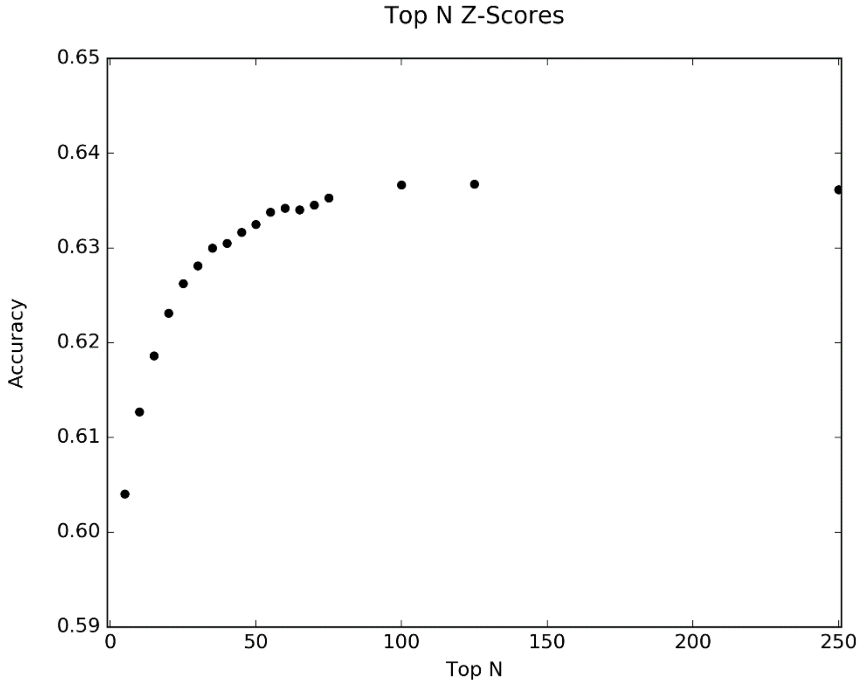


**Figure 3.6.** The effect of top N selection of scored fragments on prediction accuracy.

Resolution from crystallographic results correlates well with the accuracy of spatial location of the protein's atoms. Although resolutions <2.5 Ångstroms can typically resolve the structure of rotamers well [9], a systematic approach to determine the optimum resolution cutoff for this pipeline should exist. Thus, we performed our algorithms using several template libraries where the entries in each set were below a certain Ångstrom cutoff. Figure 3.7 displays the resulting prediction accuracy for each resolution cutoff. Although, a template library with the best resolved structures would be ideal, there isn't, currently, sufficient data to well-inform the probability matrix. For instance, the template library with a resolution cutoff of 1.5 Ångstroms contains only 4911 proteins while the library with resolution cutoff of 2.1 Ångstroms contains 19,924 entries.

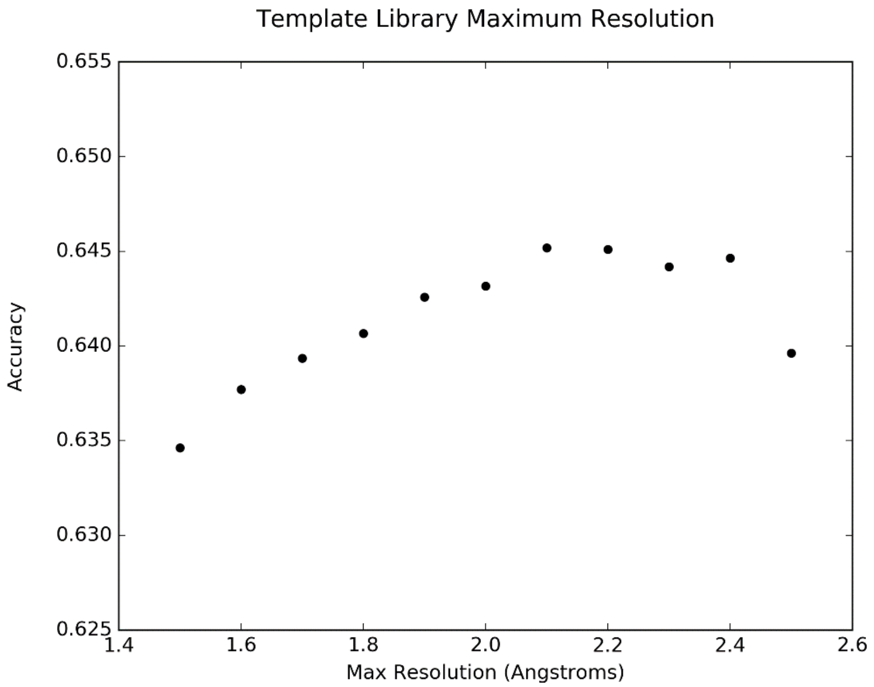


**Figure 3.7.** The effect of template library resolution cutoff on prediction accuracy.

Pseudocount plays a tremendous role in smoothing probability distributions for cases where observed frequencies are low. Two tunable variables are considered for Rot1Pred: the beta constant used to describe the relative weighting of the pseudocount and the substitution matrix where the pseudocounts are sampled from (see Eqn 5 in [53]). Figure 3.8 and 3.9 illustrate the varying beta constants and different forms of the COLLMX matrix (See Section 2.3), respectively.

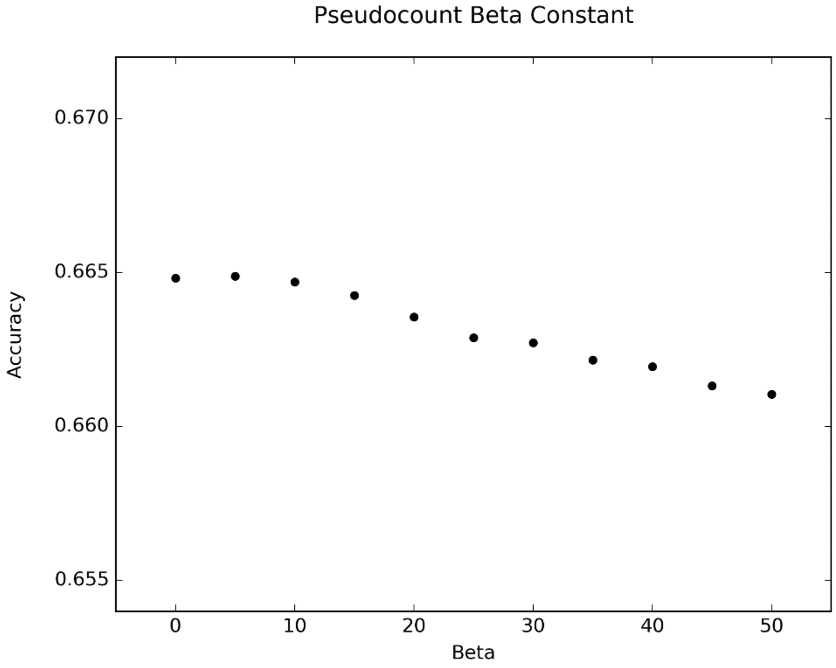


**Figure 3.8.** The effect of the pseudocount beta constant on prediction accuracy.

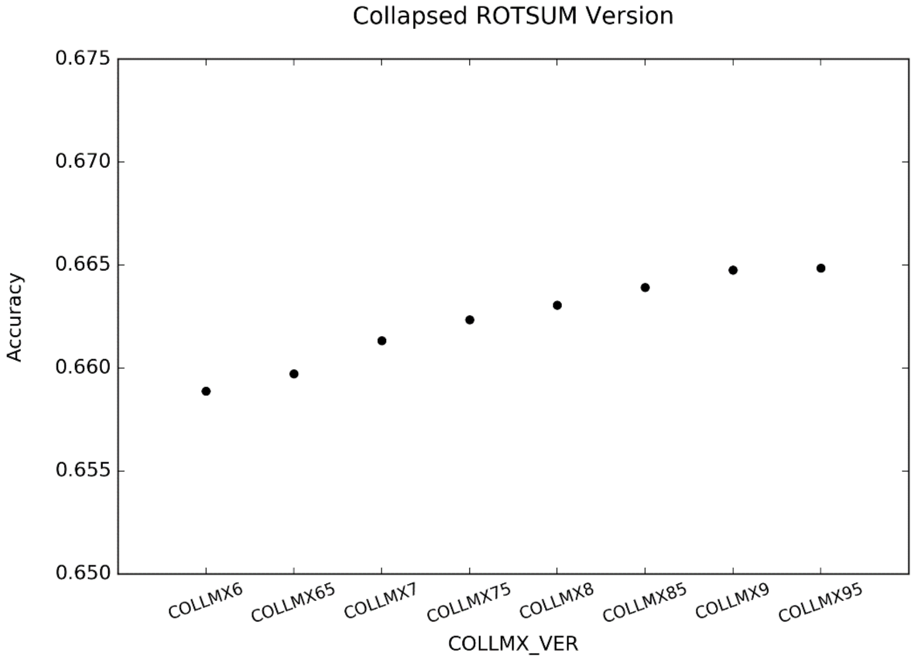


**Figure 3.9.** The effect of COLLMX version used for pseudocount generation on prediction accuracy.

### 3.3.2 Accuracy of Rot1Pred in Rotameric Classification

The evaluation of our methods' performance is quantified by the comparison between the rotamer sequences generated by approaches explained in Section 2 versus rotamer sequences constructed by a naïve approach as described later. Alanine and glycine residues are excluded from the analyses due to the lack of a sidechain gamma atom. We compute the accuracy of Rot1Pred's classification of a side chain's rotamer state from the test set of 300 proteins mentioned previously. Since local sequence alignments are used to determine the significance of fragments, for each query sequence, we remove homologous templates whose sequence identity are over 30%.

Naturally, the probability for each $\chi 1$ rotamer type for a given amino acid type is not equally distributed. One can naturally construct a rotamer sequence as the combination of the most observed rotamer type in the Protein Databank for each amino acid type in a query. This approach is labelled as 'naïve' in the following analyses and discussions. The naïve approach yields a prediction accuracy of 56.8% on the testing set, and we use this result as a baseline when considering Rot1Pred's performance.

### 3.3.3 Chi-1 Rotamer Sequence Prediction by Threading Algorithms

As shown in Figure 3.2, all fold-recognition techniques yielded accuracies well-above that of the naïve approach. Profile-profile alignments (PPA) provided the highest reliability (63.4%) compared to its other counterparts. The strength of PPA perhaps derives from its high threading coverage allowing rotamer types to be inferred from relevant, related structures.

Approaches from Section 2.3 and 2.4 were combined to provide a threading-assisted technique for $\chi 1$ rotamer prediction labelled as *Rot1PredT* throughout this study. The performance for both

Rot1Pred and Rot1PredT are summarized in Figure 3.3. Our benchmark results show that this approach detailed in this study can correctly identify 64.1% or 68.3% of the $\chi 1$ rotamers with homologous templates removed. For the purposes of using this approach with the entire template library, 88.72% accuracy was achieved.

We explored the effects of fusing probabilities from sequence/secondary structure fragments and threading alignments across different levels of sidechain solvent exposure. In Figure 3.4, the fraction of correct rotamers were determine for a sidechain's relative solvent accessibility ranging from 0.0 to 1.0 in 0.1 bin increments. For most amino-acid types, as expected the prediction capability degrades as the sidechain becomes more solvent-exposed. Considering all amino-acid types in the 'ALL' plot, the correlation is smoother perhaps due to sufficient sample number in the solvent accessible bins. Here, residues that typically reside deep in the core are much easier to predict compared to naturally solvent-exposed residues. This phenomenon is not new or unique. Other sidechain prediction algorithms like SCRWL, which even when a native backbone is supplied, cannot determine the correct conformational isomer for highly exposed residue types compared to others.

|  | Rot1Pred[a] | Naïve[b] (*p*) | Rand_Within[c] (*p*) | Rand[d] (*p*) | SCRWL4[e] (*p*) |
|---|---|---|---|---|---|
| CYS | 45.95° | 48.00° (4.7E-16) | 82.62° (3.2E-24) | 89.27° (1.7E-31) | 59.60° (7.4E-07) |
| ASP | 43.09° | 46.03° (4.4E-76) | 87.58° (9.0E-162) | 91.03° (8.6E-151) | 59.60° (1.9E-30) |
| GLU | 51.51° | 54.47° (1.6E-105) | 91.15° (4.8E-153) | 89.78° (5.2E-138) | 62.26° (8.5E-19) |
| PHE | 40.26° | 41.78° (3.1E-25) | 86.49° (1.5E-119) | 90.47° (1.7E-116) | 59.17° (5.6E-28) |
| HIS | 45.00° | 46.70° (7.6E-17) | 87.35° (2.8E-56) | 88.37° (2.5E-53) | 58.41° (1.3E-08) |
| ILE | 29.55° | 30.63° (4.3E-22) | 90.78° (2.2E-234) | 89.28° (2.5E-225) | 50.89° (7.3E-66) |
| LYS | 49.06° | 51.69° (3.6E-79) | 91.45° (6.4E-146) | 89.81° (5.0E-138) | 62.48° (9.8E-27) |
| LEU | 37.77° | 40.53° (2.0E-142) | 89.74° (1.0E-304) | 91.16° (4.4E-293) | 54.43° (1.1E-61) |
| MET | 44.71° | 47.50° (1.1E-27) | 88.95° (4.7E-46) | 88.09° (3.7E-38) | 60.89° (4.0E-09) |
| ASN | 47.33° | 50.30° (2.3E-55) | 88.95° (1.3E-94) | 91.05° (5.8E-94) | 61.06° (2.1E-16) |
| PRO | 22.72° | 22.83° (4.8E-01) | 56.69° (1.6E-136) | 90.45° (8.6E-223) | 26.68° (7.3E-13) |
| GLN | 46.09° | 48.86° (6.6E-53) | 88.61° (7.4E-87) | 92.84° (3.5E-100) | 61.45° (8.1E-22) |
| ARG | 50.13° | 52.92° (3.8E-65) | 90.10° (1.0E-106) | 88.64° (2.0E-96) | 64.25° (7.4E-23) |
| SER | 62.60° | 63.15° (3.6E-07) | 77.00° (1.6E-56) | 91.02° (9.2E-61) | 77.77° (1.2E-18) |
| THR | 43.13° | 42.50° (4.1E-16) | 74.53° (4.3E-127) | 88.74° (5.2E-130) | 62.70° (1.2E-17) |
| VAL | 35.21° | 35.92° (5.7E-25) | 82.07° (1.0E-262) | 91.01° (8.7E-246) | 55.80° (8.3E-49) |
| TRP | 47.16° | 48.92° (3.0E-09) | 87.05° (7.6E-30) | 92.90° (2.3E-30) | 61.43° (4.7E-07) |
| TYR | 44.60° | 45.85° (1.4E-16) | 87.68° (1.2E-80) | 84.18° (1.5E-64) | 57.45° (9.2E-11) |
| ALL | 43.37° | 45.19° (<1.0E-309) | 85.23° (<1.0E-309) | 90.05° (<1.0E-309) | 58.81° (<1.0E-309) |

