TR-0859

Protein 3D Structure Prediction Based on Multi-Level Description

by
K. Onizuka & K. Asai (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 3D Structure Prediction Based on Multi-Level Description

KENTARO ONIZUKA[1]† KIYOSHI ASAI‡
onizuka@icot.or.jp asai@etl.go.jp

MASATO ISHIKAWA†
ishikawa@icot.or.jp

†Institute for New Generation Computer Technology(ICOT)
1-4-28, Mita, Minato-ku, Tokyo, 108 Japan

‡Electrotechnical Laboratory(ETL)
1-1-4, Umezono, Tsukuba Ibaraki, 305 Japan

Abstract

We propose a novel scheme for protein 3D structure prediction using the Multi-level Description scheme (MLD). In this prediction scheme, a local conformation is not only determined by the primary structure at that region (i.e., primary constraints) but is also constrained by the neighboring or surrounding local conformations (i.e., geometric constraints).

The MLD describes a protein conformation with multiple levels of different scales and degrees of abstraction. This scheme facilitate to model the geometric constraints between the neighboring local conformations by analyzing the frequency of overlapping patterns of the local conformations. The primary constraints are modeled by analyzing the relationship between the primary structure and the local conformation at that region.

The MLD representing a real protein conformation must satisfy most of the constraints above. Thus, a protein conformation can be predicted by searching for the optimal MLD that bset satisfies the constraints. This problems is formulated as a combinatorial optimization problem.

[1]鬼塚健太郎
(財) 新世代コンピュータ技術開発機構
〒108 東京都港区三田 1-4-28 三田国際ビル 21 階

1 Introduction

The prediction of a protein 3D structure (i.e., the tertiary structure) from its amino-acid sequence (i.e., the primary structure) is one of the most important yet unsolved problems in molecular biology. In conventional prediction schemes, the sequence of secondary structures is predicted from the primary structure of a protein[Chou and Fasman 74], and then, the secondary structures are packed into the tertiary structure[Cohen et al 82]. Since the perfect secondary structure prediction is assumed in these prediction schemes, the predicted secondary structures are fixed during the tertiary structure formation. It is, however, reported that a local conformation is not directly determined by the primary structure at that region but strongly constrained by the environment in which the local conformation forms.

Thus the local conformation should not be determined only by the primary structure at that region (i.e., primary constraints) but should also be constrained by the environment generated by the folded global structure, and by turn, the global structure is not only determined by the global property of the primary structure but also geometrically constrained by the substructures (i.e.,the geometric constraints).

To include these constraints in a prediction method, we proposed a novel description scheme of protein conformation that models the constraints of protein folding[Onizuka et al 93]. This scheme, MLD (Multi-Level Description), describes a protein conformation with multiple levels of different scales and abstractions. At each level, a protein conformation is represented by a symbolic sequence each symbol of which denotes a local conformation type of the level size. The sequence at low levels represents the fine conformational structures with fairly high resolution. The sequence at high levels represents the abstracted large scale topologies. The MLD is reconstructable into the 3D structure because the MLD has approximately whole information on its tertiary structure.

MLD models two kinds of important constraints. The geometric constraints between the neighboring local conformations are modeled by analyzing the overlapping patterns of local conformations. The primary constraints are modeled by analyzing the relationship between the local conformation type and the primary structure at that region. The primary constraints of the short fragments are considered to represent the local factors of structure formation, and those of long fragments are considered to represent the global factors. Thus, MLD models both local and global factors.

In order to represent local conformations with a symbolic sequence at multiple levels, the classification of local conformations is required. The classification of large conformations has almost never been tried so far, even though the classification of small local conformation has been frequently proposed [Unger et al 89, Miller et al 93, Zhang et al 93]. It is because a large conformation has many degrees of freedom. We, however, solved this problem by abstracting the topology of large conformations. We linearly expand the series of coordinates of C^{α} atoms in a local conformation into the expansion coefficients, where the expansion is cut at an appropriate fixed order. Thus, a local conformation is represented by a set of fixed number of expansion coefficients. We call this set of coefficients LTPs (Linear Topological Parameters). The local conformations represented by LTPs are classified by the statistic clustering technique.

