Inverse Protein Folding using Multi-Scale Structure Description

Kentaro Onizuka

Matsushita Research Institute Tokyo, Inc.
3-10-1 Higashimita, Tama-ku, Kawasaki 214, Japan
E-Mail: onizuka@mrit.mei.co.jp
Fax: +81-44-934-3363

Kiyoshi Asai

Electrotechnical Laboratory (ETL)
1-1-4, Umezono, Tsukuba,
Ibaraki, 305 Japan
asai@etl.go.jp
Fax: +81-298-58-5939

Abstract

The applicability of the Multi-Scale Structure Description (MSSD) scheme to the inverse-folding problems was investigated. An MSSD represents a 3D protein structure with multiple symbolic sequences. Each symbol in the symbolic sequence denotes a type of local structure of the level scale. The structure fragments are, thus, classified at each scale level respectively according to the shape and the environment around the fragments: how the structure is exposed to the solvent or buried in the molecule. I modeled the propensity of an amino-acid sequence to the structure fragment type (i.e., primary constraint) at each scale level. The local propensity is, therefore, modeled at small scale levels, while the global propensity modeled at large scale levels. Thus, superposing all the primary constraints, a 3D protein structure yields an amino-acid sequence profile. Evaluating the fit of an amino acid sequence to the profile derived from the known 3D protein structure, we can identify which 3D structure the given amino-acid sequence would fold into. I checked whether a sequence identifies its own structure over two hundred protein sequences. In many cases, an amino acid sequence identified its own 3D protein structure.

1 Introduction

With the recent rapid increase in the number of known 3D protein structures, the method to identify protein sequences that fold into a known 3D structure would be more promising than the 3D structure prediction *ab initio*. After the Chothia's shocking declaration that *"There would be no more than thousand protein families!"* [Chothia 91], the inverse protein folding problem has ever been attracting. In any method for the inverse folding problem, it is necessary to define a scoring function to evaluate the fit of an amino-acid sequence (1D being) to protein conformations (3D being). Some focused on the compatibility of each amino-acid type to the environment around the residue [Bowie et al. 91], some on the empirical potential derived from the known 3D protein structure [Sippl and Weitckus 92], and other on the statistical potential based on Bayesian principle [Goldstein et al. 94].

This paper discusses on the applicability of MSSD scheme and the structure-sequence propencity evaluation method to inverse folding problem. An MSSD represents a protein conformation at mul-

tiple scale levels. At each level, the conformation is described by a symbolic sequence, each symbol of which denotes a type of local structure of the level scale. Local structures are classified into several types at each level respectively according to their shape and the environment. The classification is, therefore, closely related to the secondary structures particularly at the small scale levels. The description at middle scale level is considered to represent the supersecondary structures, and that at high levels represents the global topology.

Since I classified the structures according not only to their shape but to their environment, two structures with similar shapes but in the different environments are classified into different types: the helix exposed to the solvent is classified into a different type from those buried in the molecule. Let us call the compatibility of the structure type to the amino acid sequence "primary constrains" which we regard as the constraints from the primary sequence to the choice of structure types. Hence, given an amino acid sequence fragment, we can roughly estimate which type of local structure it would form. The 3D structure prediction method based on the MSSD scheme is discussed in the literature [Onizuka et al. 94].

To apply the MSSD scheme to the inverse protein folding problem, the primary constraints are used inversely. Given a fragment of amino-acid sequence, we can evaluate its fit to the structure types of the fragments. Or rather, given a structure type at a level, we can obtain an amino-acid sequence profile attached to the structure type under my model. The fit of a given amino-acid sequence to this profile is, therefore, equivalent to the fit to the structure type. Since the structures are classified according to their shape and environment, my approach is, in some sense, the extension of the method proposed in the literature [Bowie et al. 91], where the compatibility of an amino-acid sequence to the secondary structure type and the environment around each residue in the sequence is considered to evaluate the fit. The extension, here, indeed concerns the multiple scale evaluation of the fit. The sequence profile is calculated by superposing all the subprofiles derived from the structure fragment types in the given MSSD. The fit of a sequence to the whole 3D structure is not only evaluated at the small scale level in the MSSD, but at all scale levels available. Chances are that even though a given sequence does not fit to a MSSD at low levels, the sequence may well fit at high levels. Thus, we can identify a sequence that fold into an unknown 3D structure but similar to a known 3D conformation, even though the local fine structures of the unknown one would be quite different from those of the known one: the fine structures may differ even if the amino-acid sequence of the two protein is very similar to each other.