**Table 3.1.** MAE of torsion angle prediction on I-TASSER modelled backbones.
*Torsion Angle Determinations:*
[a]*Rot1Pred:* Rot1PredT algorithm
[b]*Naïve*: Most observed probability for the predicted chi-1 category type
[c]*Rand_Within*: Random angle within the given range of the predicted chi-1 category type (See section 2.1)
[d]*Rand:* Random angle between 0.0° inclusive and 360.0° exclusive.
[e]*SCWRL4*: SCWRL4 algorithm

### 3.3.4 Accuracy of Real Value Dihedral Prediction

The performance comparison between Rot1Pred and SCWRL4 on I-TASSER-modelled backbones is displayed in Table 3.1 and Table 3.2. The performance criteria are based on two calculations: mean absolute error (Eqn. 6) and the root mean squared deviation (Eqn. 3.7) of the $\chi$1 torsion angle and gamma atom prediction between the prediction and native values, respectively.

$$MAE = \frac{1}{\sum_{i=1}^{M} L_i} \sum_{i=1}^{M} \sum_{j=1}^{L_i} |P_{ij} - N_{ij}| \tag{3.6}$$

$$RMSD = \sqrt{\frac{1}{\sum_{i=1}^{M} L_i} \sum_{i=1}^{M} \sum_{j=1}^{L_i} d_{\gamma}^2} \tag{3.7}$$

In Eqns. 3.6 and 3.7 *M* is the number of proteins in the dataset, $L_i$ is the total number of residues in protein *i*, and *j* is an index in protein *i*. In Eqn. 3.6, *P* and *N* represent the predicted and native torsion angle, respectively, and are in the range of [0, 360). In Eqn. 3.7, the distance between the predicted and native gamma atoms after the backbone atoms (N, Cα, C) and the beta-carbon (Cβ) have been superposed is represented as $d\gamma$. Since angles can assume multiple representations for the same orientation (e.g. -20 and 340), we calculate the absolute distance of the smallest angle by Eqn. 3.8:

$$D' = abs(X1 \; mod \; 360 - X2 \; mod \; 360) \tag{3.8a}$$

$$D_{smallest} = \min(D', 360 - D') \tag{3.8b}$$

Construction of the sidechain beta atom is done by equations outlined previously [75]. Sidechain τ angles and the Cα-β atom distance used in these equations are defined by averaging these values for each chi-1 rotamer type in the PDB.

Rot1Pred's prediction for the χ1 dihedral is more accurate when tested on a dataset of I-TASSER modelled backbones. Statistical values calculated in this section were performed via Wilcoxon signed-rank test. The mean absolute error for the dihedrals predicted for all amino acids are 37.13°

and 50.35° for Rot1Pred and SCWRL4, respectively, and for each amino acid type, the dihedral was predicted by Rot1Pred much more significantly than SCWRL4. Rot1Pred's predicted torsion angles were also compared to randomly generated angles under three conditions: most probable angle of the $\chi 1$ category (Naïve), random angle within the $\chi 1$ category predicted by Rot1Pred (*Rand_Within*), and a random angle selected within the range [0.0, 360.0) (*Rand)*. The comparison between Rot1Pred and Naïve highlights the quality of the real-value determination given the $\chi 1$ category. Even though the mean absolute error between these two conditions only differ by only less than two degrees, the mean difference is extremely significant (p-value <1.0E-309). As expected, protein core-occupying residues (leucine, isoleucine, and valine) generally exhibit a lower MAE than other types. Proline's MAE result is unique compared to other types due its ability to only adopt two $\chi 1$ rotamer states with average dihedral angles of +/- 25°. Rot1Pred accurately predicts the location of the gamma atom significantly across all amino acid types. Across all query positions in the testing test, the average gamma atom RMSD is 0.95Å or 1.25Å ($p < 1.0E\text{-}309$) for Rot1PredT and SCWRL4, respectively. This result suggests that Rot1Pred can better identify the position of the native gamma atoms than an algorithm which already has all backbone coordinate information. Notably, the method described in this study does not require the backbone coordinates to determine the $\chi 1$ torsion angle, but instead uses similar fragments found in the PDB. Interestingly, however, these fragments are not necessarily related especially since homologous templates (defined by sequence identity greater than 30 percent) were removed before prediction. One could expect even greater accuracy with all templates allowed. The major tradeoff for accuracy, however, is the time calculation since sequence fragments need to be searched compared to the near instant prediction by SCWRL.

|  | Rot1Pred | Naïve (*p*) |
|---|---|---|
| CYS | 1.18 Å | 1.50 Å (5.6E-09) |
| ASP | 0.95 Å | 1.25 Å (3.9E-39) |
| GLU | 1.12 Å | 1.32 Å (7.0E-12) |
| PHE | 0.87 Å | 1.24 Å (1.3E-38) |
| HIS | 0.96 Å | 1.22 Å (7.5E-13) |
| ILE | 0.69 Å | 1.13 Å (1.6E-96) |
| LYS | 1.08 Å | 1.32 Å (5.7E-12) |
| LEU | 0.83 Å | 1.16 Å (2.5E-77) |
| MET | 0.97 Å | 1.29 Å (3.1E-11) |
| ASN | 1.04 Å | 1.29 Å (1.3E-17) |
| PRO | 0.62 Å | 0.69 Å (7.4E-03) |
| GLN | 1.01 Å | 1.29 Å (6.9E-18) |
| ARG | 1.09 Å | 1.35 Å (7.4E-19) |
| SER | 1.24 Å | 1.51 Å (1.4E-10) |
| THR | 0.90 Å | 1.26 Å (6.9E-17) |
| VAL | 0.81 Å | 1.23 Å (5.2E-112) |
| TRP | 0.99 Å | 1.27 Å (9.5E-10) |
| TYR | 0.96 Å | 1.21 Å (2.1E-17) |
| ALL | 0.95 Å | 1.25 Å (<1.0E-309) |

**Table 3.2**. RMSD of Predicted Gamma Atom.

### 3.4 Discussion

We developed a sequence fragment-based Rot1Pred, for *ab initio* prediction of rotamer classes and/or real value torsion angles of the χ1 angle. The purpose of this work is two-fold: (1) predict χ1 rotamer sequences for purposes of sequence-structure alignments in potential threading programs, and (2) generate real value χ1 dihedrals for sidechain orientation determination in tertiary protein structure prediction. More details about the prospect of generating these alignments is currently being explored. The prediction time for this algorithm takes approximately 1.5 CPU seconds per amino acid in the sequence on an x86 Intel 3.6 GHz processor. The webserver, executables, and source of the algorithms discussed here are found on the Rot1Suite and are available for users online at our webserver: https://zhanglab.ccmb.med.umich.edu/Rot1Pred.

We investigated two ways of validating this algorithm. First, the accuracy of rotameric classification were compared to naïve approaches a backbone-dependent sidechain program. Next, I-TASSER generated models were constructed and partial sidechains were generated by Rot1Pred and SCWRL4. Performance of the structural prediction was measured as the RMSD between native gamma atoms.

Although determination of the $\chi1$ category provides an idea of the orientation of the side-chain, the detailed determination of the angle must also be considered. Without this consideration, one could simply choose the most probable angle for a given predicted $\chi1$ category; however, as shown in Table 3.1, even a very minute alteration in the angle may confer a significant sidechain conformational change. This also supports the fact that $\chi1$ angles are discretized and their torsion angle probability distribution is extremely tight.

Rot1Pred performs well in predicting the real value $\chi1$ rotamer compared to backbone-dependent methods. However, since the approached explained here does not consider rotamer-pair interactions explicitly, some steric clash can be expected. It is suggested that Rot1Pred should be applied to predicted models rather than native ones and further tuning by sidechain refinement strategies may be necessary. If the user has information of the native backbone already at-hand, SCWRL4, or similar algorithms, are suggested to be used instead. Users should also consider run-time effects in predicted $\chi1$ rotamers with Rot1Pred. Since a fragment-based approach is orders of magnitude slower than conventional backbone-dependent methods, one should consider the context for which the prediction algorithm is being used. Especially in cases where a full protein structure is predicted from sequence alone, Rot1Pred would be highly suitable for sidechain prediction. As the gap between entries submitted in sequence and structure databases continues to widen, it is becoming more important to construct algorithms to provide structure information from sequence alone.

## 3.5 References

1. Apweiler, R., et al., *UniProt: the Universal Protein knowledgebase.* Nucleic Acids Res, 2004. **32**(Database issue): p. D115-9.
2. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
3. Zhang, Y., *Progress and challenges in protein structure prediction.* Curr Opin Struct Biol, 2008. **18**(3): p. 342-8.
4. Ramachandran, G.N., C. Ramakrishnan, and V. Sasisekharan, *Stereochemistry of polypeptide chain configurations.* J Mol Biol, 1963. **7**: p. 95-9.
5. Dunbrack, R.L., Jr. and M. Karplus, *Backbone-dependent rotamer library for proteins. Application to side-chain prediction.* J Mol Biol, 1993. **230**(2): p. 543-74.
6. Dunbrack, R.L., Jr. and F.E. Cohen, *Bayesian statistical analysis of protein side-chain rotamer preferences.* Protein Sci, 1997. **6**(8): p. 1661-81.
7. Wu, S. and Y. Zhang, *ANGLOR: a composite machine-learning algorithm for protein backbone torsion angle prediction.* PLoS One, 2008. **3**(10): p. e3400.
8. Song, J., et al., *TANGLE: two-level support vector regression approach for protein backbone torsion angle prediction from primary sequences.* PLoS One, 2012. **7**(2): p. e30361.
9. Krivov, G.G., M.V. Shapovalov, and R.L. Dunbrack, Jr., *Improved prediction of protein side-chain conformations with SCWRL4.* Proteins, 2009. **77**(4): p. 778-95.
10. Miao, Z., Y. Cao, and T. Jiang, *RASP: rapid modeling of protein side chain conformations.* Bioinformatics, 2011. **27**(22): p. 3117-22.
11. Berjanskii, M.V., S. Neal, and D.S. Wishart, *PREDITOR: a web server for predicting protein torsion angle restraints.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W63-9.
12. Zhang, Y., *I-TASSER server for protein 3D structure prediction.* BMC Bioinformatics, 2008. **9**: p. 40.
13. Wang, G. and R.L. Dunbrack, Jr., *PISCES: a protein sequence culling server.* Bioinformatics, 2003. **19**(12): p. 1589-91.
14. Heinig, M. and D. Frishman, *STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W500-2.
15. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J Mol Biol, 1970. **48**(3): p. 443-53.
16. Yan, R., et al., *A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction.* Sci Rep, 2013. **3**: p. 2619.
17. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
18. Yang, Y., et al., *Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates.* Bioinformatics, 2011. **27**(15): p. 2076-82.
19. Wu, S. and Y. Zhang, *MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information.* Proteins, 2008. **72**(2): p. 547-56.
20. Soding, J., *Protein homology detection by HMM-HMM comparison.* Bioinformatics, 2005. **21**(7): p. 951-60.
21. Wu, S. and Y. Zhang, *LOMETS: a local meta-threading-server for protein structure prediction.* Nucleic Acids Res, 2007. **35**(10): p. 3375-82.
22. Rotkiewicz, P. and J. Skolnick, *Fast procedure for reconstruction of full-atom protein models from reduced representations.* J Comput Chem, 2008. **29**(9): p. 1460-5.
23. Parsons, J., et al., *Practical conversion from torsion space to Cartesian space for in silico protein synthesis.* J Comput Chem, 2005. **26**(10): p. 1063-8.

# Chapter 4: Evolutionary Protein Design using Chi-1 Rotamer Statistical Potentials

## 4.1 Introduction

Protein design can be considered as the inverse of protein structure prediction, where instead of starting from an amino-acid and predicting the lowest energy fold, a scaffold tertiary fold is used as input to search the “best-fit” sequence to that scaffold according to a given force-field [42, 76]. The vastness of the sequence-space presents itself as a major challenge in computational protein design. On the amino-acid level, the total number of unique sequences available is $20^L$. However, for most computational design applications, a rotamer library [77] is utilized to provide representation of a sidechain’s three-dimensional orientation required for most energy functions in a force field, and thus, the total number of possible rotamer combinations are even higher.

Most successful protein structure prediction and design force-fields incorporate some form of a knowledge-based potential or statistically effective energy function (SEEF) [47]. Rosetta’s REF15 [78] leverages potentials based on probabilities observed in the Protein Databank (PDB) [9]. One of the REF15 energy components derives an energy from the amino acid probability given backbone $\varphi$ and $\psi$ torsion angles. REF15 also implements the rotamer version of this energy function. A further study performed in the ABACUS protein design pipeline [79] refined this energy function by predicting the energy for a $\chi^1$ rotamer given backbone torsion angles and solvent accessibility.

