Finding Conserved Regions in Protein Structures Using Support Vector Machines and Structure Alignment

Tatsuya Akutsu[1,a)] Morihiro Hayashida[1,b)] Takeyuki Tamura[1,c)]

Abstract: In this technical report, we propose a novel method for finding conserved regions in three-dimensional protein structures, which combines support vector machines (SVMs), feature selection and protein structure alignment. For that purpose, a new feature vector is developed based on structure alignment for fragments of protein backbone structures. The results of preliminary computational experiments suggest that the proposed method is useful to find common structural fragments in similar proteins.

1. Introduction

Classification of protein structures and identification of common protein structural patterns are major topics in structural bioinformatics. Multiple alignment of protein structures is a powerful approach to identification of common structural patterns. However, existing methods have some problems in computation time and/or accuracy. Therefore, other approaches should also be studied. Indeed, several methods based on *support vector machines* (SVMs) and *kernel methods* [4] have been proposed. Dobson and Doig developed a feature vector based on various information on proteins [5], which includes secondary-structure content, amino acid propensities, surface properties and ligands. Their feature vector was applied to classification of proteins into enzymes and non-enzymes. Borgwardt *et al.* developed kernel functions based on graph kernels [3], where each protein structure is represented as a graph using secondary structure information. In order to improve the prediction accuracy, they also used additional features similar to those used by Dobson and Doig. Qiu *et al.* proposed a kernel for protein structures using a structure alignment algorithm [8]. Though these methods are very useful for predictions, it is difficult to extract structural information or common regions of proteins from the results of SVM learning. Therefore, it is desirable to develop a method with which structural information and/or common regions can be extracted.

In this article, we propose a simple feature vector for finding common regions of protein structures. The proposed feature vector is based on the concept of *spectrum kernel* for sequence data [7], which is based on the numbers of occurrences of substrings of fixed length. Instead of substrings, our proposed feature vector uses a set of template fragments of protein backbone structures. And then, occurrences of similar fragments are taken into account in the feature vector. Different from the spectrum kernel, we use longer fragments each of which consists of several tens of Cα atoms. Moreover, similarities between fragments are measured by means of structural alignment [1] because gaps cannot be ignored for such long fragments.

In this short article, we describe the method only. The results of computational experiments can be found in [2],

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611–0011, Japan
[a)] takutsu@kuicr.kyoto-u.ac.jp
[b)] morihiro@kuicr.kyoto-u.ac.jp
[c)] tamura@kuicr.kyoto-u.ac.jp

2. Method

In the proposed method, each protein structure P in training and test data sets is transformed into a feature vector $\Phi(P)$ and then SVM learning and classification are performed in a usual manner. Furthermore, feature selection is performed in order to extract common structural fragments. In the following, we describe outlines of computation of a feature vector and feature selection.

2.1 Feature Vector

Each protein structure is represented by a sequence of positions of Cα atoms. Let $P = (p_1, p_2, \ldots, p_n)$ be a sequence of positions of Cα atoms. In the proposed method, a feature vector $\Phi(P)$ for protein structure P is defined as follows (see also **Fig. 1**).

Let L be the length of a structural fragment, where a fragment is a consecutive sequence of positions of Cα atoms, and $L = 40$ was employed in this work based on several trails. Let $\mathcal{T}$ be a set of template structures. Let $Q = (q_1, \ldots, q_m)$ be a template structure in $\mathcal{T}$. A set of fragments $frag(Q)$ from Q is defined by

$$frag(Q) = \{ (q_{i\Delta+1}, q_{i\Delta+2}, \ldots, q_{i\Delta+L}) \mid i = 0, 1, 2 \cdots \text{ and } i\Delta + L \leq m \},$$

where $\Delta = 10$ was used in this work. Then, a set of template fragments $\mathcal{F}$ is defined as

$$\mathcal{F} = \bigcup_{Q \in \mathcal{T}} frag(Q).$$

That is, a set of template fragments contains several fragments

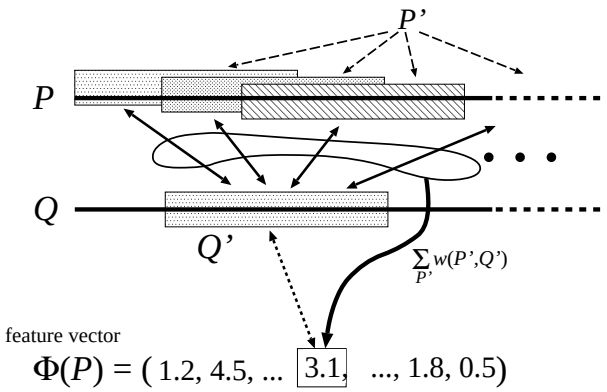

Fig. 1 Computation of a feature vector. Each coordinate in a feature vector corresponds to template fragment Q', where the coordinate value is defined by the sum of the scores for fragments in P against Q'.

from each template structure, where template structures are selected from positive and negative classes (but not included in training or test data set).

