Journal of Theoretics
Volume 5-4, Aug-Sept 2003

A Comparative Analysis of Protein Homology Detection Methods

N. M. ZAKI*, S. DERIS
Department of Software Engineering,
Faculty of Computer Science & Information System,
University Technology Malaysia, Malaysia.
*nazar@siswa.utm.my

R. M. ILLIAS
Department of Bioprocess Engineering,
Faculty of Chemical and Natural Recourses Engineering,
University Technology Malaysia, Malaysia

Abstract: Functional annotation of new gene sequences is an important challenge for computational biology systems. While much progress has been made towards improving experimental methods for functional assignment to putative genes, most current genomic annotation methods rely on computational solutions for homology modeling via sequence or structural similarity. With the increasing number of computer methods available for protein remote homologies detection, a comparative evaluation of the methods from biological prospective is warranted. This study uses benchmark SCOP dataset to test and compare the ability of five different computational methods for protein homologies detection. The results provide insight to biologist as to usage, value, and reliability of the numerous methods available.
Keywords: Homology detection, hidden Marko model, protein classification, support vector machines

1. INTRODUCTION

The last decade has witnessed a consistent effort in sequence biological information retrieval, caused in part by technological breakthroughs in large-scale sequencing and the human genome project. The main challenge facing modern biology is to interpret this newly generated sequence data, and perhaps most significantly, in the short term, to assign function to many putative gene predictions. Many approaches have been presented for the protein classification problem, including methods based on pairwise similarity of sequences profiles for protein families (Gribskov, 1987), BLAST (Altshul et al. 1997), Fasta (Pearson & Lipman 1998), consensus patterns using motifs (Bairoch, 1991; Attwood, 1998) and hidden Markov models SAM (Hughey, 2000), (Krogh, 1994), (Jaakkola, 2000). Most of these methods are one of the following:

- Generative approaches: the methodology involves building a model for a single protein family and then evaluating each candidate sequence to see how well it fits the model.

- Discriminative approaches take a different point of view: protein sequences are seen as a set of labeled examples {positive if they are in the family and negative otherwise} and a learning algorithm attempts to learn the distinction between the different classes. Both positive and negative examples are used in training for a discriminative approach, while generative approaches can only make use of positive training examples. One of the most successful discriminative approaches to protein classification is the work of the fisher kernel method (Jaakkola et al. 1999) for detection of remote protein homologies. In this paper we use benchmark SCOP dataset to test and compare the ability of five different computational methods for protein homologies detection.

2. PROTEIN HOMOLOGIES DETECTION METHODS

Protein sequences are very difficult to understand and model due to their complex random length nature. Various statistical models have been developed to measure the similarity between two sequences. This has been driven by the goal of attempting to group proteins with similar function together. In this study five methods were included. Brief descriptions of each program are provided below:

BLAST (Altschul, 1997), is an approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straightforward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. In addition to its flexibility and tractability to mathematical analysis, BLAST is an order of magnitude faster than existing sequence comparison tools of comparable sensitivity.

HMMER (Eddy 1995), is an implementation of profile Hidden Markov Model (HMM) methods for sensitive database searches using multiple sequence alignments as queries. Basically, you give HMMER a multiple sequence alignment as input; it builds a statistical model called a "hidden Markov model" which you can then use as a query into a sequence database to find (and/or align) additional homologues of the sequence family.

SAMT98, (Hughey 2000), is a linear HMM that implements the Baum-Welch algorithm. The estimated parameters are the transition and observation probabilities. Once the method converges, a multiple alignment can be created and the homologies detected.

SVM- Fisher, (Jaakkola 2000), this method is mainly designed to find all the proteins which belong to a particular superfamily. A generative HMM is used as a way of extracting features from the variable length protein sequences. This HMM represents the super family of interest and is trained on sequences from that family. Positive and negative training sequences are then run through the model, and the feature vectors produced which represent the original protein sequences can then be modeled in Euclidean space. A general discriminate method in this case was Support Vector Machine SVM is then used to classify the data points into the super family of interest. Test sequences can then be run through the model and the discriminate method can then be used to classify the protein sequence.

SVM-Motif (Logan, 2001), this method relies on combining probabilistic modeling and supervised learning in high dimensional feature spaces. The system uses a transformation

that converts protein domains to fixed-dimension representative feature vectors, where each feature records the sensitivity of each protein domain to a previously learned set of 'protein motifs' or 'blocks.' Subsequently, the system utilizes SVM classifiers to learn the boundaries between structural protein classes. Several methods are not included in this study. We focused our study on state-of-art methods.