The primary constraints are modeled by analyzing relationship between the distribution of amino acids in a primary structure fragment and the local conformation type at that region.

A real protein conformation must satisfy both the primary and geometric constraints. We assume that such a MLD that satisfies most of the constraints would represent the most probable conformation of a protein. Thus, we can predict the conformation from a primary structure by searching such a good MLD that satisfies most of the constraints. This problems is normally formulated as a combinatorial optimization problem where many optimization algorithms such as genetic algorithm, simulated annealing and integer programming are available.

2 Multi-Level Description

In this section, we shall roughly describe the Multi-Level Description of protein conformation (Detailed in [Onizuka et al 93]). Let us consider the position of the C^{α} atom of each residue as the representative position of the residue. $3N-6$ parameters are required for the complete representation of a local conformation with N residues in a 3D space. This is almost proportional to the number of residues in the local conformation. In our case, however, the degree of abstraction for the representation changes according to the local conformation size because the MLD represents a conformation with multiple levels of different scales where the low levels represent the fine structures and high levels represent the abstracted large scale topologies. Thus, the number of parameters is fixed so that it is sufficient for the complete representation of the smallest local conformation at the lowest level. In our study, the number of parameters is always fixed to nine $(= 3 \times 5 - 6)$ where five is the number of residues in the local conformations of lowest level.

In order to obtain a fixed number of parameters from the local conformations of any size, *we linearly expand the coordinate representation of a local conformation and obtain the expansion coefficients as the parameters of the local conformation.* The number of parameters is fixed by cutting the expansion at the corresponding fixed order. These LTPs (Linear Topological Parameters) may be reverse-transformed into the original coordinate representation. Thus, the abstracted topology of a local conformation is reconstructable from LTPs. We extract the fixed number of

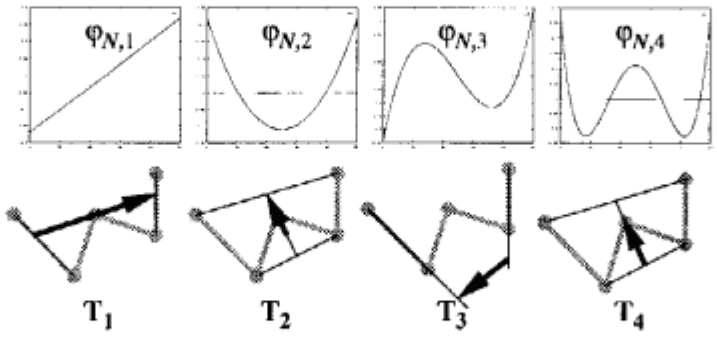


Figure 1: The Four Bases and Corresponding Topological Vectors

LTPs as follows. First, a set of orthonormal bases, $\varphi_{N,ki}$, for linear expansion is

provided, where N is the number of residues in the local conformation, k is the order, and i is the index of the component. Then, we define the topological vectors $\mathbf{T}_k$ as the expansion coefficients. These are obtained by expanding the coordinate representation, $\mathbf{S}_i$, of a local conformation, where $\mathbf{S}_i$ is the positional vector of the ith residue in the local conformation.

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

The orientation of the topological vectors obviously depends on the orientation of the local conformation. To normalize the orientation, a set of unit vectors specific to the local conformation is defined. Two of the topological vectors, normally $\mathbf{T}_1$ and $\mathbf{T}_2$, determine the direction of these unit vectors. The set of LTPs are calculated as the scalar products of unit vectors and topological vectors. These parameters are naturally invariant of the position and the orientation of the local conformation. Thus, the local conformations are classified by clustering the sets of LTPs.

