A Multi-Level Description Scheme of Protein Conformation

Kentaro Onizuka,

Masato Ishikawa and **Stephen T.C. Wong**
Institute for New Generation Computer Technology (ICOT)
1-4-28, Mita, Minato-ku, Tokyo, 108 Japan
onizuka@icot.or.jp,
ishikawa@icot.or.jp, wong@icot.or.jp
FAX +81-3-3456-1618

Kiyoshi Asai
Electrotechnical Laboratory (ETL)
1-1-4, Umezono, Tsukuba Ibaraki, 305 Japan
asai@etl.go.jp
FAX +81-298-585939

Abstract

We propose a novel description scheme of protein backbone conformation that can model the important factors of protein structure formation, such as global interaction and geometric constraints. This description scheme represents a protein conformation with several symbolic sequences of multiple levels of abstraction. Each symbol in the sequence denotes the class of abstracted topology of subconformation with the size specific to the level. Low level sequences of this description represent fine structures of high resolution, and high level sequences represent the abstracted topologies of large scale.

The classification of protein backbone subconformations of various sizes is the most important base for this description scheme. This has never been tried so far due to the complexity in dealing with the number of degrees of freedom in subconformations. However, the proposed technique solved this problem by abstracting the topology of middle and large scale subconformations. This linear expansion technique extracts a fixed number of parameters as the expansion coefficients from the coordinate representation of subconformations. In this case, the simple reverse-transformation from the expansion coefficients reconstructs the three-dimensional topology of a subconformation.

The analysis of the relation between primary structure of a region and the subconformation of that region at each level in this description helps to model both local and global interactions of protein structure formation. Further, the statistic analysis of overlapping patterns of two subconformations models the geometric constraints important for a structure prediction system in generating a conformation which is geometrically sound.

Keywords

Multi-level Description, Abstruction of Subconformation, Resolution, Subconformation of Various Sizes, Linear Expansion, Fixed Number of Parameters, Classification of Subconformation, Clustering, Protein Structure Prediction.

Introduction

One of the most significant aspects of protein structure prediction is the description scheme of protein conformation. The conventional prediction methods usually adopt the sequence of secondary structures as the description of protein conformation (Chou and Fasman 74). Despite its popularity, such a description suffers two major problems.

1. Since the description is only concerned with the local conformation, it is impossible for the prediction methods to include the global interaction between

distant sites in the primary structure (Branden and Tooze 91; Asai et al 93; Metfessel and Saurugger 93).

2. The description does not have the sufficient information for reconstructing the three-dimensional conformation. This means that the description does not have complete information of three-dimensional conformation. It is impossible to model three-dimensional geometric constraints.

Thus, it is difficult for the secondary structure prediction schemes to include the important factors and constraints of protein structure formation. The accuracy of the secondary structure prediction is, therefore, too poor to meet the practical demands in spite of all the improvements proposed so far (Garnier et al 78; King and Sternberg 90; Qian and Sejnowski 88; Lim 74; Bohr et al 88; Cohen et al 86). Also, a method to pack the sequence of secondary structures into tertiary one is necessary to predict the three-dimensional conformation besides the secondary structure prediction. The tertiary structure prediction from the sequence of secondary structures is yet another difficult problem (Cohen et al 82; Fasman 89).

We propose a novel description scheme for protein backbone conformation that can model the aforementioned significant factors or constraints. In this scheme, *a protein conformation is described by multiple levels of abstraction.* At each level, a symbolic sequence represents the protein conformation, where each symbol denotes the class of subconformation of that level size. A low level sequence represents the fine conformational structures, such as secondary structures with fairly high resolution, and a high level sequence represents the abstracted large-scale or global topologies with low resolution.

The most important base of our scheme is the classification of subconformations of various sizes. Despite a lot of classifications proposed so far, they are, in general, concerned with small subconformations of a fixed size, such as those of two or three residues (Miller et al 93). Since the number of degrees of freedom in the topology is almost proportional to the number of residues in a subconformation, it is hard to formalize a uniform method to classify the subconformations of different sizes.

A large-scale subconformation is complex because of the large number of degrees of freedom. Some abstraction on the subconformation's topology is needed to reduce this complexity. Our technique does this by linearly expanding the coordinate representation of subconformations. A fixed series of expansion coefficients is obtained as the result of the expansion. With slight modification, these coefficients would become numerical parameters that represent the topology of the subconformation. Further, we can control the number of parameters by cutting the expansion at an appropriate order. Hence, the whole conformation of a protein is described with a set of symbolic sequences where each symbol denotes the class of a subconformation and the different sequence represents the topology at the different level of abstraction. There are three advantages in such a multi-level description.

1. The analysis of the relation between the topology of a subconformation and the primary structure of that region at each description level models both the local, intermediate, and global interactions of protein structure formation.
2. The statistic analysis of the overlapping patterns of neighboring subconformations models the three-dimensional geometric constraints that leads the prediction method to generate a protein conformation which is geometrically sound.
3. The reconstruction from the multi-level description into the original three-dimensional coordinate representation is mathematically possible with an tolerable error. This means that the multi-level description can approximate the whole information of the three-dimensional conformation.

Thus, based on this description scheme, we can include most factors of protein structure formation in the prediction.

The organization of this paper is as follows. First, a numerical representation of subconformations is described in detail. Second, a clustering technique for the classification of subconformations is briefly mentioned. Third, a modeling technique of geometric constraints of subconformations is illustrated. Fourth, we show the examples of protein conformation described. Fifth, we show the examples of the statistically modeled geometric constraints. Finally, we conclude with our vision of protein tertiary structure prediction using this description scheme.