Evolutionary protein design strategies which leverage conventional bioinformatics techniques have also shown promise. EvoDesign [58], a predecessor to the studies performed here, utilizes the PDB to construct a structure-derived sequence based position specific scoring matrix, or profile, in the form of an L×20 matrix, where L is the length of the input structure and 20 represents the number of amino acid types. The profile describes a probability-related distribution of amino-acids across the query structure. This distribution is generated proceeding a structural similarity search which extracts templates whose structures are similar to that of the input structure (usually with a TM-score threshold of 0.7 or greater) [26]. As expected, the profile serves as a mechanism to preserve conserved residues especially in cases where residues are imperative for a protein's structure or function. One drawback to this method, as well as other potentials that uses amino-acid probabilities to derive energy, is the lack of explicit structural information. In state-of-the-art computational protein design algorithms, rotamers are selected and judged based on the energy force field. Using this profiles in this form, thus, cannot discern energetically favorable rotamers from forbidden ones. This profile is, of course, supplemented by other energetic components [66, 80] to ensure reliable rotamer choices.

Here we propose a new evolutionary protein design algorithm, EvoDesign++, which contains a novel force field. The force field is a linear combination of seven weighted energy components:

$$\begin{gathered} E = w_1 E_{RotProf} + w_2 E_{Coev} + w_3 S2 + w_4 LJ + w_5 E_{PPI} + \\ w_6 E_{betaSelf} + w_7 E_{betaIntra} + w_8 E_{betaInter} \end{gathered} \tag{4.1}$$

Each function in the force-field has been designed specifically for EvoDesign++, except for *S2* which has been adapted by the ABACUS [79] energy function.

### 4.1.1 Framework for Statistical Potentials

Assuming a canonical ensemble in thermodynamic equilibrium, each microstate *i* can be assigned a Boltzmann factor:

$$e^{\frac{-E_i}{kT}} \tag{4.2}$$

where E*, k, and T,* represent the energy of microstate *i*, the Boltzmann constant, and temperature, respectively. The total number of possible microstates *Ni* that are accessible by the ensemble is described by the canonical partition function *Z*:

$$Z = \sum_{n=1}^{N} e^{\frac{-E_n}{kT}} \tag{4.3}$$

Thus, we can conveniently assign a probability for each microstate, $P_i$ by taking its corresponding Boltzmann factor and dividing it by the partition function.

$$P_i = \frac{e^{\frac{-E_i}{kT}}}{Z} \tag{4.4}$$

From each microstate probability, we can rearrange the formula to explicitly express its energy in the system.

$$E_i = -kTln(P_i) - \ln(Z) \tag{4.5}$$

After selection of any arbitrary reference microstate $E_{ref}$, the energetic transformation from the reference state to the target state can be represented as ΔE.

$$E_{ref} = -kTln\left(P_{ref}\right) - \ln(Z) \tag{4.6}$$

$$\Delta E_{ref \to i} = E_i - E_{ref} = -kTln\left(\frac{E_i}{E_{ref}}\right) \tag{4.7}$$

Most statistical potentials or statistically effective functions often take this form; however, the choice of a reference state is usually non-trivial and can make a considerable difference in the energy outcome.

### 4.1.2 Challenges of Beta Strand Design

Predicting beta strand contacts have posed an interesting challenge in the protein-structure prediction field [81]. Unlike alpha helices, the hydrogen bonding network found in beta sheets have no explicit relationship to the relative sequence locality of the included strands. For instance, the backbone of a beta strand on an N-terminus of a protein may hydrogen bond with another found on the C-terminus end. Thus, algorithms that attempt to enumerate all possible strand contacts in proteins that contain a considerable number of predicted beta strands may prove unfeasible. Fortunately, in a fixed-backbone design situation, the strand contacts are apparent in the input structure; however, *de novo* design of beta sheets produces a new set of challenges. The major goal of designing beta sheets is to prevent disruption of the beta sheet hydrogen bonding network. Slight changes in the amino-acid and rotamer composition can disturb these bonds, usually through changes in backbone torsion angles [62]. Exposed beta strands with unfulfilled hydrogen bonds confer molecules with a “sticky” property that promotes protein aggregation [82].

## 4.2 Methods

### 4.2.1 Evolutionary Chi-1 Rotamer-based profile

Structural homologs of the user-provided backbone structure are searched in a non-redundant monomeric structure database via TM-align [55]. Query-anchored alignments whose TM-score is above a selected cutoff are used to generate the multiple sequence-alignment. In the default case, a TM-score cutoff of 0.7 or is selected; however, for cases where the number of alignments does not successfully reach the threshold, we slowly decrement the cutoff by 0.05 until the number of sequences that compose the multiple sequence alignment is sufficient. For EvoDesign++, ten sequences are used as a minimum to construct a well-informed profile. These parameters were adopted from the parent version and can be further investigated for further tuning. Upon generation of the MSA, we convert each sequence to its rotamer-sequence form to generate the rotamer multiple sequence alignment (rMSA). Finally, we generate the position specific scoring matrix, or profile for the given scaffold. The details of the matrix construction are similar to that performed in PSI-BLAST [53]. The probability *Q* of a rotamer *i* at position *pos* is defined by:

$$Q_{pos,i} = \frac{\left(\alpha + p_{obs_{pos,i}}\right) + \left(\beta + p_{pseudo_{pos,i}}\right)}{(\alpha + \beta)} \tag{4.8}$$

$Q_i$ is a weighted average of the raw probability and the pseudocount probability, $p_{pseudo}$, derived from a rotamer substitution matrix. Constants $\alpha$ and $\beta$ represent the relative weighting of the probabilities, respectively. Constant $\alpha$ is defined by the unique number of rotamers in each position *i* minus one, and $\beta$ was set as 5. The pseudocount probability $p_{pseudo}$ is defined by:

$$p_{psuedo_{pos,i}} = \sum_{j=1,i\neq j}^{55} \frac{p_{pseudo_{pos,j}}}{P_j q_{ij}} \tag{4.9}$$

where *j* is the rotamer idx, $P_j$ is the background probability of rotamer *j* and $q_{ij}$ is the observed pair probability of *i* and *j* inherit in the substitution matrix. Specifically, for ROTSUM, this is defined as:

$$q_{ij} = fas \tag{4.10}$$

However, for the $\chi^1$ rotamer-based profile, we employ the ROTSUM72 matrix to calculate the pseudocount. Finally, the position-specific score of the profile $E_{RotProf}$ is defined by the equation:

$$E_{RotProf(pos_i)} = -\ln\left(\frac{Q_{pos,i}}{P_i}\right) \tag{4.11}$$

The comparison of the canonical pseudocount generation (a) is compared with the ROTSUM-based approach (b).

### 4.2.2 Sequence-based Coevolution profile

Sequence-based profiles, including PSI-BLAST or similar, also have the potential to effectively describe appropriate residue distributions across the design protein. Currently, EvoDesign is currently studying the effects of adding PSI-BLAST profiles to augment structure-derived ones. In EvoDesign++, however, different information about the protein family is extracted from the same matrix.

Instead of using the PSI-BLAST profile directly, we are most concerned with the generated multiple sequence alignment anchored by the input sequence. If the user provides a protein structure whose complete protein sequence is known during submission, EvoDesign++ uses it as the input for PSI-BLAST. Alternatively, if no sequence information is provided alongside the structure, the sequence of the top template hit in the previous scoring function is used as input if the TM-score is greater than or equal to 0.7. If neither criterion is satisfied, this scoring function is turned off.

From the original protein structure query, we determine all residue pairs in the protein. An interacting pair is defined by two residues whose $C_\beta$ atoms are less than or equal to 8 Ångstroms. For the purposes of calculating this distance with no *a priori* knowledge of sidechain placement, we project a pseudo $C_\beta$ atom as if each position was mutated to an alanine residue. After all pairs are enumerated, we generate a two-dimensional matrix for each pair to describe the propensity of two amino-acid types coexisting.

$$E_{coev} = -kT\ln\left(\frac{p_{obs_i}, p_{obs_j}}{p_{bg_i} p_{bg_j}}\right) \tag{4.12}$$

The disadvantage to this scoring function is the lack of structural information presented by sequence databases. However, this is somewhat ameliorated by the relative size of the database yielding distributions which can effectively provide some information of pertinent interactions found throughout a protein family.

### 4.2.3 Beta-Topology Statistical Potentials

#### 4.2.3.1 Construction of Beta Protein Database

Starting from SCOPe [15] classification of all-beta protein structures, we culled this list to remove redundant structures by 70% sequence identity and structures with resolutions ≥ 2.5 Ångstroms using the PISCES [68] server. From the remaining single 2,689 chains, 84,520 beta strand-occupying residues served as a dataset to calculate statistics explained further. For analyses related to this section, we describe the background probability $p_{bg}$ as the probability for a $\chi^1$ rotamer in this dataset.

We adopted the five topological states reported in [83] and summarized in Figure 4.1. Two topologies consider a residue flanked by a single parallel (A) or antiparallel (B) neighboring beta strand. For the remaining topology types, the target residue is flanked by two neighboring strands that are both parallel (C), one parallel and one antiparallel (D), or both antiparallel (E) to the direction of the target residue's strand.

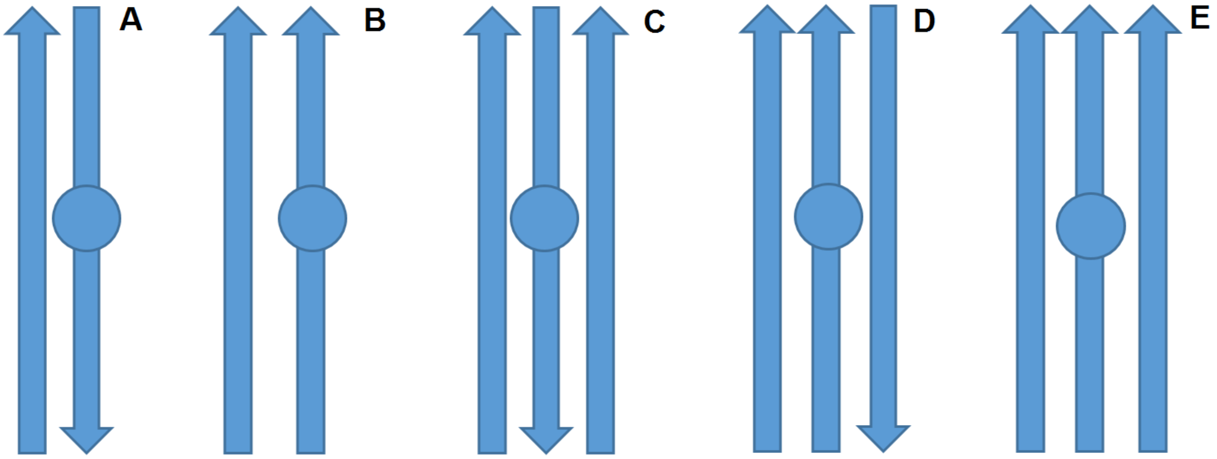


**Figure 4.1.** Topology environments for beta sheet-occupying residues. Target design residue designated as a blue circle. Five different beta sheet environment types are defined: (A) double anti parallel, (B) double parallel, (C) triple antiparallel, (D) triple mixed, (E) triple parallel

### 4.2.3.2 Self-Energy

First, we considered leveraging the distribution of the $\chi^1$ rotamers across the defined topological states in this dataset. From previous knowledge of beta sheet propensities, one may expect that rotamers belonging to different amino-acid types would reflect observed propensities.

$$E_{beta\_self} = -kTln\left(\frac{p_{obs_i}|T_i, StrPos_i}{p_{bg_i}}\right) \tag{4.13}$$

Symbols $T_i$ and $StrPos_i$ represent the beta strand topology type and the relative position on the strand (e.g. N- or C-terminus of the strand—See Section 4.3.1.2). The reference state for this scoring function was determined as the background probability of a $\chi^1$ rotamer across the entire PDB.

### 4.2.3.3 Intra-Strand Pairwise Energy

The first type of pairwise interactions considered are those consisting of two rotamers emplaced upon the same strand.

The energy component that describes this type of interaction is of the following form:

$$E_{beta\_intra} = -kT\ln\left(\frac{p_{obs_i}, p_{obs_{i+2}}|T_i}{p_{bg_i}p_{bg_j}}\right) \tag{4.14}$$

The reference energy for a pair-wise $\chi^1$ rotamer interaction assumes independent interaction described as the product of the self-energies.

### 4.2.3.4 Inter-Strand Pairwise Energy

Next, we also consider pairwise rotamer interactions from different strands but close in spatial proximity:

$$E_{beta\_inter} = -kT\ln\left(\frac{p_{obs_i}, p_{obs_j}|T_iT_j, relRSMSD}{p_{bg_i}p_{bg_j}}\right) \tag{4.15}$$

This scoring function's reference state is similar to the aforementioned state; however, both rotamers are not required to belong to the same strand type, and the relative orientation of the sidechains are considered when looking for similar oriented pairs in the PDB.

#### 4.2.3.5 Beta Strand Termini Analysis

From a design perspective, signaling for beta strand initiation and termination can provide control over its length and locality. From the dataset, we calculate the negative log likelihood of each $\chi^1$ rotamer type:

$$NLL = -\ln\left(\frac{p_{obs_{i,term}}}{p_{bg_i}}\right) \tag{4.16}$$

Here, we express this equation as a log likelihood rather than an SEEF for consistency with previous studies that express residue propensities in this form. Practically, either form is suitable for use in the EvoDesign++ force field.

