Protein Sequence-Structure Alignment Using 3D-HMM

Masashi Fujita fujita@kuicr.kyoto-u.ac.jp
Hiroyuki Toh toh@kuicr.kyoto-u.ac.jp
Minoru Kanehisa kanehisa@kuicr.kyoto-u.ac.jp

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

Keywords: protein structure prediction, alignment accuracy, 3D-HMM

1 Introduction

Tertiary structure of proteins provides valuable information on their biochemical functions. But experimental structure determination is still labor-intensive, in spite of technological advances. Therefore, rapid and accurate protein structure prediction method is not only interesting from a theoretical point of view but also practically useful.

Currently, homology modeling, or comparative modeling, is the most reliable structure prediction method and widely used. It is applicable when structure of at least one homolog of the target protein is solved, and predicts 3D coordinates by aligning the target sequence and the template structure. It is noted that the accuracy of homology model depends largely on the accuracy of target-template alignment. But the alignment accuracy of conventional alignment algorithms declines sharply when the sequence identity between the target and the template proteins is lower than 45% [4].

Here we introduce a novel sequence-structure alignment method to improve the alignment accuracy between distantly related proteins. It is based on hidden Markov model (HMM), which have been successfully applied to many bioinformatics problems. The distinguishing feature of our method is the explicit consideration of 3D coordinates of each amino acid residue. Each state of the HMM has a coordinate in 3D space. Alexandrov and Gerstein first introduced such idea in the field of protein structure classification, and abbreviated it as 3D-HMM [1].

In this study, we adapted 3D-HMM for the protein sequence-structure alignment problem and tested its ability in alignment accuracy. These spatial coordinates allow us to introduce structure-based transition probabilities such as restriction on the distance between two consecutive Cα atoms. Moreover, an alignment path generated by 3D-HMM has explicit 3D coordinates. Thus, energy of the model can be directly calculated without additional model building procedure. This can greatly enhance the speed of model quality evaluation when we have a large number of alternative models.

2 Method and Results

2.1 Model Architecture

Hidden states of our HMM consist of Cα atoms of template structure and lattice points around them. The purpose of lattice points is to allow insertion only at the solvent exposed region of the template.

Emission probabilities of Cα state were calculated based on the position-specific scoring matrix generated by PSI-BLAST [2]. Emission probabilities of lattice states were set to the average amino acid frequency at the protein surface region. Transition probabilities between two hidden states were defined by spatial distance between two states and disturbance in the template structure.

2.2 Parameter Training and Performance Evaluation

Data set for parameter training and performance evaluation was obtained from structural alignment database HOMSTRAD [5]. 95 distantly homologous protein pairs whose pairwise sequence identity was below 25% were selected and divided into 60 and 35 pairs. The former was used for parameter training and the latter was for performance evaluation.

To define alignment accuracy, the structural alignment of HOMSTRAD was used as the reference alignment.

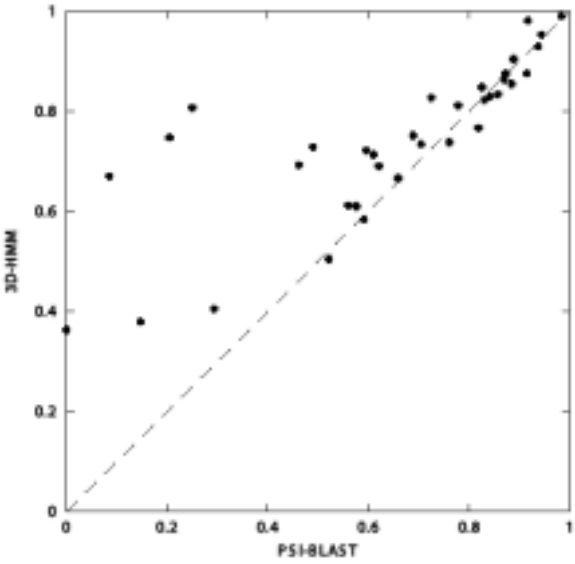

Figure 1: Alignment accuracy

2.3 Energy-based Filtering

In addition to the optimal alignment, we generated 1000 suboptimal alignments and compared accuracy of 100 alignments that had lower energy with the optimal alignment. The suboptimal alignments were generated by stochastic sampling traceback algorithm [3]. A statistical energy function whose interaction center was Cα atom was used as the energy.

2.4 Results

Figure 1 shows comparison of alignment accuracy between our method and PSI-BLAST. When the accuracy of PSI-BLAST was low (i.e. <80%), our method significantly improved alignment accuracy. Furthermore, in most cases, 100 suboptimal alignments that had lower energy contained more accurate alignments compared with the optimal one. The average improvement through suboptimal alignment on the optimal one was 10%.

3 Discussions

From the point of view of 3D-HMM, conventional profile methods such as PSI-BLAST can be considered as an attempt to improve the emission probabilities. Although 1D-HMM can take account of the position-specific indel propensity, less attention has been paid to improve transition probabilities.

In this study, we showed 3D-HMM could align distantly homologous protein pairs quite accurately. Our method is still crude and further development is inevitable, however, we believe this method is potentially useful in not only structure prediction but also various analyses of protein sequence-structure relationship.

References

[1] Alexandrov, V. and Gerstein, M., Using 3D Hidden Markov Models that explicitly represent spatial coordinates to model and compare protein structures, *BMC Bioinformatics*, 5:2, 2004.

[2] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* 25:3389-3402, 1997.

[3] Durbin, R., Eddy, S., Krogh, A. and Mitchison, G., *Biological sequence analysis*, Cambridge University Press, 1998

[4] Jaroszewski, L., Li, W. and Godzik, A., In search for more accurate alignments in the twilight zone, *Protein Sci.,* 11:1702-1713, 2002.

[5] Mizuguchi K, Deane C.M., Blundell T.L. and Overington J.P., HOMSTRAD: a database of protein structure alignments for homologous families, *Protein Science* 7:2469-2471, 1998.