Numerical Representation of Subconformation

This section describes how to extract a fixed number of numerical parameters representing the topology of a backbone subconformation of any size. Let us consider the position of the C^α atom of each residue as the representative position of the residue in a protein conformation. Let N be the number of residues in a subconformation. It requires $3N$ parameters (X,Y,Z coordinates for N C^α atoms) for the complete representation of the subconformation's position, orientation and topology in the three dimensional space. *The number of parameters is equal to the number of degrees of freedom,* regardless of the physical or chemical constraints. The three degrees of freedom for the subconformation's position and the other three for its orientation must be subtracted from the total number of degrees of freedom. Thus, $3N-6$ *is the number of parameters for the complete representation of the subconformation's topology.*

This is, obviously, almost proportional to the number of residues in the subconformation.

We, however, *can admit the abstraction of representation of the topology to be changed according to the size of the subconformation.* This is because low level sequences represent fine conformational structures almost without abstraction, and, thus, high level sequences only have to represent abstracted topologies of large scale conformations. The number of parameters may be so fixed that it is sufficient for the complete representation of the smallest subconformations at the lowest level. Thus, when the number of residues in a subconformation of the lowest level description is five, $3 \times 5 - 6 = 9$ is the number of parameters.

To obtain the fixed number of parameters from the topology of subconformations of any size, *certain linear expansion is applied to the coordinate representation of a subconformation.* The set of the expansion coefficients obtained becomes, with slight modification, the set of parameters representing the subconformation's topology. The number of parameters may be controlled by cutting the expansion at an appropriate order because *the significant coefficients usually appear at the lower orders in the linear expansion.* The simple reverse-transformation from these parameters reconstructs the three-dimensional topology of the subconformation.

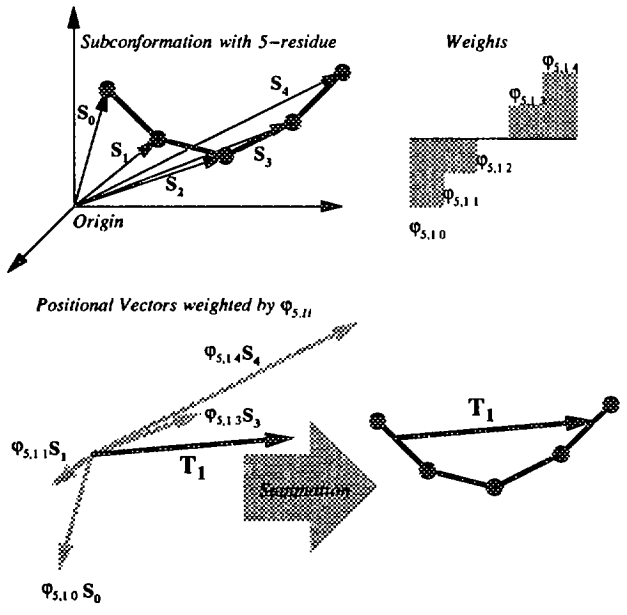

Figure 1: Abstraction of Subconformation's Topology

The procedure of extracting the target parameters is as follows. First, a set of orthonormal bases for linear expansion is provided. Second, the set of topological vectors is introduced as the abstracted form of the subconformation's topology by expanding the coordinate representation of a subconformation. To normalize the orientation of the subconformation, a set of unit vectors specific to the subconformation is determined by two of topological vectors. Finally, the set of target parameters X_k, Y_k, Z_k is calculated as the scalar products of unit vectors and topological vectors.

In the following subsections, we shall mention the procedure of extracting the parameters in detail. First, we discuss the conditions for the orthonormal bases for linear expansion. Secondly, we illustrate abstracted features of a subconformation's topology. Thirdly, we discuss how to normalize the orientation of a subconformation and how to obtain a fixed number of target parameters from that subconformation. In the subsection, we further prove that the total number of non-trivial parameters representing the topology of a subconformation with N residues is at most $3N - 6$. Finally, we briefly describe how to reconstruct the subconformation even there is an insufficient number of subconformation parameters.

Orthonormal Bases

The set of bases for the linear expansion in this study must be orthonormal *in the discrete system.* A special set is thus required. One of the simplest set of bases is defined by polynomials. Let N be the number of components of the base. Let $\varphi_{N,ki}$ denote the ith component of the base of kth order. This is simply defined by a kth order polynomial of x, $\varphi_{N,k}(x_i) = c + c_1 x_i + c_2 x_i^2 + c_3 x_i^3 + \cdots + c_k x_i^k$. The orthonormal condition for this set is,

$$\begin{cases} 0 = \sum_{i=0}^{N-1} \varphi_{N,ji}\varphi_{N,ki} & j \neq k \\ 1 = \sum_{i=0}^{N-1} (\varphi_{N,ki})^2 \end{cases} \quad (1)$$

Please note $x_{i+1} - x_i$ is not necessarily constant. Since the maximum number of bases is equal to the number of components of the bases, we can define a set of N bases with N components from 0th to $(N-1)$th order.