### 4.2.4 Inter-chain Chi-1 Rotamer Interaction Potential

We derive another *SEEF* to describe the interaction between dimers at the rotameric level.

$$E_{PPI} = -kTln\left(\frac{p_{obs_i}, p_{obs_j}|SS_i, SS_j, SA_i, SA_j, relRMSD}{p_{bg_i}p_{bg_j}}\right) \tag{4.17}$$

#### 4.2.4.1 Dataset

To study the effects of drawing statistics from interfaces or protein monomers, we construct two databases which sample from these environments. The first dataset included rotamer pairs that were observed in native dimers. Starting from non-redundant interfaces clustered by PIFACE [84], structures whose resolution was < 2.5 Ångstroms were removed. From the resulting 12,799 protein interfaces, we considered rotamer pairs whose $C_\beta$ (or simulated $C_\beta$) distances were within 8.0 Ångstroms.

### 4.2.5 Monte Carlo

#### 4.2.5.1 Uniform Pseudo-Random Number Generation

In 2011, the C++ standardization committee released a library for generation of pseudo-random numbers [85]. Previous uses of C programming language's *rand* function were determined to not produce a uniform distribution of integers due to smaller entropy in the lower bits returned. Currently, the C++ standard *random* library provides utilities and interfaces to produce a reliably uniform series of values (as integer or floating-point representations). Herein, the design pipeline that requires pseudo-random number generators uses an engine based on the Mersenne Twister algorithm [86].

#### 4.2.5.2 Trajectory generation

Before initialization of the Monte Carlo process, a random $\chi^1$ rotamer sequence is generated with identical size to the query. For each Monte Carlo step, a candidate sequence is proposed by a $\chi^1$ rotamer mutation at a random position in the sequence. Due to the unequal number of $\chi^1$ states across amino-acid types, we first randomize the amino acid selection then randomly choose one of its rotamer states. Moreover, to reflect the benefits of simulated annealing, the temperature is slowly decremented based on an exponential decay function.

$$T_i = T_0 \times 0.9^i \tag{4.18}$$

where *i* is the iteration number of the Monte Carlo simulation and initial temperature $T_0 = 150$.

Next, the energy of the candidate sequence is then calculated by the $\chi^1$ -dependent force-field and compared to the previous accepted sequence. For all cases where the energy *E* of the candidate sequence *i* is lower than its accepted predecessor ($E_i < E_{i-1}$) we accept the mutation. EvoDesign++

employs a Metropolis criterion [36] to occasionally accept mutations whose mutation confers a non-significantly greater energy.

Following 20,000 iterations, SPICKER [87] clusters the final 25% of sequences of the trajectory and yields five sequences which represents the cluster centers of the largest five clusters. Alternatively, lowest energy sequences determined by the inherit force field are also reported by EvoDesign++. The advantages of using sequences from either approach have yet to be fully investigated with wetlab experiences for EvoDesign++ and its predecessor.

### 4.2.6 Rotamer Library

Although a full rotamer library is not theoretically required for this design algorithm, Van der Waal interactions complement the statistical functions well in creating a well-packed protein core. The size of the library should be seriously considered. One with low number of entries will provide sufficient coverage of rotamers at each position; however, the accuracy to represent the actual protein's sidechain's configuration may not be satisfactory. The converse is true when the number of rotamers in the library are many (typically in the tens of thousands). To address which sidechain orientations are significantly different from others, a SPICKER-based clustering approach of sidechains from the PDB is proposed.

### 4.2.7 Energy Function Validation and Weight Training by Native Amino Acid Recovery

A well-trained function should discriminate native-like sequences from random. Thus, we utilize the following objective function to tune various parameters and weights of our force field. We maximize this function which describes the probability of the native amino-acid $p(aa_{nat})$ seen when mutation the residue to the 19 other amino acid types [33]:

$$P(aa_{nat}) = \frac{\exp(-E(aa_{nat}))}{\sum_{i}^{20} \exp(-E(aa_i))} \tag{4.19}$$

In theory, if only the native amino-acid type were to perfectly occupy the top rank across all proteins' positions, the design function would return naturally-found sequences; however, in practice, this hasn't been observed.

### 4.2.8 Structural Predictions of Designs

Tertiary structure prediction of designed sequences served as an *in-silica* method to validate the efficacy of the design algorithm. Moderate use of evolution techniques during the design procedure led the decision to utilize *ab initio* structure prediction methods such as QUARK [88] over template-based ones such as I-TASSER [40]. To further reduce structural bias of template models, all structure templates greater than 30% global sequence identity to the designed sequence were removed before running QUARK. For all predicted models shown in this chapter, the refined model of the largest cluster center is chosen.

## 4.3 Results

### 4.3.1 Beta Statistical Potentials

#### 4.3.1.1 Self-Energy Analysis

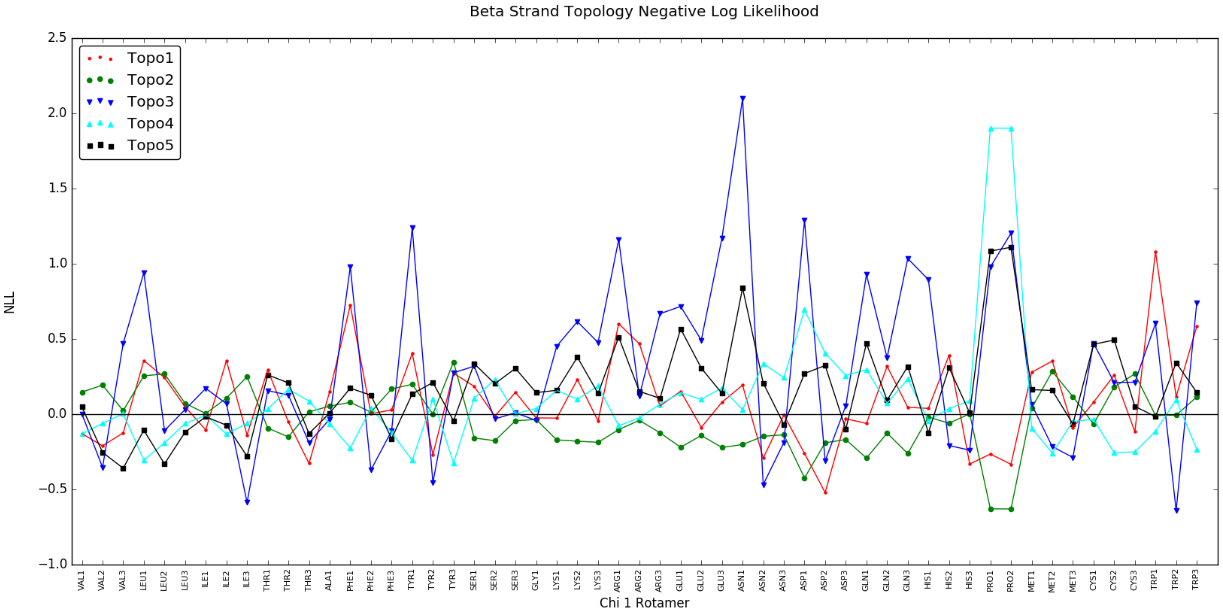


**Figure 4.2.** Negative log-likelihood of chi-1 rotamers on beta strand topology types. The negative log-likelihood (NLL) calculated by Eqn 4.16 are determined for the 55 chi-1 rotamers on an all-beta protein dataset.

Linglin Yu et al. illustrated that across all topology types, the OPUS-Beta self-packing energy for all residues exhibit the same pattern. The general trend shown by OPUS-Beta is the relative propensity of amino-acid types in beta strands. According to their functions, hydrophobic residues (valine, leucine, and isoleucine) are preferred across all beta sheet types. There are major implications with these results. First, the beta strand contact type is relatively independent of the amino acid type composition in the strands, and the strand packing is only influenced by backbone hydrogen bonding and/or external forces. This also implies that in a protein design situation, the selection of an amino-acid type in a beta strand should be independent of the strand's surrounding environment with regards to its beta sheet configuration. Results shown in this study, however, exhibit a stark contrast to these implications. Valine chi-1 rotamer class 3 (VAL3) are preferred in triple parallel beta sheets while generally disfavored in triple anti parallel beta sheets.

#### 4.3.1.2 Beta Strand Cap Analysis

We also investigated the $\chi^1$ rotamers' propensity to occupy the N- and C-terminal caps of beta strands. Previous studies examined the propensities of amino acid types [89] on beta strand termini. FarzardFard *et al*. performed the analysis on a dataset that wasn't exclusively all-beta protein structures. Figure 4.3 shows similar analysis but performed on the beta-protein dataset described earlier. Both results agree that the major beta strand terminators are aspartic acid, asparagine, and proline. However, also suggested by this figure, beta sheet initiators are often glycine and charged residues. Figures 4.4 and 4.5 show the negative log-likelihood distribution when the amino acids are categorized in their respective $\chi^1$ rotamer type. Interestingly, different observations which seem to contradict previous results. After the disambiguation, there is a strong preference for charged residues of class 3 (e.g. ASP3, GLN3, and LYS3) to occupy the N-terminus of beta strands while, generally, other classes of the same amino acid type plus all proline rotamers are disfavored here.

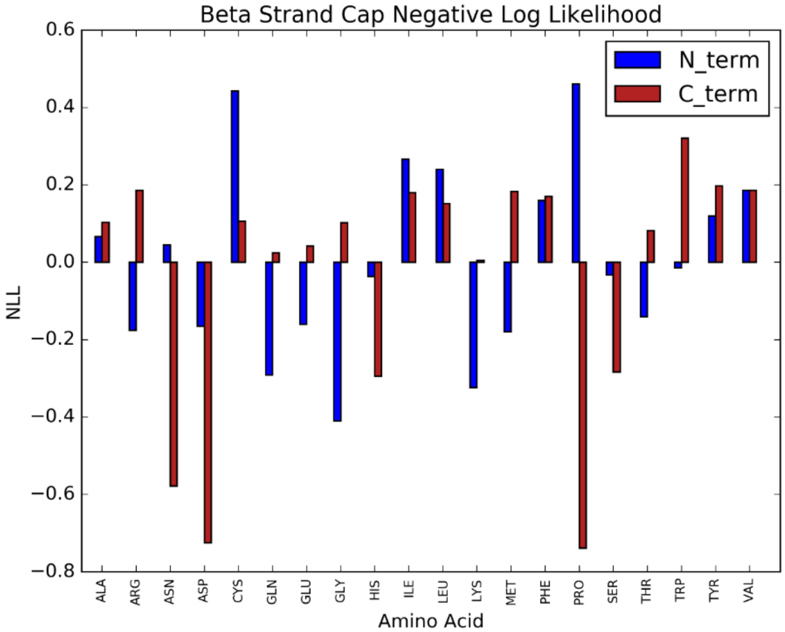


**Figure 4.3.** Amino acid propensity for beta strand initiation (N-terminus position of beta strand) and beta strand termination (C-terminus position of beta strand).

For C-terminus ends of beta strands, nearly the opposite trend is observed. $X^1$ rotamer classes 1 and 2 of asparagine and aspartate and all proline rotamers primarily dominate the population of C-terminus occupying $\chi^1$ rotamers. While aspartate and asparagine consist of similar phyiochememical properties as their glutamate/glutamine counterparts, neither glutamine nor glutamate are preferred generally for the C-terminus ends. A reason for this phenomenon, suggested by FarzardFard *et al*., was due to glutamate's longer sidechain (additional methylene), specific hydrogen-bonding patterns were not as readily available compared to aspartate.

