TR-0858

Protein Structure Prediction Based on Multi-Level Description

by
K. Onizuka, H. Tsuda, M. Ishikawa, A. Aiba,
K. Asai (ETL) & K. Ito (ETL)

ICOT
Mita Kokusai Bldg. 21F
4-28 Mita 1-Chome
Minato-ku Tokyo 108 Japan
(03)3456-3191～5

Institute for New Generation Computer Technology

Protein Structure Prediction Based on Multi-Level Description

Kentaro Onizuka, Hiroshi Tsuda, Masato Ishikawa, Akira Aiba
Institute for New Generation Computer Technology (ICOT)
1-4-28, Mita, Minato-ku, Tokyo, 108 Japan
onizuka@icot.or.jp, tsuda@icot.or.jp,
ishikawa@icot.or.jp, aiba@icot.or.jp
FAX +81-3-3456-1618

Kiyoshi Asai, Katunobu Ito
Electrotechnical Laboratory (ETL)
1-1-4, Umezono, Tsukuba Ibaraki, 305 Japan
asai@etl.go.jp, kito@etl.go.jp
FAX +81-298-58-5939

Abstract

We propose novel prediction schemes for protein 3D structure prediction that include both local and global factors of protein structure formation.

We have developed a powerful description scheme for protein conformation, MLD (Multi-Level Description), in order to model the protein structure formation. In this scheme, the description is reconstructable into the three-dimensional conformation with a tolerable error. The MLD scheme facilitates the modeling of 1) the relation between the local conformation and the primary structure at that region at various scales (i.e., primary constraints), and 2) the geometric constraints between the neighboring local conformations. Hence, in our prediction schemes, the problem of protein 3D structure prediction is formulated as a combinatorial optimization problem: the 3D conformation of a protein is predicted as the optimal MLD that satisfies most of the constraints.

We implemented several schemes to solve this problem. We proved that the degree of prediction accuracy is much improved by introducing the geometric constraints.

1 Introduction

The prediction of protein 3D conformation from the primary structure (i.e., amino-acid sequence) is one of the most important unsolved problems in molecular biology. To formalize prediction schemes, we considered that the description scheme of protein conformation would be the most significant aspect.

We proposed the MLD scheme (Multi-Level Description) [Onizuka et al 93] which represents both the fine conformation and the global topology of a protein conformation at multiple levels. The global topology is represented at the high levels, and the fine conformational structures are represented at the low levels. We designed the MLD scheme such that the description would be approximately reconstructable into 3D conformation. In this sense, MLD is the compact form of the coordinate representation of a protein's 3D structure.

To formalize the problem of protein 3D structure prediction, we modeled two important kinds of constraints in protein structure formation based on MLD. A primary constraint is the relation between a local conformation type and the primary structure at that region. Many pattern recognition techniques are applicable [Chou and Fasman 74, Asai et al 93A, Mamitsuka and Yamanishi 93] to the modeling of primary constraints. A geometric constraint is a stochastic constraint between two neighboring local conformations. Here, the geometric constraints have two important roles in 3D structure prediction. One is to avoid geometric inconsistency in the predicted conformation, and the other is to include global or long-rage interaction in the structure formation.

Geometric constraints are regarded as the interactions between neighboring local conformations during the process of structure prediction. In our prediction schemes, thus, a local conformation is determined not

only by the primary structure at that region but is also strongly influenced by the neighboring local conformations. In our prediction schemes, thus, chances are that the region which has a strong tendency to form a helix may form a strand in the final result if strongly recommended by the geometric constraints from the neighboring local conformations.

In the following section, we briefly explain the MLD scheme, and define the notation of the primary constraints and the geometric constraints. Section 3 overviews the prediction schemes based on MLD. In Section 4, several modeling techniques for the primary constraints are discussed. Hidden Markov Models (HMMs), property-based modeling, and neural networks have been implemented so far. Motif-based modeling is also proposed. Section 5 details the method for protein structure prediction. Two systems are actually working. One is based on the stochastic propagation model and the other uses dynamic parsing with grammar. Integer programming and other combinatorial optimization algorithms, as well as the folding simulation, are also proposed as prediction schemes. Section 6 presents a resultant prediction example.

2 Multi-Level Description

In this section, we briefly illustrate the main objectives and features of the MLD scheme. For further details, refer to [Onizuka et al 93].

There are two important objectives of the MLD scheme. One is to include the global factors of protein structure formation in the structure prediction schemes, and the other is to design a description scheme such that the description would be reconstructable into coordinate representation. An MLD represents a protein conformation at multiple level of scales, where the global topologies are represented at the high levels, and the fine structures at the low levels. Thus, the global or long-range interactions are modelable using descriptions at the high levels, and the short-range ones can be modeled by those at the low levels.