3. NUMERICAL EXPERIMENTS

In this test, remote homology is simulated by holding out all members of a target SCOP (Murzin, 1995) family from a given superfamily. SCOP is a publicly accessible database over the Internet, this database stores a hand classified set of protein sequences. We investigated the performance of homology detection method on the SCOP database version 1.37 PDB90 (Murzin, 1995). The benchmark datasets used (Fig 1) is designed by Jaakkola *et al* (Jaakkola, Diekhans and Haussler, 2000). The training and testing sets used in this previous work are available online from:
http://www.cse.ucsc.edu/research/compbio/discriminative/.

This data is organized so that each experiment had a positive testing and training set for each model and each superfamily had a common negative testing and training data. A file was also provided which lists all the experiments that can be performed. The SCOP 1.37 Database was used, so all the identifiers refer to that version. A typical experiment would be the classification of the *G proteins family* which uses the training sequences from the *nucleotide triphosphate hydrolases* SCOP superfamily. Two other families were used to provide the positive training data, these were: - the *Nucleotide and nucleoside Kinases family* and the *Nitrogenase iron protein-like family*, so in the directory which relates to the G Protein family, there will be two models for each of the training families and corresponding training sequences for these two families. The test sequences for the two models will be the same and these are sequences from the *G proteins family*. In the directory below there are two sets of negative test and training sequences which are split up into two folds, this is to allow cross validation of the negative training data.

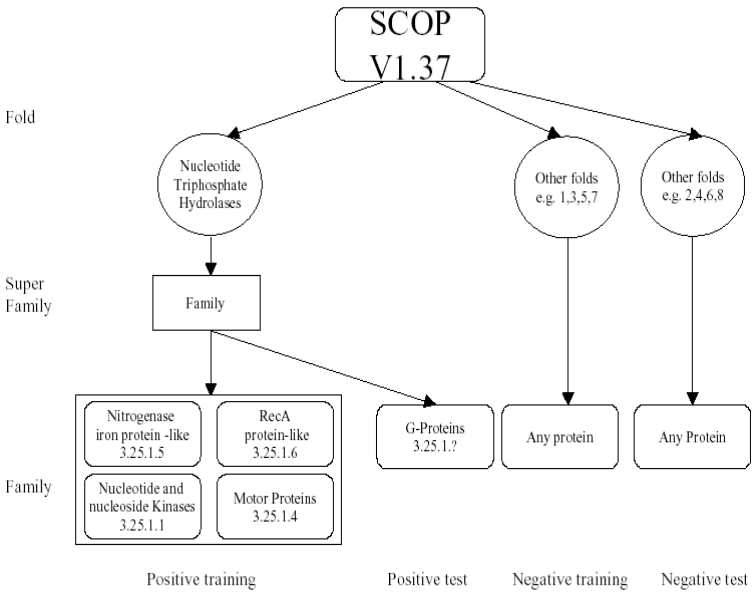

Fig 1: Layout of SCOP sets used

4. RESULTS

We follow Jaakkola's way to compare the results of all methods. Since each method produces a probability score on a different scale they cannot be directly compared, so the rate of false positive (RFP) is used. This is defined as the fraction of negative test sequences that score as high or better than, the positive sequence we are testing. So therefore a score of zero is very good.

For the comparison of the overall performance for the five methods on the 33 test families, we computed the median RFP for the family, as shown in table 1. Values for the median RFP are shown on the Y-axis. On X-axis we plot the number of SCOP families, out of the 33 families that we tested, for which the given method achieves that median RFP performance or better. We included results from all the methods (Table 1). These include the original experimental results from Jaakkola, HAMMER, SAM-T98 iterative HMM, SVM-motif, and BLAST on the same data. To approximate a family-based homology detection method, BLAST is run using a randomly selected training set sequence for one iteration, with the positive training set as a database and a very low E-value inclusion threshold. The resulting matrix is then used to search the test set for a maximum of 20 iterations using the default E-value inclusion threshold. The results for all 33 SCOP families are summarized in (Table 1) and (Figure 2). Each series corresponds to one homology detection method. Qualitatively, the SAM-T98 and Fisher-SVM methods perform slightly better than the rest. However, if we evaluate the statistical significance of these differences using a two-tailed signed rank test (Heniko, 1997; Salzberg, 1997), including a Bonferroni adjustment for multiple comparisons only the SVM-Fisher method does better than any other method: SVM-Fisher's performance is better than that of BLAST with a p-value of 0:000045. The results show that the SVM-Fisher framework is a considerable superior over the previous methods, only 27 out of the 33 families had a better score for the maximum RFP over the SAM-T98 method and 32 had a better medium RFP over the SAM-T98 method. The SVM-Fisher method classified all families better than using BLAST, HMMER., SAM-98, and SVM–motif methods.

