Published in final edited form as:

Proc WRI World Congr Comput Sci Inf Eng. 2012 January 24; 125: 781–786. doi:
10.1007/978-3-642-25789-6_106.

TERPRED: A Dynamic Structural Data Analysis Tool

Karl Walker¹, Carole L. Cramer², Steven F. Jennings³, and Xiuzhen Huang⁴

¹University of Arkansas at Little Rock, Little Rock, AR 72204 USA

²Arkansas State University, State University, AR 72467 USA

³University of Arkansas at Little Rock, Little Rock, AR 72204 USA

⁴Arkansas State University, State University, AR 72467 USA

Abstract

Computational protein structure prediction mainly involves the main-chain prediction and the side-chain confirmation determination. In this research, we developed a new structural bioinformatics tool, TERPRED for generating dynamic protein side-chain rotamer libraries. Compared with current various rotamer sampling methods, our work is unique in that it provides a method to generate a rotamer library dynamically based on small sequence fragments of a target protein. The Rotamer Generator provides a means for existing side-chain sampling methods using static pre-existing rotamer libraries, to sample from dynamic target-dependent libraries. Also, existing side-chain packing algorithms that require large rotamer libraries for optimal performance, could possibly utilize smaller, target-relevant libraries for improved speed.

Keywords

protein structure; protein side-chain; rotamer library

I. Introduction

Proteins, such as enzymes, carrier proteins, receptors and antibodies, play important roles in the cell. There are four levels of protein structures: primary structure (protein sequence), secondary structure, tertiary structure and quaternary structure. Protein tertiary structure is essential for its correct function in the cell. Protein tertiary structure prediction is a very important research area. Its applications include drug design in medicine and the design of enzymes in biotechnology. To make fast and accurate predictions of protein tertiary structures remains very challenging. Computational protein structure prediction mainly involves the main-chain prediction and the side-chain confirmation determination.

As more structures are determined and deposited in the Protein Data Bank, it becomes more evident that proteins are composed of regular secondary structures. Secondary structures are conformations formed over short segments of the polypeptide chain. Examples of secondary structures include helices, beta-sheets, and random coils. Chou and Fasman calculated

conformational parameters for residues in various secondary structures from known protein structures (Chou and Fasman, 1974). These parameters quantify the propensity of an amino acid to be in a helix, sheet, or random coil. A method was later developed for predicting secondary structures of proteins from amino acid sequences (Chou & Fasman, 1978). However, the original Chou-Fasman parameters are now considered to be unreliable (Kyngas & Valjakka, 1998), and have since been updated using current data and a modified Chou-Fasman algorithm (Chen, Gu, & Huang, 2006). Currently, there are several secondary structure prediction methods that have been developed. Secondary structure prediction has become an integral part of many tertiary prediction methods.

Although a large number of tertiary prediction methods have been developed, most of these can be grouped into one of the two categories, that is, methods that use predetermined structures as parameters (template-based modeling) and those that do not (*ab initio* modeling). Template-based modeling has produced some of the fastest and most accurate server predictions. The accuracy in predicting a target is often dependent upon how well it can be aligned to template structures. To identify structure templates from the PDB is still challenging and there are techniques such as sequence profile-profile alignments, hidden Markov models, machine learning, and others (Zhang, 2008). When structural templates do not exist or cannot be identified by threading, *ab initio* methods are used for structure prediction. These methods do not rely on predetermined structures, but seek to determine structure from scratch. The core is to develop a knowledge-based scoring function to discriminate native structures from nonnative ones and sample this potential in search of a global minimum (Bowman & Pande, 2009).