Also, information on the global topology at the high levels is indispensable to reconstructing the precise 3D structure. In conventional prediction schemes, the single level description of protein conformation, such as the sequence of secondary structures and others [Miller et al 93, Zhang et al 93] are widely used in protein structure prediction. This kind of single level description scheme suffers, however, from an inevitable problem that the description cannot be precisely reconstructable into the 3D conformation. Since the local conformations dealt with by these schemes are normally of small peptide fragments, the diversity in each local conformation type inevitably accumulates during the reconstruction. This means that the topology of the reconstructed conformation would be largely different from the original. Multiple levels in MLD, however, solve the problem of error accumulation, where any accumulated error, can be naturally adjusted by the description at the higher levels.

To design the MLD scheme, it is first necessary to classify the local conformations of various sizes. The classification of local conformations with many residues is generally difficult, since the number of numerical parameters required for the complete representation of a local conformation is almost proportional to the number of residues in the conformation. A kind of linear transformation, however, may extract a fixed number of characteristic parameters from local conformations of any size by cutting the linear expansion at an appropriate fixed order, although large local conformations would be abstracted during the parameterization.

We classified the local conformations at each level into several types by clustering. Thus, an MLD represents a protein conformation with multiple symbolic sequences, each symbol of which denotes the local conformation type of the level size. The MLD scheme models two kinds of significant constraints in structure formation, the primary constraints and the geometric constraints. These are detailed in the following subsections.

The MLD example below represents the conformation of Trypsin Inhibitor (8PTI). In this description, the local conformations at each level are classified into sixteen types, each being denoted by a letter from **A** to **P**. It is observed that **A** at the 5,9,17-residue level roughly corresponds to an α-helix, while **P**,**F** and **C** at the 5 residue-level and **D** at the 9-residue level roughly correspond to a β-strand. This obviously means that the symbols at low levels are closely tied to the secondary structures and those at high levels are considered to represent the structural motifs such as helix-turn-helix conformation.

2.1 Primary Constraints

A primary constraint represents the relation between a primary structure and the local conformation type at that region. In our study, the relation is considered as the propensity of a primary structure fragment to form a certain type of local conformation.

Level	Level Size	Description
3	33	NNNNGHHHHHHKKOGGGGGMMMFFOO
2	17	IIGMMNNLEFFHIGGJJJLEEFFDMMMMNPPPOOOOABBPPO
1	9	FCFLHDIPPNNNPLHOOOIIKMGGLHHOIPPNNNNNEJJLFKMEBBAACF
0	5	BAAGIHLFPOEJEIECFOFOFPDANEILOFPDGGMGJHIBGJCDJBAAAAAAGM

Figure 1: MLD representing the conformation of BPTI(8PTI)

For further discussions, we define the notations used for the primary constraints. In our present study, the local conformations are classified into sixteen types at each level, with letters from $\mathbf{A}^k$ to $\mathbf{P}^k$ denoting the types in MLD, where k denotes the level in MLD.

Let γ_t^k denote a local conformation type, where normally $\gamma_1^k = \mathbf{A}^k, \gamma_2^k = \mathbf{B}^k, \ldots, \gamma_{16}^k = \mathbf{P}^k$. Let σ^k denote a primary structure fragment at the kth level. And we further denote w^k as the number of residues in the fragment or local conformation at the kth level. We denote $\Gamma_i^k \in \{\mathbf{A}^k, \mathbf{B}^k, \ldots, \mathbf{P}^k\}$ as the variable that takes the local conformation type, where i denotes the position. We also denote Σ_i^k as the variable that takes the primary structure fragment. Note that the position i here denotes the position of the first residue of a local conformation in the primary structure.

The probability of a primary structure fragment, σ^k, forming a type of local conformation, γ_t^k, is represented as $P_P(\Gamma_i^k = \gamma_t^k | \Sigma_i^k = \sigma^k)$. Since we assume that the primary constraint is independent of its absolute position in the primary structure but only depends on the conformation type and the primary structure fragment at that region, it may be represented as $P_P(\Gamma^k | \Sigma^k)$.

2.2 Geometric Constraints

The co-occurrence of two neighboring local conformations is restricted by the two local conformation types involved and the relative positions of the two in the primary structure. When a local conformation overlaps another local conformation, the shared region of both local conformations must be identical. This means that if the substructure of a local conformation type is similar to that of another local conformation type, it would be possible for those two local conformation types to overlap each other at the certain possible offset.

Such constraints would be modeled by analyzing the frequencies of the overlapping patterns of two local conformations. Here, the patterns are defined by 1)the type of the preceding local conformation, 2)the type of the following local conformation, 3)the offset of the overlap. Both intra-level and inter-level overlapping patterns are considered.

The geometric constraint between two local conformation types, $\gamma_{t_1}^{k_1}, \gamma_{t_2}^{k_2}$, is denoted by $P_G(\Gamma_{i_1}^{k_2} = \gamma_{t_1}^{k_1}, \Gamma_{i_2}^{k_2} = \gamma_{t_2}^{k_2})$, where the the position of the first residue in $\gamma_{t_1}^{k_1}$ is left to that of the first residue in γ^{k_2} in the primary structure. Since we assume that, as we also do for primary constraint, a geometric constraint is independent of the absolute position of the two fragments in the primary sequence, depending only on the relative positions (i.e., the offset), $d = i_2 - i_1$, of the two, the geometric constraint is, thus, represented simply as $P_G(\Gamma^{k_1}, \Gamma^{k_2}, d)$.

