國立交通大學

生物資訊研究所

碩士論文

藉由序列側寫比對偵測遙遠的蛋白質同源關係

Detection of remote homologues by sequence profile-profile comparison

研 究 生：劉力彰

指導教授：黃鎮剛　教授

中華民國九十四年一月

藉由序列側寫比對偵測遙遠的蛋白質同源關係

學生：劉力彰　　　　　　　　　　　　　　　　　　　　指導教授：黃鎮剛

國立交通大學生物資訊研究所碩士班

中文摘要

我們的研究提供了一個通用的側寫比對程式。我們所提出的方法稱為PROF^2，這個方法採用兩種不同的側寫(profile)資訊: 序列側寫資訊以及結構側寫資訊，資訊的相互結合而增加側寫比對整體的成功率，同時，我們的方法相當的具有彈性，不只可以單純利用序列的側寫檔案，我們更可以使用任何形式類似的側寫檔案，並且准許使用者自訂各種必需要的參數，以用來偵測蛋白質遙遠的同源關係。我們將加入二級結構資訊的程式用另一個名字命名加以區分，PROF$^{2'}$ 。

我們的方法在所有的實驗數據上，都大幅的領先之前一些公認較佳的程式，同時，我們也發現加入二級結構的資訊可以有效的增加成功率，尤其是在錯誤較少的情況下。

Detection of remote homologues by sequence profile-profile comparison

Student : Li-Chang Liu   Advisor : Jenn-Kang Hwang

Institute of Bioinformatics

National Chiao Tung University

Abstract

This work presents a method for profile-profile comparison. The proposed method uses two kinds of profiles, sequence profile and secondary structure profile, to increase the accuracy of profile comparison. Our method is $PROF^2$. Besides the sequence profile-profile comparison, we make it can input any profile style to improve the flex of program. The proposed profile-profile comparison tool allows for gaps, global alignment and local alignment to detect weak similarities between protein families. In comparison of two sequence profiles, the parameters of the local alignment have been optimized to produce alignments that are greater than original one. On the other way, the combination program $PROF^{2'}$ which represents the combination of sequence and secondary structure information has much accuracy then the output of $PROF^2$ which just include sequence information when we implement a practical tool.

The proposed results show that this tool detects more similarities between protein fold of distant homology than the previous methods. At the same time, we found that including the secondary structure information could increase the accuracy before much false positive occurred. It could be useful for creating general tools.

誌謝

在人生所有的旅程中，相信有許多難忘而深刻的回憶，對我來說，在交大研究所的日子，將是我這輩子永難忘懷的回憶之一，這些日子以來，體驗了研究的辛苦及喜悅，也學習到許多難得寶貴的經驗。

首先要感謝的，是指導教授黃鎭剛老師，老師提供了一個自主自發獨立而自由的研究環境，而當我的研究碰到瓶頸的時候，老師總是給予鼓勵與啓發，從老師身上，不僅學習到科學研究上的知識，更可以得到研究及生活的準則與態度。

此外，還要感謝實驗室所有的伙伴：詹鎭熊學長、尤禎祥學長、林勇欣學長、游景盛學長以及梁涵堃學長，在專業領域及日常生活上的照顧，還有志杰以及少偉所給予技術上的支援。有了你們的協助，這篇論文才得以順利完成，同時，也感謝實驗室裡所有的學長姊、學弟妹，生活中有了你們的陪伴，才讓生活更加的充實，謝謝你們。

最後要感謝我的父母和女友，在這碩士生的生涯中，不斷給予我鼓勵及支持，使我能順利完成學業。僅以此論文作爲我碩士生涯的結語，並將此獻給所有關心我及我所誠摯感謝的人。

目錄

Introduction

The automated methods for determining the relationship of two proteins become increasingly important because of the rapid growth of the number of known protein sequences. In the traditionally, the method that the unknown sequence is compared with each known sequence in the database, one at a time, is called a pair-wise alignment. In the previous literatures, some researches indicating two similar sequences may imply a common evolutionary ancestry between these two corresponding proteins. Homologous sequences usually have the same fold and the close or related biological structure or function[1; 2; 3; 4].

The pairwise sequence alignment is useful to detect the similarity of two proteins. In the coarse evolution, protein sequences may have mutations or insertions. Moreover, in many cases, proteins may still have high sequence similarity but not share the similar structures. According to previous literatures, if the identity of these two sequences beyond the 25% ~ 30%, called the twilight zone, sequence-sequence comparison is hard to detect the relationship[5; 6]. A great deal of work has been done to develop tools that can detect such fields. The methods involving profile-sequence comparisons include several widely accepted searching protocols. PSI-BLAST[7] and IMPALA[8] use the same profile schema and scoring system. PSI-BLAST constructs a profile using an iterative method from the hits obtained after the previous iteration step. Then PSI-BLAST inputs such profile and searches the database. Nonetheless, PSI-BLAST may miss weak sequence similarities and include the unrelated sequences into the profiles. In such case, the iterative against method will contain more unrelated sequences and produce a worse profile. IMPALA is a sequence-profile comparison tool based on PSI-BLAST. It is designed to search a database of profiles with a given sequence. SAM-T99[9] is another successful approach for the profile-sequence comparison. It uses the alignment to predict the secondary structure and to build a hidden Markov models (HMM)[10; 11] that is then used to search the PDB for similar proteins. As a further step in the use of the alignment information, several methods have been developed for the comparison of multiple alignments to multiple alignments. The COMPASS[12] (comparison of multiple protein alignments with assessment of statistical significance) method involves the construction of local profile-profile alignments allowing gaps by means of a dynamic programming algorithm.

Our method bases on the Jeasen-Shannon divergence between probability dirstributions[13] which is used in the other program prof_sim[14]. In the proposed method, we use the different parameters with prof_sim for local alignment and modify the alignment score to get higher accuracy in the begging.

Methodology

Definition of sequence profile

The sequence profiles are created by the PSI-BLAST. Generally, a profile is a representation of a group of related protein sequence which is usually based on multiple sequence alignment. First, PSI-BLAST searches the database and finds out some sequences, then creates general architecture of the score matrix. Second, it will use pairwise alignment to construct a multiple sequence alignment result. Third, PSI-BLAST gives weights for sequences within the multiple sequence alignment and evaluates the effective number of independent observations which the multiple sequence alignment constitutes. Fourth, PSI-BLAST also can estimate target frequencies and construct the matrix score. Fifth, PSI-BLAST uses the profile into next iteration. After a lot of iterating searches, we get the profile in which reflects the likelihood of observing any amino acid *k* at position *i*.

The profile is defined as series of probability distributions $P = p_1 p_2 p_3 \cdots p_n$ when n is the length of the sequence and p_i is a probability distribution over the 20 amino acids at position i. We can think a profile is a $20 \times n$ matrix.

Definition of secondary structure profile

In the proposed method, we combined two different profiles to make the accuracy increase. One is the sequence profile created by PSI-BLAST and the other is secondary structure profile. The secondary profile is based on PSI-BLAST position specific scoring matrix. We use the database search to get some position specific scoring matrices in which are created by each sequence appear in the form of a $20 \times M$ position-specific scoring matrix from five iterations of a PSI-BLAST search, where M is the length of the target sequence. We use the PSIPRED[15] procedure to predict the secondary structures. The scoring matrix for a window of 15 positions, centered on the target residue, is used as the input to the SVM[16] (Support vector machine). Then the calculated probabilities of each secondary structure will be transferred from $\pi/2$ to $-\pi/2$. The equation is as below:

$$p = \frac{[\arctan(vote \times 2) + \frac{\pi}{2}]}{\pi} \qquad (1)$$

where the p is probability for each secondary structure and the *vote* is the internal output given from SVM. The $8 \times M$ profile is the secondary structure profile.