The summation of all components of each base, except that of 0th order, must be equal to zero so that the topological vectors generated by these bases should be independent of the subconformation's position. This condition is automatically satisfied because of the orthogonal condition with the base of 0th order, whose components have the constant value as is defined by the 0th order polynomial.

$$0 = \sum_{i=0}^{N-1} \varphi_{N,0i}\varphi_{N,ki} = \frac{1}{\sqrt{N}} \sum_{i=0}^{N-1} \varphi_{N,ki} \quad k \neq 0,$$

$$\text{therefore,} \quad 0 = \sum_{i=0}^{N-1} \varphi_{N,ki} \quad k \neq 0. \quad (2)$$

Topological Vectors

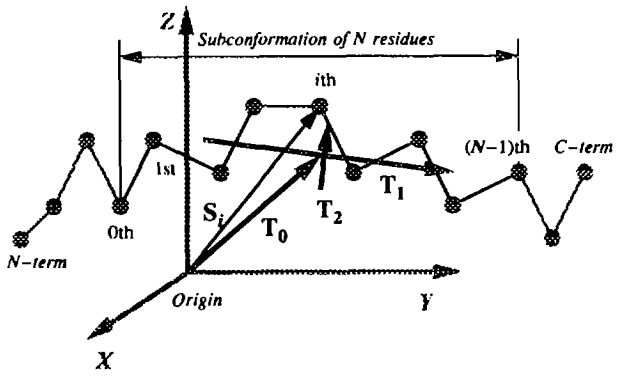

Figure 2: A Subconformation and Its Major Topological Vectors

Let $\mathbf{S}_i$ be the positional vector representing the position of ith residue in a subconformation. The operation of the orthonormal base $\varphi_{N,ki}$ to the series of the positional vector $\mathbf{S}_i$ generates the topological vector $\mathbf{T}_k$ as the expansion coefficients of the linear expansion.

$$\mathbf{T}_k = \sum_{i=0}^{N-1} \varphi_{N,ki}\mathbf{S}_i. \tag{3}$$

The neglect of $\mathbf{T}_0$ is just the subtraction of the degrees of freedom for translational transformation, because the position of the subconformation or the translational transformation does not influence the other topological vectors as below. For $k \neq 0$,

$$\begin{aligned}\mathbf{T}_k = \sum_{i=0}^{N-1} \varphi_{N,ki}(\mathbf{S}_i + \mathbf{t}) &= \sum_{i=0}^{N-1} \varphi_{N,ki}\mathbf{S}_i + \mathbf{t}\sum_{i=0}^{N-1} \varphi_{N,ki} \\ &= \sum_{i=0}^{N-1} \varphi_{N,ki}\mathbf{S}_i.\end{aligned} \tag{4}$$

where $\mathbf{t}$ is an arbitrary vector representing a translational transformation.

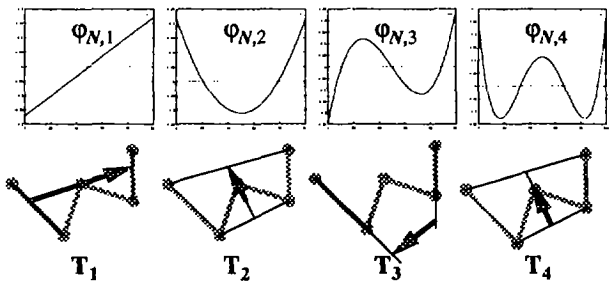

Figure 3: The Four Bases and The Corresponding Topological Vectors

It is important to know what these topological vectors represent. Considering the properties of the bases used to generate these vectors,

1. $\mathbf{T}_1$ represents approximately the difference between the mean positions of those residues near the first residue and those near the last one in a subconformation. The length of this vector may be, therefore, considered as the abstracted length of the subconformation.
2. $\mathbf{T}_2$ represents approximately the difference between the mean position of those residues at the middle and the mean position of those near the first and last ones in a subconformation. Thus, the the length of this vector represents how the subconformation curves.
3. $\mathbf{T}_3$ obviously represents the twist.
4. $\mathbf{T}_4$ represents the meander.

Hence, this set of topological vectors are the abstracted form of the topology of a subconformation.

Target Parameters

The direction of the topological vectors depends on the orientation of the subconformation. To obtain a set of parameters invariant of rotational transformation, the orientation of the subconformation must be normalized. Let us introduce a set of unit coordinate vectors, $\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_k$, for the subconformation to be represented. The conditions for these unit vectors are so determined that the direction of $\mathbf{T}_1$ is always the same as that of $\mathbf{e}_z$, and $\mathbf{T}_2$ is in the plane defined by $\mathbf{e}_x$ and $\mathbf{e}_z$, as below.

$$\begin{aligned}&\mathbf{T}_1 \cdot \mathbf{e}_z > 0, \quad \mathbf{T}_1 \cdot \mathbf{e}_x = 0, \quad \mathbf{T}_1 \cdot \mathbf{e}_y = 0, \\ &\mathbf{T}_2 \cdot \mathbf{e}_z > 0, \quad \mathbf{T}_2 \cdot \mathbf{e}_x > 0, \quad \mathbf{T}_2 \cdot \mathbf{e}_y = 0.\end{aligned} \tag{5}$$

These unit coordinate vectors must satisfy the conditions that 1) their length is equal to one, 2) they are mutually orthogonal, and 3) the vector product of $\mathbf{e}_x$ and $\mathbf{e}_y$ is $\mathbf{e}_z$, as always.