Regardless of the methods used for prediction, finding the lowest-energy, side-chain conformation is an extremely important step. This step in protein structure prediction is actually considered as a separate problem: the protein side-chain packing problem. The major difficulty with predicting side-chain conformations is that there are an extremely large number of possible conformations for even small residues. In order to reduce the potential search space, the number of allowed conformations for each type of side-chains is restricted to a limited number of configurations called rotamers. The typical approach is to generate a collection of these rotamers for each type of residue, called a rotamer library, from structural bioinformatics or other statistical analyses of side-chain conformations in experimentally determined protein structures (Dunbrack, 2002). These libraries generally contain information about the conformation, its frequency, and the variance about the mean dihedral angle. Rotamer libraries can be backbone independent, secondary structure dependent, or backbone dependent.

In this paper, we present a new structural data analysis tool, called TERPRED, which implements a hybrid idea to protein structure prediction. The approach we describe applies the idea of template-based modeling in that it uses predetermined protein structures as a parameter. However, it differs in that it does not require direct alignment of the target sequence onto template sequences or structures. It also applies the idea of *ab initio* methods in that it uses physiochemical properties of amino acids to drive the prediction towards the lowest free energy. It differs in that it does not require computationally expensive simulations to predict structure. We present an algorithm that generates a tertiary prediction

search space from the amino acid sequence of a target protein. Compared to various modern rotamer sampling methods (Dunbrack, 2002; Lovell *et al.*, 2000; Xiang & Honig, 2001; Peterson, Dutton, & Wand, 2004), our work is unique in that it provides a method to generate a rotamer library dynamically based on small sequence fragments of a target protein.

II. Dynamic Structural Data Analysis

Our new structural data analysis tool called TERPRED is composed of two components: the Database and the Rotamer Generator, which are discussed in the following two subsections. TERPRED can perform well even with very low sequence identity (<10%). The algorithm accepts various customizable parameters including amino acid sequence, a set of predetermined structures (global search space), secondary structure prediction, motifs, and domains.

A. The TERPRED Database

The purpose of the construction of the TERPRED platform was to enable detailed analysis of structural data in order to determine the factors that define the structure of a protein. In order to do this, a database was constructed to hold tertiary structure data in a relational format so as to facilitate the querying of structures based upon various combinations of attributes. Over thirty database tables were constructed with interconnectivity at the forefront of the design. Database tables were designed to hold sequence data, Cartesian coordinate data, polar coordinate data, secondary structure information, binding sites information, amino acid properties, and metadata about a protein such as its type, function, organism, etc. There are also tables to hold data about the encoding and translation of a protein such as mRNA transcript, genetic code, and codon-usage tables. These tables are provided in order to facilitate queries about the co-translational effect on protein structure.

Data to be loaded into the tables were downloaded from the Protein Data Bank, NCBI, and other sources. In order to facilitate loading data into the TERPRED database, several Perl scripts were designed to read, extract, and store data into the appropriate tables. The data from a single PDB file were stored across several database tables.

Predetermined structures in the Protein Data Bank were downloaded to the TERPRED server. The file organization structure used on the PDB server was retained after the data was migrated. Initial testing of the TERPRED system was first done with approximately twenty randomly-selected proteins, and benchmark testing was later done using the Lindahl dataset (Lindahl and Elofsson, 2000), which is a set of 976 non-redundant protein structures widely used to test prediction algorithms. The PDB record IDs corresponding to each of the records in the Lindahl dataset were extracted and each of the structure files were loaded directly from the original PDB files. This was done so that all of the metadata or descriptive information could be extracted in addition to the structural information.

Other information needed for the TERPRED system was derived from data stored within PDB records. Secondary structure assignments were made using the algorithm developed by Wolfgang Kabsch and Chris Sander called DSSP (Define Secondary Structure of Proteins).

The standard method for assigning secondary structures to the amino acids of a protein is the DSSP algorithm using the atomic-resolution coordinates of the protein. By means of an electrostatic definition, the hydrogen bonds of the protein structure are first identified. A hydrogen bond is identified by the DSSP algorithm if the energy E is less than -0.5 kcal/mol. The torsion (or dihedral) angles of bonds were also derived from the atomic coordinates within the PDB files.