2 Method

This section describes the methods used in my inverse-folding scheme. On MSSD scheme, reference the literature [Onizuka et al. 93]. We extract a set of numerical parameters from structure fragments, where paramters represent the fragment shape and the environment around it. Then the structure fragments are classified using Learning Vector Quantization.

The first subsection describes the classification of the environment arond the structure fragment. The second subsection shall define the primary constraints between the structure types and the primary sequence fragments. And then I formalizes the scoring function for the inverse folding problem. The last subsection shall illustrate the dynamic programming with A^* algorithm applied to the alignment between the sequence profile derived from the 3D structure and the amino-acid sequence.

2.1 Classification of The Environment around Structure Fragments

Here, I discuss how we can incorporate the solvent accessibility of structure fragments into the structure classification.

More and more biologists are aware of the importance of hydrophobic interaction between the residues during the folding process. A protein chain folds into a tertiary structure so that the hydrophobic residues would be buried inside the molecule, whereas the hydrophillic ones exposed to the solvent. The hydropathy of each residue must be a strong factor determining the environment around the residue. When a structure fragment is deeply buried in the molecule, most residues in the fragment should be hydrophobic, while hydrophillic when exposed to the solvent. Indeed is it that, when the fragment is half buried and half exposed, the residues around the buried region

should be hydrophobic and other residues hydrophillic. The propensity of each amino acid type to the environment is considered even stronger than that to the secondary structure [Saito et al. 93]. Considering the propensity from the primary sequence, we can estimate how the the structure fragment would be buried or exposed . In order to characterize the environment around a structure fragment, I introduce a new parameter attached to each residue in the structure, the Quasi Buried Depth (QBD), which takes positive value when the residue is buried inside the molecule while takes negative value when exposed to the solvent. The dimension of the parameter is length so that the calculation with the topological parameters physically would make sense. First, I give the definition of QBD, and then I illustrate how to parameterize the solvent accessibility of a structure fragment.

A residue deeply buried inside the molecule is surrounded by more residues than those exposed to the solvent. The number of residues nearby a given residue within a certain distance can be considered to measure how the residue is buried or exposed. This number is given by counting the number of residues in a sphere with certain radius centered at the position of a given residue. The predictability of this number from a given primary sequence is discussed in the literature [Saito et al. 93]. The Quasi Buried Depth is derived from the number, and has the dimension of length.

¿From the investigation of the maximum number of residues M in a sphere whose radius is r, I found that M is almost proportional to the $r^{2.45}$, and is calculated as $M = 0.15r^{2.45}$. This suggests, the residues are not optimally packed but are suboptimally packed in the sense of fractal dimension. We can consider that when the actual number of residues N in the sphere with radius r centered at a given residue would be equal to M, the depth from the surface of the protein molecule to the residue would be estimated greater than r, while the depth would be estimated around zero when N is a half of M. The number N can be, therefore, transformed into the quasi depth of the residue from the surface. The Quasi Buried Depth d^Q is, therefore, calculated as $d^Q = (2N/M - 1)r$, where $M = 0.15r^{2.45}$. When d^Q takes a positive value, the residue is considered buried, while considered exposed for the negative d^Q.

Likewise the topological parameters which are obtained by linear transformation, the set of environmental parameters representing how a structure fragment is buried or exposed is calculated by transforming the set of residues' QBD in a structure fragment. The environmental parameter of kth order E_k is calculated as below.

$$E_k = \sum_{i=0}^{N-1} \varphi_{N,ki} d_i^Q, \tag{1}$$

where d_i^Q is the QBD of ith residue in the fragment. Since the physical dimension of environmental parameters is length, these parameters can be used with topological parameters. In my study, the maximum order of expansion is five, and five environmental parameters are to represent the solvent accessibility of the structure fragment.

The parameters obtained from QBD alanysis are marged to other parameters representing the fragment shape. And then the structures are classified according to the parameters using Learning Vector Quantization.