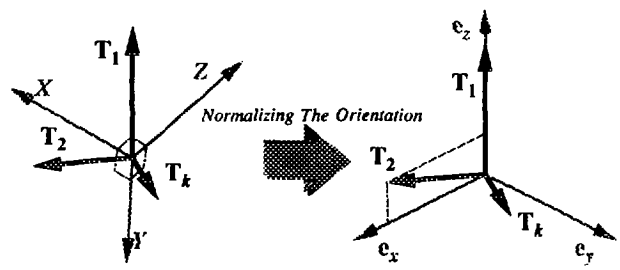

Figure 4: The Unit Coordinate Vectors and Topological Vectors

The target parameters X_k, Y_k, Z_k representing the abstracted topology of subconformation are calculated as below,

$$X_k = \mathbf{T}_k \cdot \mathbf{e}_x, \; Y_k = \mathbf{T}_k \cdot \mathbf{e}_y, \; Z_k = \mathbf{T}_k \cdot \mathbf{e}_z, \tag{6}$$

obviously, X_1, Y_1, and Y_2 are always equal to zero. They shall be neglected.

The subconformation can be represented by $3M-3$ parameters using M topological vectors. Since the number of possible bases are at most equal to the number of residues N in a subconformation, at most $N-1$ topological vectors excluding $\mathbf{T}_0$ for the subconformation may be generated. The total number of possible parameters is, therefore, $3M-3 = 3(N-1)-3 = 3N-6$. This is equal to the number of degrees of freedom in the topology of subconformations with N residues as mentioned before.

A fixed number of target parameters is obtained by cutting the linear expansion at the appropriate fixed order for the subconformations of any size. In our case, nine parameters Z_1, X_2, Z_2, X_3, Y_3, Z_3, X_4, Y_4, Z_4 representing the topology of a subconformation are obtained from four topological vectors, $\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3, and \mathbf{T}_4$, which are in turn generated by four bases $\varphi_{N,1i}, \varphi_{N,2i}, \varphi_{N,3i}, \varphi_{N,4i}$.

Reconstruction of Subconformation

The reverse-transformation from this numerical representation of a subconformation into its coordinate representation is:

$$\mathbf{S}_i = \sum_{k=1}^{N-1} \varphi_{N,ki}\mathbf{T}_k, \tag{7}$$

where $\mathbf{S}_i$ represents the vector from the mean position of the residues to the position of ith residue in the subconformation since the position of this subconformation has already been normalized to the mean position of the residues in this subconformation by neglecting $\mathbf{T}_0$.

The topology of a subconformation with more than five residues is abstracted by the insufficient number ($< 3N-6$) of parameters. Thus, it cannot be reverse-transformed exactly from its own insufficient parameters. The reconstruction of those large subconformations is, however, possible provided that its component subconformations are known. It is because the relative position and orientation of the component subconformations against the parent subconformation can be calculated as the three topological vectors of the component subconformations from 0th to 2nd.

Let A be a subconformation with N residues. Let B be the subconformation with $L(< N)$ residues. Suppose A comprises B at the offset $l(< N-L)$, B's topological vector $\mathbf{T}_{B,j}$ is calculated by,

$$\mathbf{T}_{B,j} = \sum_{i=l}^{l+L-1} \varphi_{L,j(i-l)}\mathbf{S}_i. \tag{8}$$

Since the exact $\mathbf{S}_i$ cannot be calculated from the insufficient number ($< N-1$) of A's topological vectors $\mathbf{T}_{A,k}$, the approximated positional vector $\mathbf{S'}_i$ directly reverse-transformed from the insufficient number of topological vectors is, instead, used. Thus, approximated $\mathbf{T'}_{B,j}(\cong \mathbf{T}_{B,j})$ is calculated as below,

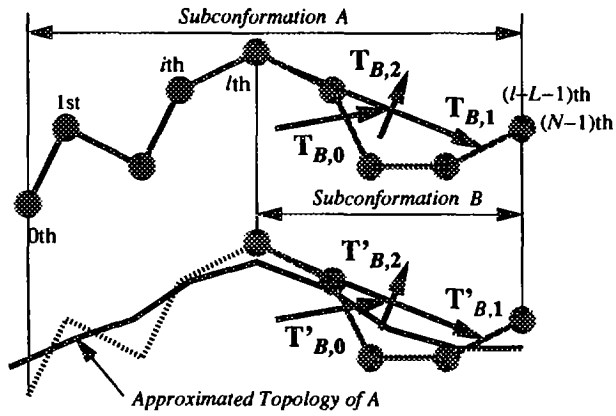

Figure 5: Topological Vectors of The Component Subconformation

$$\mathbf{T'}_{B,j} = \sum_{i=l}^{l+L-1} \varphi_{L,j(i-l)}\mathbf{S'}_i$$
$$\text{where} \quad \mathbf{S'}_i = \sum_{k=1}^{M<N-1} \varphi_{N,ki}\mathbf{T}_k. \tag{9}$$

The low order $\mathbf{T'}_{B,j}$ is the good approximation of corresponding $\mathbf{T}_{B,j}$ as long as M is not so much smaller than N, because the $\varphi_{L,j(i-l)}$ of low order may be well approximated by the superposition of $\varphi_{N,ki}$. Thus, the relative position of B against A is calculated as $\mathbf{T'}_{B,0}$ the relative orientation is calculated as $\mathbf{T'}_{B,1}$ and $\mathbf{T'}_{B,2}$. The exact $\mathbf{T}_{B,j}$ may be calculated by iterative approximation. Hence, the topology of high level subconformation may be reconstructed by the information how it is built up of its low level subconformations.