They can be modeled in other ways. Stochastic models on an intra-level like N-gram (a type of Markov chain), or the frequencies of the types of local conformations on the inter-level can both be used.

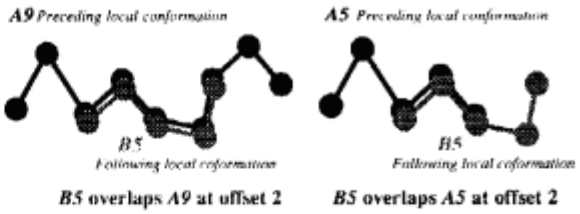


Figure 2: Overlapping Patterns

Note that two kinds of information are included in the statistics of overlapping patterns. One is the restriction of the geometrically possible combinations of local conformation types, and the other is the combination propensity of local conformations. This suggests that an geometrically possible overlapping pattern would not always be frequent under certain conformational conditions. In this paper, however, we consider these two kinds of information together as geometric constraints, since it is difficult to distinguish one from the other.

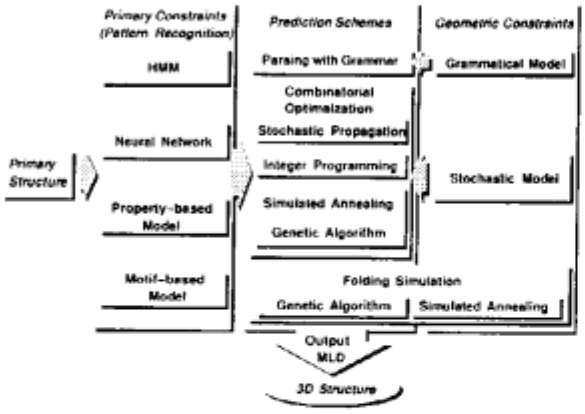


Figure 3: Overview of Prediction Schemes

3 Overview of Prediction Schemes Based on MLD

The prediction schemes that we are proposing, are *completely different from conventional methods for tertiary structure prediction.* In many cases of tertiary structure prediction, when the predicted secondary structures are packed into a tertiary structure, perfect secondary structure prediction is assumed. The secondary structures are, therefore, fixed and shall not change during the tertiary structure prediction phase [Cohen et al 82]. It is, however, observed that *the secondary structures easily change during the folding process.* Chances are that even if a fragment of primary structure has a strong propensity to form a helical conformation, and even if it forms a helical conformation at an early stage of the foling process, it may become a strand at the final stage due to the environment formed by the folding process. We must, therefore, consider *the interactions between local conformations.* We can, then, merge the phase of predicting local conformation and that of global conformation into a single phase where feedback on the structure formation is possible. Considering structure formation factors (i.e., the primary structure's propensity to certain conformations and the interactions between the local conformation) as stochastic constraints of structure formation, *the problem of protein structure prediction is formulated as a combinatorial optimization problem,* where a protein conformation is predicted as the optimal MLD that satisfies most of the constraints.

Many optimization algorithms are already available to predicting the structure. Stochastic propagation model provides a fairy rapid means of searching for a good MLD, where the MLD is changed iteratively so as to satisfy the geometric constraints by taking a hill-climbing approach. Sophisticated algorithms, such as integer programming, genetic algorithm and simulated annealing, are also available. Multi-level parsing is the extension of the parsing technique in the field of speech recognition. Here geometric constraints are used as a grammar representing the conformational rules [Asai et al 93B]. Folding simulation is another way of searching for the best 3D conformation, where the conformation is folded so as to satisfy most of the primary constraints.

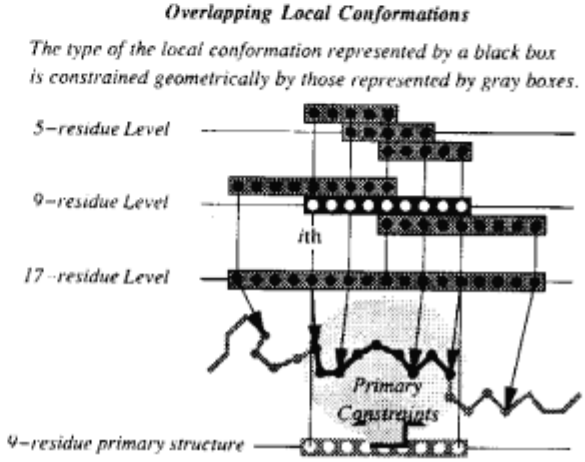


Figure 4: Constraints

To model the primary constraints, any of several pattern recognition approaches are available. Some of these have already been applied to secondary structure prediction. In our study, HMMs (Hidden Markov Model), neural networks, property-based model are implemented. Motif based model is now being developed. The geometric constraints are modeled by analyzing the frequency of the overlapping patterns of local conformations.

4 Modeling of Primary Constraints

The MLD scheme is particularly suitable for modeling both the local and global factors of structure formation. The primary constraints for short structure fragments naturally represent the local factors, and those for long ones represent the global or long-range factors.