2.2 Primary Constraints

The primary constraints relate the primary sequence and the structure type at each region. MSSD scheme is particularly suitable to model both local and global factors of structure formation. The primary constraints for short structure fragments naturally represent local factors, and those for long ones represent global or long-range factors. For further discussion, I define several notations here.

Let γ_t^k denote a structure 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 sequence fragment at the kth level. And we further denote w^k as the number of residues in the structure fragment at the kth level. We denote $\Gamma_i^k \in \{\mathbf{A}^k, \mathbf{B}^k, \ldots, \mathbf{X}^k\}$ as the variable that takes a structure type, where i denotes the position in the primary sequence. We also denote Σ_i^k as the variable that takes a primary sequence fragment. Note that the position i here denotes the position of the first residue of the structure fragment in the primary sequence.

The probability of a primary sequence fragment σ^k forming a type of structure 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 invariant of its absolute position in the primary sequence but only depends on the structure type and the primary sequence at that region, it may simply be represented as $P_P(\Gamma^k|\Sigma^k)$.

In the field of molecular biology, the sequence profiles are frequently used to analyze the relationship between a sequence pattern and the structure or function at that region, where the frequency of each amino-acid type is counted with respect to the position. This technique is directly applicable to model the primary constraints at small scales, though it requires large number of parameters, again, for the primary constraints at large scale. For example, at five-residue level, the number of parameters representing the frequency is $100 = 20 \times 5$ where 20 is the number of amino-acid types, and 5 is the number of residues in the structure fragment at that level. At the large scale levels, where the number of residues are more than 100, more than 2000 parameters are required. In this case, however, we can compress the sequence profile using the same technique as I applied to the structure abstraction. We can always reduce the number of parameters into 100 using linear expansion again.

2.3 Inverse-folding Scheme

Given an MSSD representing a 3D protein structure, we can estimate the most probable sequence from the MSSD using the inverse primary constraints $P_I(\Sigma|\Gamma)$, which is simply given by calculating the fit of a sequence to a profile. $P_P(\Gamma|\Sigma)$ is calculated by applying the prior $P(\Gamma)$ to $P_I(\Sigma|\Gamma)$.

Let i denote a position in the sequence. Let t^A denote an amino-acid type, and let T^A be a variable that takes one of the amino acid type t^A. We can derive the probability $P(T_i^A = t^A)$ of the amino-acid type occurring at the position i, from the structure fragment type covering the position i. Let $P_I(T_i^A = t^A|\Gamma_j)$ denote the probability of the amino-acid type t^A occurring at the position i in the fragment. To superpose the $P_I(T_i^A)$, we have to divide this value by the prior $P(T_A = t^A)$, because the prior is doubly or triply calculated. Thus, $P_I(T_i^A)$ is calculated as below.

$$P(T_i^A = t^A) = P(t^A) \prod_{All\ \Gamma_j\ covering\ i} \frac{P_I(T_i^A = t^A|\Gamma_j)}{P(t^A)} \tag{2}$$

In this case, however, the prior $P(t^A)$ does something unpreferable. The probability $P(T_i^A = t^A)$ almost always suggests that Alanine is the most probable amino-acid type at any position. This means that the inverse primary constraint $P_I(T_i^A|\Sigma_j)$ is much weaker than the prior. Hence, I adopt $C_I(T_i^A = t^A) = P_I(T_i^A = t^A)/P(t^A)$ instead of $P_I(T_i^A = t^A)$. This value is greater than 1.0 when the amino-acid type stochastically occurs more than random level.

The superposition of all the inverse primary constraints from the MSSD derived from the given 3D structure yields a stochastic sequence profile. The fit of a sequence to this profile is considered the fit to the given conformation represented by the MSSD and by turn the fit to the given 3D structure.

2.4 3D-1D Alignment

The alignment between the sequence and the profile is carried out simply by dynamic programming. The dynamic programming searches for the optimal alignment that minimize the score E below. Some appropriate gap penalty should be used when we permit gaps.

$$E = -\sum_i \log C_I(T_i^A = t^A) + gap\,penalty. \tag{3}$$

We consider the resultant score E as the fit of amino-acid sequence to the sequence profile $C_I(T_i^A)$ derived from the MSSD representing 3D structure. Hence, given a primary sequence of a protein whose 3D structure is unknown, we can search for the most compatible 3D structure in the protein structure database. This is far simpler than that of those schemes using Sippl potential [Sippl and Weitckus 92, Jones et al. 92, Yuke and Dill 92, Skolnick and Kolinski 92], where it is necessary to apply the double dynamic programming that requires large amount of calculation.