The data set

The SCOP 1.50 classification of protein structures as a test set to calculated some parameters for $PROF^2$ and $PROF^{2'}$ programs, the $PROF^2$ is profile-profile comparison tool just using sequence information and the $PROF^{2'}$ means the

$PROF^2$ using combinational information which combine the sequence information and secondary structure information; this manually created database contains 24186 protein domains classified into 1296 protein families, 820 superfamilies, 548 folds and seven classes. We select some seed family sequences to run PSI-BLAST against their own family. Some families for which there is only one member or for which PSI-BLAST failed to generate a profile were represented by a profile generated directly from the seed sequence using the original BLOSUM62 frequency matrix. The flow charts are listed on the Figure 1 and Figure 2.

We reduce some seed profiles which are all profiles within families that contain only one sequence. Finally, we get 1075 seed sequence profiles and select subset of 563 families. Those are all families within superfamilies that contain at least two other families. Using the same conditions, there are 2155 protein profiles chosen from SCOP 1.63 as our data set.

Profile-profile comparison tool

The proposed profile-profile comparison is applied using dynamic program which is in the same way sequence-sequence alignment. The difference of sequence-sequence alignment is that the sequence-sequence comparison uses BLOSUM62 to give the score for different pairs of aligned amino acids, but the profile-profile comparison uses the similarity score. The similarity score will be introduced in the following.

The divergence score

We define two profiles $P = p_1 p_2 p_3 ... p_n$ and $Q = q_1 q_2 q_3 ... q_m$ where n and m are the lengths of the profiles and p_i, q_j are probability distributions over the 20 letter alphabets of amino acid. We will define the similarity score based on this statistical feature.

A common method used to measure the statistical similarity between two probability distribution $p_i(x)$ and $q_i(x)$ is the Kullback-Leibler (KL) divergence[17] :

$$D^{KL}\left[p_i \parallel q_j\right] = \sum_{k=1}^{20} p_{ik} \log_2 \frac{p_{ik}}{q_{jk}} \qquad (2)$$

This measure has disadvantage which is asymmetric and unbounded. We can find a better method from literature issued by Jensen-Shannon (JS) divergence:

$$D^{JS}\left[p_i \parallel q_j\right] = \lambda D^{KL}\left[p_i \parallel r\right] + (1-\lambda) D^{KL}\left[q_j \parallel r\right] \qquad (3)$$

in where

$$r = \lambda p_i + (1-\lambda) q_j \qquad (4)$$

where the λ is a prior weight and it is suggested to be $1/2$.

The significant score

After a calculated statistical score, we would consider if it is significant enough. Image that we have two random sequences and let PSI-BLAST create their profiles each resemble the overall distribution of amino acid in the database. In such case, are those profiles significantly similar? Obviously not, they may be similar by chance. For this problem, we use significant score to judge:

$$S = D^{JS}\left[r \parallel P_0\right] \qquad (5)$$

in where P_0 is background probability which defined as the overall amino acid distribution in a large database such as SWISSPROT. r is the source distribution from two comparing profiles. We can use these two score to create out dynamic program scoring function:

$$Score(p_i,q_j)=\frac{1}{2}(1-D)(1+S)$$

$$=\frac{1}{2}\left(1-D^{JS}\left[p_i \,\|\, q_j\right]\right)\left(1+D^{JS}\left[r \,\|\, P_0\right]\right) \qquad (6)$$

This $Score(p_i,q_j)$ is called “column score”. According to the equation (3), the measure is symmetric and ranges between 0 and 1 where the divergence for identical distributions is 0. There are four situations in the column score. (i) (D->0)(S->0) pair means this scoring scheme distinguishes two distributions that each one is similar to the background distribution. (ii) (D->0)(S->1) pair means that those profiles are very similar and they are far from the background distribution. (iii) (D->1)(S->0) pair means that those profiles are dissimilar and the common source of those profiles is similar to the background distribution. (iv) (D->1)(S->1) pair means that those profiles are dissimilar and they are similar to the background distribution. No matter which situation, those scores locate on the 0 to 1.

Optimization of parameters

The primary requirement is that the $PROF^2$ program can input not only PSI-BLAST profiles but also other protein profiles created by many different protein features from our laboratory when we implemented this profile-profile comparison tool. By referring to many previous famous alignment programs, we know that a few parameters are very important and they also affect the alignment results a lot. In order to get these parameters, we adapt SCOP 1.50 as our test set. The profile similarity scores range from 0 to 1 are defined in the previous section. For local alignments, the similarity function, *Score(a ,b)*, must satisfies two requirements: one is the mean value of *Score(a ,b)* which must be negative (Otherwise, the extension of a random match would tend to increase its score, which is contradicting the idea of local similarity.). The other is the maximum value of *Score(a ,b)* must be positive (That means that the possibility of the match with a positive score.). These criteria are satisfied by all standard scoring matrices, such as BLOSUM and PAM matrices. Our profile similarity scores must be adjusted to meet these requirements. Therefore, we divide our program to two parts for our substitution matrix to satisfy those requirements. The first part is to use the shift value to transfer all scores to fit the rules. The second part is the find-tuning.

First, the shift value must be determined because the substitution matrix scores are reasonable if those must satisfy those requirements. In order to satisfy such rules, we choose the top 100 families of the SCOP database. Each family has a seed sequence described above. We state the amino acid at position i of the seed sequence as the seed amino acid of the $i-th$ profile column. Two seed amino acids are defined as similar, neutral, or dissimilar based on their BLOSUM62 scoring matrix, with positive, zero and negative substitution scores. There are four classifications of column pairs: (1) a column with itself (we called the identical columns), (2) different columns that are associated with the similar seed amino acids (similar columns), (3) different columns with mutually neutral seed amino acids (neutral columns), (4) different columns with dissimilar seed amino acids (dissimilar columns). The Figure 3 shows the smallest value of distribution of neutral columns locates in 0.42, so the shift should be higher than 0.42. There is no threshold which can clearly distinguish two distributions form the gray area. We hope the inferences of dissimilar columns are the smallest, thus the point 0.5 will be the upper bound of shift values. For the combinational profiles of primary information and secondary structure information, we use the same procedure to determine the range of shift value.

Second, the adjusting method which makes the results better is different from other alignment programs. When we use the adjusting linear equation, we find the different values of the substitution matrix. Because of the different values of the substitution matrix, the values of gaps, gap extensions, and shift values comparatively become better. The adjusting linear equation is described as below:

$$S(x) = ax + b \tag{7}$$

where *a* and *b* are constants chosen from the results of SCOP 1.50 and *S(x)* is one of the value of substitution matrix.

According to previous requirements, we can imply that:

$$\overline{X} < \frac{-b}{a} \tag{8}$$

$$X_i > \frac{-b}{a} \tag{9}$$

The $\overline{X}$ is the mean of those values calculated from equation (6) after shifting and *a* and *b* are the same constant with equation (7). Such equation also indicates that there are infinite combinations for constants for *a* and *b*. After shifting, the mean is -0.0055, average maximum is 0.081, and average minimum is -0.0582 gotten from the test set, SCOP 1.50. Hence, the range of the ratio is -0.081 < $\frac{b}{a}$ < 0.0055. In this case, one of the parameters could be assigned arbitrarily by us. We make the value of a from 22 to 24, because the results of ROC curve, is introduced on the below section, are better. Then the value of b will locate between -1.782 and 0.132. In the other way, such method is just a fine-tuning for the gap, gap-extension and shift values. We use the best result in the true positive numbers in the 200th false positive as our final parameters. According to the average value of new substitution, the parameter, *a*, will not influence much for result, therefore, we adjusted the parameter, *a*. Actually, the parameter “a” fine-tunes to gap and gap extension and the parameter “b” does for the shift value. Fortunately, such method could discover parameters of gap, gap extension, and shift values better than the parameters which Yona suggested. The Figure 4 exhibits the influence of different parameters. Doing little adjustment in the parameter “b” will cause the huge difference in the whole ROC curve. Finally, we choose the triangle line in our chart and the pair of the parameters is 23 and -0.3.