4.1 Hidden Markov Model

Hidden Markov Model (HMM) is a framework of stochastic model widely used in the field of speech recognition and becoming more and more popular in the field of molecular biology. The applicability of this technique to protein secondary structure prediction has been discussed in [Asai et al 93A].

We applied the same kind of HMMs to modeling the primary constraints at multiple levels in MLD. As is discussed in [Asai et al 93A], the degree of accuracy is higher when the adjacent amino-acid pair is fed to HMM as an output signal. The same approach is, therefore, adopted for the primary constraints at the 5-residue level. To the primary constraints at higher levels, we applied normal type HMMs for which the output signal is an amino-acid type. We fixed the number of states at five for all levels, except for the 5-residue level, in order to avoid the over-learning. At 5-residue level, the number of states is four, because the pair of adjacent amino-acid is fed as the output signal.

The performance of HMMs for the primary constraints is evaluated by the degree of prediction accuracy without geometric constraints. The degree of accuracy normally differs with the level. At the 5,9, and 17-residue levels, the degree of accuracy is around 25%. That at the 33 and 65-residue level is, however, around 15%, though that at the 129-residue level is higher than 20%. This suggests that a super-secondary structure is not directly determined by the primary structure at that region. The good performance at the 129-residue level should be thought of as resulting from over-learning, since the data set available to model the primary constraints at that level is much smaller than those at the other levels. The table below shows the performance of HMMs for primary constraints. The degree of accuracy below 5* is the result achieved by HMM of the 4-state 2-letter type to which an adjacent amino-acid pair is fed. The others results are for 5-state 1-letter type.

Level	5*	5	9	17	33	65	129
Accuracy	29%	25%	24%	22%	14%	17%	25%

4.2 Neural Network

Neural networks have been one of the most popular techniques in the field of pattern recognition since the algorithm of back propagation learning was devised. The applicability of neural networks to protein secondary structure prediction has been frequently discussed [Qian and Sejnowski 93, Burkhard and Sander 93].

We have been attempting to apply neural networks to primary constraint modeling. The performance is slightly lower than that obtained by HMMs. Some features of sequence patterns can be learned effectively, though there are many patterns the neural networks find difficult to learn. Further investigation to determine the optimal network topology is desired.

4.3 Property-based Model

The classification of primary structure fragments leads us to model the primary constraints statistically. The frequency of each combination of the local conformation type and primary structure type is simply a statistical constraint, where the set of primary constraints is represented by a contingency table, each cell of which denotes the relative frequency of the combination of primary structure type and local conformation type.

Classify the primary structure fragments is not simple, however. Since a primary structure fragment is a sequence of twenty types of amino-acid residues, the number of permutations of the primary structure fragments becomes quite large, even if the fragment is short. Here, we adopted the strategy to extract numerical parameters from primary structure fragments, classifying them by some clustering technique.

In our study, four physico-chemical properties, hydropathy, charge, volume of side chain, and molecular weight of side chain are used to parameterize the fragment. These properties are usually considered as the important factors driving the structure formation. These properties are, in our case, evaluated as numberical parameters. The distribution of these numerical properties in the sequence can be, therefore, abstracted by operating the orthonormal bases $\varphi_{N,ki}$ to the distribution.

We provide an orthonormal complete set $\{\varphi_{N,ki}\}$, having N components, to linearly expand the distribution, where N is the number of residues in the fragment, k is the order of the base, and i is the index of the component, as shown.

$$P_k = \sum_{i=0}^{N-1} \varphi_{N,ki} p_i, \qquad (1)$$

where p_i is the numerical property of the ith residue in the fragment. By fixing the number of bases in this transformation, we can obtain the fixed number of parameters $\{P_k\}$, from a fragment of any length. In this way, we only extract twenty numerical parameters from the fragment of any length, using five bases and four properties.

The resultant degree of accuracy is much worse than that obtained by HMMs, even when the primary structure segments are classified into more than one hundred types. One significant reason for this poor result might be that the primary structure fragments are classified independently of the conformational types.

4.4 Motif-based Model

A protein primary sequence often has several motif patterns [Staden 88] that characterize its conformation and functions. The known motif patterns have been accumulated into Prosite database.

It is frequently the case that the similar functions of a protein are derived from the similar conformations. Since conformational features are described as structure motifs such as those in MLDs, the relations between the structural motifs and sequence motifs may be considered as the primary constraints.

These relations between sequence motif patterns and structure motifs have already been investigated [Rooman et al 90].
The relations concerning the binding proteins, such as Zinc Finger [Klug and Rhodes 87], Leucine Zipper [Landschulz and Johnson 88] and so forth are particularly investigated and established.

The MLD scheme facilitates the investigation of the relation between structure motifs and sequence motifs, because the techniques for sequence analysis, such as multiple sequence alignment [Barton and Sternberg 87] are applicable to searching the structural motifs. If we succeed in building up a database of large structural motifs, we can obtain sufficient primary constraints to determine most protein local conformations.