Expt.	SCOP Family	HMMER	BLAST	SAM T98	SVM MOT	SVM
1	Phycocyanins	0.471	0.391	0.450	0.528	0.364
2	Long-chain cytokines	0.375	0.721	0.446	0.092	0.035
3	Short-chain cytokines	0.386	0.407	0.109	0.035	0.002
4	Interferons/interleukin-10	0.511	0.324	0.289	0.054	0.004
5	Parvalbumin	0.000	0.000	0.000	0.000	0.000
6	Calmodulin-like	0.808	0.023	0.000	0.000	0.000
7	Immunoglobulin V dom	0.595	0.135	0.000	0.006	0.000
8	Immunoglobulin C1 dom	0.738	0.033	0.000	0.110	0.000
9	Immunoglobulin C2 dom	0.181	0.119	0.000	0.232	0.000
10	Immunoglobulin I dom	0.680	0.007	0.000	0.135	0.000
11	Immunoglobulin E dom	0.723	0.168	0.178	0.568	0.073
12	Plastocyanin/azurin-like	0.885	0.016	0.039	0.753	0.013
13	Multidomain cupredoxins	0.040	0.342	0.003	0.504	0.002
14	Plant virus proteins	0.063	0.641	0.088	0.504	0.133
15	Animal virus proteins	0.698	0.750	0.204	0.407	0.066
16	Legume lectins	0.312	0.278	0.278	0.276	0.083
17	Prokaryotic proteases	0.652	0.080	0.000	0.052	0.000
18	Eukaryotic proteases	0.317	0.000	0.000	0.000	0.000
19	Retroviral protease	0.394	0.238	0.012	0.029	0.003
20	Retinol binding	0.281	0.475	0.165	0.169	0.051
21	alpha-Amylases, N-term	0.095	0.630	0.007	0.086	0.000
22	beta-glycanases	0.131	0.517	0.009	0.440	0.008
23	type II chitinase	0.145	0.350	0.110	0.346	0.031
24	Alcohol/glucose dehydro	0.465	0.041	0.019	0.022	0.008
25	Rossmann-fold C-term	0.351	0.121	0.015	0.224	0.005
26	Glyceraldehyde-3-phosphate	0.412	0.315	0.009	0.024	0.002
27	Formate/glycerate	0.474	0.022	0.001	0.019	0.002
28	Lactate&malate dehydro	0.362	0.530	0.024	0.002	0.002
29	G proteins	0.359	0.378	0.007	0.001	0.000
30	Thioltransferase	0.540	0.000	0.000	0.002	0.000
31	Glutathione S-transfer	0.834	0.311	0.273	0.292	0.238
32	Fungal lipases	0.210	0.044	0.000	0.014	0.000
33	Transferrin	0.162	0.875	0.007	0.389	0.026

Table 1: Rate of false positives for all 33 families, HMMER, BLAST, SAM-T98, SVM MOT (SVM-motif method), SVM (SVM-Fisher method).