Integrating secondary structure with primary structure

The Jeasen-Shannon divergence can detect the similarity of two different distributions, so we will treat the primary and secondary information separately. The new profiles extended from the original sequence profiles. We augment the profile columns of sequence information to make a probability distribution over 28 values (the 20 amino acids plus 8 secondary structures defined in DSSP.

We give the weight θ to secondary structure information, and $1-\theta$ for sequence information where θ form 0 to 1:

$$SCORE(p_i, q_j) = (1-\theta) \times SCORE_{1D}(p_i, q_j) + \theta \times SCORE_{2D}(p_i, q_j) \qquad (10)$$

where the $SCORE_{1D}(p_i, q_j)$ is the similarity score of primary profile for the row, p_i, in the first profile and the row, q_j, in the secondary profile. The Figure 7 shows the different true positive in the 50th false positive. It implies the best θ is 0.03.

Statistical significance

For the sequence alignment, the distribution of alignment scores should be like the extreme value distribution because the true positive pairs are minorities in whole pairs. In our method, we estimated the E-values through the distributions of scores obtained during the local alignment and used them as criteria to determine the pairs which are true positive or not. Comprehensively speaking, the pairwise profile comparison is applied to use parameters which we defined previously. One profile compares with others except itself and such procedure will produce a lot of scores recorded on a file, called a score list. This score list indicates where the position of the profile will be, and then the evd-fit program which was written by Chen-hsiung in our laboratory referred to the fit function in gnuplot (www.gnuplot.info) using the nonlinear least-squares (NLLS) Marquardt-Levenberg algorithm. The Figure 5 shows the differences of the ROC curves whether the evd-fit program can be run. The evd-fit procedure is shown on the Figure 6. We collect all scores which compared with other profiles to be an input for evd-fit program, then get the λ and k for the E-value.

Results and Discussion

ROC analysis

The quality measure is the receiver operating characteristic (ROC)[18] which is evaluated by means of a plot of the true positive fractions versus the false positive fractions using a continuously various decision threshold. In order to plot such curve for a specific method, we first sort the results by their E-values calculated by using the evd-fit program and count the numbers of true positive until the 50th false positive occur. The result of ROC_{50} using the SCOP 1.50 as the database is represented on the Figure 8. We can observe when the first false positive occurs, the $PROF^{2'}$ has the best result. Actually, it is more practical when we implement a general tool for remote homologues detection. The other concern is the numbers of true positive in the high false positive. The Figure 9 exhibits the result of ROC_{1000}, and it presents the higher accuracy of $PROF^2$ and $PROF^{2'}$ when the 1000th false positive happened.

Obviously, the proposed method is the preferable choice than others, and this method indicates our parameter is better than other programs. Even though Yona and we sample in the same dataset, why can we get better results from $PROF^2$ than from prof-sim? We believe the fine-tuning for the gap penalty and shift value is the key point. From the Figure 4, it displays the different results in the different parameters and we could find that they will cause huge change. Besides, the evd-fit procedure is also noticeable. The prof-sim program always uses the same parameters, λ and k for the E-value calculation, but our program uses different parameters for different comparison pairs. Because our procedure could make better parameters for E-value to different dataset, such result of variance in the SCOP1.63 exhibited in Figure 10 and Figure 11 became more apparently.

Lindahl's benchmark

In order to test the performance of our program, we adopt the Lindahl's benchmark[19]. There are 976 protein sequences in this dataset. Also we use all against all pair-wise alignment and plot ROC curve and sens-spec plot for each SCOP levels.

Sens-spec plots analysis

Sens-spec plots[20; 21] describes how many of the possible true positive in whole pairwise alignment are detected at a given confidence level. It defines two values:

$$Sensitivity = \frac{TP}{TP + FN} \tag{11}$$

$$Specificity = \frac{TP}{TP + FP} \tag{12}$$

where TP being the number of correct hits having a score above threshold, FN being the number of correct hits with a score less than threshold and FP being the number of false hits that have a score above threshold.

The Figure 12, Figure 13 and Figure 14 exhibit different results for fold, superfamily and family in the Lindahl dataset with three different programs, *PSI-BLAST*, *IMPALA* and $PROF^2$. As we know, the architecture of SCOP database is class, fold, superfamily and family. According to the original paper of Lindahl benchmark, when we calculate the true positive of superfamily level of a pair, we ignore it which is the same family. We call that is superfamily only. Those three charts show that the $PROF^2$ has the best result than other programs. The other hand, we also compare our program with prof_sim. We can find that our programs are better than others and especially in the fold level (Figure 14), there are significant differences.

Conclusions

In the proposed study, it shows that our program can get better result than previous research and it means that we can find out a better template for an unknown sequence exactly. At the same time, we find that to add secondary structure information is indeed increasing the accuracy before the much false positive occurred.

Our program is a good profile-profile comparison tool, because it does not only input sequence profiles but also use another profile as the input. We can use our program to make a general website for detecting remote homologues and combine other information into the profile to increase its sensitivity.

References

1. Chothia C, L. A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823-826.
2. Sander, C., Schneider, R. (1991). Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins* **9**, 56-68.
3. Brenner, S. E., Chothia, C., Hubbard, T. J. P. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA* **95**, 6073-6078.
4. Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85-94.
5. Doolittle, R. F. (1981). Similar amino acid sequences: change or common ancestry? *Science* **214**, 149-159.
6. Vogt, G., Etzold, T. & Argos, P. (1995). An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. *J. Mol. Biol.* **249**, 816-831.
7. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.
8. Schaffer, A. A., Wolf, Y.I., Ponting, C. P., Koonin, E. V., Aravind, L. & Altschul, S. F. (1999). IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* **15**, 1000-1011.
9. Karplus, K., Barrett, C., Cline, M., Diekhans, M., Grate, L. & Hughey, R. (1999). Predicting protein structure using only sequence information. *Proteins: Struct. Funct. Genet.* **37**, 121-125.
10. Krogh, A., Brown, M., Mian, I. S., Sjolander, K. & Haussler, D. (1994). Hidden Markove models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235**, 1501-1531.
11. Durbin, R. E., Krogh, A., Mitchison, G. & Eddy, S. (1999). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK.
12. Ruslan Sadreyev, N. G. (2003). COMPASS: A Tool for Comparison of Multiple Protein Alignments with Assessment of Statistical Significance. *J. Mol. Biol.* **326**, 317-336.
13. Lin, J. (1991). Divergence measures based on the Shannon enropy. *IEEE Trans. Info. Theory* **37**, 145-151.
14. Golan Yona, M. L. (2002). Within the Twilight Zone: A Sensitive Profile-Profile Comparison Tool Based on Information Theory. *J. Mol. Biol.*

315, 1257-1275.

15. Jones, D. T. (1999). Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. *J. Mol. Biol.* **292**, 195-202.

16. Joachims, T. (1999). Making large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press*.

17. Kullback, S. (1959). *Information Theory and Statistics*, John Wiley and Sons, New York.

18. Michael Gribskov, N. L. R. (1996). Use of reciver operating characteristic (ROC) analysis to evaluate sequence matching. *Computers Chem.* **20**, 25-33.

19. Erik Lindahl, A. E. (1999). Identification of Related Proteins on Family,Superfamily and Fold Level. *J. Mol. Biol.* **295**, 613-625.

20. Hargbo, J. E., A. (1999). A study of hidden Markov models that use predicted secondary structures for fold recognition. *Proteins* **36**, 68-87.

