The Classification of Local Structures for Modeling Protein Structure Geometric Constraints

Kentaro Onizuka, Kyosuke Saito, Stephen T. C. Wong
1: Institute for New Generation Computer Technology
2: Electrotechnical Laboratory
3: Lehigh University

1 Abstract

In this paper, we describe a new method to classify three-dimensional local structures of protein and to model the geometric constraints of the protein tertiary structures. These constraints would allow us to predict tertiary structure more accurately than existing techniques.

2 Introduction

The prediction of a protein's three-dimensional structure (i.e., its tertiary structure) from its amino acid sequence (i.e., primary structure) is an important yet unsolved problem in molecular biology. Many methods have been proposed but none meet practical demands. A majority of these methods predict the site of stable structures such as helices, strands and turns (i.e., secondary structures) from the primary structure of each site [Fasman 89], and pack the predicted secondary structures into a tertiary structure [Cohen et al 82].

Obviously, such methods work if the assumption that the local primary structure determines the secondary structure of that region is valid. So far, however, they generate poor prediction results. Recently, it is known by the molecular biologists that such an assumption ignores the critical factor that a local or a secondary structure is constrained not only by the primary structure of that region but also by the environment of the global structure [Branden and Tooze 91].

To include this factor into protein structure prediction, one must be able to model the constraints between local structures of various sizes (i.e., geometric constraints) as well as the constraints between primary structures and local structures (i.e., primary constraints). Our research goal is to devise a reliable prediction method that incorporates such constraint-based reasoning.

There are two crucial elements to this prediction method. 1) the precise classification of local three-dimensional structures of various sizes for modeling the geometric constraints between neighboring local three-dimensional structures, and 2) the precise representation of local primary structures of various lengths for modeling the primary constraints between the tertiary and primary structures. In this paper, we propose a novel technique to classify local structures and to describe the geometric constraints between the local structures of various sizes based on the data from Protein Data Bank (PDB).

The organization of this paper is as follows. Section 3 discusses how to obtain a set of representation parameters for the local structures. Section 4, briefly, describes how to classify the local three-dimensional structures using a clustering technique. Section 5 illustrates how to model geometric constraints. The final section presents our conclusion.

3 Local Structure Representation

Several techniques of local three-dimensional structure classification have been proposed so far [Matsuo 90], but none can be used for our purpose. They are inherently restricted to structures with a small fixed number of residues because large numbers of parameters are required to represent a local three-dimensional structure with many residues.

Existing methods usually requires $3N - 6$ or more parameters to represent a local three-dimensional structure of $N$ residues even when we consider the position of the C' atom as the
representative position of each residue in that structure. The reason is that we require a total of $3N$ parameters to represent $N$ positions in three-dimensional space, and we subtract the degrees of freedom in rotational and translational transformation, $\delta$, from the total number of parameters. In contrast, our new technique extracts a fixed number of parameters from a local three-dimensional structure of any number of residues (i.e., size of the structure). Thus, we can represent local structures of various sizes within a single technique.

If we allow the resolution of the representation to be changed according to the structure's size, that is, small structures are represented in high resolution, and large structures with many residues are represented in low resolution, the number of parameters can then be further reduced. Furthermore, we can restore a large three-dimensional structure with many residues from its approximate form when we know how the structure is built up out of its component smaller structures. We call this description of large structure in terms of its smaller components hierarchical description.

Our technique to obtain the fixed number of parameters from a local three-dimensional structure is as follows. First, we define several topological vectors that describe the topological properties of local structures. We name a topological vector as length vector, curve vector, twist vector, and meander vector according to the feature that it represents. We then define unit coordinate vectors of the structure to be represented using two of these topological vectors. For XYZ coordinates, we set the length vector on positive Z coordinate and a curve vector on X-Z positive plane. The total number of non-zero parameters representing the set of topological vectors is $3M - 3$, where $M$ is the number of topological vectors. In this way, we represent a local structure of $N$ residues with $3M - 3$ parameters.

Wavelet transforms is used to determine the topological vectors. Roughly speaking, wavelet transforms are techniques to obtain the local behavior of a function with a localized wave packet (i.e., wavelet) [Combes et al 89]. A topological vector $T$ is determined as the sum of the coordinates of the residues' position weighted by a discrete wavelet, that is:

\[ T = \sum_i w_i X_i, \]  

where each $X_i$ is the vector of fixed origin representing the position of each $C^\alpha$ atom in a local structure and $w_i$ is the corresponding wavelet.

Figure 1 shows the four wavelets that we use in this study. The length vector is determined by applying wavelet type 1 to a local structure, the curve vector by applying type 2, the twist vector by applying type 3, and the meander vector by applying type 4.

![Wavelet Types](image)

**Figure 1:** Wavelets

Typically, a wavelet function $\psi(x)$ satisfies the following conditions.

\[ \begin{align*}
& \int_{-\infty}^{\infty} \psi(x) dx = 0 \\
& \int_{-\infty}^{\infty} (\psi(x))^2 dx = 1
\end{align*} \]  

The function $\psi(x)$ has both positive regions and negative regions. Wavelet transform in this case is defined as a transformation that uses the function $\psi((x - b)/a)$ as a base with base parameters $a$ and $b$:

---

114 ---
\[ F(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} w \left( \frac{z-b}{a} \right) f(z) dz, \]  

(3)

where \( f(x) \) is the function to be transformed. Here, \( a \) determines the width and the local frequency, and \( b \) determines the position of the wavelet.

We developed a technique to generate a reasonable wavelet in a discrete system (i.e., a discrete wavelet). Hence, the wavelet consists of a set of weights \( w_i \). The conditions in (2) then become the following:

\[
\begin{align*}
\sum_i w_i &= 0 \\
\sum_i w_i^2 &= 1
\end{align*}
\]  

(4)

4 Classification of Local Structures by Clustering

![Figure 2: Clusters of five-residue local structures](image)

The local structures can be classified by clustering the parameters of local structures' topological vectors. We used a method involving a hypercube histogram. Briefly, in this method, we first count the number of data in each hyper lattice cube in \( n \) dimensional space where \( n \) is the number of data components. We then merge the hypercubes to form a cluster according to the number of data these cubes and the Euclidian distance between the cubes. This method is powerful especially in dealing with complex forms of cluster and a large amount of data. Figure 2 illustrates the result of clustering the local structure of five residues.

5 Formation of Geometric Constraints

As shown in Figure 3, there are two classes of geometric constraints: hierarchical constraints and overlapping constraints. A hierarchical constraint relates a class of local structures with several classes of smaller structures (i.e., components). One can also represent a hierarchical constraint with a set of component constraints that relates individual components with the larger structure under a certain offset position of its beginning reside. An overlapping constraint states that a class of local structures should overlap a class of local structures of the same size. Likewise, this constraint can also represented as a constraint between two class of local structures with an offset.
These geometric constraints are derived from the protein tertiary structures recorded in PDB. Since they are based on known tertiary structures, they may not necessarily hold for unknown proteins. Well-predicted tertiary structures, however, are not expected to violate many constraints. In other words, we can gauge the accuracy of a predicted tertiary structure by the number of constraints the tertiary structure that it satisfies. The more constraints that a candidate structure is satisfied, the more precise is its prediction.

6 Conclusion

In this paper, we have proposed a new method that modifies the naive assumption of existing prediction methods of protein structures with more updated biological knowledge. This method can be used to classify local structures of various sizes, and to model the geometric constraints of protein structure. We have illustrated its practicality with experimental results. The next stage of this research is to study the constraints between a local protein structure and the primary structure of that region.

References