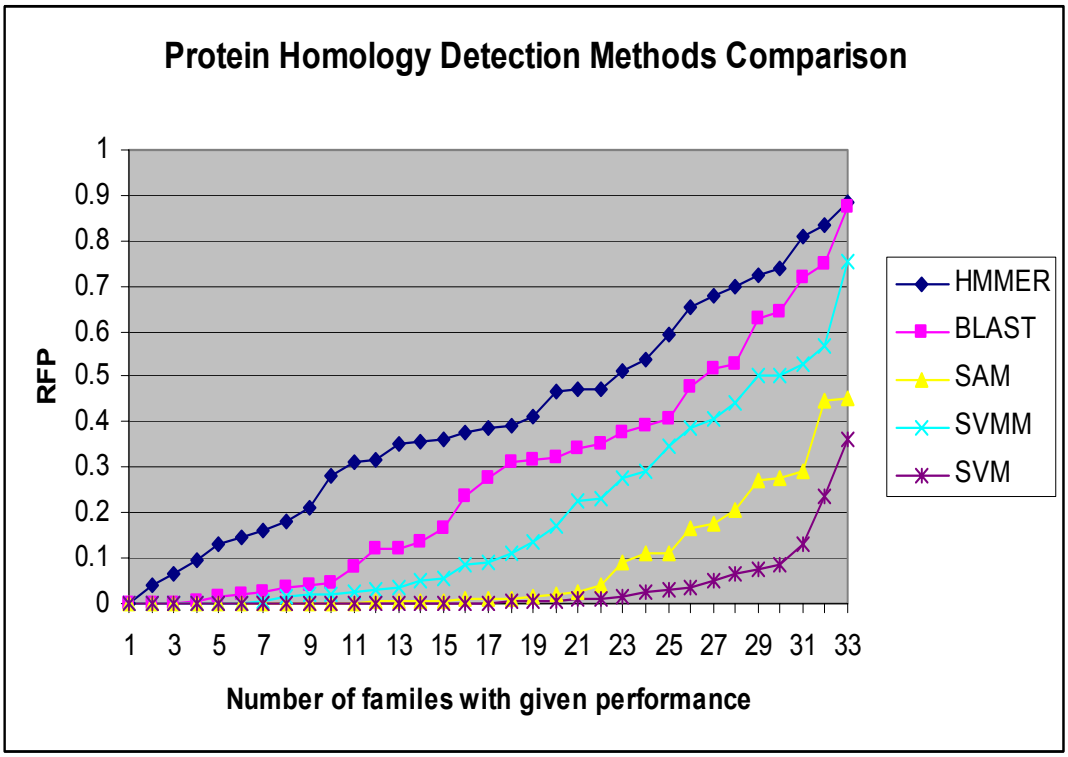

Figure 2. Comparison of the overall performance for the five methods on the 33 test families. For each family, we computed the median RFP for the family, as shown in table 1. Values for the median RFP are shown on the X-axis. On Y-axis we plot the number of SCOP families, out of the 33 families that we tested, for which the given method achieves that median RFP performance or better.

5. CONCLUSION AND FUTURE DIRECTIONS

The purpose of this study is to fine the most reliable method for protein homologies detection currently available. While all methods analyzed were able to detect the protein homologies, methods based on Hidden Markov Model and the combination of generative and discriminative model such as support vector machines, were superior. Our study indicated that SVM-Fisher method (Jaakkola, 2000) is a superlative method currently available for homologies detection. One of our future researches is to look in depth on the methods based on discriminative model as they shown to be more reliable and efficient. The comparison will be based on the ability of those methods in protein sequences feature extraction. Note that currently there are few new methods for protein homology detection show more successful results. Examples of these methods are Mismatch String Kernels (Christina 2002), SVM-Pairwise (Liao 2002) and SVM-String Kernel method (Zaki 2003). In the future all this methods will be compared and in depth analyzed.

REFERENCES

- Murzin A. G., S. E. Brenner, T. Hubbard and C. Chothia, (1995). SCOP: A structural classification of proteins database for the investigation of sequences and structures. JMB, 247:536.

- Logan B., P. Moreno, B. Suzek, Z. Weng and S. Kasif, (2001) "A Study of Remote Homology Detection" CRL Technical Report 2001/5.

- Hughey R., K. Karplus and A. Krough (2000), SAM, Sequence Alignment and Modelling Software System Manual and implementation, University of California.

- Altschul S. F., T.L. Madden, A. A. Schaer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research, 25:3389.

- Heniko, S. and J. G. Heniko, (1997). Embedding strategies for effective use of information from multiple sequence alignments. Protein Science, 6(3): 698.

- Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. Data Mining and Knowledge Discovery, 1:371.

- Eddy, S. R. (1995). Multiple alignment using hidden Markov models. In ISMB, pages 114-120. *AAAI* Press.

- T Jaakkola, M Diekhans, and D Haussler, (2000). A discriminative framework for detecting remote protein homologies. Journal of Computational Biology.

- Liao C., W. S. Noble. (2002). Combining pairwise sequence similarity and support vector machines for remote protein homology detection. Proceedings of the Sixth Annual International Conference on Research in Computational Molecular Biology, (RECOMB), 225-232

- Christina L., E. Eskin, J. Weston and W. S. Noble. (2002). Mismatch String Kernels for SVM Protein Classification. NIPS 2002.

- Zaki, N. M., S. Deris, R. M. Illias, Y. L. Chin. (2003). Application of String Kernel on Protein Sequence Classification. AIAI 2003.

Received February 2003

Journal Home Page