5 3D Structure Prediction

In this section, we discuss the 3D structure prediction schemes based on MLD. Since a protein 3D structure can be approximately reconstructed from the MLD, MLD prediction from a primary structure is simply a type of 3D structure prediction.

The MLD of a protein is predicted as the optimal MLD that satisfies most of the primary constraints and the geometric constraints. To predict the MLD from the primary structure of a protein, we developed two systems. One system is based on stochastic propagation model which is a sort of combinatorial optimization algorithm, and the other uses dynamic parsing with grammar. The applicability of integer programming, simulated annealing, and genetic algorithm are also be discussed. In the last subsection, we briefly illustrate how to apply MLD to protein folding simulation.

5.1 Stochastic Propagation Model

The stochastic propagation model, here, searches for a good (not always the best) MLD pattern by taking a hill-climbing approach. At the initial state, the probability of the conformation type at each site and level is directly derived from the primary constraints at that region. The probability of the next state are calculated as a linear combination of the primary constraint term and the geometric constraint term. The probabilities are carried by the geometric constraints and propagate during the iteration so that the MLD satisfies the geometric constraints better than that in the previous step.

Let $\psi(\gamma_i^k)$ denote the probability of the type of conformation γ_i^k at each state, where i denotes the position of the local conformation and k denotes the level. The initial $\psi_0(\gamma_i^k)$ is equal to $P_P(\Gamma_i^k = \gamma^k | \Sigma_i^k = \sigma^k)$, where σ^k is the primary structure at that region. The probability in the next step is calculated as below.

$$\begin{aligned}\psi_{t+1}(\gamma_i^k) &= W_I \psi_t(\gamma_i^k) \\ &+ W_P P_P(\Gamma_i^k = \gamma_i^k | \Sigma_i^k = \sigma^k) \\ &+ W_G \sum_{l=k-1,k+1} \sum_j \sum_{\Gamma_j^l} \psi_t(\Gamma_j^l) P_G(\gamma^k, \Gamma_j^l).\end{aligned}$$

where W_I is the weight of the inertia term, W_P is that of the primary constraints ,and W_G is that of the geometric constraints. These weights shall be so determined that the degree of prediction accuracy is the best. It is observed that the solution converges quite rapidly in about ten steps, when W_I is small. The experimental result are given in the last page of this paper.

5.2 Dynamic Parsing with Grammar

The applicability of dynamic parsing with grammar for secondary structure prediction was well discussed in [Asai et al 93B]. This approach is close to the parsing of speech signals in continuous speech recognition, where the terminal symbols are words and non-terminal symbols are phrases or sentences.

For protein structure prediction, however, it is difficult to obtain the appropriate representation for the

grammar. Neither non-terminal symbols representing super-secondary structures nor grammatical rules between the non-terminal symbols can be obtained easily. The MLD scheme, however, provides both non-terminal symbols (local conformation types) and grammars (geometric constraints).

We are going to explain dynamic parsing at multiple levels. Here only two levels, the 5 and 17-residue level, are considered. This algorithm is naturally extensible for parsing at more than two levels.

The protein conformation is parsed using the score of the fragments at the 5 and 17-residue levels. The geometric constraints between the adjacent fragments and those between the four consecutive 5-residue level fragments and the 17-residue level fragment at that region are used as the grammar. To avoid a searching space explosion, a threshold to prune poor scoring combinations is set at each level.

From each consecutive 5 residues, σ_i^5, the primary constraint at that region at 5-residue level returns the score, $L_i^5 = \log P_P(\Gamma_i^5|\Sigma_i^5)$, the logarithm of probability for σ_i^5 to be a certain conformation type γ_i^5. We use $\sigma_0^5\sigma_4^5\sigma_8^5\ldots$ for parsing, where only one residue is shared by two adjacent 5-residue level conformation types (Here we fix the segmentation at 5-residue level. There is a choice of overlapping segmentation). Four consecutive 5-residue level fragments $\sigma_0^5\sigma_4^5\sigma_8^5\sigma_{12}^5$ form a fragment at 17-residue level, σ_0^{17}, where the primary constraint at that region at the 17-residue level returns the 'score', $L_0^{17} = \log P_P(\Gamma_0^{17}|\Sigma_0^{17})$.

Parsing proceeds from left to right. The parsing begins from σ_0^5, then moves onto σ_4^5. The number of candidates becomes 16×16, which is the number of combinations of the local conformation types, σ_0^5, σ_4^5. The score for each candidate is the sum of the scores from the primary constraint (i.e., L_0^5 and L_4^5) and the scores of the geometric constraints between adjacent fragments, which are the logarithms of the frequency, as are those in a Markov chain. Then, the parsing moves onto σ_8^5, then σ_{12}^5, adding the score L_8^5 and L_{12}^5 each time. Here, we have a total score of 17-residue fragments, and two types of scores are added. One is L_0^{17}, which is the score of the primary constraints for this fragment at the 17-residue, σ_0^{17}. The other is the score of the grammar derived from the geometric constraints between conformation types at the 17-residue level and those at the 5-residue level. These stochastic rules are actually the frequency of the pattern of four consecutive local conformation types at the 5-residue level comprised by the local conformation types at 17-residue level. When the parsing reaches the right end of the amino-acid sequence, we have several candidates with their score as the reliability.