Classification of Subconformation

The subconformation represented by the aforementioned numerical parameters shall be classified by the statistic clustering. The purpose of the classification is not for the chemical or biophysical analysis of subconformations but for the classification of their topology. Any clustering technique is applicable as long as the resultant classification satisfies the condition that the well populated region in the parameter space should not be divided.

We provide an original method which involves a hypercube histogram so as to detect the well populated regions in the parameter space. In this method, the population of the data is counted in each hyper lattice cube in n dimensional parameter space where n is the number of data components. After this process, the well populated and poorly populated regions shall be detected by the population in each cube just like a histogram. The clusters are formed by merging the hypercubes according to the population in the cubes and

the Euclidian distance between the cubes. The position of each cluster's center determined as the mean position of the data in each cluster is influenced by the size of hyper cubes and the condition under which the clusters are formed such as the thresholds in merging. It should be, therefore, refined by iterative improvement as follows.

1. The center of each cluster is calculated as the mean position of its initial members.
2. The whole data are discriminated into the clusters by their Euclidian distance from the clusters' center in the sense that the data should belong to the cluster whose center is the nearest of all the others.
3. The center of clusters is again calculated from its new members.

This procedure is iterated until the position of each cluster center converges. The poorly populated clusters is discarded during the iteration. the discarded data shall be discriminated into the nearest active cluster at the next iteration. The number of clusters can be controlled by monitoring this number during the iteration. We adopted sixteen as the number of clusters for each subconformation size.

Modeling Geometric Constraints

The geometric constraints shall be modeled by analyzing all possible overlapping patterns of two classes of subconformations. If two subconformations can geometrically share several residues or overlaps, such an overlapping pattern should appear frequently in the real protein conformation. Thus, the frequency of each overlapping pattern in the real protein conformation may be considered as a stochastic constraint of protein conformation.

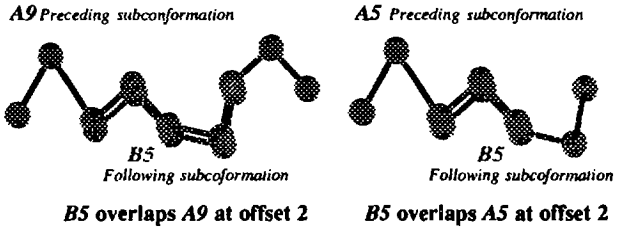

Figure 6: Overlapping Patterns

We define each overlapping pattern by 1) the class of preceding subconformation, 2) the class of following subconformation and 3) the offset in the overlapping. The offset of 3) denotes the relative position of the initial residue of the following subconformation against the position of the initial residue of the preceding subconformation in the primary structure. For instance, the a 5-residue subconformation of class **B** overlaps a 9-residue subconformation of **A** and the initial residue of **B** is the second residue of **A** can be described as "**B5** overlaps **A9** at offset 2." We define initial residue of a subconformation as "0th residue" not "first residue". Note that the offset in the pattern of small subconformations is strongly constrained by the two classes of subconformations involved in the pattern since the small subconformations are classified with high resolution and this distribution of the offset is peaky. On the contrary, the offset of the pattern of large subconformations is not so much constrained as that of smaller ones since the classification of large subconformation is rough and the distribution of the offset is fuzzy.

The complete representation of the overlapping patterns of large subconformations requires large number of parameters because the number of possible offset is almost proportional to the size of the subconformations. The number of parameters, however, can be reduced remarkably like the aforementioned abstraction technique applied to the abstracted topology of large subconformation. The linear expansion abstracts the distribution of the frequency with respect to the offset. Thus, only five coefficients or parameters $g_l (0 \leq l \leq 4)$ represent the distribution even when the number of possible offset is large. Let L be the number of possible offset. Let f_i be the frequency of overlapping pattern with respect to the offset i. The same orthonormal set $\varphi_{L,li}$ as we defined in subsection 2.1 may be applied for this expansion as below.

$$g_l = \sum_{i=0}^{L-1} \varphi_{L,li} f_i. \tag{10}$$

Let N be the number of residues in the preceding subconformation, and M be that of the following one. Let us further denote j as the class of the preceding subconformation and k as of the following one. Let $l (0 \leq l \leq 4)$ denotes the order of the coefficient of those representing the abstracted distribution of frequency. The stochastically modeled geometric constraints of overlapping pattern shall be described by the table of probability coefficients $G_{NM,jk,l}$.

For those overlapping patterns whose two mutually overlapping subconformations have the same number of residues N, the possible offset i is restricted to those satisfy the condition $0 \leq i \leq (N-1)$. On the other hand, when N is greater than M, the possible offset i is restricted to satisfy the condition $0 \leq i \leq (N-M)$, since we restrict the pattern to be those whose larger subconformation completely comprise the smaller one.

It is important to note that a very frequent class of subconformations has many chance to overlap any class though a rare class has few chance. Thus, the frequency of overlapping pattern should be normalized by the frequency of the class of the subconformations in the patterns. When the probability of the following subconformation in a pattern is requested, the pattern's frequency must be divided by the frequency of the class of preceding subconformation. When the probability of the preceding one is requested, the pattern's frequency

must be divided by the that of the following one. If the requested probability is concerned with the pattern itself, the frequency of the pattern should be divided by both frequency of preceding and following ones and then, the frequency of the pattern should be multiplied by the total number of subconformations involved in the analysis. This normalized frequency divided by both frequency of classes and multiplied by the total number of subconformations evaluates how frequently it occurs regardless of the frequency of the classes.