21. Rice, D. E., D. (1997). A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J. Mol. Biol.* **267**, 1026-1038.

Figures

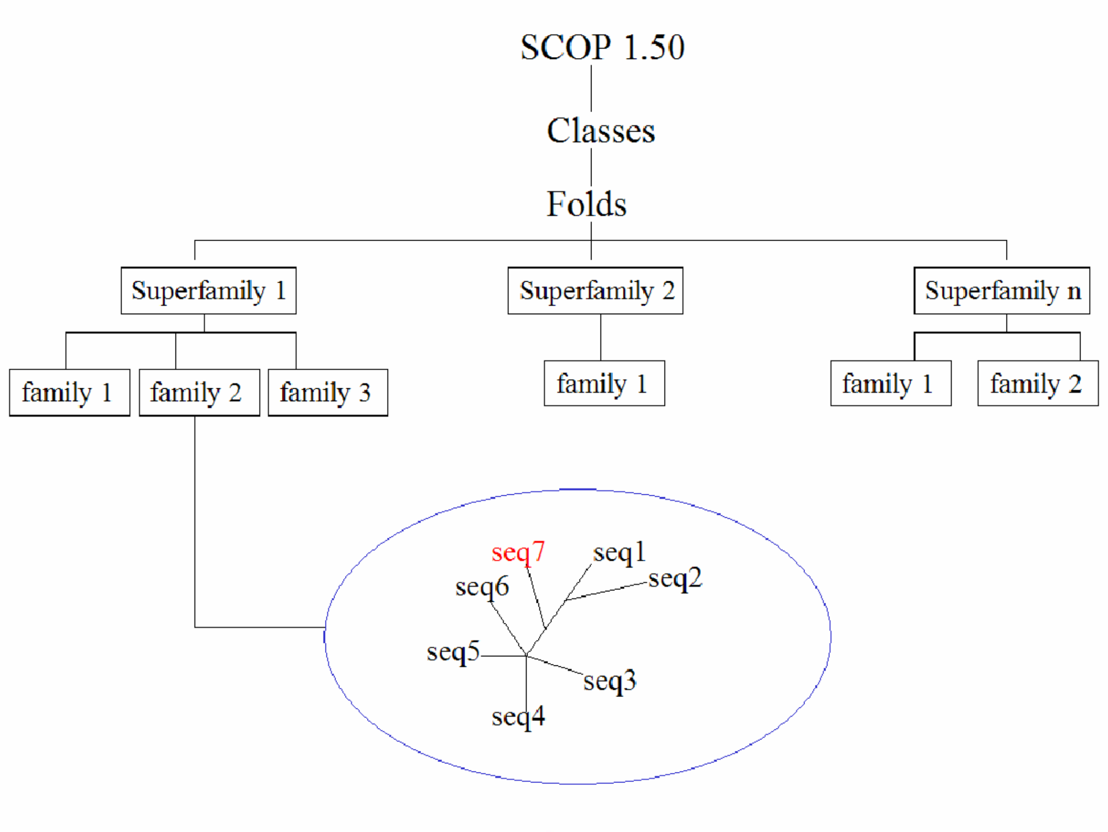


Figure 1

The SCOP hierarchy architecture is classes, folds, superfamilies and families. There are several sequences in each family. We chose the seed sequence which has the nearest distance with the other sequences in the one family. For example, there are seven sequences in the “family 2” which belonged to “superfamily 1”, and the “seq 7” is the seed sequence.

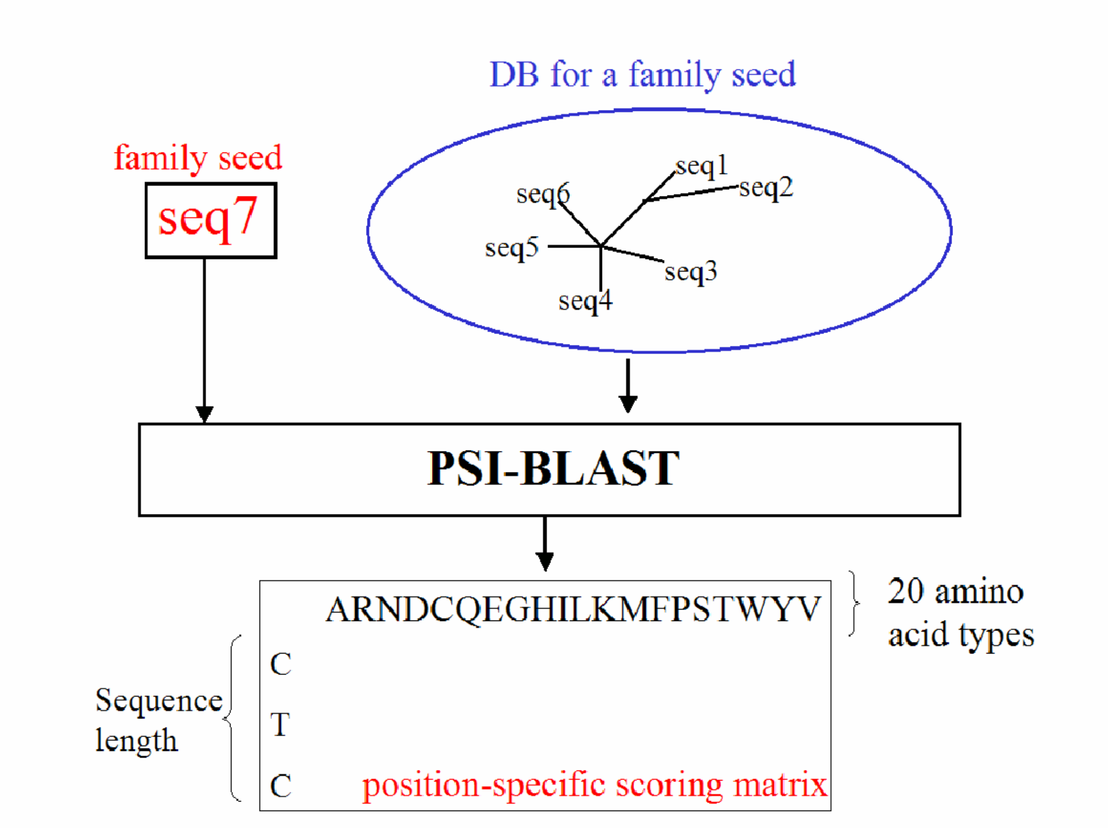


Figure 2

The seq 7 is the query sequence, and the other sequences which belong to the same family with seq 7 are the database for the PSI-BLAST. After the PSI-BLAST, we can get a sequence profile (position-specific scoring matrix) which represents the family. The profile is defined as series of probability distributions $P = p_1 p_2 p_3 \ldots p_n$ when n is the length of the sequence and p_i is a probability distribution over the 20 amino acids at position i.

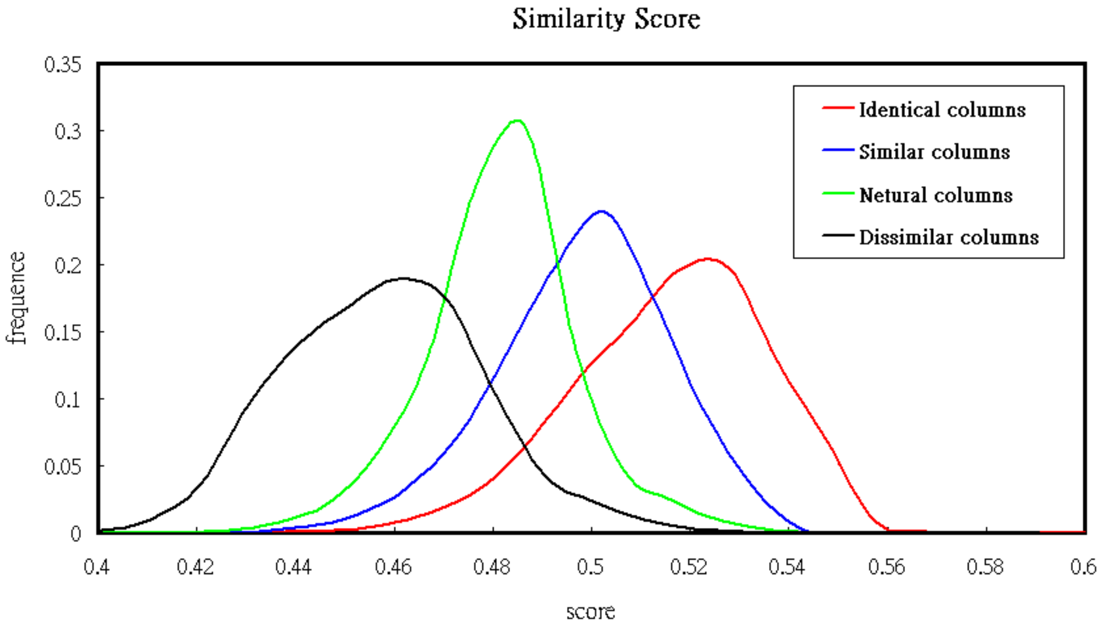


Figure 3

There are four classifications of column pairs: (1) a column with itself (we called the identical columns), (2) different columns that are associated with the similar seed amino acids (similar columns), (3) different columns with mutually neutral seed amino acids (neutral columns), (4) different columns with dissimilar seed amino acids (dissimilar columns).

