Protein Three Dimensional Structure Prediction on Object Oriented Database

Akira Shimada
shimada@asahi-kasei.co.jp

Hideki Takehara
takehara@asahi-kasei.co.jp

Kazunori Toma
toma@asahi-kasei.co.jp

Computer Science Department, Asahi Chemical Industry Co., Ltd.
2-1, Samejima, Fuji, Shizuoka 416, Japan

Abstract

We defined the protein inverse folding problem with an object oriented database and an empirical hydrophobic penalty function, which was derived from the number of residues around each residue in a protein three dimensional structure. Under the database management system, we compiled the known structures of proteins and the evaluation function into one functional database. In order to compare our approach with the methods proposed by other groups, the functional database was applied to the problem of globin family recognition. Although the penalty function itself is simple and non-optimized, it gave considerably good results.

1 Introduction

The function of predicting protein 3D structures from amino acid sequences is one of the most important requirements of the genome-informatics system. Although there have been much efforts in the field over 30 years, no practical method has been found. One of the emerging methods is the inverse-folding approach(1). We would like to present a novel way to implement the problem on an integrated database of genome-related information.

2 Methods

Protein 3D structures were taken from the Protein Data Bank(2). The choice of protein structures followed the study by Sippl and Weitckus(3).

We used the RIS penalty function(4) as the evaluation function for the inverse-folding approach. The RIS of a sphere size was calculated as the number of residues in the sphere of a defined radius around a given residue. Radii from 6 to 14 Å, with a 1 Å increment, were used to generate the RIS values for each sphere size. The penalty value for each radius size is defined by the following formula.

$$PenaltyValue = \sum_{i=1}^{N} \{|RIS(cal,i) - RIS(stand,aa)|/sd(aa)\}$$

島田 章、竹原 英毅、戸澗 一孔：旭化成工業株式会社コンピューターサイエンス室，〒416 静岡県富士市鮫島 2-1

where *RIS(cal,i)* denotes the calculated RIS value of the i-th position on a given structure, *RIS(stand,aa)* does the standard RIS value defined as the average over real values of the data set proteins, and *sd(aa)* does the standard deviation of the RIS value for each amino acid.

We constructed the database using the ONTOS object-oriented database management system on a Sun Sparc Station 10/51. Protein amino acid sequences were chopped into residue units, and stored with coordinates and residue numbers. Also, all penalty values corresponding to 20 amino acids in each position was pre-calculated and stored into the database as the information linked to each residue data. In the evaluation, only structures which had larger residue number than the amino acid sequence were considered, and we fitted the amino acid sequence on each structure without any gaps. The starting point was shifted one by one until the C terminal of the sequence met that of the structure.

3 Results and Discussion

We applied our method to the problem of the globin structure identification, whether seven globin sequences can identify their own structures or not. We threaded globin sequences through all protein structures in the database. RIS of large radii generally gave better results. While RIS of radius 6 Å has weak discriminatory power, three globin was scored at the top by 13 and 14 Å. All of the seven sequences found their own structures in the top 1% at radii more than 11 Å.

The fact that RIS of 6 Å has little ability to detect native structure indicates that local interactions can not determine the global structure of a protein. The performance over 7 Å shows that such contacts explain most of the important interactions within a protein.

We can propose several reasons why some of the globin sequences could not identify their native structures by the top score. One of the important defects of the present implementation is the treatment of gaps in the sequence vs. structure alignment. Another important defect is the treatment of oligomers.

However, it is noteworthy that this simple non-optimized method showed considerable discrimination. We think this type of database approach can be extended toward dealing with other heuristic problems related with biological data starting from genome sequence up to biological functions.

Acknowledgement

This work was performed as a part of Research and Development Project for Innovative Technology for the Earth supported by New Energy and Industrial Technology Development Organization.

References

[1] Wodak, S. J. and Rooman, M. J., *Curr. Biol.*, 3, 247-259 (1993)

[2] Bernstein, F. C. et al., *J. Mol. Biol.*, 112, 535-542 (1977)

[3] Sippl, M. J. and Weitckus, S., *Proteins*, 13, 258-271 (1992)

[4] Toma, K., *J. Mol. Graphics*, 11, 222-232 (1993)