Detection of Remote Homologue Using Predicted Structural Information

Daisuke Ishibe[1]
daisuk-i@is.aist-nara.ac.jp
Katsunori Uehara[1]
k-uehara@is.aist-nara.ac.jp
Takeshi Kawabata[1]
takawaba@is.aist-nara.ac.jp
Nobuhiro Go[1,2]
ngo@is.aist-nara.ac.jp

[1] Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, Nara 630-0101, Japan
[2] CCSE, Japan Atomic Energy Research Institute, 8-1 Umemidai, Kizu-cho, Souraku, Kyoto 619-0215, Japan

Keywords: remote homologue detection, PSI-BLAST, secondary structure prediction, support vector machine

1 Introduction

Detection of homologues for a protein sequence is important for deducing its function and tertiary structure, many researchers have made efforts to develop sensitive method to detect homologues. Recently, PSI-BLAST [1] becomes the standard tool for finding remote homologue, however, many homologous relationships still exist which PSI-BLAST cannot detect. To improve the performance of PSI-BLAST, we develop a filter choosing true homologues among homologue candidates using predicted structural information. The candidates are picked up using PSI-BLAST with the high threshold E-value. The filter used many other features, such as secondary structure prediction, solvent accessibility prediction. Final decision is made by simple centroid classification or support vector machine.

2 Data and Methods

In order to evaluate our method, we use the 3999 representative protein domains in SCOP databases with sequence identity 30 % or less. The relationship "family" and "superfamily" defined in SCOP are considered as the correct homologues.

PSI-BLAST outputs 40,475 protein pairs whose E-values are less than 10. This threshold E-value is rather high, they include both homologous and analogous relationships. According to the SCOP database, 13,000 pairs are really homologous, the rest 27,475 pairs are not homologous. We tried to make a filter separating these 13,000 homologous pairs from the others.

To design the filter program, we tried several kinds of input features: E-value of PSI-BLAST, structure matching score of predicted secondary structure and solvent accessibility. Secondary structure prediction(SSP) was done using Psipred [2], and solvent accessibility prediction(SAP) was predicted using the in-house neural network program.

The outputs of the solvent accessibility is two states:exposed or buried. In order to evaluate statistical significance of matches of secondary structure and solvent accessibilities, we introduce the Zscore assuming random matches can be approximated by binomial distribution.

```
            HHHccccccccccceeeeccccHHHHHHccCCHHhh--------HHHHHHHHHHHHcCCCCC
Query: 121 AIHDVDHPGVSNQFLINTNSELALMYNDESVLEN--------HHLAVGFKLLQEEHCDI 171
            IHD+ + G+S+ F       L ++YN+     N        +  A+GF+LL++   +I
Sbjct: 46  LIHDLKYLGLSDFFQDEIKEILGVIYNEHKCFHNNEVEKMDLYFTALGFRLLRQHGFNI 104
            HHHhhhhhCChhhhhhhhhhhHHHHHhhCCHHccccccccchHHHHHHHHHHHhCCCCC
```

Figure 1: Secondary structure predictions are added to the PSI-BLAST alignment.Matched structures are shown in **bold**.

$$Zscore = \frac{M - Np}{\sqrt{Np(1-p)}} \tag{1}$$

where M is the number of residues with the same structure, N is the number of compared residues in PSI-BLAST alignments, and p is the probability that query residues randomly has the same structure as subject residues.

We used two kinds of machine learning algorithms : simple centroid classification and support vector machine(SVM). The simple centroid classification used the value of inner product between the feature vector $\boldsymbol{x}$ and the projection vector $\boldsymbol{w}$, as the final decision score. The vector $\boldsymbol{w}$ is defined as the difference between the means of two classes $\boldsymbol{w} = \frac{1}{N_+}\sum_{\boldsymbol{x}\in+}\boldsymbol{x} - \frac{1}{N_-}\sum_{\boldsymbol{x}\in-}\boldsymbol{x}$. Support vector machine is the powerful supervised machine learning algorithm for two class discrimination, which is proposed by Vapnik [3]. We employed the standard radial basis function (RBF) as the kernel.

The performance was evaluated by the five-fold cross validation method. The dataset are divided five groups, the learning was done using one of the group, and performance was tested using the rest of data.

3 Results and Discussion

Table 1 summarizes the performance of our methods. From Table 1, Combination of SSP and E-value is better than the original PSI-BLAST. Therefore, Secondary Structure Prediction is effective features for recognizing protein remote homologue. SAP is not effective on this case. The performance of SVM and centroid is not so much different.

For future work, we plan to introduce other features, such as other predicted structure or ProSite motif.

Table 1: Performance of homologue detection.

		Accuracy	F-measure
PSI-BLAST		0.8776	0.7873
PSI-BLAST SSP	SVM	0.8953	0.8298
	Centroid	0.8974	0.8323
PSI-BLAST SSP SAP	SVM	0.8956	0.8298
	Centroid	0.8979	0.8330

SSP means zscore from Secondery Structure Prediction
SAP means zscore from Solvent Accessibility Prediction

References

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

[2] Jones, D.T., Protein secondary structure prediction based on position-specific scoring matrices, *J. Mol. Biol.*, 292:195–202, 1999.

[3] Vapnik, V.N., *Statistical Learning Theory*, A Wiley-Interscience Publication, 1998.

[4] `http://cl.aist-nara.ac.jp/~taku-ku/software/TinySVM/`