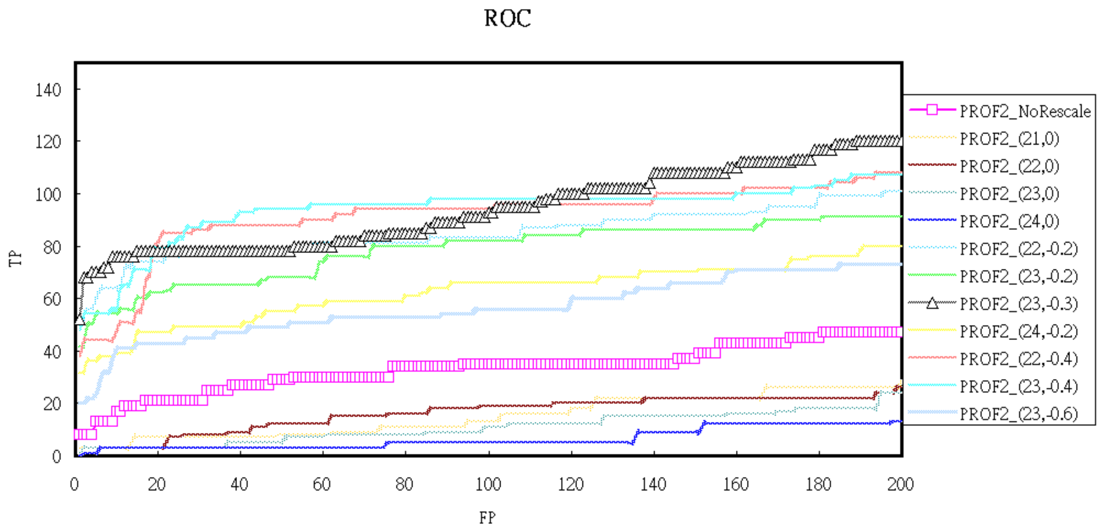


Figure 4

The numbers in the brackets are the different parameters for $PROF^2$. It implies that a little change in the parameters will cause huge different result in the ROC curve. The pink square is the original result which adjusted by any parameters and the dark triangle is the better result in the 200th false positive.

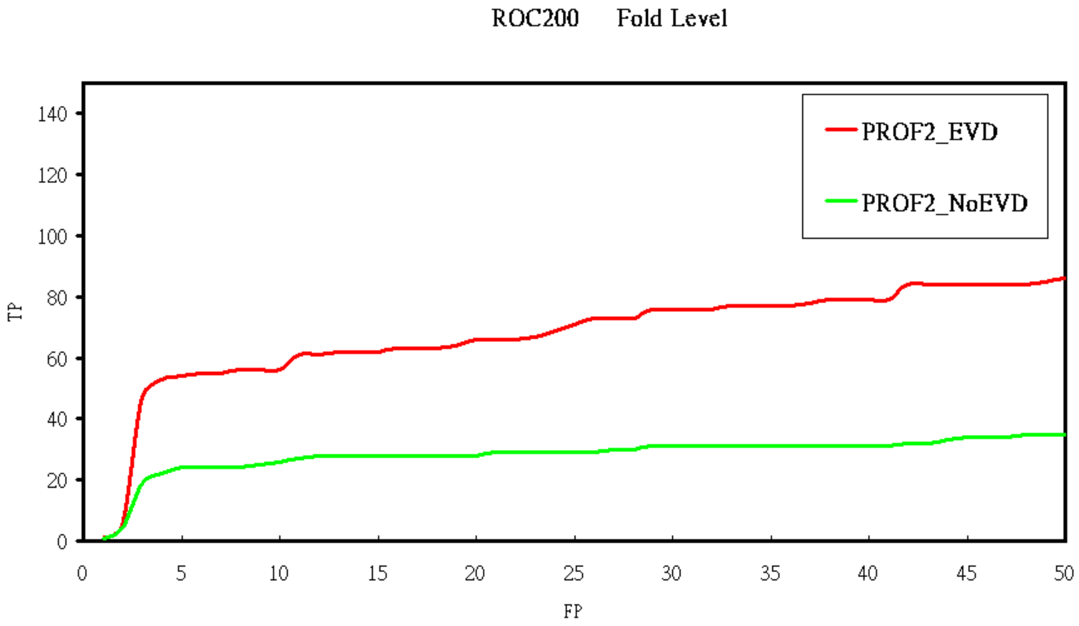


Figure 5

The green line is the result created by $PROF^2$ without the EVD fit procedure, and the red line is the result created by $PROF^2$ using the EVD fit procedure.

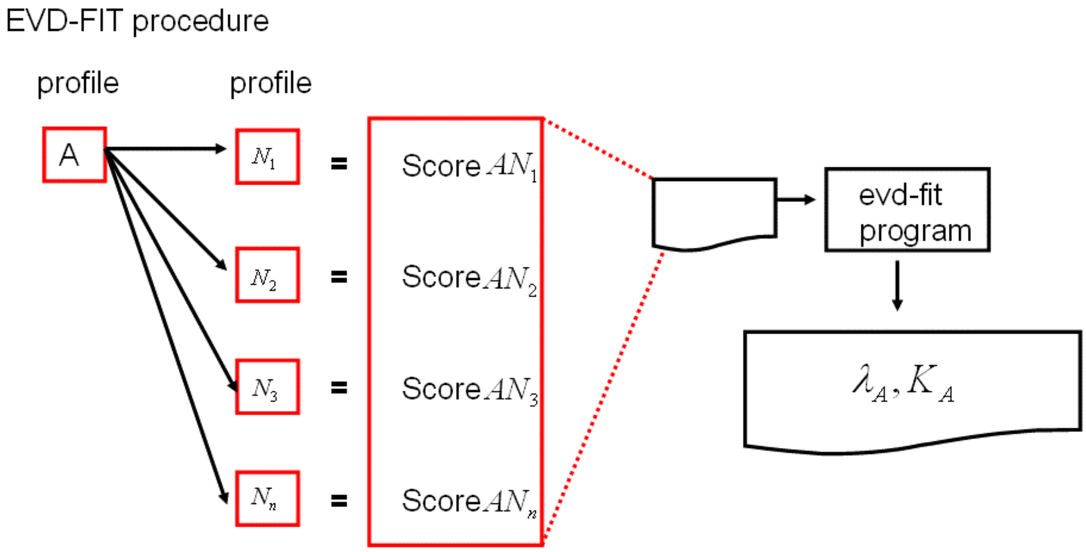


Figure 6

The profile “A” belongs to the subset separated from SCOP1.50 dataset chosen by the rules of mentioned before and the profiles, $N_1 \sim N_n$ are all profiles in that dataset. After the profile-profile comparison, the $ScoreAN_1 \sim ScoreAN_n$ will be collected as the input of evd-fit program. The λ_A and K_A are specific parameters for profile “A” to calculate e-value.