After the data were successfully downloaded, derived, and loaded into the system, queries could be run against it. The following are several example queries, “What proteins have a specific pattern within their sequences?” or “What are the various sequence segments that fold into a right-handed alpha-helix?” The answers of the queries are limited by the quality and quantity of the data within the database. As a result, TERPRED was designed to hold a large number of predetermined structures while providing the capability of various structures to be grouped into datasets that allow users to query a subset of the total available records. Users may utilize all of the records loaded into the system, a public dataset such as the Lindahl dataset, or create their own custom dataset (collection of PDB identifiers, model numbers, and chain identifiers). This allows users to generalize the search or limit the scope of the search to a set of non-redundant structures or to a set of specific structures like globular proteins, enzymes, receptors, etc. based upon information known (or hypothesized) about a target protein. There are various preloaded datasets within the TERPRED Rotamer Generator, which will be discussed in the next section. We provide a few queries to the TERPRED database in the following.

First, a simple query to extract all of the conformations for a particular torsion angle of an amino acid will be constructed. In other words we want to know the answer to the question, “How many different states can the chi torsion angle of Valine be observed?” To construct this query, we issue the command: `SELECT chi FROM pdb_angles WHERE aa = 'V'`. A similar query for each amino acid can be submitted for other amino acids, and a probability distribution can be computed from the results. Now that we have the total distribution for each amino acid, we can see how this distribution changes as we constrain the search results even further. We accomplish this by relating the torsion angles to data in other tables such as `pdb_dssp` which holds the secondary structure assignments of amino acids. For example, now that one has seen the total probability distribution of Valine (Figure 1A), one may want to see how the distribution differs when Valine is in an alpha helix (Figure 1B). Inside the TERPRED database, this search can be conducted by using the secondary structure annotations extracted from the PDB files (`pdb_helix` table) or by using the secondary structure assignments automatically generated by DSSP (`pdb_dssp` table). We will use the latter in our example: `SELECT a.chi FROM pdb_angles a, pdb_dssp d WHERE a.recname = d.recname AND a.seqno = d.resnum AND a.aa = 'V' AND d.ss = 'H'`;

The two queries shown above demonstrate how the possible torsion angles may be constrained based upon the amino acid and secondary structure of the amino acid. What about a particular sequence pattern? Selecting the possible chi torsion angles based upon a particular sequence pattern is a more complex SQL statement. However, the process of retrieving torsion angles based upon a sequence pattern has been encapsulated into a stored procedure within the TERPRED database. Therefore, the question “what are the possible

configurations of the torsion angles for each residue in the sequence 'KDEL', can be answered (based on the dataset) by a call to the stored procedure getangles: CALL getangles('KDEL'); The getangles procedure accepts a single parameter: the sequence pattern to search for as a regular expression. The getangles procedure is one of several procedures or functions developed as part of the TERPRED database to simplify complex querying. The getangles procedure actually depends upon two of these functions to select angles based on a sequence pattern, rlocate and rendpos. These functions find the first and last position, respectively, of the sequence fragment matched by the regular expression passed to the parameter. The reextract function uses the rlocate and rendpos to extract a matching sequence fragment. This function is especially useful when an inexact pattern is supplied as the regular expression pattern. For example, instead of strictly searching for 'KDEL', one may want to extend the search to other Endoplasmic Reticulum (ER) retention signals like 'QDEL', for example: CALL getangles('[KQ]DEL'); The brackets "[]" instruct the interpreter to match any one of the characters between them. Therefore, either 'KDEL' or 'QDEL' will be matched by the statement above, and the reextract function is used to find out which of the possible patterns it matched.

