Protein Seer: A Web Server for Protein Homology Detection

B.T. Logan, U. Karaoz[1], P.J. Moreno, Z. Weng[1], S. Kasif[1]
Cambridge Research Laboratory
HP Laboratories Cambridge

HPL-2004-131
July 30, 2004*

protein family, classification, remote homology, sequence similarity, Support Vector Machines, kernel methods

We present and evaluate a publicly available web server which classifies protein sequences into SCOP 1.63 PDB95 structural superfamilies. The website returns ranked lists of likely superfamilies and hence implicit structural predictions according to three computational techniques: BLAST, HMMER and a discriminative classifier SVM-BLOCKS. It is the first website to provide predictions using SVM-BLOCKS. In addition to the ranked lists, the website displays alignment information and a web services interface is also available for computationally intensive use. We conduct a large-scale evaluation which mimics the predictions returned by the website. The study indicates that the site provides valid predictions and that SVM-BLOCKS approach can outperform BLAST and HMMER when sufficient examples are available to learn the SVM classifiers.

* Internal Accession Date Only

[1] Bioinformatics Program, Department of Biomedical Engineering, Boston University, Boston, MA, USA

Approved for External Publication

I. INTRODUCTION

Structural genomics initiatives [1] have greatly expanded the collection of experimentally determined protein structures. The traditional and still the most reliable ways to determine the 3-dimensional (3-D) structure of a protein are X-ray crystallography and NMR, which are time-consuming, costly, and currently infeasible for some protein families. Thus, computational approaches remain, thus far, the only resort for deducing the structure information for many sequences

Evolutionary selection and functional pressures forces the retention of sequence features important for structure and function. This has been the main impetus behind homology-based methods, which infer homology from computed sequence similarity. Alignment tools such as BLAST and PSI-BLAST [2] have been widely used to provide evidence for homology by matching a new sequence against a database of previously annotated sequences such as Pfam [3].

Some homologous proteins are evolutionarily divergent, that is they do not exhibit significant sequence similarity. In order to detect such weak or remote homologies, we can rely on the concept of protein family which denotes a group of sequences sharing the same evolutionary origin. In particular, a statistical model can be built for each protein family or superfamily. To classify a new sequence, we compute the probability (or likelihood) it was generated by each model in a library of trained statistical models. Computational methods that relate a sequence to a superfamily-specific model often outperform pairwise sequence comparison methods.

Examples of statistical protein models include sequence profiles (also known as Position Specific Scoring Matrices or PSSMs) [2] and Hidden Markov Models (HMMs) e.g. [3,4]. These probabilistic models are often called *generative* because they induce a probability distribution over protein sequences that can subsequently be used to "generate" members of the family using stochastic simulation. Generative probabilistic models can be contrasted to *discriminative* frameworks, which focus on learning the combination of features that discriminate most effectively between families. Support Vector Machines (SVMs) (e.g. [5]) and Neural Networks are two popular discriminative methods. A discriminative framework is typically implemented using a classifier that learns a boundary between two or more classes.

We have developed Protein Seer, a website that can make predictions of structural family membership for submitted protein sequences. The site provides predictions based on three techniques: sequence-similarity, probabilistic models and discriminative frameworks that combine statistical models and SVMs. Specifically, we classify submitted sequences as belonging to SCOP 1.63 structural superfamilies [6] according to BLAST [2], HMMER [4] and our previously introduced SVM-BLOCKS approach [7,8]. The site is found at http://genomics15.bu.edu/protseer.html.

In this paper, we describe the operation and capabilities of the website and an experimental evaluation of the three techniques it supports. We have previously reported classification results on SCOP 1.37 [7,8]. Specifically, we investigated the classification accuracy of BLAST, HMMER, SVM-BLOCKS and another SVM-based technique on a small subset of SCOP. We found that the two SVM techniques had comparable performance and both outperformed BLAST and HMMER.

In this paper, we experiment with a larger and more recent version of SCOP, SCOP 1.63. Additionally, rather than conducting classification experiments, we perform an information-retrieval style study in which we assess how well a structurally unlabelled protein sequence can be classified as belonging to the correct superfamily and hence receive the correct structural assignment. Such a study mimics the operation of the website and is a more appropriate indicator of performance than classification experiments.

