ネイティブ配列とランダム配列の比較に基づいた遺伝的アルゴリズムによるアミノ酸インデックスの抽出

金井　理[†,☆]　藤　博幸[†]

球状タンパク質は、立体構造の折れたたみに関する情報を、その一次構造中のアミノ酸の配置として持っていると考えられる。したがって、折れたたむネイテイブ配列をランダムにシャッフルする（すなわちランダム配列）と、その情報は失われると期待される。この仮定に基づいて、われわれは、折れたたみに影響を及ぼすと考えられる要因をアミノ酸インデックスの形で抽出するために、以下のような自己組織的手法を開発した。まず、ネイテイブ配列とランダム配列を、インデックスにより配列プロフィールに変換する。さらに、AR 解析および LPC ケプストラム解析を行い、配列プロフィール間の差異を LPC ケプストラム距離として表現する。そして、その距離が大きくなるように、遺伝的アルゴリズムでインデックスを進化させる。この手法により得られたランダム配列に対しネイティブ配列を特徴づけるインデックスは、疎水性インデックスと高い相関を持つことがわかった。このことは、アミノ酸の性質を親水性・疎水性で見たとき、その配置が、タンパク質の折れたたみに関連していることを示唆している。また、このインデックスは、タンパク質がどのような二次構造を多く含むかで、若干の違いがあることが認められた。

GA Generates New Amino Acid Indices through Comparison between Native and Random Sequences

SATORU KANAI[†,☆] and HIROYUKI TOH[†]

The amino acid sequence of a protein carries its folding information. If the information is encoded by the arrangement of the amino acid residues along the primary structure, the random shuffling of the residues would degrade the information. We developed a new method to compare the native sequence with random sequences generated from the native sequence, in order to extract such information. First, amino acid indices were randomly generated. That is, the initial indices have no significance on the feature of residues. Next, using the indices, the averaged distance between a native sequence and the random sequences was calculated, based on the autoregressive (AR) analysis and the linear predictive coding (LPC) cepstrum analysis. The indices were subjected to the genetic algorithm (GA) using the distance as the fitness, so that the distance between the native sequence and the random sequences becomes larger. We found that the indices converged to hydrophobicity indices by the GA operation. The AR analysis with the converged indices revealed that the autocorrelation in the native sequence is related to the secondary structure.

1. Introduction

The amino acid sequence of a native protein folds into a globular structure to exert its biological activity. A statement known as Anfinsen's dogma[1] maintains that the information about the folding of a globular protein is carried by the amino acid sequence. If we extracted such information from a sequence, we could predict the tertiary structure of the sequence. However, we do not fully understand the relationship between amino acid sequences and the corresponding 3D-structures of proteins.

One of the approaches to tackle this problem is to find the orders or rules held by the amino acid sequences. In the analyses, the sequences are transformed into a series of numerical data. Such transformation can be performed using amino acid indices. An index is a set of numerical values, each of which corresponds to a residue. Each residue of a given sequence is replaced with the corresponding numerical value of a given index. Then, the sequence of a protein is expressed as one-dimensional numerical value data, like time series data. Hereafter, the series of numerical values correspond-

[†] 生物分子工学研究所
Biomolecular Engineering Research Institute
[☆] 現在，富士通株式会社
Presently with Fujitsu Limited

ing to a sequence is called "profile". The profile has been analyzed by signal processing technique in order to find periodicity or autocorrelation in a given sequence. Some people insist that residues are randomly arranged in the sequences of native proteins[2)~4)]. However, other people have found periodicity or autocorrelation in the sequences of native proteins[5)~10)]. Thus, the results obtained from the various approaches are still controversial.

Our method discussed in this paper is regarded as solving an inverse problem against the current signal processing approaches. We did not use any of the known amino acid indices for the study. Instead, we made a following assumption, in order to obtain indices a related to the structural information carried by sequences; if the information about the folding of a protein is carried by the arrangement of the residues along the primary structure, then the information is degraded by the random shuffling of the sequence. Therefore, it is expected to extract information related to protein folding through comparison of sequences of native proteins and the randomly shuffled sequences. We connected the sequences of native proteins to generate a long sequence, which we call the "native sequence" in this paper. For comparison, each sequence constituting a native sequence was randomly shuffled, and then was connected in the same order as in the native sequence. The long sequence composed of the shuffled sequences is called "random sequence". Besides, we prepared a large number of indices, whose elements were randomly generated. Using each index, the native sequence and the random sequences were transformed into profile data sets. The former is called "native profile", while the latter is called "random profiles". Both profiles were subjected to a autoregressive (AR) analysis[11)]. Then, the distance between the native and random profiles was calculated, which is known as the linear predictive coding (LPC) cepstrum distance[12)]. Using the distance as the fitness of the index, the population of indices was subjected to optimization by subjected to a genetic algorithm (GA)[13)] as follows. The index with the highest fitness in the final generation is expected to distinguish the native sequence from the random sequences efficiently.