To investigate the relationship between the amino acid sequence and the tertiary structure, as is shown one can constrain the many possible conformations of an amino acid's side-chain by limiting it to a particular secondary structure, and even further by limiting it to within a particular pattern. A logical question to investigate further is how adjacent amino acids effect the possible torsion angle conformations. Do adjacent amino acids impose physical constraints on a side-chain due to size or shape? What about constraints imposed by chemical properties like charge or hydrophobicity? In other words, what influence do adjacent amino acids have, or what is the local influence on structure? Using the getangles procedure mentioned earlier, one can look at how Valine responds when in a sequence between two Proline amino acids: CALL getangles('PVP'); Figures 1D and 1G show the distribution of Valine between two Prolines (according to the Lindahl dataset). One may also search for less specific patterns. For example, one may want to see how Valine responds when between two basic amino acids: CALL getangles('[RK]V[RK]'); In the example above, either of the two basic amino acids, Arginine and Lysine, can be matched on either side of Valine. If one assumes that the native pH is 6.0 or lower, one could also include Histidine as a basic amino acid: CALL getangles('[RKH]V[RKH]'). As demonstrated, possible angle conformations can be extracted for specific or more general patterns.

B. The TERPRED Rotamer Generator

The concept of finding the local amino acid influence has been integrated into an algorithm that generates a tertiary prediction search space from an amino acid sequence. The TERPRED Rotamer Generator is primarily designed to address the search space representation aspect of the protein side-chain packing problem. However, since torsion angles for both side-chain and main-chain bonds were computed from the atomic coordinates in PDB files and integrated into the TERPRED Database, it is flexible enough to use in virtually any project that involves structural analysis: protein main-chain prediction, protein design, homology modeling, and the protein docking problem. As the free energy

landscape limits the potential folding hyperspace *in vivo* to those energetically reachable, TERPRED seeks to limit the prediction search space *in silico* to those configurations observed in proteins with similar sequence, structural elements, or function.

To accomplish this goal, TERPRED provides several parameters that may be customized based upon known properties of a target protein. These parameters facilitate the dynamic search and retrieval of structural data relevant to a specific target. The customizable parameters allow a tertiary predictor to select a subset of model structures and extract conformational data based on the influence of local amino acids, secondary structure predictions, and key words (motifs) or phrases (domains) found within an amino acid sequence. TERPRED provides several parameters whose use can be customized to search the database of known structures using more information than just the amino acid sequence of the target protein. For example, if the target protein is a known enzyme, then one can select “enzymes” as a subset of model structures. Or if one knows that the target protein is a dehydrogenase enzyme, one can further limit the set of model structures by supplying “dehydrogenase” in the keyword parameter. This allows one to find models with a functional relationship to the target. Perhaps one may not know the function, but may have an idea of the type of protein it is (globular, trans-membrane, disordered), then “globular” or “trans-membrane” proteins may be selected as a subset of model structures. There are also options to select preloaded datasets with non-redundant model structures such as the Lindahl dataset. One can also generate his or her own non-redundant set of model structures to be used with TERPRED using tools like SCOP or CATH (Csaba et al., 2009).

The window of influence that neighboring amino acids have on the conformational space can be set to an odd value between 3 (± 1) and 7 (± 3). This type of influence may be considered as general influence because it is not based upon a specific pattern of amino acids known to be associated with a motif or domain. Within TERPRED, the influence of neighboring amino acids is captured using a sliding window approach. In this approach, a window of the selected size is scanned across the amino acid sequence of the target beginning at the first residue and incrementing by one until the last window of that size is reached. In our example, a window size of 3 is chosen to capture the influence of the amino acids on either side (± 1) (Figure 2a). Each pattern extracted from the window scan is searched for in the database (Figure 2b). After the patterns are found, the associated dihedral angles are extracted and assigned to the appropriate residue of the target sequence (Figure 2c). With a window size of three, each residue (except for the first and last two) will appear in three positions: as the first, second, and last residue of a window. The angle conformations of a particular residue are stored selecting all the conformations or only those where the residue is in the center of the pattern (position 2 in Figure 2c), for example.