Description Examples

The data set used in this study was obtained from PDB in July 1992. The total number of entries is 1252 and the total number of protein chains is 1836. We selected 466 backbone chains whose mutual homology of their amino acid sequence is less than 80% after the pairwise alignment.

The data set of N-residue level is obtained by calculating the nine parameters of all the possible subconformations with N residues in all the selected protein chains. We classified the subconformations of 5,9,17,33,65 and 129 ($N = 2^n + 1$) residues and obtained the sixteen classes for each size of subconformation by the aforementioned clustering technique. The alphabets from **A** to **P** denote the classes of subconformations. The population of each class is shown in this table below:

Class	Size of Subconformation					
	5	9	17	33	65	129
A	23150	14669	10427	3637	3291	3006
B	3801	3990	5092	5192	2628	156
C	3995	3387	4481	4658	3468	1506
D	4050	6608	4166	4092	3142	2440
E	2960	3818	3093	4674	3170	1489
F	5751	3447	3250	4259	3875	2343
G	4227	3237	4881	3897	3772	1573
H	3239	5102	4070	3992	2997	1745
I	2851	4495	3965	2915	3172	1970
J	3412	3578	3509	4880	3925	1639
K	2630	4386	3664	3317	3241	1909
L	3215	4179	3916	3575	3276	1438
M	2572	2715	4223	2973	963	2245
N	2740	4328	3088	3557	2911	1750
O	3780	4750	5702	4399	3735	2160
P	7190	4492	5124	4444	3229	2662
Total	81715	79324	74784	66599	52943	33069

The conformation of a Protein 6TIM-B (B chain of TRIOSEPHOSPHATE ISOMERASE) is described in this scheme as below, where S. S. means secondary structure. We can see that at the 5-residue level, the position of symbol **A** usually corresponds to that of helix denoted by **h**, and those of **P**,**F** and **C** usually corresponds to that of sheet denoted by **s**. This means that the sequence of five-residue level well corresponds to that of secondary structures. Likewise, we may assume that the sequence of 17- or 33-residue level would correspond to the sequence of *super secondary structures*. Also, the sequence of 65- or 129-residue level would correspond to the sequence of *domain or global structure*.

Level	0	1	2	3
129				
65				EEEEEJJ
33			FFFOOBNNDDDDDFFFOONNNNN	
17		KKGGMMMMOOOOOAAAAAADBBPPPLDDDHI		
9		DDDDDDIPPPNKEBBAAAAAAAAFFFLPPNNLHDD		
5		HLFPLPPPPMGCCHBAAAAAAAAAAAGJFPDIECFPP		
S. S.		sssss	hhhhhhhhhhh	ss

	4	5	6	7
129				PPJAAAAAAAAAAAA
65	JJJJJJJJHHHFFFFFFFFFFFFFFFJJJLLLLLLLLAA			
33	DDDDHHHHHKKKKLLLLLLBBBBDDDIIIIIHHHNNLLM			
17	IGGJMNNCOOOBBBPPLLEFFHHIKKGGGJJJPPLLEOOO			
9	DOIKEBBAAAAAFFLNKMGLHDDOOPPPPNNNNMMNLLKE			
5	PFCHBAGMAAAAAGKFIBGJOFPODKOHDJEIGJDGJHOB			
S. S.	sss	hhhhhhh	sssss	

	8	9	10	11
129	AAAAEEEEEEEEPPPPPPPPPAAAAAAAIJJJJFFFFFFF			
65	AAFFFFFFFJJJGGGLLLJJJJJJJJLLLHHHFEEEEEEF			
33	MGGGOOOONNNNDDDDDHHHHFFOOBBBDDDDDMMMOOLL			
17	ABBPPLEFFDHIIGGGBBPPPCCCCOCBAAAAABBBPPLE			
9	BBACCJGLPHOOIIEEBBACJGGBKEBBAAAAAAAACCJG			
5	AAAAANEPMKCLDKCBAAAAANEPCBAAAAAAAAAAAANE			
S. S.	hhhhh	ssss	hhhhh	hhhhhhhhhhhh

	12	13	14	15
129	FJAAAAAAAAAALLLLPPPPPPPPPPPPPPFFFFAAAAAA			
65	JGGGGGGGGKKKKKKLLLLLLLLLKKKKEEEOOOKKDDDD			
33	LLBJJJPPPIIIIIHFFOOBBBJJJPPPCCCCHFFOOLLL			
17	EFFHHKKKGGGJMNNLCCOOAAAAAAAAAABBBCCPPLDD			
9	LPHDDDDDIKEBBCCJGBEFAAAAAAAAAAACFBNEBBAF			
5	CLPLPFFCHOBAAAANEMMAAAAAAAAAAAAAGJOBAAAA			
S. S.	ssssss	hhhhh	hhhhhhhhhhhh	hhhh

	16	17	18	19
129	AAAAAAAFFFFFFFFFFFFFPPPAAA			
65	DDDKKKKKKKKKKNNNDDDLLLLKKKKKKKKKKKNNNE			
33	BJJJJPPPPIIIHHFFOOBBBBJJJEPPPPCCCCHFFOOL			
17	DHHHKKGGGGMNLLLOOOOMMAAAAAAAAAABBBBPPPCC			
9	FHHDDDIKEBBCCJGLHIKEBBAAAAAAAAAAAAAACJGGE			
5	GJOPPFCOBAGMGIEJCLHBAAAAAAAAAAAAAAAAANHB			
S. S.	hhssss		hhhhhhhhhhhhhhhhh	