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

We classified the local conformation with 5,9,17,33,65, and 129 ($N = 2^n + 1$) residues and obtained sixteen types at each level. The letters from **A** to **P** denote the types. The five-residue local conformation type **A** corresponds helices, and type **P**,**F**, and **C** usually corresponds strands in our present study. This means, the types at five-residue level well corresponds secondary structures. We assume that the types at higher levels would correspond super secondary structures or domains of protein conformation.

3 Primary Constraints

The relation between the primary structure and the local conformation at that region are modeled statistically by analyzing the relationship between a local conformation type and the distribution of amino acids in the primary structure fragment at that region.

Let γ_t^k denote a type of local conformation, 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 fragment of primary structure at the level k. Let w^k denote the number of residues in the fragment or the local conformation at the level k. The variable of the local conformation type at the position i at the level k is denoted by $\Gamma_i^k \in \{\mathbf{A}^k, \mathbf{B}^k, \ldots, \mathbf{P}^k\}$, and the variable of the fragment of the primary structure at the that position and level is denoted by $\Sigma_i^k \in \{\mathbf{a}^k, \mathbf{b}^k, \ldots, \mathbf{x}^k\}$. 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 the local conformation type γ_t^k is represented by $P_P(\Gamma_i^k = \gamma_t^k | \Sigma_i^k = \sigma^k)$. Since we assume that the primary constraint is independent of the position, it is simply represented as $P_P(\Gamma^k | \Sigma^k)$.

Hence, given a protein sequence, we can directly predict the local conformations type at each region at any level without considering the geometric constraints.

In the present study, We apply Hidden Markov Models (HMMs) to the modeling of primary constraints. HMM is a popular framework in the field of speech recognition. There is a work applying HMMs to secondary structure prediction [Asai et al 93].

We applied the same kind of HMMs to modeling the primary constraints at multiple levels in MLD. As is discussed in [Asai et al 93], 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 Geometric Constraints

The geometric constraints are modeled by analyzing all possible overlapping patterns of two types of local conformations. If it is possible for certain two local conformations to share some residues geometrically or overlap each other, such an overlapping pattern would be frequently found in the real protein conformation. Hence, the frequency of the occurrence of each overlapping pattern is considered as the stochastic constraint of protein conformation.

Each overlapping pattern is defined by 1) the type of the preceding local conformation, 2) the type of the following local conformation, and 3) the offset in the overlapping, where the offset denotes the relative position of the initial residue of the following local conformation from the position of the initial residue of the preceding local conformation in the primary structure. For instance, "**B5** overlaps **A9** at offset 2" represents an overlapping pattern whose preceding conformation type is **B** at the five-residue level, whose following type is **A** at the nine-residue level, and where the initial residue of **B** is the second residue of **A**. Here, we define initial residue of a local conformation as the "0th residue" not the "first residue."

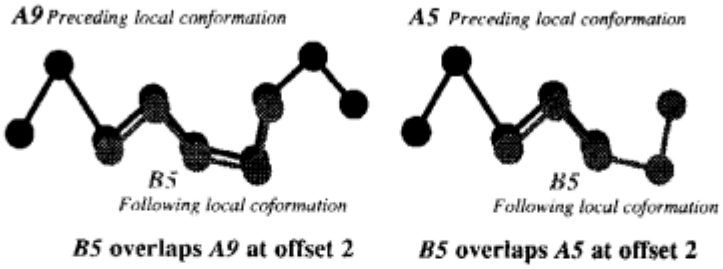


Figure 2: Overlapping Pattern

The complete representation of the frequency of overlapping patterns of large local conformations requires a large number of parameters because the number of possible offsets is almost proportional to the size of the local conformations. The parameters, however, can be reduced remarkably. Here, we also apply linear expansion. The distribution with respect to the offset is approximately represented by only five coefficients. This abstraction is natural because the local conformations involved in an overlapping pattern are already abstracted.