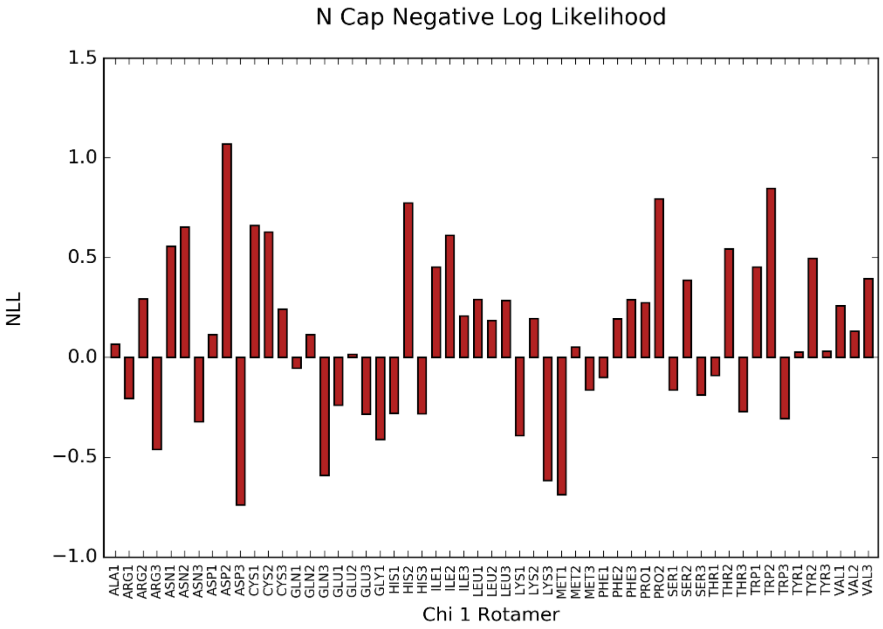


**Figure 4.4.** Chi-1 angle propensity for beta strand initiation (N-terminus position of beta strand)

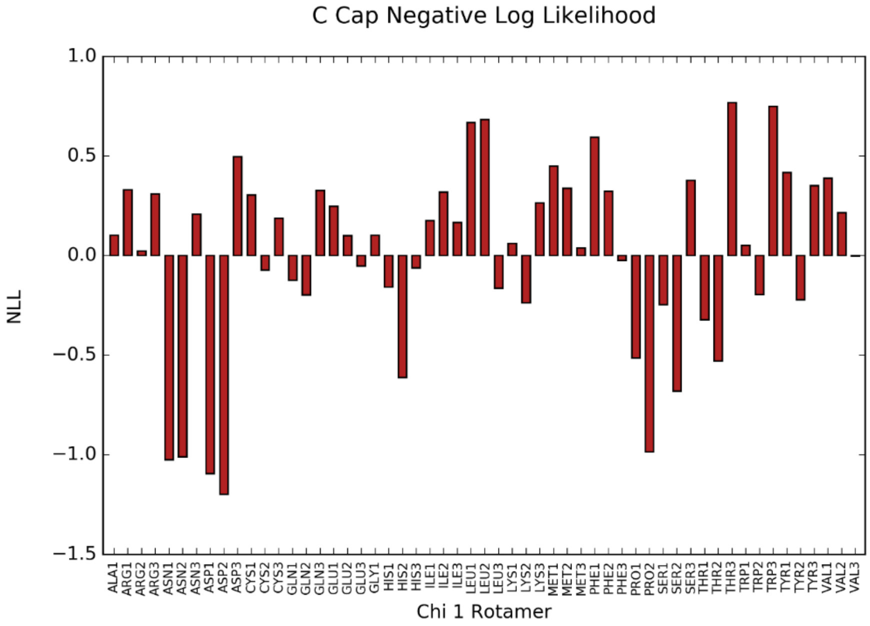


**Figure 4.5.** Chi-1 angle propensity for beta strand termination (C-terminus position of beta strand).

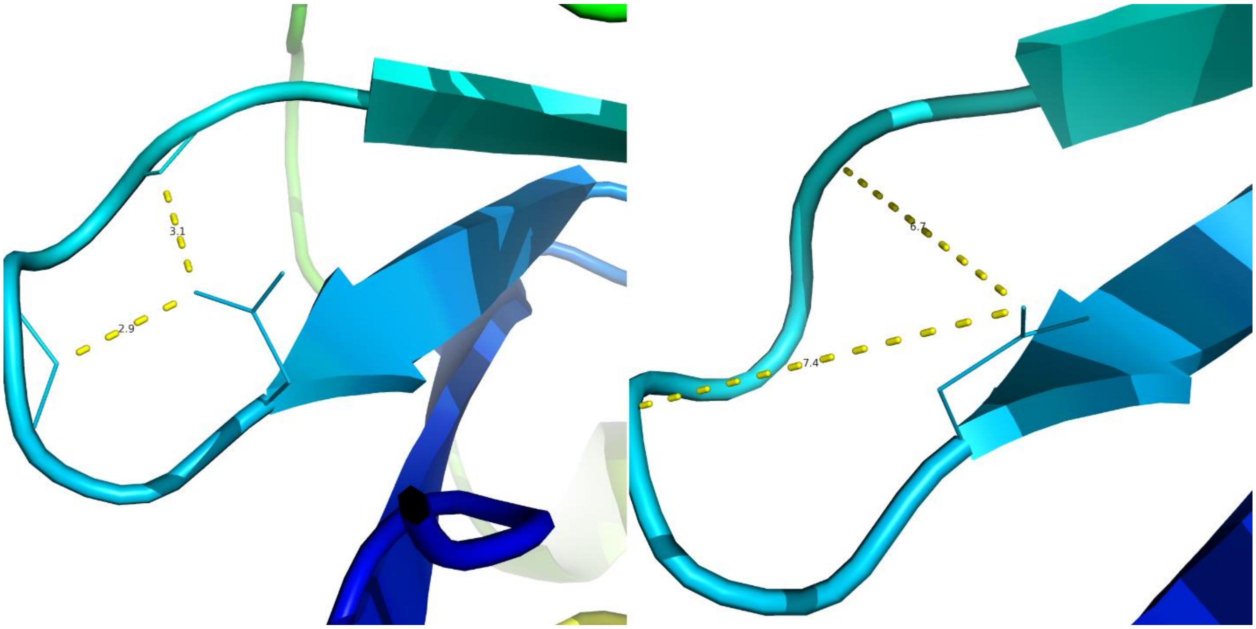
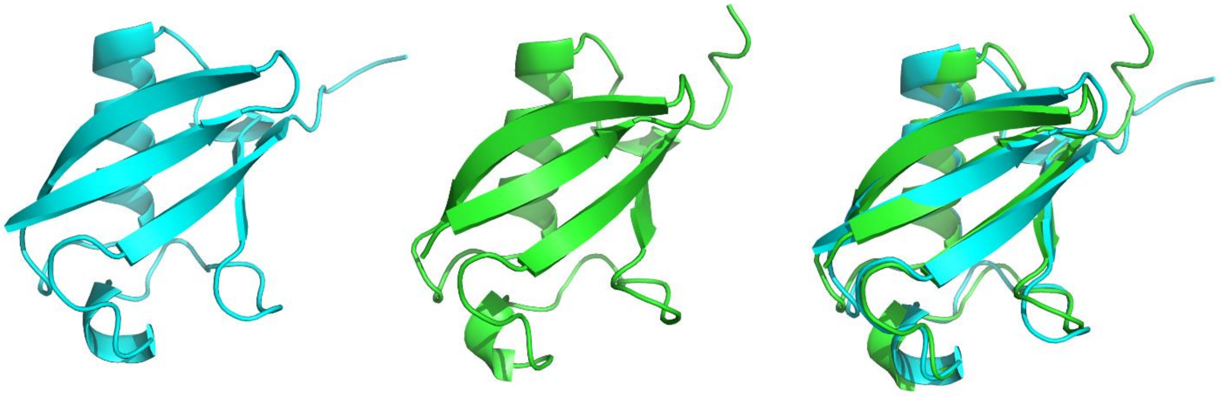


**Figure 4.6.** Example of hydrogen bonding patterns of ASP1 versus ASP3 at the beta strand-terminating position (PDBID: 1a1xA; D24).

We illustrate the $\chi^1$ rotamer dependence of beta strand termination in Figure 4.6. In this example, we specifically analyze a strand-terminating aspartate's hydrogen bonding patterns. The left image shows two hydrogen bonding instances (H-bond distances: 2.9Å and 3.1Å) involving the sidechain's anionic oxygen atom and two backbone nitrogen atoms. These types of interactions could be crucial in close positioning the subsequent strand. Rotating the $\chi^1$ dihedral to -60°, which conforms to a class 3 aspartate rotamer state (ASP3; right image), results in a disruption of these bonds (new distances: 7.4Å and 6.4Å, respectively). Although the frequency ratio of ASP3 to ASP1 in all beta strands are 5:1, ASP3 rotamers are rarely seen in the strand-terminating position compared to ASP1 and ASP2.

### 4.3.2 Design Case Studies

#### 4.3.2.1 Design of Ubiquitin Domain



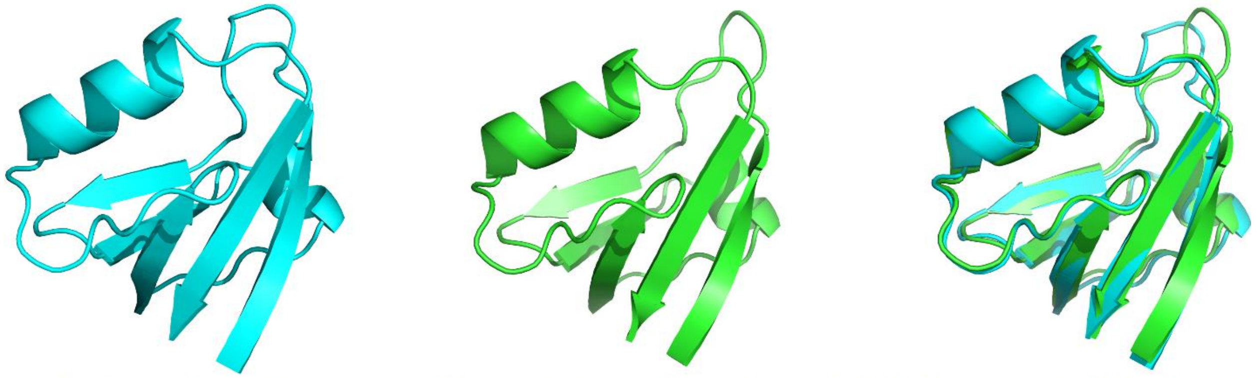
```
WT:  MQITVSTMDGQEITIQLSREMTVKQLKQHIQKRWGLPHEQQMLIWSGKWLEDHKTLQDYQVQDNSMVHLMTRRMSR
     ::: : :  :  ::        :    :  ::   : :  :: ::  :: :::  :: ::  :  :  ::  :
Des: MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGG
```

**Figure 4.7.** Structural comparison of native and designed ubiquitin (1ubqA). Crystal structure of ubiquitin (cyan), QUARK-predicted model of design generated by EvoDesign++'s Rotamer Profile term only (green), and the pair superimposed by TM-align (rightmost). Wildtype and design NW-align sequence alignments are shown below the structures.

Ubiquitin offers a convenient test target for computational *de novo* design. Especially from an evolutionary perspective, there are a vast number of homologs, both close and diverse, that can be used as a quality set of templates for multiple sequence alignment and profile construction. PDB entry 1UBQ:chain A is a 76-mer protein with a near 50/50 composition of alpha helices and beta-sheet residues.

The crystal structure of wildtype ubiquitin (cyan), the QUARK-predicted structure of a designed sequence (green), and their TM-align superposition, are shown in Figure 4.7. Wildtype and designed ubiquitin share a sequence identity of 42% and a sequence similarity of 64.5% as determined by EMBOSS [90]. The structural similarity between the crystal structure and the QUARK-predicted model share a TM-score of 0.87. These findings suggest that the folding patterns are extremely similar. To further support this fact, we performed structure modeling on the wildtype 1ubqA sequence and compared the result to its crystal structure counterpart. The TM-score for the structure pair was 0.88; this suggests that the structure of the design is well-predicted and could fold identically to the crystal structure, considering the model prediction error.

#### 4.3.2.2 Design of PDZ Domain

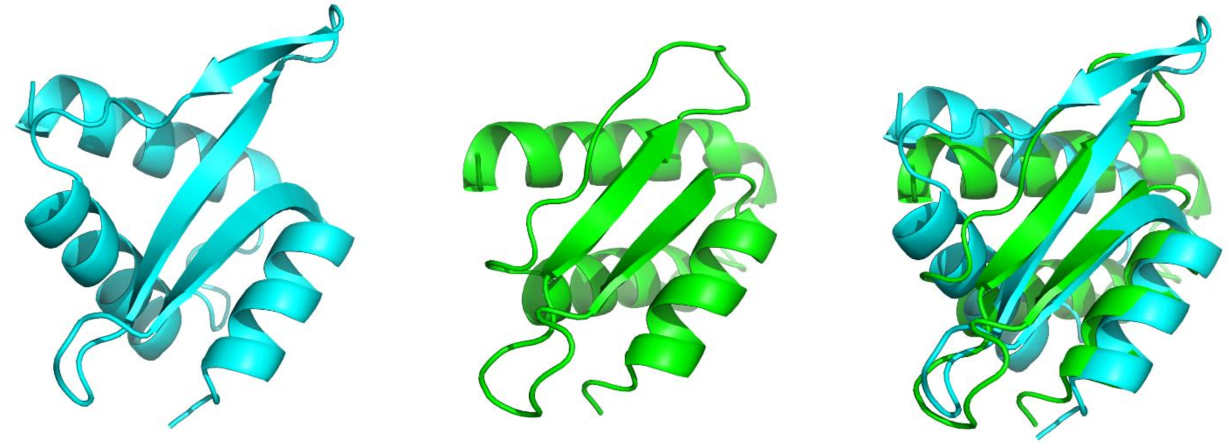


```
WT:  PRTITMHKDSTGHVGFIFKNGKITSIVKDSSAARNGLLTEHNICEINGQNVIGLKDSQIADILSTSGTVVTITIM
      ::    : :    ::      : : :    : : ::     : : ::  : :         :   :  ::   :
Des: VRTVELEKSSSEGLGFSIAGIYISSVVPGGPAERAGLQVGDQILEVNGVSVEGMTHEEAVELLKSAGSKVTLVVM
```

**Figure 4.8.** Structural comparison of native and designed PDZ domain (1obyA). Crystal structure of PDZ domain (cyan), I-TASSER-predicted model of design generated by EvoDesign++'s full monomer forcefield (green), and the pair superimposed by TM-align (rightmost). Wildtype and design NW-align sequence alignments are shown below the structures.

The crystal structure of wildtype PDZ domain, the QUARK-predicted structure of a designed sequence, and their TM-align superposition, are shown in Figure 4.8. The sequence identity and sequence similarity between wildtype PDZ and its design are 32.0% and 54.7%, respectively. Compared to ubiquitin, PDZ's fraction of beta sheet-occupying residues are higher. Initial attempts to predict the model proved difficult. Removal of templates over 30% sequence identity for the fragment generation in QUARK led to a poor structural alignment against the wildtype PDZ domain (TM-score 0.31). Using I-TASSER to predict its model allowing the full template library, however, yielded an extremely strong structural similarity (TM-score 0.92) even with a low sequence identity.

#### 4.3.2.3 Design of Cas9 Protein Inhibitor AcrIIA4



```
WT:  SMNINDLIREIKNKDYTVKLSGTDSNSITQLIIRVNNDGNEYVISESENESIVEKFISAFKNGWNQEYEDEEEFYNDMQTITLKSE
       ::    :         : : :        : :     :    ::  :              : :       :       :  ::
Des: MMNLEEFIEFLRELGISVTLTGDPPEKKPTLTINFCGNGHTITISHTEKSTLFQEYREQYRDGPNGKDQNVQELFRKLIEICKKSM
```

**Figure 4.9.** Structural comparison of native and designed AcrII4A domain (5VW1B). Crystal structure of PDZ domain (cyan), QUARK-predicted model of design generated by EvoDesign++'s full monomer forcefield (green), and the pair superimposed by TM-align (rightmost). Wildtype and design NW-align sequence alignments are shown below the structures.