5.3 Integer Programming

In this subsection, we formalize the structure prediction based on MLD in terms of integer programming, which is a general framework for combinatorial optimization [Schrijver 86].

Let $X_i (i = 1, \ldots, n)$ be a variable that takes a local conformation type ($X_i \in \gamma_1, \ldots, \gamma_{16}$). Each $X_1, \cdots, X_N$ corresponds to $\Gamma_0^1, \Gamma_1^1, \cdots \Gamma_0^2, \cdots \Gamma_0^3, \cdots$. The most probable MLD maximizes the following formula.

$$W_P \sum_{i=1}^{N} C_P(X_i) + W_G \sum_{i=1}^{N}\sum_{j=1}^{N} C_G(X_i, X_j) \qquad (2)$$

Here, C_P and C_G are primary and geometric constraints, respectively.

Next, we introduce binary variables $x_{ik} \in \{0, 1\} (i = 1, \ldots, N, k = 1, \ldots, 16)$ that satisfy the following condition.

$$\sum_{k=1}^{16} x_{ik} = 1 \quad (i = 1, \ldots, N). \qquad (3)$$

C_P and C_G are formalized as follows.

$$C_P(X_i) = \sum_{k=1}^{16} P_P(X_i = \gamma_k) x_{ik} \qquad (4)$$

$$C_G(X_i, X_j) = \sum_{k=1}^{N}\sum_{l=1}^{N} P_G(X_i = \gamma_k, X_j = \gamma_l) x_{ik} x_{jk} \qquad (5)$$

(2) is formalized as follows.

$$\sum_{i=1}^{N}\sum_{k=1}^{16} c_{ik} x_{ik} + \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\sum_{l=1}^{N} d_{kl} x_{ik} x_{jk} \qquad (6)$$

$$c_{ik} = W_P P_P(X_i = \gamma_k) \qquad (7)$$

$$d_{ijkl} = W_G P_G(X_i = \gamma_k, X_j = \gamma_l) \qquad (8)$$

(7) and (8) are given from the primary structure, primary and geometric constraints. The problem is to find the value of x_{ik} that maximizes (6) under constraints (3), that gives a 0-1 integer programming form of the tertiary structure prediction.

When there is a good linear approximation to the latter non-linear (quadratic) form of (6), many sophisticated algorithms of integer scheduling problem such as branch-and-bound method, LP-relaxation, and so forth, are available [Schrijver 86]. There is also a work on optimizing quadratic forms using the Bolzman machine model [Ackley et al 85].

5.4 Other Combinatorial Optimization Algorithms

There are many other sophisticated algorithms for optimization problems available. Simulated annealing and genetic algorithm are popular algorithms for combinatorial optimization problems, even in the field of molecular biology [Ishikawa et al 93, Konagaya and Kondou 93]. These algorithms optimize the state of a solution according to evaluation functions, such as formula 6.

In order to apply these algorithms to structure prediction based on MLD, we have to define the minimal modification (or mutation in genetic algorithm). This is simply defined as changing the type of local conformation γ^k in MLD. The crossover operation in genetic algorithm is defined naturally because MLD itself is a set of symbolic sequences.

5.5 Folding Simulation

Folding simulation using MLD predicts a 3D structure without having considering the geometric constraints explicitly; it is impossible to fold a protein chain into the geometrically impossible conformation in 3D space.

In order to formulate the folding simulation as an optimization problem, we have to define the minimal modification (or mutation in genetic algorithm) of the conformation. The conformation of a protein molecule is normally modified by changing the dihedral angle of the chemical bonds. Thus, it is natural to define a small change in a dihedral angle as the minimal modification for the folding simulation. Here, the score of the conformation to be optimized is the summation of primary constraints in the MLD, which are generated from the conformation after each modification. Both simulated annealing and genetic algorithm are applicable to folding simulation. In order to implement genetic algorithm, we adopt the same formulation as that applied in [Unger and Moult 93].

6 Experimental Results

This section presents the experimental results in our schemes. The experiment was a closed test, where the learning set analyzed to model the constraints, and the test set used to check the degree of accuracy of prediction, are the same. The primary constraints used in this experiment were modeled by HMM. The MLD given in the last page of this paper is the resultant conformation predicted by the stochastic propagation model. The upper sequences are the true description of the conformation of 4HHB's C-chain, the middle sequences are predicted only by the primary constraints, and the lower sequences are predicted by both of the primary and geometric constraints. It is observed that the MLD symbols are revised by geometric constraints at several sites, and in many cases the conformation types at the revised sites match the true description. This suggests that the geometric constraints are indispensable for accurate structure prediction.