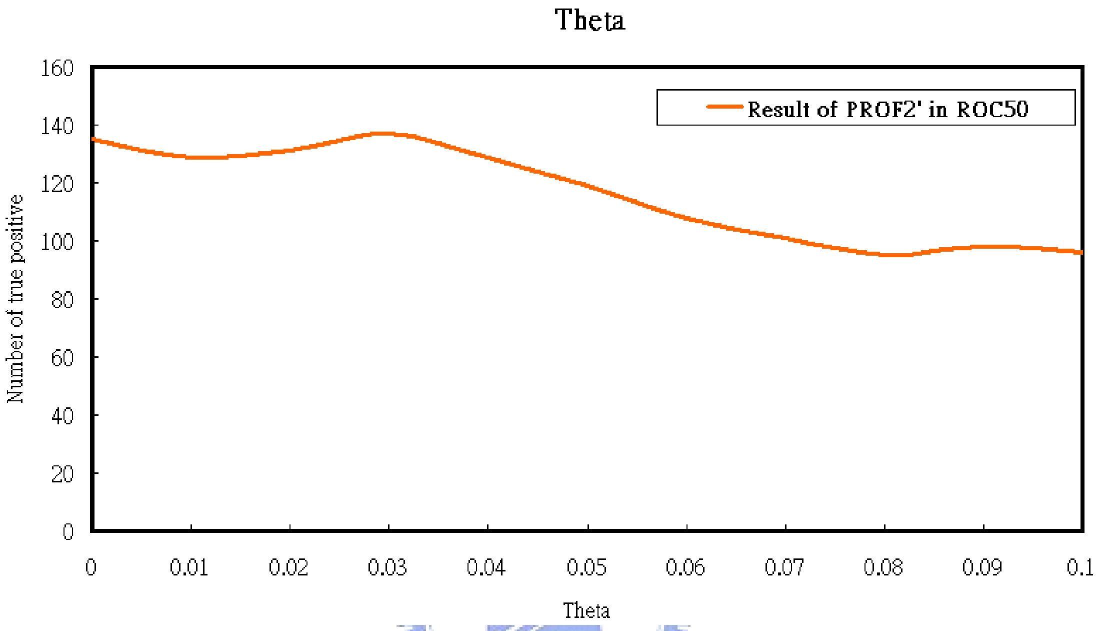


Figure 7

The effect of the mixture parameter θ on the performance. Performance is measured by the number of true relations that are detected before the 50th false positive.

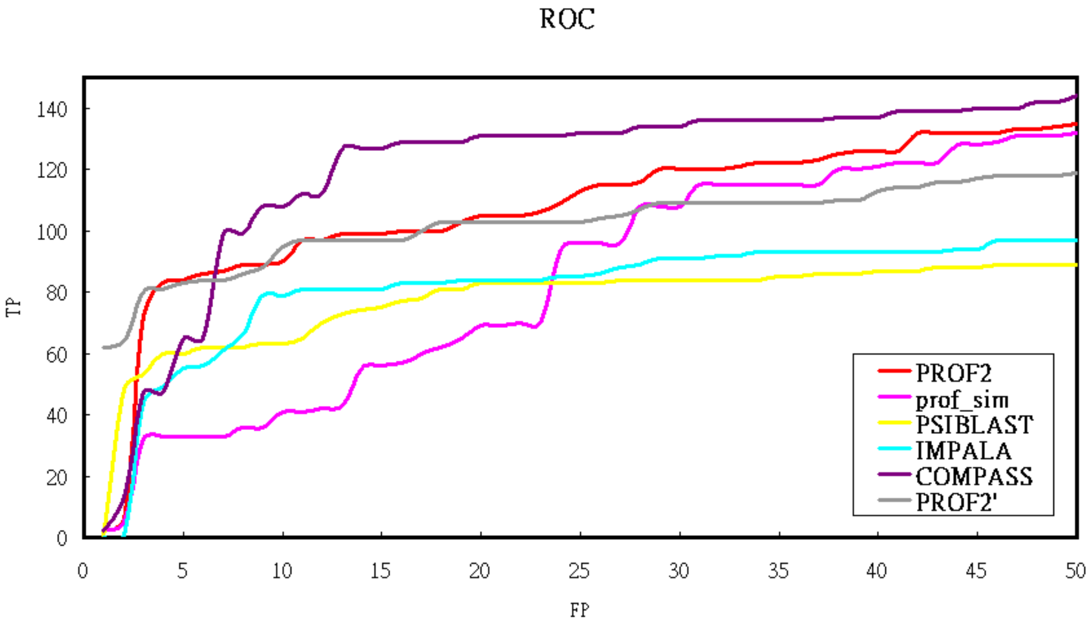


Figure 8

The result of SCOP1.50 in the ROC_{50}. The $PROF^{2}$ and $PROF^{2'}$ are our programs for profile-profile comparison. When the first false positive occurred, the $PROF^{2'}$ had the best result in all programs. A true positive id defined as a connection between families within the same fold.

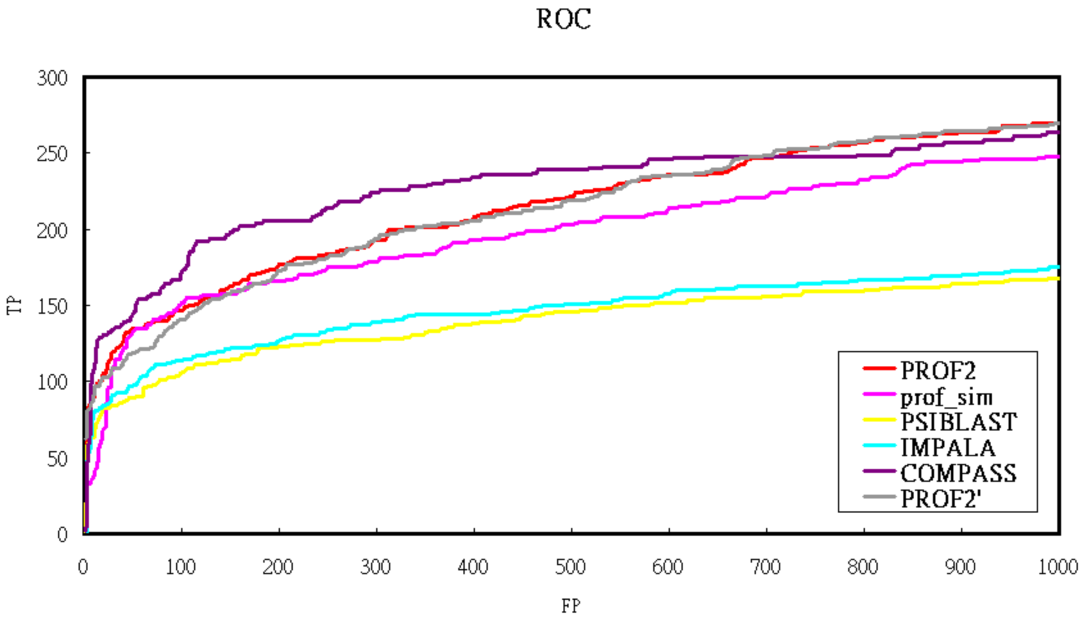


Figure 9

The result of SCOP1.50 in the ROC_{1000}. After the 700th false positive, the $PROF^{2}$ and $PROF^{2'}$ perform well then the COMPASS.