	20	21	22	23
129				
65	EEEEAAAJJJJJJJJAFF			
33	LLBBJJJDPPIIIIIIHHKKKKKLLLLNNNNDDD			
17	CODDHHHHIIGGMMNNCCOOOBPPPLEEDFDIGGMMNNPC			
9	BBAFFLHDDIIPPNKEBBAAAFACCJGLHOOIIEBBCCEB			
5	AAAAGIFPPFFIKPCBAGAAAAAGIBGJOJPLPDAAAGNB			
S. S.	hhhhhh	sss	hhhhhh	ssss hhhh

	24
129	
65	
33	
17	OO
9	BAAAAA
5	AAAAAAAG
S. S.	hhhhhh

The next partial descriptions of three protein conformations show that 129-residue level of the conformation of 5MBA (MYOGLOBIN) is very close to those

of 4HHB (HUMAN HEMOGLOBIN). We can assume from these description that the 129-residue level description "DDDD..." should be a feature of Globin Family.

However, the 129-residue level description of 4FXN (FLAVODOXIN), which does not belong to Globin Family, is also "DDDDDDDDDD". In this case, the topology is very different at the 33-residue level description. This means that, *the abstracted topology of 4FXN's global conformation is close to 4HHB-C or 4HHB-D, but the global topology is built up of the different component subconformations or super secondary structures.*

Chain	129-Residue level
5MBA	DDDDDDDDDDDDDDDDDDD
4HHB-C	DDDDDDDDDDDDD
4HHB-D	DDDDDDDDDDDDDDDDDD
4FXN	DDDDDDDDDD
	Part of 33-Residue Level
5MBA	EEEEEEEEEEEEEDCCCCCCCKKKKMMGGGGGGGG
4HHB-C	EEEEEEEEEEPPPPCCCCCHHFFFOGGGGGGGGG
4HHB-D	EEEEEEEEEPPPPCCCCCKKKKKMMGGGGGGGGG
4FXN	DDDDHFFFOOOGGGGGCCCHHHKKKKKLLLBBBDD
	Part of 5-Residue Level
5MBA	HHBAAAAAAAAAGAAAAAKDAAAAAAAAAAAAAAK
4HHB-C	OBAAAAAAAAAAAAKMAGMAAAAAAAAAAAAAAKD
4HHB-D	COBAAAAAAAAAAAAGIPDAAAAAAAAAAAAAAKDA
4FXN	PFCFPDNEMAAAAAAAAAAAAAAANEIFPOFCHBAG

In this way, a protein conformation described in this multi-level description scheme shows how the conformation is built up of the component subconformations.

Examples of Geometric Constraints

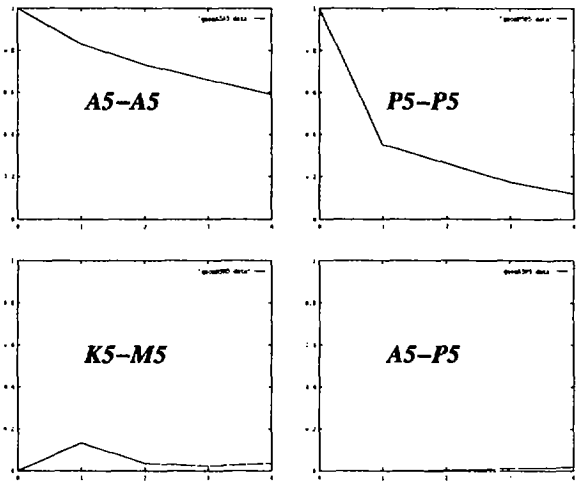

Figure 7: Geometric Constraints of 5-residue Subconformations

The data set for analyzing the geometric constraints is the same as that of the previous section. We analyzed all overlapping patterns with homogeneous geometric constraints, that is, constraints which have the same subconformation levels. On the other hand, we only analyzed overlapping patterns of adjacent levels which have heterogeneous geometric constraints, that is, constraints for different subconformation levels.

These figures show how frequently the subconformations overlap. The horizontal axis indicates the offset and the vertical axis indicates the probability, that is, the normalized frequency of the pattern. In this case, the frequency of the pattern is divided by the frequency of preceding subconformation's class. The 5-residue subconformation **A5** which corresponds to helix is a continuous subconformation. Thus, the frequency distribution of **A5** and **A5** with respect to the offset is flat. The frequency distribution of **P5-P5** is not so flat as that of **A5-A5**, though **P5** which corresponds to a kind of strand is also a continuous subconformation. This means that **A5** is more continuous than **P5**. The frequency distribution of **K5-M5** suggests that **K5** usually overlaps **M5** at offset 1, and it rarely overlaps at the other offset. **P5** and **A5** hardly overlap as shown in Figure 7, since helix is geometrically very different from strand.