7 Discussions and Future Works

We showed how the protein structure prediction based on MLD could be formulated as an combinatorial optimization problem. The primary constraints have been modeled by HMMs, neural nets, and property-based model. The applicability of motif-based models has also been discussed. We proposed several prediction schemes for protein 3D structures. Stochastic propagation model, dynamic parsing with grammar have been actually implemented. Prediction systems based on integer programming, simulated annealing, and genetic algorithm, and folding simulation are currently being developed. The experimental result for the C-chain of 4HHB (human hemoglobin) shows that the results obtained only by the primary constraints were improved by introducing the geometric constraints.

Many points should be discussed on the evaluation score optimized in the combinatorial optimization problem for structure prediction. In the present study, except for dynamic parsing, the summation of all the considered probabilities is the evaluation score to be optimized. In most cases of stochastic optimization, however, the summation of the logarithms of probabilities or co-relations is considered as the score to be optimized. Further investigation of the appropriate stochastic model is strongly required.

The most significant aspect of our prediction schemes is that the global factors of structure formation are considered as the primary constraints of long structure fragments. In this case, however, the information from the primary structure fragments is much abstracted, and thus, the direct interaction between the two small sites which are mutually distant in the primary structure is not considered directly. In our future works, we intend to model those factors that

are not considered in the present study. This suggests that the MLD scheme itself should be changed according to the new models of factors, such as the packing pattern propensity of primary structures.

References

[Onizuka et al 93] Onizuka, K.: K. Asai: M. Ishikawa: and S.T.C. Wong 1993. "A Multi-Level Description Scheme of Protein Conformation". *Proc. of ISMB-93*:301-310.

[Chou and Fasman 74] Chou, P.Y.: and G.D. Fasman 1974. "Prediction of protein conformation". *Biochemistry 13*: 222-244.

[Cohen et al 82] 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 and Tooze 91] Branden, C.: and J. Tooze 1991 *Introduction to Protein Structure*. New York: Garland Publishing, Inc.

[Asai et al 93A] Asai, K; S. Hayamizu; and K. Handa 1993. "Prediction of protein secondary structure by the hidden Markov model". *Cabios 9-2*: 141-146.

[Asai et al 93B] Asai, K.: S. Hayamizu: and K. Onizuka 1993. "HMM with Protein Structure Grammer". *Proc. of the 26th HICSS vol. 1*: 783-791.

[Miller et al 93] 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.

[Zhang et al 93] Zhang, X.: J.S. Fetrow; W.A. Rennie; D.L. Waltz; and G. Berg 1993. "Automatic Derivation of Substructures Yields Novel Structural Building Blocks in Globular Proteins". *Proc. of ISMB-93*: 438-446.

[Mamitsuka and Yamanishi 93] Mamitsuka, H.: and K. Yamanishi 1993. "Protein α-Helix Region Prediction Based on Stochastic-Rule Learning". *Proc. of the 26th HICSS vol. 1*: 659-668.

[Unger and Moult 93] Unger, H.: and J. Moult 1993. "On the Applicability of Genetic Algorithms to Protein Folding". *Proc. of the 26th HICSS vol. 1*: 715-725.

[Qian and Sejnowski 93] Qian N.: and Sejnowski T.J 1988. "Predicting the Secondary Structure of Globular Proteins Using Neural Network Models". *J.Mol.Bilo.202*:865-884.

[Burkhard and Sander 93] Burkhard, R.: and C. Sander 1993. "Prediction of Protein Secondary Structure at Better than 70% Accuracy". *J.Mol.Biol.232*:584-599.

[Schrijver 86] R. Schrijver 1986. "Theory of Linear and Integer Programming". *John Wiley & Sons*.

[Ackley et al 85] Ackley, D.: Hinton, G.: and Sejnowski, T. 1985. "A Learning Algorithm for Boltzmann Machines". *Cognitive Science, Vol.9, No.1*.

[Staden 88] Staden, R. 1988. "Methods to Define and Locate Patterns of Motifs in Sequences". *Comput. Applic. Biosci.*, Vol.4, No.1, pp.53-60.

[Rooman et al 90] Rooman, M.J.: Rodriguez, J.: and Woodak, S.J. 1990. "Automatic Definition of Recurrent Local Structure Motifs in Proteins: Relations between Protein Sequence and Structure and Their Significance". *J. Mol. Biol.*, Vol.213, 327-350.