Depending on whether one chooses to keep all the angles from the initial analysis or not, there are other options to further limit the possible conformations using concepts in probability theory and/or set theory. The major difference between the two approaches is that probability theory considers the number of occurrences of any particular value, whereas set theory does not consider this. An obvious method of selecting angles is to select the ones with the highest probability within a given distribution. The distribution from which the most probable conformations are selected may be based on the amino acid, amino acid in a

secondary structure, or an amino acid within a sequence fragment (Figures 2a-d). Methods based on probability theory seek to answer the question “What is the most likely conformation given the values of x, y, z parameters?”

Alternatively, methods based on set theory seek to answer the question “What are the possible conformations, and which ones are in common between groups?” One such method of filtering possible angle conformations is to find the mutual influence that all the surrounding residues had on a particular residue. In order to find the mutual influence, the sets of angles corresponding to each of the three positions (in the three patterns) may be intersected pair-wise, and any angle conformation that exists in at least two of the sets will be retained: $V_{\text{mutual}} = (V_{p1} \cap V_{p2}) \cup (V_{p1} \cap V_{p3}) \cup (V_{p2} \cap V_{p3})$ Using intersections is a very crude method of filtering out angle conformations. However, in our test example, it proved to be quite effective even with a small dataset of only 970 structures (see Figure 3). The accuracy of this method may improve greatly by increasing the number of structures in the database.

The secondary structure information may also be used to filter angle conformations. This requires that the sequence of the target protein be submitted to one of several secondary prediction algorithms. After a secondary structure prediction is attained, the angle conformations selected in the sliding-window analysis are filtered for those matching the secondary structure predicted for each amino acid. Those conformations that do not match are eliminated. Conceptually, the set of angles that satisfy the sequence-dependence are intersected with the set of angles that satisfy the secondary structure-dependence: $V_{p, \text{helix}} = (V_{p1} \cup V_{p2} \cup V_{p3}) \cap V_{\text{helix}}$, where V_{helix} equals the distribution of chi for all Valine side-chains in an alpha helix. If the user has chosen to keep only one of the sets in the initial analysis or filtered them by mutual influence, this filter can still be performed: $V_{p2, \text{helix}} = V_{p2} \cap V_{\text{helix}}$, $V_{\text{mutual, helix}} = V_{\text{mutual}} \cap V_{\text{helix}}$.

As shown earlier, torsion-angle conformations may also be selected for a specific pattern corresponding to a motif or domain. Rotamers selected from such specific patterns are not grouped or filtered along with the rotamers selected from non-specific patterns. They are considered separately by the algorithm and are assumed to be more likely candidates than non-specific rotamers. Selecting rotamers based upon motifs and domains first requires that these elements be found in the target sequence. For this purpose, TERPRED integrates the analysis with third-party tools designed to locate such patterns within an amino acid sequence. To search for motifs within a sequence, TERPRED uses the PROSITE scanning tool, ScanProsite (de Castro *et al.*, 2006). Domains can be located using tools like SCOP (Murzin *et al.*, 1995) or CATH (Orengo *et al.*, 1997). These tools use databases of known sequence patterns or tertiary structure elements to search for motifs or domains, respectively. The motifs or domains matched within a target sequence are then submitted to TERPRED for the extraction of rotamers based upon their associated patterns.

III. Discussion and Further Work

We have developed a web-based tool capable of selecting rotamers, not only by amino acid and secondary structure but also by factoring in the influence of neighboring amino acids.

We have presented this new structural data analysis tool and have shown ways in which it may be used to analyze the protein structures maintained in the Protein Data Bank. The initial testing of this tool has shown promising success in generating rotamer libraries containing values very close to the actual values of the test structures. The Rotamer Generator provides a means for existing side-chain sampling methods that use pre-existing, static rotamer libraries to sample from dynamic, target-dependent libraries. The filtration step of the Rotamer Generator is of great importance in reducing the computational load on protein side-chain packing algorithms. The secondary structure and mutual influence methods have shown to be quite effective at reducing the number of possibilities while retaining accurate values. A combination of filters generally applicable to any target protein is yet to be determined.