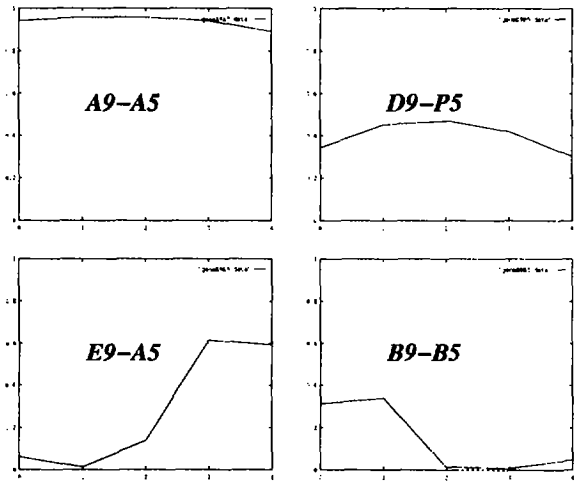

Figure 8: Geometric Constraints of 9-residue and 5-residue Subconformations

Since the 9-residue subconformation **A9** corresponds to helix, the component subconformation here should belong to **A5**. The figure shows that the frequency distribution is flat and the normalized frequency at any offset is almost 1. This means, when the class of subconformation at 9-residue level is **A9**, no other class of subconformation at 5-residue level than **A5** can occur at that region. Likewise, **D9** which corresponds to strand should be built on **P5**, and the figure shows that

the normalized frequency distribution at any offset of **D9-P5** is around 0.5. The 9-residue subconformation **E9** usually occurs at the beginning of helix. Thus, **E9** should permit **A5** at around offset 3 or 4. The probability of occurrence of **A5** at the region of **E9** shows that **A5** hardly occurs at offset 0 or 1, but often at offset 3 or 4.

In this way, the statistically modeled geometric constraints show how the subconformations overlap and how they are built on the component subconformations. Thus, these geometric constraints would greatly improve protein structure prediction.

Conclusion

In this paper, we have proposed a multi-level description scheme of protein conformation, in which the fine structures are represented at the low level with high resolution and the global structure at the high level with low resolution. A low level description corresponds well to the sequence of a secondary structure. Further, the description gives the information of how the global conformation is built up of component subconformations of various sizes.

The advantages of this scheme are that: 1) we can model the constraints between the subconformation and the primary structure at each level, from local to global, 2) we can model the geometric constraints of protein conformation, and 3) a protein conformation described in this scheme can be reconstructed into original three dimensional conformation.

Frequencies of overlapping patterns of subconformations may be used as stochastic constraints in predicting protein tertiary structures. Hence, we propose a prediction method that predicts the protein conformation in the sense that the topology of a subconformation is not only determined by the primary structure of that region but also constrained by those surrounding or comprising subconformations. A scheme of stochastic reasoning shall be used to implement this prediction method. The stochastic constraints of that scheme is derived from real protein conformation data.

Recent results of our prediction system which contains the statistically modeled geometric constraints show that the set of geometric constraints can improve the predicted conformation remarkably. This prediction system first predicts the class of subconformation at every site of every level from the primary structure at that region regardless of the geometric constraints. Then, it refines the roughly prediction by introducing the geometric constraints.

Now that, we have a mathematically well formalized description scheme of protein conformation. Pattern recognition techniques can then applied to model the relations between the primary structures and the conformations. This description scheme also enables the researchers to apply particular algorithms to include geometric constraints in their prediction systems. The multi-level description scheme shall be an important base for the protein structure prediction.

Acknowledgements

This article describes the research done at Institute for New Generation Computer Technology (ICOT). We are very grateful to M. Akahoshi at ICOT who implemented a lot of tools for this research.

References

Chou, P.Y.; and G.D. Fasman 1974. "Prediction of protein conformation". *Biochemistry 13*: 222-244.

Fasman, G.D. ed. 1989. *Prediction of Protein Structure and the Principles of Protein Conformation*. New York: Plenum Publishing Corporation.

Garnier, J.; D.J. Osguthorpe; and B. Robson 1978. "Analysis of the accuracy and implication of simple methods for predicting the secondary structure of globular proteins". *J. Mol. Biol. 120*: 97-120.

King, R.D.; and M.J.E. Sternberg 1990. "Machine learning approach for the prediction of protein secondary structure". *J. Mol. Biol. 216*: 441-457.

Qian, N.; and T.J. Sejnowski 1988. "Predicting the secondary structure of globular proteins using neural network models". *J. Mol. Biol. 202*: 865-884.

Lim, V.I. 1974. "Algorithms for prediction of α-helices and β-structural regions in globular proteins". *J. Mol. Biol. 88*: 873-894.

Bohr, H.; J. Bohr; S. Brunek; M.J.R. Cotterill; B. Lautrup; L. Norskov; H.O. Olsen; and B.S.Pertersen 1988. "Protein secondary structure and homology by neural networks". *FEBS Letters 241(1.2)*: 223-228.

Cohen, F.E.; R.M. Abarbanel; I.D. Kuntz; and R.J. Fletterrick 1986. "Turn prediction in proteins using a pattern matching approach". *Biochemistry 25*: 266-275.

Cohen, F.E.; M.J.E. Sternberg; and W.R. Taylor 1982. "Analysis and prediction of the packing of α-helices against a β sheet in the tertiary structure of globular proteins". *J. Mol. Biol. 156*: 821-862.

Branden, C.; and J. Tooze 1991 *Introduction to Protein Structure*. New York: Garland Publishing, Inc.

Asai, K.; S. Hayamizu; and K. Onizuka 1993. "HMM with Protein Structure Grammer". *Proc. of the 26th HICSS vol. 1*: 783-791.

Miller, R.T.; R.J. Douthart; and A.K.Dunker 1993. "An Alphabet of Amino Acid Conformations in Protein". *Proc. of the 26th HICSS vol. 1*: 689-698.

Metfessel, B.A.; and P.N. Saurugger 1993. "Pattern Recognition in the Prediction of Protein Structural Class". *Proc. of the 26th HICSS vol. 1*: 679-688.