2. Materials and methods

2.1 Preparation of native sequences and generation of randomly shuffled sequences

The proteins used in this study were selected based on the structure classification by CATH[14)]. 20 proteins were selected from the mainly α class. 21 proteins were selected from the mainly β class. 39 proteins were taken from the α–β class. All of the selected proteins satisfy the following structural conditions: (a) the sequence length is equal to or greater than 100 residues, (b) each protein is made of a single domain, and (c) no hetero atoms or ligands are contained in the structure. A sequence length of a single protein was too short to obtain enough samples for the AR analysis. That is the reson why we connected sequences to form a native sequence. We constructed three native sequences by connecting sequences belonging to the same structural class. The length of the three native sequences corresponding to the mainly α, the mainly β, and the α–β class were 3677, 3946, and 7050 residues, respectively. In addition, the three native sequences were connected to form one more sequence, which was referred to here as "all data". The four sequences were used as native sequences. Corresponding to each native sequence, 100 random shuffled sequences were generated, according to the procedure described in Introduction. The number of random sequences generated from a native sequence was 100.

2.2 Genetic algorithm to generate amino acid indices

To solve the problem, the standard GA algorithm was encoded into a program that can perform each operation as follows.

Chromosome representation and initialization: A chromosome indicates an amino acid index. That is, a chromosome is a set of 20 numerical values, each of which corresponds to an unknown feature of a residue. Each element of an index is restricted in a range from 0.0 to 1.0. An initial population was composed of 500 indices, whose elements were randomly generated.

Distance between native and random profiles: First, a sequence is converted to a profile, using a given index. Then, a profile is analyzed as the univariate AR model. Finally, based on the obtained AR models, the LPC cepstrum distance betweem two profiles is calculated. The AR order examined in this study ranged from 1 to 8, and the order of LPC cepstrum was 15.

Reproduction: The raw fitness of chromosome x is obtained as follows. (I) A profile of a native sequence, $\mathbf{P}(x)$, is generated using a chromosome x. Then, a set of random profiles, $\{\mathbf{P}'_1(x), \ldots, \mathbf{P}'_i(x), \ldots, \mathbf{P}'_S(x)\}$, is generated by applying chromosome x to the set of random sequences, where S is the number of random sequences generated from a native sequence. (II) The LPC cepstrum distance between the native profile and a random profile i, $D(\mathbf{P}(x), \mathbf{P}'_i(x))$, is calculated. (III) The raw fitness of the chromosome x, $RF(x)$, is then calculated as

$$RF(x) = \frac{100}{S} \cdot \sum_{j=1}^{S} D(\mathbf{P}(x), \mathbf{P}'_i(x)). \qquad (1)$$

For efficient selection, the raw fitnesses are further modified by sigma truncation.

Elitism strategy: The top 1% of the reproduced offspring are regarded as elite when the individuals of the offspring are sorted by scaled fitness. The elite population is transfered to the next generation, skipping the operation of crossover and mutation. On the other hand, the remaining non-elites are subjected to the following two operations.

Crossover: The uniform crossover operation is adopted. The crossover probability is 0.1 in actual runs.

Mutation: In this study a mutation means an increase or a decrease in the value of an element of an index by a given constant. The mutation probability is 0.1. 0.05 is used as the constant value for an increment or a decrement. A generated random integer determines whether the change is increment or decrement. When the numerical value stored in the element becomes greater than 1.0 by the increment operation, the value is re–set to be 1.0. Likewise, the value is re–set to be 0.0 when the value becomes less than 0.0 by the decrement operation.

Judgment of termination: When the GA operation is repeated by a given number, the program is terminated. In many cases, 50 generations were sufficient for the highest fitness in the population to converge to a constant value. To ensure the convergence, we added 50 more generations.

2.3 Evaluation of generated amino acid index

To compare the generated indices with each other or with known indices, the correlation coefficient between two indices was calculated. Moreover, the cluster analysis was performed using the absolute values of the correlation coefficients as the distance between two indices.A dendrogram was constructed by the unweighted pair–group method with the arithmetic mean.

3. Results and Discussion

In all of the cases of the evolution process, the fitness converged rapidly. The converged indices for the mainly α class, the α–β class, and all data, were similar to each other when the AR orders are ≥ 2. For the mainly β class, except for the case of the AR orders $= 1$ and 2, the indices were highly correlated with each other.