The geometric constraint between two local conformation types, γ^{k_1} and γ^{k_2}, is denoted by $P_G(\Gamma^{k_2}_{i_1} = \gamma^{k_1}_{t_1}, \Gamma^{k_2}_{i_2} = \gamma^{k_2}_{t_2})$. Since we assume that the geometric constraint depends only on the relative position (i.e., the offset) of two local conformations, $d = i_2 - i_1$, the geometric constraint is represented by $P_G(\Gamma^{k_1}, \Gamma^{k_2}, d)$. The geometric constraints are considered to be the probabilities of the co-occurrence of two local conformation type $\gamma^{k_1}_{t_1}$ and $\gamma^{k_2}_{t_2}$ with respect to the offset d.

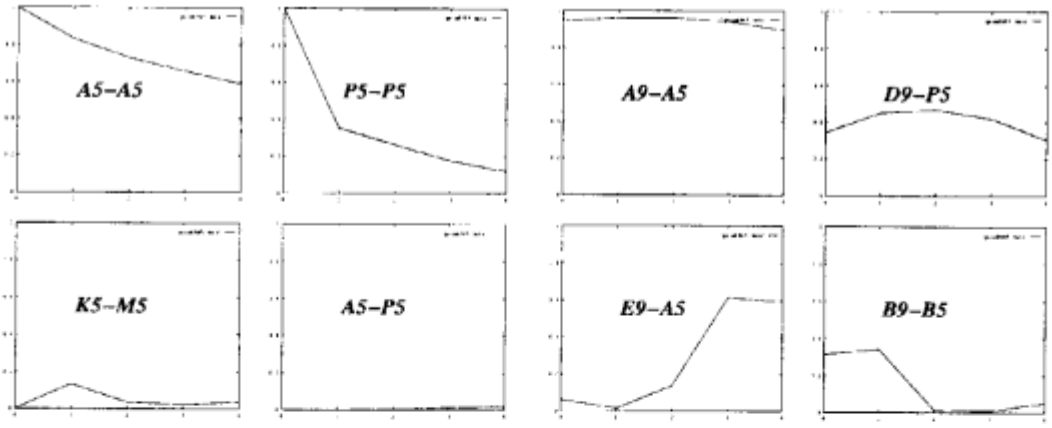


Figure 3: Geometric Constraints

The figures show the frequently of the overlapping patterns with respect to the offset. The horizontal axis indicates the offset and the vertical axis indicates the probability, that is, the normalized frequency of the patterns. In this case, the frequency of the pattern is divided by the frequency of preceding local conformation type. The five-residue local conformation **A5** which corresponds to helices is a continuous conformation. 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 conformation. 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 offsets. **P5** and **A5** hardly overlap because a helix is geometrically very different from a strand.

Since the nine-residue local conformation **A9** corresponds to helices, the five-residue local conformation at that region should be **A5**. The figure shows that the frequency distribution is flat and the normalized frequency with respect to any offset is little less than 1.0. This means that the nine-residue local conformation **A9** hardly allows other local conformation type to occur at five-residue level than **A5** at that region. Likewise, **D9** which corresponds to strands should be built up of **P5**, and the figure shows that the normalized frequency distribution with respect to any offset of **D9-P5** is about 0.5. The nine-residue local conformation **E9** usually occurs at the beginning of helix. Thus, **E9** should allow **A5** to occur at 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 occurs at offset 3 or 4.

5 Optimization Algorithm

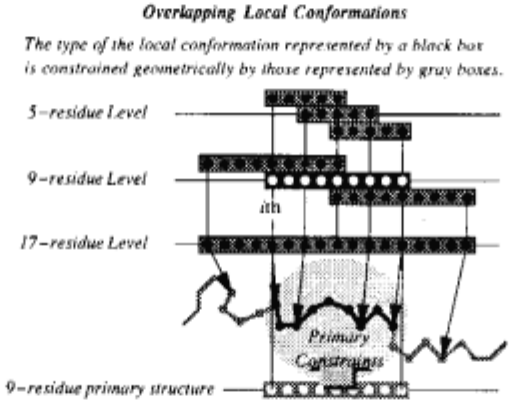


Figure 4: Primary and Geometric Constraints

In this section, we shall propose an optimization algorithm of MLD for the protein 3D structure prediction. This scheme iteratively improves the degree of satisfaction of the geometric constraints by stochastic propagation model.