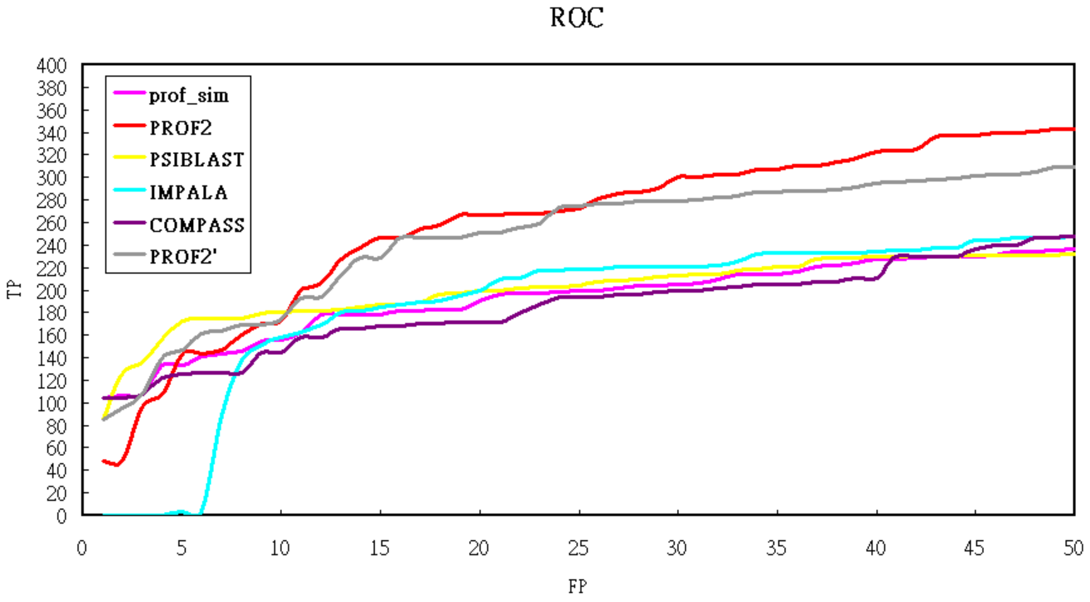


Figure 10

The result of SCOP1.63 in the ROC_{50}

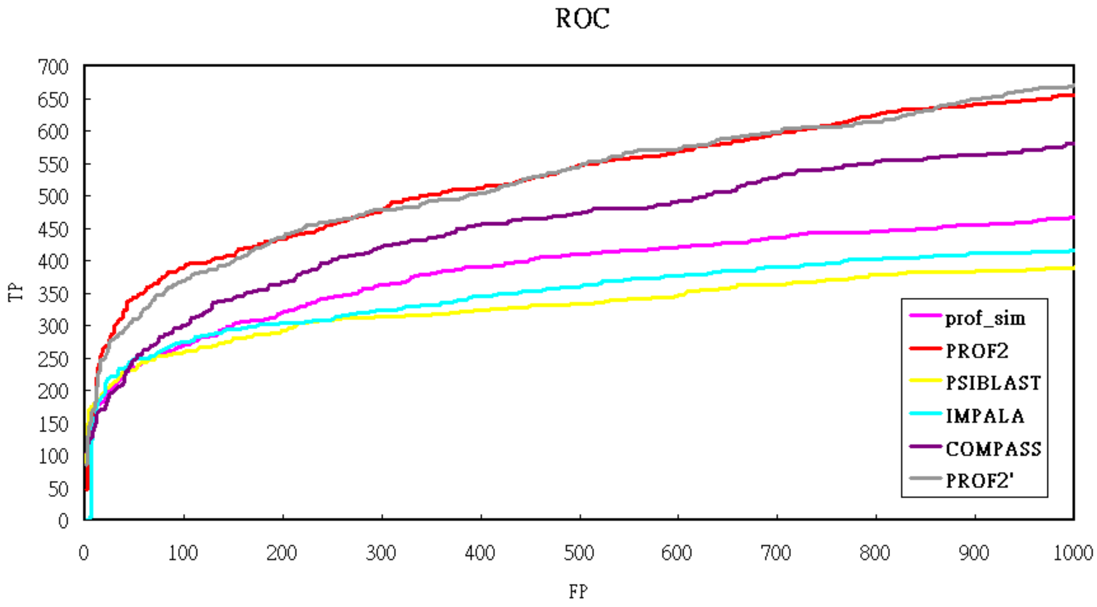


Figure 11

The result of SCOP1.63 in the ROC_{1000}

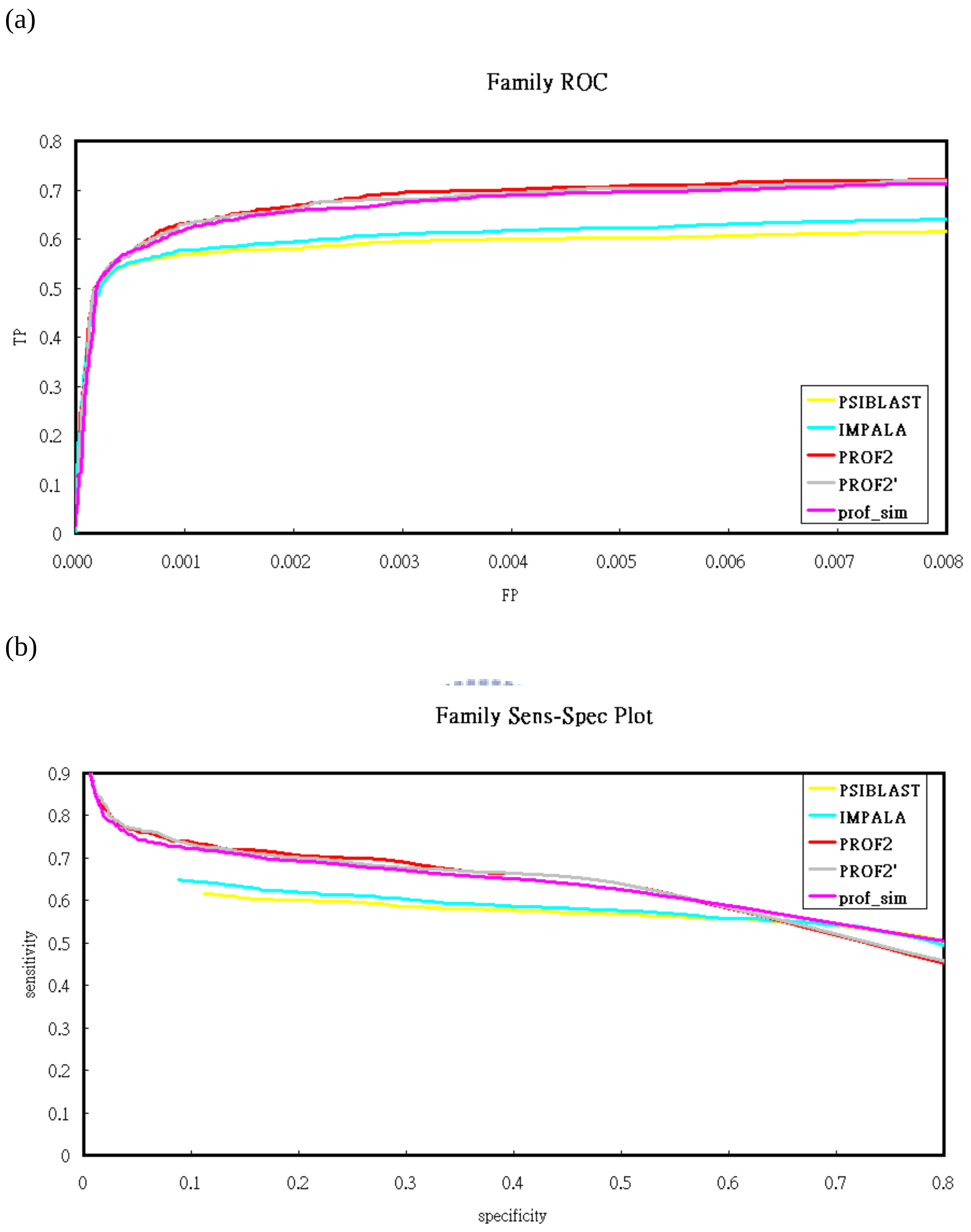
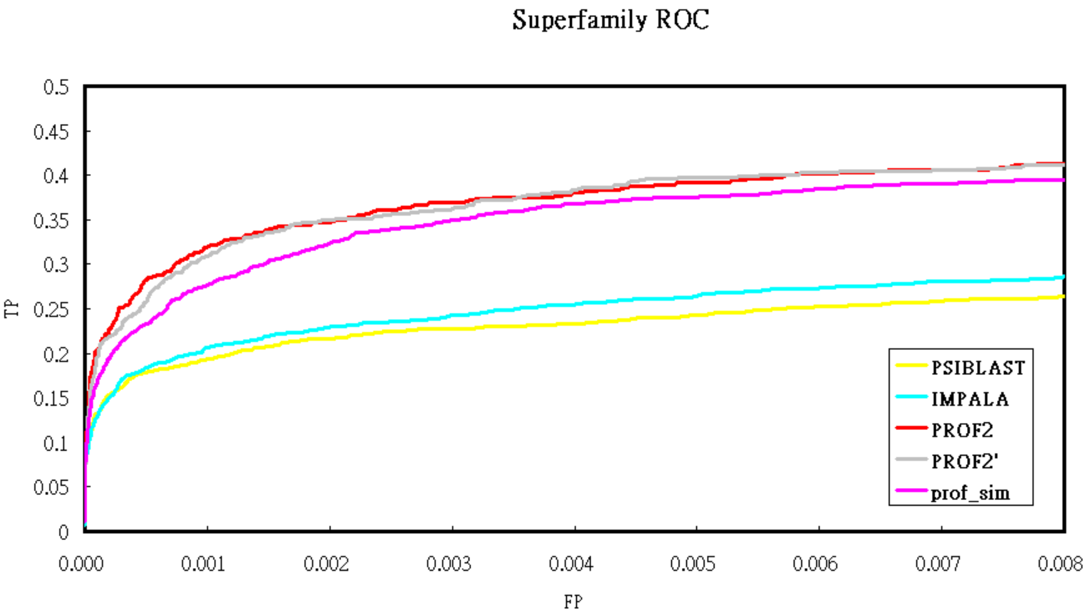