We would further work on the third component of TERPRED, the Structure Modeler. The goal of the Structure Modeler is to find the lowest-energy conformation, given the rotamers of the Rotamer Generator as parameters. However, existing side-chain packing algorithms that require large rotamer libraries for optimal performance could possibly utilize smaller, target-relevant libraries produced by the Rotamer Generator for improved speed. Furthermore, the TERPRED platform provides researchers with the tools to closely analyze the relationship between sequence and structure and also allows researchers to search for the protein folding code.

Acknowledgments

This publication was made possible partly by NSF Experimental Program to Stimulate Competitive Research (EPSCoR) Arkansas Center for Plant-Powered Production (P3) seed grant Fund# 224050 and NIH Grant # P20 RR-16460 from the IDeA Networks of Biomedical Research Excellence (INBRE) Program of the National Center for Research Resources.

References

1. Bowman GR, Pande VS. The roles of entropy and kinetics in structure prediction. *PLoS One*. 2009; 4(6)
2. Chen H, Gu F, Huang Z. Improved Chou-Fasman method for protein secondary structure prediction. *BMC Bioinformatics*. 2006; 7(Suppl 4):S14. doi:10.1186/1471-2105-7-S4-S14. [PubMed: 17217506]
3. Chou PY, Fasman GD. Prediction of protein conformation. *Biochemistry*. 1974; 13(2):222–45. [PubMed: 4358940]
4. Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Molecular Biology*. 1978; 47:45–148.
5. Csaba G, Birzele F, Zimmer R. Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. *BMC Structural Biology*. 2009; 9(23) doi:10.1186/14726807-9-23.
6. De Castro E, Sigrist CJA, Gattiker A, Bulliard V, Petra S, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research*. 2006; 34:362–365.
7. Dunbrack RL. Rotamer libraries in the 21st Century". *Current Opinion Structural Biology*. 2002; 12(4):431–440. doi:10.1016/S0959-440X(02)00344-5.
8. Kyngas J, Valjakka J. Unreliability of the Chou-Fasman parameters in predicting protein secondary structure. *Protein Engineering*. 1998; 11(5):345–348. [PubMed: 9681866]
9. Lindahl E, Elofsson A. Identification of related proteins on family, superfamily and fold level. *J Mol Biol*. 2000; 295(3):613–25. [PubMed: 10623551]

10. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*. 1995; 247:536–540.
11. Orengo CA, Michie AD, Jones DT, Swindells MB, Thornton JM. CATH: A. 1997
12. Peterson RW, Dutton PL, Wand AJ. Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. *Protein Sci*. 2004; 13:735–751. [PubMed: 14978310]
13. Xiang Z, Honig B. Extending the accuracy limits of prediction for side-chain conformations. *Journal of Molecular Biology*. 2001; 311:421–430. [PubMed: 11478870]
14. Zhang Y. Progress and challenges in protein structure prediction. *Current Opinion Structural Biology*. 2008; 18(3):342–348. doi:10.1016/j.sbi.2008.02.004.

Figure 1.

The distributions of Valine (A) in all secondary structures, (B) in a helix, (C) in a sheet, (D) between two proline residues, (E) between two proline residues in a helix, (F) and between two proline residues in an extended beta sheet. Diagrams G, H, I depict the same data as D, E, and F, respectively, but show the angle plots and the associated rose diagram (circular histogram).

Figure 2.

(a) The amino acid sequence is scanned in windows of size 3 and (b) the database is searched for each triplet of amino acids. (c) The angles found for each amino acid are assigned to the corresponding residue of the target amino acid sequence. (d) The angles are color coded according to the position they are in and graphed.

Figure 3.

(left) The angles extracted for residue V3 in all three window positions are shown as different color dots. (right) The sets of angles corresponding to each position were intersected pair-wise and those existing in at least two of the three groups were retained. The open circle is the plot of the value computed from the test protein: 2AAI (not in the database). Note that two of the rotamers retained after filtration lie within the open circle.