I applied the A* algorithm to the 3D-1D alignment, which was first applied to the protein sequence alignment in the literature [Araki et al. 93]. This algorithm finds the optimal solution

while the calculation amount is much smaller than that of conventional dynamic programming algorithms, though the implementation is much difficult.

The choice of the gap penalty has not yet established. In most cases, there are three parameters concerning the gap penalty: 1) the slide gap penalty is the cost for the offset between the two sequence; 2) the initial gap penalty is the cost to put a gap in a sequence; and 3) the incremental gap penalty is the cost for the length of each gap. When the initial gap penalty equals to incremental one, the dynamic programming turns out to be quite simple with a simple network. Thus, I adopted this penalty. The slide penalty should be zero to allow any offset between the sequence and profile without costs.

3 Results

I used the same data set of protein structures as that used for structure classification. To cross-validate the result, the data set was divided into five groups randomly so that each group would contain forty nine structure data. I obtained five sets of primary constraints, where each set was derived from the structure data in five groups. When a structure yields the sequence profile, I did not use those primary constraints that are derived from the structure group including that structure.

First, as a preliminary experiment, I investigated how a protein sequence fits its own 3D structure evaluating the Z score. Here, I did not align the profile and the sequence: the gaps are, thus, not considered. We can obtain the Z score of a sequence to a profile by normalizing the score E by the mean score $< E_{random} >$ and the deviation $\sigma_{E_{random}}$ of random sequences to that profile, where E is defined as below.

$$E = -\sum_i \log C_I(T_i^A = t^A). \tag{4}$$

Thus, Z score E_Z is,

$$E_Z = \frac{E - < E_{random} >}{\sigma_{E_{random}}}. \tag{5}$$

I investigated the fit of sequences to the structures at only one scale level, in order to see which level best corresponds the sequence. The plot below shows the mean Z score with respect to the scale level. The correspondence is the best at the lowest 5-residue level and it decreases monotonously

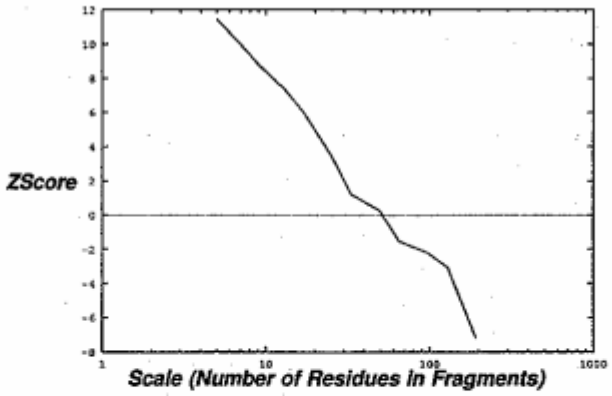

Figure 1: Z Score versus Scale

with the increase in the level. This suggests that a local sequence strongly influence the formation of the secondary structures at that region, because the classification at the 5-residue level well corresponds the secondary structures. Probably due to the over-learning, the scores at the high levels are below zero.

Second, I checked whether a sequence would identify its own structure. The hit-ratio of the self-identification direclty suggests the performance of my inverse-folding scheme. I checked whether the fit of a sequence to its own structure would scores the best among all sequence-profile combinations. We selected 188 protein structures from the data set which I used to model the primary constraints,

because the other structure data contain residue-lacks or unacceptable bond lengths. I investigated the hit-ratio of self-identification. When the compatibility score of the sequence to its own structure obtained from the 3D-1D alignment scores the best, I consider that the identification hits. I did exhausting 3D-1D alignment for 188 x 188 times. The table below shows the hit rationes.

	Total	Hit	Hit Ratio
Single Level	188	63	0.335
Multi-Level	188	90	0.478

This result actually shows that the performance of self-identification is better when many scale levels are incorporated.