Recently, a Cas9 protein inhibitor, AcrIIA4, was discovered and crystallized [91]. From simple structural analysis comparing the structure across PDB70, only one structure (PDBID) was found whose TM-score was above 0.7. Several other templates shared a TM-score of slightly above 0.5. The resulting multiple sequence alignment, therefore, could not serve as a reliable component to design the inhibitor. However, designed sequences using the full force field often yielded sequences with sequence identity slightly higher than 30%.

The crystal structure of wildtype AcrII4A inhibitor, the QUARK-predicted structure of a designed sequence, and their TM-align superposition, are shown in Figure 4.9. Although the sequence identity of the two protein sequences are 21%, they share a convincing structure similarity with a

TM-score of 0.64. The sequence similarity, as determined by EMBOSS, is approximately 48%. A large recovery of the wildtype's sidechain properties may ensure to conserve specific fold requirements. While all alpha helices are correctly accounted for and predicted, two of the three beta strands were predicted by QUARK. However, this phenomenon could be due to limitations of *ab-initio* prediction programs and not due to design issues. In the design's predicted structure, the beta strand nearest to the c-terminus is represented as coil; however, the position of this coil strand is nearly within hydrogen-bond range of the second beta strand. Also, many of the residues in the AcrII4A interface are generally hydrophilic; however, in attempts to design new residues on the inhibitor's interface, most of the residues selected by EvoDesign++ were hydrophobic. This is likely due to $E_{PPI}$ trained on monomeric rotamer interactions which primarily occur hydrophobic environments.

### 4.4 Discussion

Here, we developed an evolutionary protein design algorithm, EvoDesign++. Although the general evolutionary approach to protein design has been adopted by its predecessor algorithm, EvoDesign [92], this pipeline includes a novel set of statistical effective energy functions and structure-based scoring matrices for purposes of computational protein design with a focus of generating native-like beta sheets. The EvoDesign++ force field is a linear combination of seven energy function components plus an optional $\chi^1$ rotamer-based energy function for protein interface design. In this chapter, we mainly focus on the analysis of our beta sheet-related energy functions and its ability to detect appropriate rotamers under different beta sheet environments. Performing single site native amino acid recovery (outlined by Eqn. 4.19) on a small dataset of all-beta proteins, the native residues were identified, on average, in the top 5 out of 20 different amino acids. Moreover, the statistical potentials proposed treat strand positions specially. From structural analysis of beta strand N- and C-terminus caps, we are able to identify some rotameric propensities for these beta strand locations that amino acid distributions could not readily detect. Moreover, we also showed exemplary easy, medium, and hard design targets for the algorithm. While these designs consist of relatively low sequence identities (~30%) compared to their input structure protein scaffold, they were readily foldable by protein structure prediction methods and their corresponding the TM-scores (>0.6 for all cases) suggest that the design model structures are evolutionarily related [29]. The EvoDesign++ protein design web-server is freely available online for academic use at https://zhanglab.ccmb.med.umich.edu/EvoDesign++/.

### 4.5 Acknowledgements

University of Michigan undergraduate students Zachary Orban and Peter Tolsma assisted in the software development and initial statistics calculations for understanding rotameric preferences for beta strand topology and its relationship to structure. Undergraduate students Humza Hemani and Harsha Parchani constructed the monomer database for inter-protein rotamer-rotamer interactions and performed initial calculations to quantify these interactions. Undergraduate student Thomas Stewart led efforts to gather background information needed to design a novel protein inhibitors for Cas9.

## 4.6 References

1. Richardson, J.S. and D.C. Richardson, *The de novo design of protein structures.* Trends Biochem Sci, 1989. **14**(7): p. 304-9.
2. Gordon, D.B., S.A. Marshall, and S.L. Mayo, *Energy functions for protein design.* Curr Opin Struct Biol, 1999. **9**(4): p. 509-13.
3. Lovell, S.C., et al., *The penultimate rotamer library.* Proteins, 2000. **40**(3): p. 389-408.
4. Sippl, M.J., *Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures.* J Comput Aided Mol Des, 1993. **7**(4): p. 473-501.
5. Alford, R.F., et al., *The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design.* J Chem Theory Comput, 2017. **13**(6): p. 3031-3048.
6. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
7. Xiong, P., et al., *Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability.* Nat Commun, 2014. **5**: p. 5330.
8. Mitra, P., D. Shultis, and Y. Zhang, *EvoDesign: De novo protein design based on structural and evolutionary profiles.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W273-80.
9. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins, 2004. **57**(4): p. 702-10.
10. Schymkowitz, J., et al., *The FoldX web server: an online force field.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W382-8.
11. Miao, Z., Y. Cao, and T. Jiang, *RASP: rapid modeling of protein side chain conformations.* Bioinformatics, 2011. **27**(22): p. 3117-22.
12. Subramani, A. and C.A. Floudas, *beta-sheet topology prediction with high precision and recall for beta and mixed alpha/beta proteins.* PLoS One, 2012. **7**(3): p. e32461.
13. Dunbrack, R.L., Jr. and M. Karplus, *Backbone-dependent rotamer library for proteins. Application to side-chain prediction.* J Mol Biol, 1993. **230**(2): p. 543-74.
14. Richardson, J.S. and D.C. Richardson, *Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(5): p. 2754-2759.
15. Zhang, Y. and J. Skolnick, *TM-align: a protein structure alignment algorithm based on the TM-score.* Nucleic Acids Res, 2005. **33**(7): p. 2302-9.
16. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
17. Fox, N.K., S.E. Brenner, and J.M. Chandonia, *SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures.* Nucleic Acids Res, 2014. **42**(Database issue): p. D304-9.
18. Wang, G. and R.L. Dunbrack, Jr., *PISCES: a protein sequence culling server.* Bioinformatics, 2003. **19**(12): p. 1589-91.
19. Yu, L., *A Novel Statistical Potential for Protein Beta-Sheets Prediction*, in *BioEngineering*. 2014, Rice University. p. 48.
20. Cukuroglu, E., et al., *Non-redundant unique interface structures as templates for modeling protein interactions.* PLoS One, 2014. **9**(1): p. e86738.
21. ISO/IEC, *ISO International Standard ISO/IEC 14882:2014(E) – Programming Language C++. [Working draft]*. International Organization for Standardization (ISO), 2014.
22. Saito, M. and M. Matsumoto, *SIMD-oriented Fast Mersenne Twister: A 128-bit pseudorandom number generator*. Monte Carlo and Quasi-Monte Carlo Methods 2006, 2008: p. 607-622.
23. Hastings, W.K., *Monte-Carlo Sampling Methods Using Markov Chains and Their Applications.* Biometrika, 1970. **57**(1): p. 97-&.
24. Zhang, Y. and J. Skolnick, *SPICKER: a clustering approach to identify near-native protein folds.* J Comput Chem, 2004. **25**(6): p. 865-71.

25. Kuhlman, B., et al., *Design of a novel globular protein fold with atomic-level accuracy.* Science, 2003. **302**(5649): p. 1364-8.
26. Xu, D. and Y. Zhang, *Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field.* Proteins, 2012. **80**(7): p. 1715-35.
27. Zhang, Y., *I-TASSER server for protein 3D structure prediction.* BMC Bioinformatics, 2008. **9**: p. 40.
28. Farzadfard, F., et al., *Beta-sheet capping: signals that initiate and terminate beta-sheet formation.* J Struct Biol, 2008. **161**(1): p. 101-10.
29. Chojnacki, S., et al., *Programmatic access to bioinformatics tools from EMBL-EBI update: 2017.* Nucleic Acids Res, 2017. **45**(W1): p. W550-W553.
30. Yang, H. and D.J. Patel, *Inhibition Mechanism of an Anti-CRISPR Suppressor AcrIIA4 Targeting SpyCas9.* Mol Cell, 2017. **67**(1): p. 117-127 e5.
31. Mitra, P., et al., *An evolution-based approach to De Novo protein design and case study on Mycobacterium tuberculosis.* PLoS Comput Biol, 2013. **9**(10): p. e1003298.
32. Xu, J. and Y. Zhang, *How significant is a protein structure similarity with TM-score = 0.5?* Bioinformatics, 2010. **26**(7): p. 889-95.

# Chapter 5: Conclusions

## 5.1 Overall Conclusions

One central problem of structural biology is the prediction a protein's structure and function [10]. Currently, we are equipped with over 114 million sequences in the UniProt database [3], but structural entries lags behind by over three orders of magnitude [9, 10]. With an ever-increasing focus on developing therapeutics which targets proteins within biological pathways and elucidating pathways and protein-protein interaction networks, obtaining as much structural information as possible is crucial for medicinal progress and our fundamental understanding of biochemistry. Recently, many advancements have been made in protein structure prediction and these improvements are biannually assessed by the structural biology community [93]. Template-based modelling proves as an effective method to expose hidden structural information encoded in the protein sequence [40]. A variety of different fold recognition programs [56, 72, 73] often extract information regarding possible structural elements of the protein and use that information to find similar folds in the PDB [16]. In this study, we propose the use of $\chi^1$ rotamer information as a method for fold recognition. From our findings in Chapter 2, we observed $\chi^1$ rotamer substitutions across structures with similar fold, determined by TM-align [55]. Interestingly, rotamer classes are typically more conserved than rotamer mutations within the same amino-acid types (although there are some variable cases especially with amino acid types that are highly solvent exposed). Typically, when sequence alignments are analyzed, we often cognitively imagine amino-acid mutation events without regard for the implicit structural information for each position. We also observe that rotamer-based substitution matrices are more sensitive in identifying structurally similar folds with low sequence identity than traditional substitution matrices [18]. We thus propose the possibility that fold recognitions programs to include $\chi^1$ rotamer structure as a criteria for sequence-structure alignment.

Following backbone modelling in protein structure prediction, sidechain packing algorithms [65, 66] are often employed to guide the construction of rotamers on the modelled backbone. In Chapter 3, we observe that our proposed sequence-fragment based approach to $\chi^1$ rotamer prediction outperforms conventional methods that require a backbone structure [65] in accurate prediction in $\chi^1$ rotamer classification and real-value torsion angle. Our approach is structure-independent; the sequence information is the only input required to predict a $\chi^1$ rotamer sequence. A protein threading program which includes profile-profile and secondary structure alignments [73] is also employed to assist Rot1Pred in generation of a rotamer probability matrix to inform the prediction.

In support of our argument for considering $\chi^1$ rotamer types as evolutionary independent entities, we explore their statistical relationship with various protein environments. Here, we mathematically represent these statistical relationships in the form of statistical effective energy functions (SEEF) [47]. These types of models are often desired in force fields used for molecular modeling and design [94]. In Chapter 4, we explore these functions and their performance in designing native-like protein molecules. Here we observed that our rotamer-based SEEFs are sensitive to the preference of the $\chi^1$ rotamer's preference to occupy particular structural environments that amino acid-based potentials will miss.

## 5.2 Future Directions

### 5.2.1 Backbone Structural Flavors in ROTSUM Matrices

Currently the ROTSUM matrix provides a separation of amino acid types into their respective $\chi^1$ classifications. Theoretically, further separation is possible with additional criteria. Secondary structure (e.g. alpha, beta, and coil) and solvent accessibility (e.g. core, interface, solvent-exposed) classifications are achievable; however two concerns should be addressed. The lack of statistical data to populate these different subclasses (at least 165 total vs 20 or 55). Although pseudocounts can be considered, the manner in which they are derived should be carefully decided. Also, previous studies [30] show a correlation between $\chi^1$ rotamer type and the secondary structure element for the given position. Thus, there may exist some redundancy in the matrix. Similar rationales can be applied for amino-acid types and their observed frequency in different levels of solvent exposure.

### 5.2.2 Robust Chi-1 Rotamer-based Fold Recognition Algorithm

Preliminary examples in Chapter 2 illustrate that simple alignments guided by ROTSUM can recognize folds which BLOSUM62 misses; however, for present technology, a more sophisticated algorithm may be required, especially if it will contribute to the performance of meta-threading pipelines such as LOMETS [73]. One proposal is to extend the rotamer profile-rotamer profile alignment algorithm. Since sequences used to populate the profile requires rotamer type information, the PDB could be used as a source. However, data may not be sufficient for many templates. Alternatively an alternative database can be constructed where sequences' $\chi^1$ rotamers are predicted and the profile can thus be constructed from the predicted rotamer sequences.

### 5.2.3 Improvement of Rot1Pred Algorithm

Initial attempts at predicting $\chi^1$ rotamer type for a position in an amino acid sequences using machine learning approaches were made although unsuccessful. Even employing techniques including equal sampling and regularization, did not yield results as effective as the method explained in Chapter 3. Perhaps more robust and sophisticated deep-learning approach yields significantly higher accuracy. Although not detailed in this dissertation, features used for machine learning included PSI-BLAST profiles [53], predicted secondary structure, solvent accessibility, and backbone torsion angles. Other types of features including Shannon entropy and predicted contacts could also be explored.

### 5.2.4 EvoDesign++ Protein-Protein Interaction

Computationally designing an effective protein interface are often assisted with experimental screening to determine crucial, or “hotspot”, residues regarding the protein’s function. Some approaches will often take cycles of computational design and experiments to optimize a binding site [95]. A force-field that can identity all, or most, hotspots and their relative contributions to a target objective function (e.g. perhaps in the form of binding affinity) would be ideal. Although the protein interface library lacks enough data to inform an effective probability distribution for rotamer interactions, an atomic statistical effective energy function derived and designed specifically for protein interface design may prove useful.