For a structural fragment P' from a training or test protein structure P and a template fragment Q', we define the score $w(P', Q')$ by

$$w(P', Q') = \frac{\text{the number of superposed residue pairs}}{|P|},$$

where $|P|$ denotes the number of residues in P. We used this measure to evaluate the result of structural alignment between P' and Q' because we employed STRALIGN [1], which tries to maximize the number of superposed residue pairs within some distance threshold. Then, the feature vector $\Phi(P)$ for a training or test protein structure P is defined by

$$\Phi(P) = \left(\sum_{P' \in frag(P)} w(P', Q') \right)_{Q' \in \mathcal{F}}.$$

That is, each coordinate value corresponding to a template fragment $Q' \in \mathcal{F}$ is defined by the sum of the scores for fragments of P against Q'.

2.2 Feature Selection

In order to find conserved structural fragments, we employ Recursive Feature Elimination (RFE) [6], which is a well-known feature selection method for SVMs. Different from the original RFE [6], we use the prediction accuracy (for the training data set) as a measure for eliminating features. Moreover, pre-processing based on Pearson correlation coefficient is introduced so as to eliminate redundant features efficiently. The following is an outline of our feature selection method, where $H = 30$ and $K = 3$ were used in this work.

STEP 1: Let $\mathcal{F}_0$ be a set of all template fragments.

STEP 2: Compute Pearson correlation coefficient between each $f \in \mathcal{F}$ and the class (i.e., positive or negative).

STEP 3: Let $\mathcal{F}$ be the subset of $\mathcal{F}_0$ consisting of fragments with H highest coefficients ($H = 30$ in this work).

STEP 4: For all $Q' \in \mathcal{F}$, perform SVM training using $\mathcal{F} - \{Q'\}$.

STEP 5: Let Q'' be the feature such that the classification accuracy for $\mathcal{F} - \{Q''\}$ is the highest.

STEP 6: Let $\mathcal{F} \leftarrow \mathcal{F} - \{Q''\}$.

STEP 7: Repeat STEPS 4-6 until reaching the specified number of features K.

3. Concluding Remarks

We proposed a method for finding conserved regions in similar proteins. The method is a combination of a new feature vector based on structure alignment for fragments with two techniques in statistical learning: support vector machines and feature selection. It should be noted that, different from a common approach to identify conserved regions, the proposed method does not use multiple structure alignment though it uses pairwise structure alignment for fragments. The results of preliminary computational experiments suggest that the proposed method is useful to identify important structural fragments.

One of important future work is to perform rigorous and larger scale computational experiments, which include (i) adjustment of parameters (e.g., L, Δ, K and H) used in the method, (ii) study of the sensitivity of these parameters, (iii) comparison with other kernels for protein structures (e.g., [3], [5], [8]), and (iv) examination of other feature selection methods. It is also important to study biological meaning and/or significance of the selected fragments.

In the proposed method, configurations between fragments are not taken into account. However, configurations between fragments may play an important role in protein functions. In particular, such information seems important if we would like to predict interactions between proteins and/or interactions between proteins and chemical compounds. Therefore, a feature vector and/or a kernel function reflecting such information should also be developed.

References

[1] Akutsu, T.: Protein Structure Alignment Using Dynamic Programming and Iterative Improvement, *IEICE Trans. Inf. Syst.*, Vol. E79-D, pp. 1629–1636 (1996).

[2] Akutsu, T., Hayashida, M., and Tamura, T.: Finding Conserved Regions in Protein Structures Using Support Vector Machines and Structure Alignment, *Proc. 7th IAPR Int. Conf. Pattern Recognition in Bioinformatics*, to appear.

[3] Borgwardt, K. M., Ong, C. S., Schönauer, S., Vishwanathan, S. V. N., Smola, A. J., and Kriegel, H-P.: Protein Function Prediction via Graph Kernels, *Bioinformatics*, Vol. 21, pp. i47–i56 (2005).

[4] Cortes, C. and Vapnik, V.: Support-Vector Networks, *Machine Learning*, Vol. 20, pp. 273–297 (1995).

[5] Dobson, P.D. and Doig, A.J.: Distinguishing Enzyme Structures from Non-enzymes without Alignment, *J. Mol. Biol.*, Vol. 330, pp. 771–783 (2003).

[6] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V.: Gene Selection for Cancer Classification Using Support Vector Machines, *Machine Learning*, Vol. 46, pp. 389–422 (2002).

[7] Leslie, C., Eskin, E., and Noble, W.S.: The Spectrum Kernel: A String Kernel for SVM Protein Classification, *Proc. Pacific Symp. Biocomputing*, Vol. 7, pp. 564–575 (2002).

[8] Qiu, J., Ben-Hur, A., Vert, J-P., and Noble, W. S.: A Structural Alignment Kernel for Protein Structures, *Bioinformatics*, Vol. 23, pp. 1090–1098 (2007).