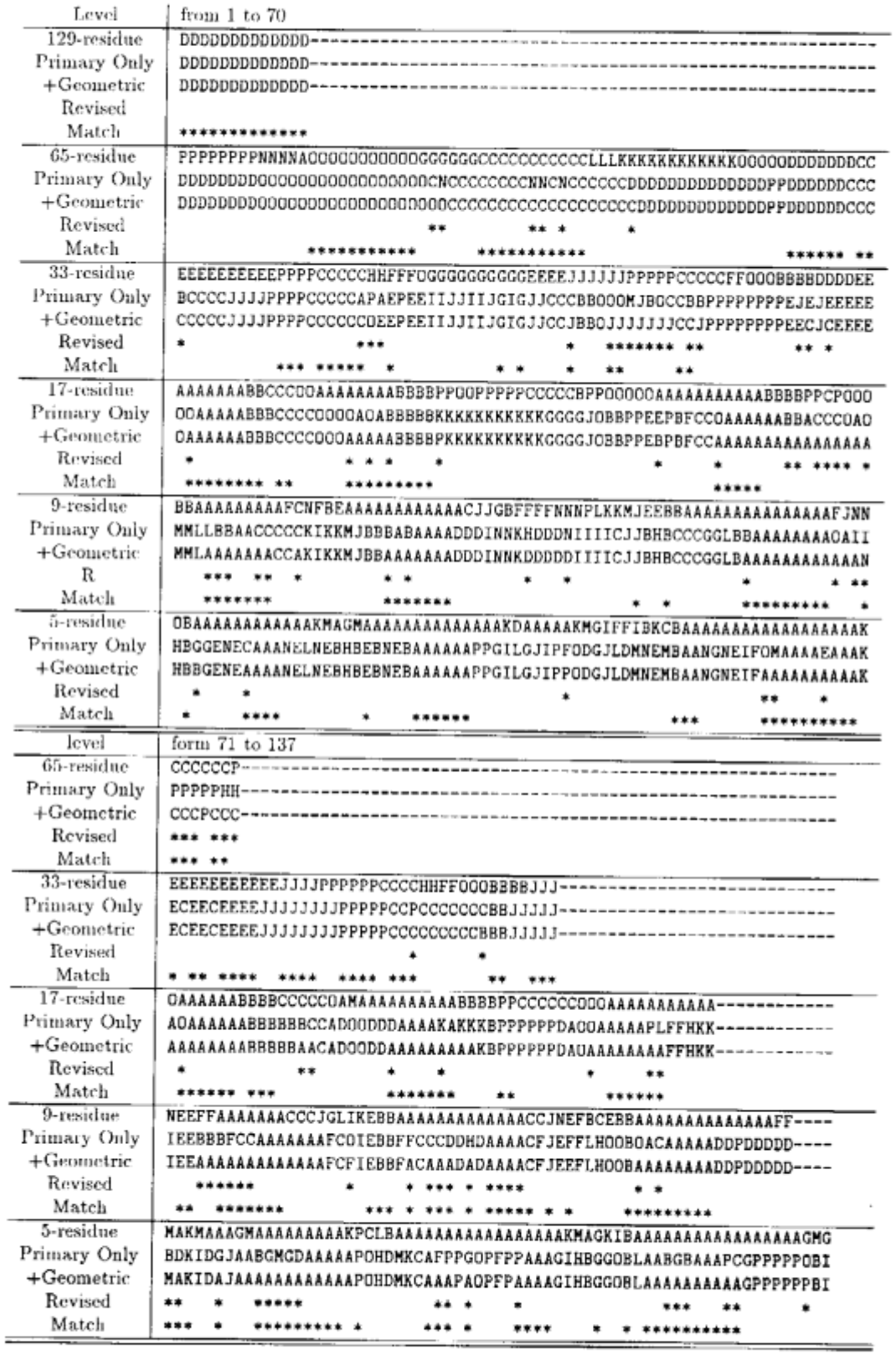
[Klug and Rhodes 87] Klug, A.: and Rhodes, D. 1987. "Zinc Fingers: A Novel Protein Motif for Nucleic Acid Recognition". *Trends in Biochemical Sciences*, Vol.12, 464-469.

[Landschulz and Johnson 88] Landschulz, W.H.: Johnson, P.F.: and McKnight S.L. 1988. "The Leucine Zipper: A Hypothetical Structure Common to a New Class of DNA Binding Proteins". *Science*, Vol.240, 1759-1764.

[Barton and Sternberg 87] Barton, G.J.: and Sternberg, M.J. 1987. "A Strategy for the Rapid Multiple Alignment of Protein Sequences: Confidence Levels from Tertiary Structure Comparisons". *J. Mol. Biol.*, 198, pp.327-337.

[Ishikawa et al 93] Ishikawa, M.: Toya, T.: Hoshida, M.: Nitta, K.: Ogiwara, A.: and Kanehisa, M. 1993. "Multiple Sequence Alignment by Parallel Simulated Annealing". *Comput. Applic. Biosci.*, Vol.9, No.3, pp.267-273.

[Konagara and Kondou 93] Konagaya, A.: and Kondou, H. 1993. "Stochastic Motif Extraction using a Genetic Algorithm with the MDL Principle". *Proc. 26th Hawaii Int'l Conf. Syst. Sci.*, Vol.1, pp.746 755.


Level
from 1 to 70
129-residue
Primary Only
+Geometric
Revised
Match
65-residue
Primary Only
+Geometric
Revised
Match
33-residue
Primary Only
+Geometric
Revised
Match
17-residue
Primary Only
+Geometric
Revised
Match
9-residue
Primary Only
+Geometric
R
Match
5-residue
Primary Only
+Geometric
Revised
Match
level
form 71 to 137
65-residue
Primary Only
+Geometric
Revised
Match
33-residue
Primary Only
+Geometric
Revised
Match
17-residue
Primary Only
+Geometric
Revised
Match
9-residue
Primary Only
+Geometric
Revised
Match
5-residue
Primary Only
+Geometric
Revised
Match