## 5.3 Final Thoughts

I am greatly optimistic for the future of computational structural biology. During the duration of these studies, we have witnessed a steady growth in the protein database (over 40,000 sequences have been submitted during my graduate studies at the University of Michigan). The explosion of sequence entries have made statistical analysis of $\chi^1$ rotamers, and thus this dissertation, possible. As we look forward to the future, more diverse and high resolution structure submissions will better inform our mathematical models to accurately design proteins and predict protein structure and function. From the observations and results obtained in this study, it is suggested that statistics based on rotamers should be preferred over ones that describe amino acid-level behavior, especially if structural information of your target's sidechain is already known. This includes amino acid substitution matrices that guide scoring functions.

Naturally, given the studies performed here, one could begin to analyze the effect of including $\chi^2$ rotamer information. Fortunately, many $\chi^2$ angles exhibit rotameric properties and can be easily separated into distinct states. However, assuming the chi-2 angle also displays three discrete states, a total of nine possible $\chi^1 - \chi^2$ rotamer combinations exist. Describing each state accurately depends on whether if the data is sufficient and also entropic effects that influence how often $\chi^2$ rotamers mutate in nature. Nevertheless, the community may soon experience enough data to realize this idea with the ever-growing PDB. Outside of rotamer-related studies, many other sub-fields of computational structural biology also rely on a diverse and vast structure repository. As the rate of the PDB growth advances, so does the birthrate of novel subfields and ideas that impact human health and our fundamental understanding of biology and biochemistry. Looking forward, it seems to be only a matter of time until computational structural biology and biology-related data science will "offer fantastic dreams of other worlds just beyond our reach"—*Fritz Lang*.

## 5.4 References

1. Shapovalov, M.V. and R.L. Dunbrack, Jr., *A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions.* Structure, 2011. **19**(6): p. 844-58.
2. Ramachandran, G.N., C. Ramakrishnan, and V. Sasisekharan, *Stereochemistry of polypeptide chain configurations.* J Mol Biol, 1963. **7**: p. 95-9.
3. Consortium, T.U., *UniProt: the universal protein knowledgebase.* Nucleic Acids Res, 2017. **45**(D1): p. D158-D169.
4. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.* Nucleic Acids Res, 2003. **31**(1): p. 365-70.
5. Edman, P., *A method for the determination of amino acid sequence in peptides.* Arch Biochem, 1949. **22**(3): p. 475.
6. Mcluckey, S.A., G.L. Glish, and K.G. Asano, *The Coupling of an Atmospheric Sampling Ion-Source with an Ion Trap Mass-Spectrometer.* Abstracts of Papers of the American Chemical Society, 1988. **196**: p. 81-Anyl.
7. Suzek, B.E., et al., *UniRef: comprehensive and non-redundant UniProt reference clusters.* Bioinformatics, 2007. **23**(10): p. 1282-8.
8. Leinonen, R., et al., *UniProt archive.* Bioinformatics, 2004. **20**(17): p. 3236-7.
9. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
10. Zhang, Y., *Progress and challenges in protein structure prediction.* Curr Opin Struct Biol, 2008. **18**(3): p. 342-8.
11. Kendrew, J.C., et al., *3-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis.* Nature, 1958. **181**(4610): p. 662-666.
12. Wuthrich, K., *Protein structure determination in solution by NMR spectroscopy.* J Biol Chem, 1990. **265**(36): p. 22059-62.
13. Jiang, W., et al., *Semi-automated icosahedral particle reconstruction at sub-nanometer resolution.* Journal of Structural Biology, 2001. **136**(3): p. 214-225.
14. Costa, T.R.D., A. Ignatiou, and E.V. Orlova, *Structural Analysis of Protein Complexes by Cryo Electron Microscopy.* Methods Mol Biol, 2017. **1615**: p. 377-413.
15. Fox, N.K., S.E. Brenner, and J.M. Chandonia, *SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures.* Nucleic Acids Res, 2014. **42**(Database issue): p. D304-9.
16. Bowie, J.U., R. Luthy, and D. Eisenberg, *A method to identify protein sequences that fold into a known three-dimensional structure.* Science, 1991. **253**(5016): p. 164-70.
17. Schwartz, R.M.D., M.O, *Matrices for detecting distant relationships.* Atlas of protein sequence and structure, 1978. **5**: p. 353-58.
18. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.
19. M.O. Dayhoff, R.M.S., B.C. Orcutt, *A Model of Evolutionary Change in Proteins.* Atlas of protein sequence and structure, 1978. **5**.
20. Pietrokovski, S., J.G. Henikoff, and S. Henikoff, *The Blocks database--a system for protein classification.* Nucleic Acids Res, 1996. **24**(1): p. 197-200.
21. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J Mol Biol, 1970. **48**(3): p. 443-53.
22. Gotoh, O., *An improved algorithm for matching biological sequences.* J Mol Biol, 1982. **162**(3): p. 705-8.
23. Chakraborty, A. and S. Bandyopadhyay, *FOGSAA: Fast Optimal Global Sequence Alignment Algorithm.* Sci Rep, 2013. **3**: p. 1746.
24. Kabsch, W., *Solution for Best Rotation to Relate 2 Sets of Vectors.* Acta Crystallographica Section A, 1976. **32**(Sep1): p. 922-923.

25. Holm, L. and P. Rosenstrom, *Dali server: conservation mapping in 3D.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W545-9.
26. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins, 2004. **57**(4): p. 702-10.
27. Kabsch, W., *Discussion of Solution for Best Rotation to Relate 2 Sets of Vectors.* Acta Crystallographica Section A, 1978. **34**(Sep): p. 827-828.
28. Siew, N., et al., *MaxSub: an automated measure for the assessment of protein structure prediction quality.* Bioinformatics, 2000. **16**(9): p. 776-85.
29. Xu, J. and Y. Zhang, *How significant is a protein structure similarity with TM-score = 0.5?* Bioinformatics, 2010. **26**(7): p. 889-95.
30. Dunbrack, R.L., Jr. and F.E. Cohen, *Bayesian statistical analysis of protein side-chain rotamer preferences.* Protein Sci, 1997. **6**(8): p. 1661-81.
31. Zanghellini, A., et al., *New algorithms and an in silico benchmark for computational enzyme design.* Protein Sci, 2006. **15**(12): p. 2785-94.
32. Dahiyat, B.I., C.A. Sarisky, and S.L. Mayo, *De novo protein design: towards fully automated sequence selection.* J Mol Biol, 1997. **273**(4): p. 789-96.
33. Kuhlman, B., et al., *Design of a novel globular protein fold with atomic-level accuracy.* Science, 2003. **302**(5649): p. 1364-8.
34. Zhang, C.T. and K.C. Chou, *Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition.* Biophys J, 1992. **63**(6): p. 1523-9.
35. Dill, K.A., et al., *The protein folding problem.* Annu Rev Biophys, 2008. **37**: p. 289-316.
36. Hastings, W.K., *Monte-Carlo Sampling Methods Using Markov Chains and Their Applications.* Biometrika, 1970. **57**(1): p. 97-&.
37. Metropolis, N., et al., *Equation of State Calculations by Fast Computing Machines.* Journal of Chemical Physics, 1953. **21**(6): p. 1087-1092.
38. Khachaturyan, A.G., S.V. Semenovskaya, and B.K. Vainshtein, *Statistical Thermodynamical Approach to the Problem of Determination of Phases of Structure Amplitudes.* Kristallografiya, 1978. **24**(5): p. 905-916.
39. Sugita, Y. and Y. Okamoto, *Replica-exchange molecular dynamics method for protein folding.* Chemical Physics Letters, 1999. **314**(1-2): p. 141-151.
40. Zhang, Y., *I-TASSER server for protein 3D structure prediction.* BMC Bioinformatics, 2008. **9**: p. 40.
41. Baker, D., *An exciting but challenging road ahead for computational enzyme design.* Protein Sci, 2010. **19**(10): p. 1817-9.
42. Gordon, D.B., S.A. Marshall, and S.L. Mayo, *Energy functions for protein design.* Curr Opin Struct Biol, 1999. **9**(4): p. 509-13.
43. Brooks, B.R., et al., *Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations.* Journal of Computational Chemistry, 1983. **4**(2): p. 187-217.
44. Cornell, W.D., et al., *A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995).* Journal of the American Chemical Society, 1996. **118**(9): p. 2309-2309.
45. Jorgensen, W.L., D.S. Maxwell, and J. TiradoRives, *Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids.* Journal of the American Chemical Society, 1996. **118**(45): p. 11225-11236.
46. Nguyen, H., et al., *Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent.* J Am Chem Soc, 2014. **136**(40): p. 13959-62.
47. Sippl, M.J., *Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures.* J Comput Aided Mol Des, 1993. **7**(4): p. 473-501.
48. Zhang, C., S. Liu, and Y. Zhou, *Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential.* Protein Sci, 2004. **13**(2): p. 391-9.

49. Zhang, J. and Y. Zhang, *A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction.* PLoS One, 2010. **5**(10): p. e15386.
50. Zhou, H. and J. Skolnick, *GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction.* Biophys J, 2011. **101**(8): p. 2043-52.
51. Thomas, P.D. and K.A. Dill, *Statistical potentials extracted from protein structures: How accurate are they?* Journal of Molecular Biology, 1996. **257**(2): p. 457-469.
52. BenNaim, A., *Statistical potentials extracted from protein structures: Are these meaningful potentials?* Journal of Chemical Physics, 1997. **107**(9): p. 3698-3706.
53. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
54. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
55. Zhang, Y. and J. Skolnick, *TM-align: a protein structure alignment algorithm based on the TM-score.* Nucleic Acids Res, 2005. **33**(7): p. 2302-9.
56. Wu, S. and Y. Zhang, *MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information.* Proteins, 2008. **72**(2): p. 547-56.
57. Stormo, G.D., et al., *Use of the Perceptron Algorithm to Distinguish Translational Initiation Sites in Escherichia-Coli.* Nucleic Acids Research, 1982. **10**(9): p. 2997-3011.
58. Mitra, P., D. Shultis, and Y. Zhang, *EvoDesign: De novo protein design based on structural and evolutionary profiles.* Nucleic Acids Res, 2013. **41**(Web Server issue): p. W273-80.
59. S. Kullback, R.A.L., *On Information and Sufficiency.* The Annals of Mathematical Statistics, 1951. **22**(1).
60. Panchenko, A.R., *Finding weak similarities between proteins by sequence profile comparison.* Nucleic Acids Res, 2003. **31**(2): p. 683-9.
61. Apweiler, R., et al., *UniProt: the Universal Protein knowledgebase.* Nucleic Acids Res, 2004. **32**(Database issue): p. D115-9.
62. Dunbrack, R.L., Jr. and M. Karplus, *Backbone-dependent rotamer library for proteins. Application to side-chain prediction.* J Mol Biol, 1993. **230**(2): p. 543-74.
63. Wu, S. and Y. Zhang, *ANGLOR: a composite machine-learning algorithm for protein backbone torsion angle prediction.* PLoS One, 2008. **3**(10): p. e3400.
64. Song, J., et al., *TANGLE: two-level support vector regression approach for protein backbone torsion angle prediction from primary sequences.* PLoS One, 2012. **7**(2): p. e30361.
65. Krivov, G.G., M.V. Shapovalov, and R.L. Dunbrack, Jr., *Improved prediction of protein side-chain conformations with SCWRL4.* Proteins, 2009. **77**(4): p. 778-95.
66. Miao, Z., Y. Cao, and T. Jiang, *RASP: rapid modeling of protein side chain conformations.* Bioinformatics, 2011. **27**(22): p. 3117-22.
67. Berjanskii, M.V., S. Neal, and D.S. Wishart, *PREDITOR: a web server for predicting protein torsion angle restraints.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W63-9.
68. Wang, G. and R.L. Dunbrack, Jr., *PISCES: a protein sequence culling server.* Bioinformatics, 2003. **19**(12): p. 1589-91.
69. Heinig, M. and D. Frishman, *STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W500-2.
70. Yan, R., et al., *A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction.* Sci Rep, 2013. **3**: p. 2619.
71. Yang, Y., et al., *Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates.* Bioinformatics, 2011. **27**(15): p. 2076-82.
72. Soding, J., *Protein homology detection by HMM-HMM comparison.* Bioinformatics, 2005. **21**(7): p. 951-60.

73. Wu, S. and Y. Zhang, *LOMETS: a local meta-threading-server for protein structure prediction.* Nucleic Acids Res, 2007. **35**(10): p. 3375-82.

74. Rotkiewicz, P. and J. Skolnick, *Fast procedure for reconstruction of full-atom protein models from reduced representations.* J Comput Chem, 2008. **29**(9): p. 1460-5.

75. Parsons, J., et al., *Practical conversion from torsion space to Cartesian space for in silico protein synthesis.* J Comput Chem, 2005. **26**(10): p. 1063-8.

76. Richardson, J.S. and D.C. Richardson, *The de novo design of protein structures.* Trends Biochem Sci, 1989. **14**(7): p. 304-9.

77. Lovell, S.C., et al., *The penultimate rotamer library.* Proteins, 2000. **40**(3): p. 389-408.

78. Alford, R.F., et al., *The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design.* J Chem Theory Comput, 2017. **13**(6): p. 3031-3048.

79. Xiong, P., et al., *Protein design with a comprehensive statistical energy function and boosted by experimental selection for foldability.* Nat Commun, 2014. **5**: p. 5330.

80. Schymkowitz, J., et al., *The FoldX web server: an online force field.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W382-8.

81. Subramani, A. and C.A. Floudas, *beta-sheet topology prediction with high precision and recall for beta and mixed alpha/beta proteins.* PLoS One, 2012. **7**(3): p. e32461.

82. Richardson, J.S. and D.C. Richardson, *Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(5): p. 2754-2759.

83. Yu, L., *A Novel Statistical Potential for Protein Beta-Sheets Prediction*, in *BioEngineering*. 2014, Rice University. p. 48.