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

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

where W_I is the weight for inertia term, W_P is that for the primary constraints, and W_G is that for the geometric constraints. These weights shall be so determined

that the degree of prediction accuracy is the best. The solution converges quite rapidly in about ten steps, when W_I is small.

6 Examples of Result

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 HMMs. 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 Conclusion

In this paper, we have proposed a novel scheme for predicting protein tertiary structures using MLD. Many factors of protein structure formation, which were usually neglected in conventional secondary structure prediction schemes, are statistically included in this prediction scheme based on stochastic propagation model. The relationships or constraints between the primary structure and the local conformation are included as the primary constraints where both global and local factors are considered at the multiple levels. The constraints between the neighboring local conformations are included as the geometric constraints. Both constraints are modeled by analyzing the frequency of co-occurrence of local conformation or primary structure types. Thus, each local conformation is so determined that it stochastically satisfies the constraints.

Many points should be discussed on the evaluation score optimized in the combinatorial optimization problem for structure prediction. In the present study, the sum of all the considered probabilities is the evaluation score to be optimized. In most cases of stochastic optimization, however, the sum of the logarithms of probabilities or co-relations is considered as the score to be optimized. In this case, the prediction scheme is formalized in terms of Markov Random Field [Geman and Geman 84], which is a stochastic framework devised to model the restoration scheme of noisy image in the field of computer vision.

One of the problem which remain unsolved is that 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, *"A Multi-Level Description Scheme of Protein Conformation"*, in *Proc. of ISMB'93*, 1993.

[Asai et al 93] Asai, K., S. Hayamizu and K. Onizuka, "HMM with Protein Structure Grammar", in *Proc. of the 26th HICSS*, 1993, pp. 783-791.

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

[Cohen et al 86] Cohen, F.E., R.M. Abarbanel, I.D. Kuntz and R.J. Fletterrick, *"Turn prediction in proteins using a pattern matching approach"*, in *Biochemistry 25*, 1986, pp. 266-275.

[Cohen et al 82] Cohen, F.E., M.J.E. Sternberg and W.R. Taylor, *"Analysis and prediction of the packing of α-helices against a β sheet in the tertiary structure of globular proteins"*, in *J. Mol. Biol. 156*, 1982, pp. 821-862.

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

[Unger et al 89] Unger, R., D. Harel, S. Wherland, and J.L. Sussman, "A 3D building blocks approach to analyzing and predicting structure of proteins" *PROTEINS: Structure, Function and Genetics 5*, 1989, pp. 355-373.