Figure 12

Lindahl’s benchmark for finding same family-only relationships. (a) The percentage of same family relationship (true positives) is plotted as a function of different family relationships (false positives). (b) Same data are plotted in terms of specificity (TP/(TP+FP)) versus sensitivity (TP/(TP+FN)).

(a)



(b)

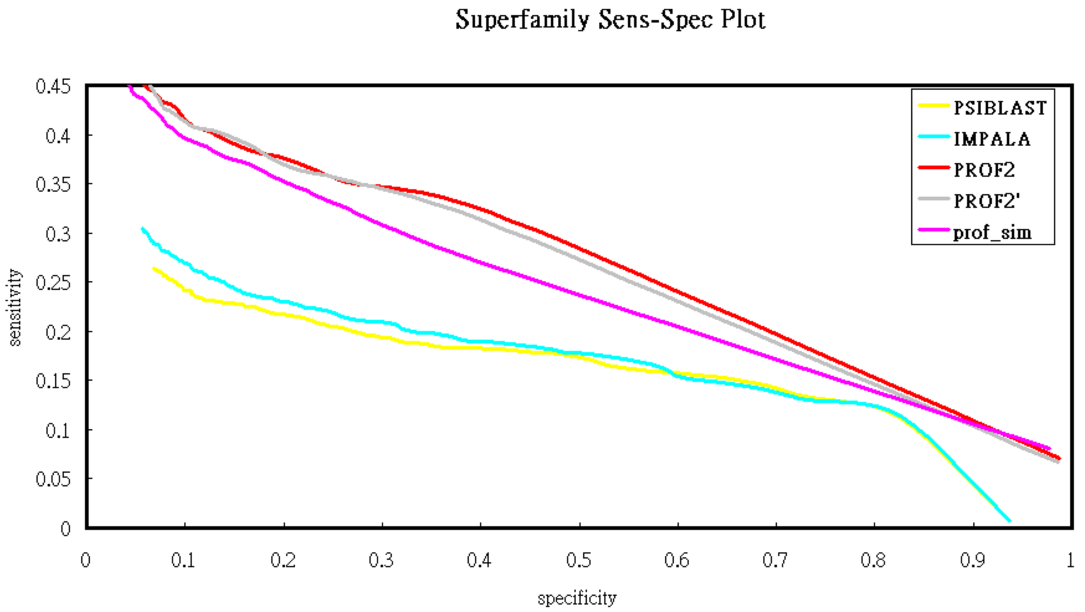


Figure 13

Lindahl’s benchmark for finding same superfamily-only relationships. Curves are as described for Figure 12, but true positives are defined as same superfamily relationships and false positives are defined as different superfamily relationships.

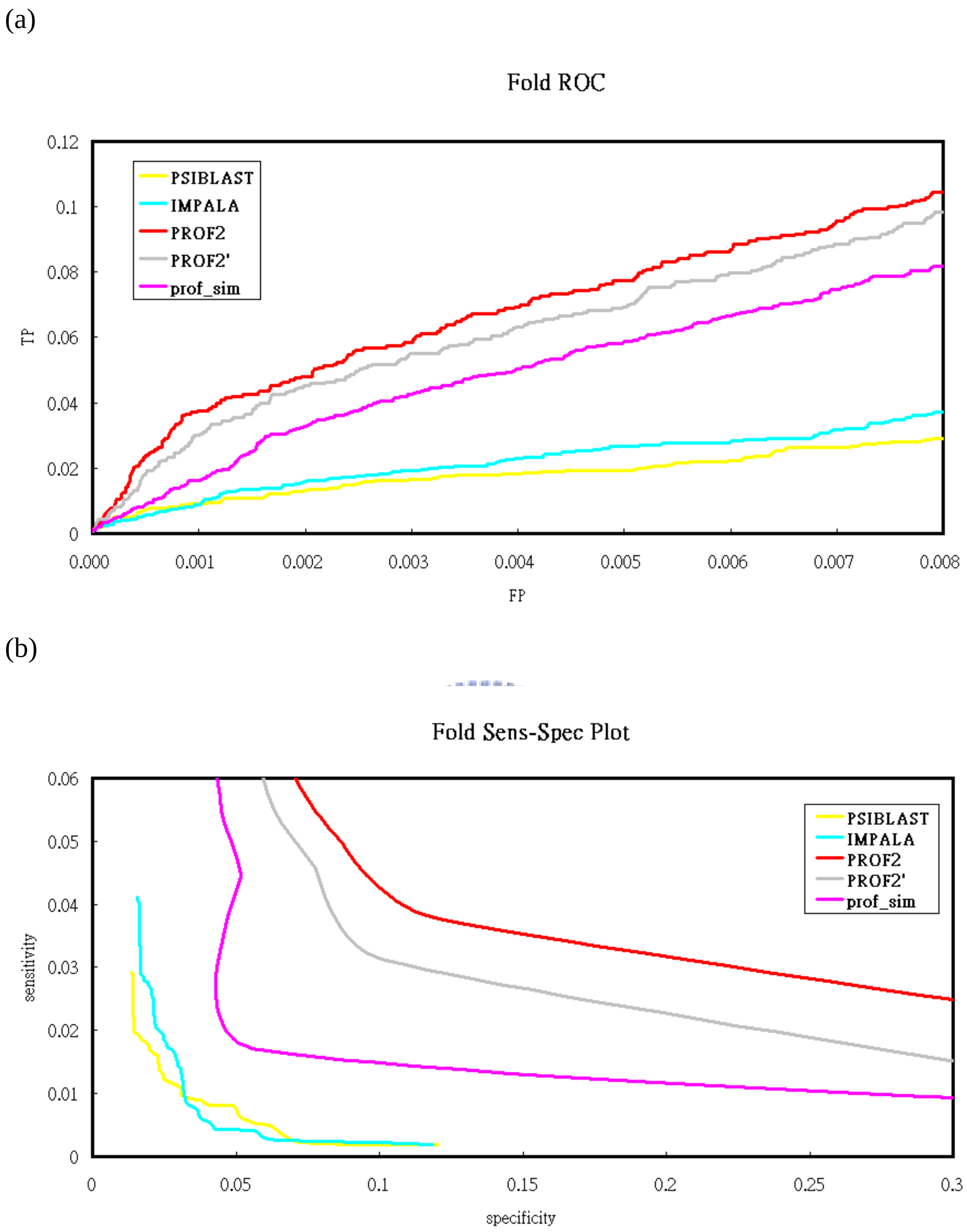


Figure 14

Lindahl’s benchmark for finding same fold-only relationships. Curves are as described for Figure 12, but true positives are defined as same fold relationships and false positives defined as different fold relationships.