The correlation coefficients were calculated between every pair of the indices and 434 known indices available in an amino acid index database, AAindex1[15]. All of the indices, except for four indices, showed high correlations not only to each other, but also to the known indices, which are classified into a group of hydrophobicity indices. The four exceptional indices included the one for the AR order $= 1$ from the mainly α class, the one for the AR order $= 2$ from the mainly β class, the one for the AR order $= 1$ from the α–β class, and the one for the AR order $= 1$ from all data. They did not show prominent similarity to any of the known indices. The relationship among the indices and the known 149 hydrophobicity indices was examined by a cluster analysis. The indices on a structural class were similar to each other, despite the difference in the AR orders. Reflecting the similarity, such indices formed a cluster corresponding to each structural class in the dendrogram. The cluster on the mainly α class was distinct from the other clusters, and showed high correlation with four known indices classified into a group of hydrophobicity indices. The cluster of the mainly β class was also distinct from the other clusters. The 19 known indices belonging to the group of hydrophobicity indices were close to the indices from the mainly β class in the dendrogram, which were different from the indices closely related to the cluster of the mainly α class. The cluster of the α–β class occupied a position between those of the mainly α and mainly β classes. This cluster was distinct from the other two clusters, although it was relatively close to the cluster of the mainly α class, rather than that of the mainly β class. The cluster of the α–β class included that of all data.

The question addressed in this study was what is the difference between native sequences and randomly shuffled ones. We considered that such a difference is related to the folding information within the native sequences. Our GA operation generated indices related to the hydrophobicity. The results suggested that the amino acid residues of proteins are arranged in the primary structures with autocorrelation in hydrophobicity.

The next question is how native sequences are designed in hydrophobicity. The sequences were expressed by the AR models, each of which is characterized by the corresponding AR coefficients. We expressed the native sequences as AR models using the converged indices, and examined the relationship between the AR coefficients and the structural class of the sequences. As a result, for the mainly α class, the pattern of the AR coefficients is consistent with the periodicity of an α helical structure. Likewise, the AR coefficients of the native profiles of mainly β class represented the preiodicity of β strands. Moreover, for the α–β class and all data, the pattern seemed to be a mixture of that for the mainly α class and that for the mainly β class.

References

1) Anfinsen, C. B.: Principles that govern the folding of protein chains, *Science*, Vol. 181, pp. 223–230 (1973).

2) White, S. H. and Jacobs, R. E.: Statistical distribution of hydrophobic residues along the length of protein chains. Implications for protein folding and evolution, *Biophys. J.*, Vol. 57, pp. 911–921 (1990).

3) White, S. H. and Jacobs, R. E.: The evolution of proteins from random amino acid sequences. I. Evidence from the lengthwise distribution of amino acids in modern protein sequences, *J. Mol. Evol.*, Vol. 36, pp. 79–95 (1993).

4) Rahman, R. S. and Rackovsky, S.: Protein sequence randomness and sequence/structure correlations, *Biophys. J.*, Vol.68, pp.1531–1539 (1995).

5) Pande, V. S., Grosberg, A. Y. and Tanaka, T.: Nonrandomness in protein sequences: evidence for a physically driven stage of evolution?, *Proc. Natl Acad. Sci., USA*, Vol. 91, pp. 12972–12975 (1994).

6) Sun, S. and Parthasarathy, R.: Protein sequence and structure relationship ARMA spectral analysis: application to membrane proteins, *Biophys. J.*, Vol.66, pp.2092–2106 (1994).

7) Irbäck, A., Peterson, C. and Potthast, F.: Evidence for nonrandom hydrophobicity structures in protein chains, *Proc. Natl Acad. Sci., USA*, Vol. 93, pp. 9533–9538 (1996).

8) Makeev, V. J. and Tumanyan, V. G.: Search of periodicities in primary structure of biopolymers: a general Fourier approach, *Comput. Appl. Biosci.*, Vol. 12, pp. 49–54 (1996).

9) Rackovsky, S.: "Hidden" sequence periodicities and protein architecture, *Proc. Natl Acad. Sci., USA*, Vol. 95, pp. 8580–8584 (1998).

10) Weiss, O. and Herzel, H.: Correlations in protein sequences and property codes, *J. Theor. Biol.*, Vol. 190, pp. 341–353 (1998).

11) Wei, W. W. S.: *Time Series Analysis : Univariate and Multivariate Methods*, Addison-Wesley (1990).

12) Rabiner, L. and Juang, B. H.: *Fundamentals of Speech Recognition*, Prentice Hall (1993).

13) Holland, J. H.: *Adaptation in Natural and Artificial Systems*, University of Michigan Press (1975).

14) Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. and Thornton, J. M.: CATH–a hierarchic classification of protein domain structures, *Structure*, Vol. 5, pp. 1093–1108 (1997).

15) Nakai, K., Kidera, A. and Kanehisa, M.: Cluster analysis of amino acid indices for prediction of protein structure and function, *Protein Eng.*, Vol. 2, pp. 93–100 (1988).