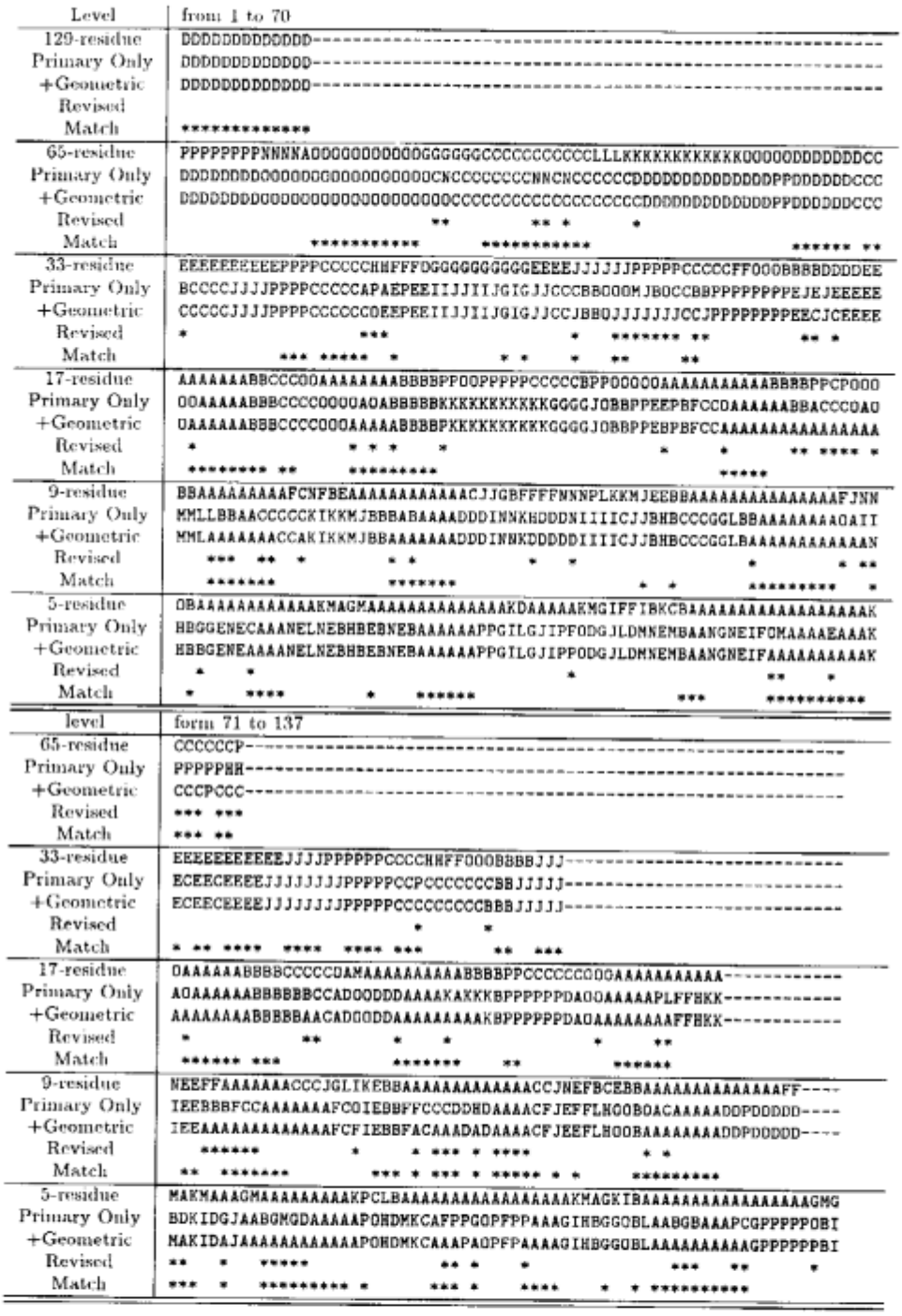
[Miller et al 93] Miller, R.T., R.J. Douthart and A.K. Dunker, *"An Alphabet of Amino Acid Conformations in Protein"*, in *Proc. of the 26th HICSS*, 1993, pp. 689-698.

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

[Metfessel and Saurugger 93] Metfessel, B.A. and P.N. Saurugger *"Pattern Recognition in the Prediction of Protein Structural Class"*, in *Proc. of the 26th HICSS*, 1993, pp. 679-688.

[Onizuka et al 94] Onizuka, K., K. Asai, K. Ito, H. Tsuda, M. Ishikawa, and A. Aiba *"Protein Structure Prediction based on Multi-Level Description"*, submitted to the 27th HICSS, 1994.

[Geman and Geman 84] Geman, S., and D. Geman, "Stochastic Relaxation, Gibbs Distributions, and the Bayesian". *IEEE PAMI 6*, 1984, pp. 721-741.


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
Revised
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