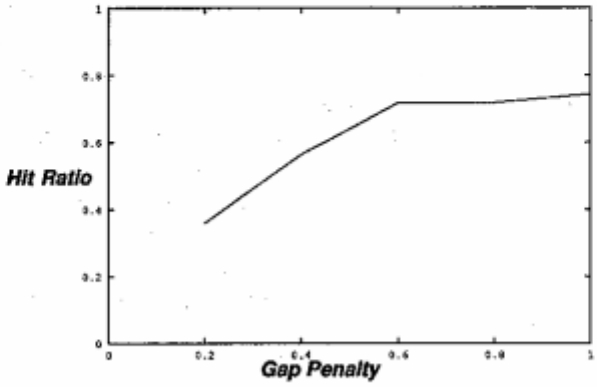

Figure 2: Hit Ratio versus Gap Penalty

Third I investigated how the gap penalty influence the hit ratio. In this case, I used only first group of data set which contains thirty nine proteins. This graph shows that the higher the gap penalty is, the better is the hit ratio.

4 Discussion

In this paper, I proposed the multi-scale evaluation scheme to solve the inverse protein folding problem. I incorporated the compatibility of sequences to 3D structures not only at the small scale level but also at the large scale levels.

The results show that the multi-scale compatibility scoring works better than the single scale one, even though the compatibility scores at large scale levels poorly corresponds the fit between the structures and sequences better than those at small scale levels. Considering size of data set containing 188 protein structures, the result is not so bad.

Considering the poor mean Z score at high levels, the 3D-1D correspondence at high levels does not seem to be stochastically modelable. Thus, we should not use those levels in order to obtain better self-identification hit-ratio.

I investigated the applicability of MSSD scheme to the inverse folding problem, and found that the multi-scale scoring works far better than single scale scoring. This means that the score at high levels does a great deal to enhance the performance.

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.

[Chothia 91] Cyrus Chothia 1992. "One thousand families for the molecular biologist". *Nature 357*: 543-544.

[Bowie et al. 91] Bowie, J.U., R. Lüthy, and D. Eisenberg 1991. "A Method to Identify Protein Sequence That Fold into a Known Three-Dimensional Structure" *SCIENCE 253*: 164-170.

[Sippl and Weitckus 92] Sippl, M., and S. Weitckus 1992. "Detection of Native-like Models for Amino Acid Sequences". *PROTEINS 13*: 258-271.

[Goldstein et al. 94] Goldstein, R. A., Z. A. Luthey-Schulten, and P. G. Wolynes 1994. "A Bayesian Approach to Sequence Alignment Algorithm for Protein Structure Recognition". *Proc. of 27th HICSS 5*: 306-315.

[Onizuka et al. 94] Onizuka, K., K. Asai, H. Tsuda, K. Ito, M. Ishikawa, and A. Aiba 1994. "Protein Structure Prediction Based on Multi-Level Description". *Proc. of 27th HICSS 5*: 355-354.

[Saito et al. 93] Saito, S., T. Nakai, and K. Nishikawa 1993. "A Geometrical Constraint Approach for Reproducing the Native Backbone Conformation of a Protein". *PROTEINS 15*: 191-204.

[Hobohm et al. 92] Hobohm, U., M.Scharf, R.Schneider, C.Sander 1992. "Selection of a representative set of structures from the Brookhaven Protein Data Bank". *Protein Science 1*: 409-417.

[Jones et al. 92] Jones, D., W. Taylor, and J. Thornton 1992. "A New Approach to Protein Fold Recognition". *Nature 358*: 86-89.

[Yuke and Dill 92] Yuke, K., and K. Dill 1992. "Inverse Protein Folding Problem". *Proc. Natl. Acad. Sci. USA 89*: 4163-4167.

[Skolnick and Kolinski 92] Skolnick, J., and A. Kolinski 1992. "Topology Fingerprint Approach". *Science 250*: 1121-1125.

[Araki et al. 93] Araki, S., M. Goshima, S. Mori, H. Nakashima, S. Tomita, Y. Akiyama, and M. Kanehisa 1993. "Application of Parallelized DP and A* Algorithm to Multiple Sequence Alignment". *Proc. of Genome Informatics Workshop V*: 94-102.