84. Cukuroglu, E., et al., *Non-redundant unique interface structures as templates for modeling protein interactions.* PLoS One, 2014. **9**(1): p. e86738.

85. ISO/IEC, *ISO International Standard ISO/IEC 14882:2014(E) – Programming Language C++. [Working draft]*. International Organization for Standardization (ISO), 2014.

86. Saito, M. and M. Matsumoto, *SIMD-oriented Fast Mersenne Twister: A 128-bit random number generator.* Monte Carlo and Quasi-Monte Carlo Methods 2006, 2008: p. 607-622.

87. Zhang, Y. and J. Skolnick, *SPICKER: a clustering approach to identify near-native protein folds.* J Comput Chem, 2004. **25**(6): p. 865-71.

88. Xu, D. and Y. Zhang, *Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field.* Proteins, 2012. **80**(7): p. 1715-35.

89. Farzadfard, F., et al., *Beta-sheet capping: signals that initiate and terminate beta-sheet formation.* J Struct Biol, 2008. **161**(1): p. 101-10.

90. Chojnacki, S., et al., *Programmatic access to bioinformatics tools from EMBL-EBI update: 2017.* Nucleic Acids Res, 2017. **45**(W1): p. W550-W553.

91. Yang, H. and D.J. Patel, *Inhibition Mechanism of an Anti-CRISPR Suppressor AcrIIA4 Targeting SpyCas9.* Mol Cell, 2017. **67**(1): p. 117-127 e5.

92. Mitra, P., et al., *An evolution-based approach to De Novo protein design and case study on Mycobacterium tuberculosis.* PLoS Comput Biol, 2013. **9**(10): p. e1003298.

93. Kryshtafovych, A., et al., *Evaluation of the template-based modeling in CASP12.* Proteins, 2018. **86 Suppl 1**: p. 321-334.

94. Li, Z., et al., *Energy functions in de novo protein design: current challenges and future prospects.* Annu Rev Biophys, 2013. **42**: p. 315-35.

95. Tinberg, C.E., et al., *Computational design of ligand-binding proteins with high affinity and selectivity.* Nature, 2013. **501**(7466): p. 212-6.

# APPENDICES

## APPENDIX A

| 3AA+class[α] | 1AA+class[β] | RotChar[γ] | | 3AA+class[α] | 1AA+class[β] | RotChar[γ] |
|---|---|---|---|---|---|---|
| ALA1 | A1 | A | | ASN1 | N1 | d |
| CYS1 | C1 | B | | ASN2 | N2 | e |
| CYS2 | C2 | C | | ASN3 | N3 | f |
| CYS3 | C3 | D | | PRO1 | P1 | g |
| ASP1 | D1 | E | | PRO2 | P2 | h |
| ASP2 | D2 | F | | GLN1 | Q1 | i |
| ASP3 | D3 | G | | GLN2 | Q2 | j |
| GLU1 | E1 | H | | GLN3 | Q3 | k |
| GLU2 | E2 | I | | ARG1 | R1 | l |
| GLU3 | E3 | J | | ARG2 | R2 | m |
| PHE1 | F1 | K | | ARG3 | R3 | n |
| PHE2 | F2 | L | | SER1 | S1 | o |
| PHE3 | F3 | M | | SER2 | S2 | p |
| GLY1 | G1 | N | | SER3 | S3 | q |
| HIS1 | H1 | O | | THR1 | T1 | r |
| HIS2 | H2 | P | | THR2 | T2 | s |
| HIS3 | H3 | Q | | THR3 | T3 | t |
| ILE1 | I1 | R | | VAL1 | V1 | u |
| ILE2 | I2 | S | | VAL2 | V2 | v |
| ILE3 | I3 | T | | VAL3 | V3 | w |
| LYS1 | K1 | U | | TRP1 | W1 | x |
| LYS2 | K2 | V | | TRP2 | W2 | y |
| LYS3 | K3 | W | | TRP3 | W3 | z |
| LEU1 | L1 | X | | TYR1 | Y1 | 0 |
| LEU2 | L2 | Y | | TYR2 | Y2 | 1 |
| LEU3 | L3 | Z | | TYR3 | Y3 | 2 |
| MET1 | M1 | a | | | | |
| MET2 | M2 | b | | | | |
| MET3 | M3 | c | | | | |

**Appendix Table A**. Chi Rotamer Mappings

α : Three letter amino acid type representation + chi-1 rotamer class

β : One letter amino acid type representation + chi-1 rotamer class

γ : Single-byte representation of chi-1 rotamer

# Appendix B

| | | | | | |
|---|---|---|---|---|---|
| 4uucA | 4n67A | 3zdsA | 1suuA | 3gt5A | 4xzfA |
| 7odcA | 4j27A | 2pefA | 4p3vA | 4zv5A | 5enqA |
| 1io7A | 3ub6A | 3gr4A | 4iwbA | 3q1xA | 4iciA |
| 2qudA | 4exkA | 3h36A | 4x2pA | 1pm4A | 4yusA |
| 4lx3A | 4eetB | 2yxnA | 3u9qA | 4ybrA | 3t0hA |
| 1wbaA | 1kmvA | 5bvuA | 3mz2A | 3e05A | 1dz3A |
| 1ihjA | 2zb4A | 5idqA | 2xkiA | 1tuvA | 3mmyA |
| 3t47A | 4qkwA | 3l8dA | 2cdcA | 2petA | 3ixlA |
| 1d2sA | 4x9kA | 5dm2A | 4yg0A | 3c6aA | 5iqnA |
| 3zsjA | 4wriA | 3l4eA | 1w2wB | 1lucA | 1yb3A |
| 4i4tA | 1mj5A | 4kruA | 4a0dA | 3dwgC | 4mxtA |
| 3l1nA | 5hmtA | 4cnnA | 4ne3B | 3isaA | 4jgiA |
| 4apxB | 3znvA | 4icvA | 4lm8A | 1g8kA | 2qzuA |
| 4e0aA | 3wa2X | 3vgzA | 1viaA | 1k5nA | 3kf6B |
| 3e2qA | 2pr7A | 4ijnA | 1tkeA | 4tx5A | 1iujA |
| 4onwA | 1p9hA | 4c3sA | 3bonA | 3x38A | 1o7iA |
| 5it3A | 5a35A | 4m0nA | 5cxxA | 5ivkA | 2vzcA |
| 2aefA | 5i32A | 5hb7A | 4cj0A | 3qufA | 2q40A |
| 3tioA | 4n4uA | 1z9lA | 2rk3A | 2fnuA | 4zv0A |
| 3bsoA | 4txwA | 3hkwA | 5cfaA | 2gdmA | 1jf8A |
| 3m0zA | 3mqqA | 4pwoA | 4zx2A | 3gr3A | 2x46A |
| 5c2nA | 4jkzA | 3t8jA | 4n0kA | 3powA | 3w20A |
| 2gpiA | 2rb7A | 4f01A | 4uj7A | 1d8wA | 2qrlA |
| 3ha2A | 4trkA | 1xkpC | 4ke2A | 4h08A | 3eipA |
| 2wz1A | 4yucA | 2oemA | 4xraA | 2pq7A | 3iq0A |
| 3ak8A | 2gb4A | 2a14A | 3e23A | 4rl3A | 4xfmA |
| 3ii7A | 4yleA | 4q3kA | 3lwcA | 3n3mA | 3ly0A |
| 3gydA | 2r2dA | 3h7hA | 2pv2A | 2yjgA | 4kk7A |
| 5dofA | 4pkmA | 2iqyA | 3rkgA | 1sx5A | 4pdyA |
| 4zldA | 3d3zA | 4p3hA | 4v3lC | 4nbpA | 3dfrA |
| 2x5pA | 5cajA | 5aogA | 2ah5A | 5f1sA | 4p5nA |
| 3ebtA | 2bsyA | 3cc8A | 4z39A | 3r8yA | 1o98A |
| 2waoA | 3hnyM | 2qsaA | 4jjaA | 4h5iA | 3webA |
| 4g6tB | 1qftA | 2qguA | 1oygA | 3go2A | 4jaqA |
| 2nwrA | 1eaqA | 4hz2A | 2eabA | 4inkA | 4x8qA |
| 4uc1A | 2iruA | 5jh8A | 3nd1A | 4bqyA | 4eunA |
| 3zg9A | 2o1qA | 4ntkA | 3qzxA | 4uqwA | 4f0wA |
| 4hvtA | 1lfkA | 4zh5A | 4osnA | 2genA | 1kq1A |
| 5it6A | 1xakA | 2hlyA | 3gfaA | 1agjA | 3sjmA |
| 1jh6A | 3pa6A | 4bpsA | 3oqpA | 2bfeA | 1qsaA |
| 2olrA | 2uwaA | 5a61A | 3q23A | 4d05A | 3u9jA |
| 1t9hA | 3sqzA | 4c2vC | 2id3A | 2ejxA | 4ku0D |
| 4q8rA | 4k37A | 5cegA | 1iktA | 4jndA | 3ctzA |
| 3wydA | 3fdhA | 1xttA | 4plzA | 1oruA | 4od6A |
| 1tifA | 3fssA | 4q7qA | 3l32A | 4jp0A | 3hhyA |
| 3c4sA | 5iu1A | 3bmzA | 4bqhA | 3rpwA | 3bhnA |
| 2cviA | 3i45A | 4wu0A | 2anxA | 4jduA | 5elbA |
| 3lxpA | 4v0hA | 1qhqA | 4v1kA | 3f4mA | 2j8bA |
| 1yd9A | 1jcdA | 3zjaA | 3laeA | 5b2pA | 1nszA |
| 1p0hA | 4hatB | 4oh7A | 2xnqA | 4s12A | 5fpzA |

**Appendix Table B**: Rot1Pred Dataset

# Appendix C

| | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 6 | -1 | -4 | -2 | -6 | -3 | -5 | -5 | -3 | -5 | -3 | -4 | -3 | -3 | -4 | -1 | -2 | -2 | -8 | -6 |
| C | -1 | 12 | -7 | -7 | -5 | -6 | -5 | -4 | -7 | -4 | -3 | -4 | -7 | -6 | -6 | -2 | -2 | -2 | -7 | -5 |
| D | -4 | -7 | 7 | 0 | -9 | -4 | -3 | -9 | -3 | -8 | -7 | 0 | -4 | -2 | -4 | -2 | -4 | -8 | -9 | -7 |
| E | -2 | -7 | 0 | 7 | -8 | -5 | -3 | -7 | -1 | -6 | -5 | -2 | -4 | 1 | -2 | -2 | -3 | -5 | -8 | -6 |
| F | -6 | -5 | -9 | -8 | 8 | -9 | -3 | -3 | -7 | -2 | -1 | -7 | -8 | -6 | -7 | -6 | -6 | -4 | -1 | 1 |
| G | -3 | -6 | -4 | -5 | -9 | 6 | -6 | -10 | -5 | -9 | -8 | -3 | -6 | -5 | -6 | -3 | -6 | -9 | -9 | -9 |
| H | -5 | -5 | -3 | -3 | -3 | -6 | 10 | -6 | -2 | -5 | -4 | -1 | -5 | -1 | -2 | -3 | -4 | -6 | -4 | -1 |
| I | -5 | -4 | -9 | -7 | -3 | -10 | -6 | 7 | -6 | 0 | 0 | -7 | -7 | -5 | -6 | -6 | -3 | 1 | -7 | -5 |
| K | -3 | -7 | -3 | -1 | -7 | -5 | -2 | -6 | 7 | -5 | -4 | -2 | -4 | 0 | 1 | -2 | -2 | -5 | -7 | -5 |
| L | -5 | -4 | -8 | -6 | -2 | -9 | -5 | 0 | -5 | 6 | 1 | -7 | -7 | -4 | -5 | -6 | -5 | -1 | -5 | -4 |
| M | -3 | -3 | -7 | -5 | -1 | -8 | -4 | 0 | -4 | 1 | 10 | -5 | -6 | -2 | -4 | -5 | -3 | -1 | -4 | -4 |
| N | -4 | -4 | 0 | -2 | -7 | -3 | -1 | -7 | -2 | -7 | -5 | 8 | -5 | -2 | -3 | -1 | -2 | -6 | -7 | -5 |
| P | -3 | -7 | -4 | -4 | -8 | -6 | -5 | -7 | -4 | -7 | -6 | -5 | 8 | -4 | -5 | -3 | -4 | -5 | -8 | -8 |
| Q | -3 | -6 | -2 | 1 | -6 | -5 | -1 | -5 | 0 | -4 | -2 | -2 | -4 | 8 | -1 | -2 | -3 | -4 | -6 | -5 |
| R | -4 | -6 | -4 | -2 | -7 | -6 | -2 | -6 | 1 | -5 | -4 | -3 | -5 | -1 | 8 | -3 | -3 | -5 | -6 | -5 |
| S | -1 | -2 | -2 | -2 | -6 | -3 | -3 | -6 | -2 | -6 | -5 | -1 | -3 | -2 | -3 | 7 | 0 | -5 | -7 | -5 |
| T | -2 | -2 | -4 | -3 | -6 | -6 | -4 | -3 | -2 | -5 | -3 | -2 | -4 | -3 | -3 | 0 | 7 | -2 | -7 | -6 |
| V | -2 | -2 | -8 | -5 | -4 | -9 | -6 | 1 | -5 | -1 | -1 | -6 | -5 | -4 | -5 | -5 | -2 | 6 | -7 | -5 |
| W | -8 | -7 | -9 | -8 | -1 | -9 | -4 | -7 | -7 | -5 | -4 | -7 | -8 | -6 | -6 | -7 | -7 | -7 | 11 | -1 |
| Y | -6 | -5 | -7 | -6 | 1 | -9 | -1 | -5 | -5 | -4 | -4 | -5 | -8 | -5 | -5 | -5 | -6 | -5 | -1 | 9 |

**Appendix Table C**. COLLMX95 Matrix.