II. METHODOLOGY

A. Structural Prediction Techniques

The SCOP database provides a detailed and comprehensive classification of all known protein structures. The unit of classification is the protein domain. The classification is into four hierarchical levels: class, fold, superfamily and family. Family and superfamily levels describe near and far evolutionary relationships. Fold levels describe geometrical relationships. In our investigations and on the website, we study classification at the superfamily level.

We use the ASTRAL database [9] and SCOP 1.63 to obtain a set of protein domain sequences with less than 95% identity to each other. For this and all versions of SCOP, there is a great deal of variation in the number of sequences available in each superfamily. To avoid studying families with very little data, we only consider superfamilies which contain at least 5 sequences. This reduces the number of superfamilies from 1437 to 738 and the number of sequences studied from 9498 to 7694.

i. Blast

Our simplest structural prediction technique uses BLAST. Specifically, we BLAST a given query sequence against the sequences for each superfamily in the database. We then rank the superfamilies according to the best (lowest) E-value returned for each superfamily. The ranking implicitly provides a structural prediction for the query protein sequence.

We are also considering providing predictions using PSI-BLAST but are still investigating the best set of parameters and methodology to use, in particular whether to run PSI-BLAST separately on each superfamily or on all of SCOP. For the small superfamilies, it may be necessary to bring in additional sequences from another database.

ii. HMMER

The website also provides structural predictions using an HMM-based classifier. We use the HMMER 2 package [4]. We build a HMM model for each of the 738 superfamilies as follows. First we align all the domains in the training set constructed for each superfamily using the multiple alignment tool CLUSTALW [10]. Then, using the *hmmbuild* tool, we build HMM models based on these multiple alignments. We use the default parameters of *hmmbuild*. Given a query sequence we use *hmmsearch* to score it against the HMM for each superfamily. We then rank the superfamilies by E-value returned, similar to BLAST.

iii. SVM-BLOCKS

The novel component of our server is that in addition to BLAST and HMMER, the website also returns structural predictions according to the SVM-BLOCKS technique [7,8]. We have previously reported promising results with this approach on SCOP 1.37.

The SVM-BLOCKS technique consists of two major steps. First each protein sequence or subsequence of interest is converted to a feature vector of fixed dimensionality. Each vector is created by scoring a set of pre-learnt biological motifs against the protein sequence. Each dimension of these feature vectors thus represents the sensitivity of the protein to a particular motif. For the case where a motif is detected multiple times in a sequence, we take the maximum of all scores for that motif in that sequence. There are a number of databases of short protein motifs available. We use the BLOCKS database [11] and the associated BLIMPS tool to score each block in BLOCKS against each given protein sequence. The number of dimensions of each vector thus corresponds to the number of BLOCKS, currently around 12,000.

Once we convert the proteins in each superfamily to the fixed dimensional representation we form a set of positive examples for each superfamily. We also construct negative examples for each superfamily using sequences in other folds. Given the feature vectors, we then learn SVM classifiers [5] to separate each superfamily from “the rest of the world” represented by the negative examples. SVM classifiers have been used successfully in number of biological applications since they have been shown to be effective tools for learning in sparsely sampled high dimensional spaces [12].

B. Website Implementation

The website provides both an html interface and a web services interface to the structural analysis techniques studied. The html interface is suitable for casual use while the web services interface is provided for more intensive analyses.

The html interface is implemented using *PISE* [13]. *PISE* allows rapid prototyping of websites for molecular biology applications. It produces websites which provide a familiar interface to analysis engines and is customizable via XML descriptions of the tools.

The homepage at http://genomics15.bu.edu/protseer.html is shown in Figure 1. Here, a user can submit sequences for analysis either by cutting and pasting text or by specifying a file. The user can also set options for the analysis. Currently, the options are the maximum number of top scoring superfamilies to display and which superfamilies to use for the analysis. The default is to compare the sequences to all 738 superfamilies.

Once the options have been set, the user hits “Run protseer” to start the analysis. This runs BLAST, HMMER and SVM-BLOCKS on the submitted sequences for the specified superfamilies. If the analysis takes longer than 1 minute, a page is returned notifying the user that they will be emailed when the results are ready. Otherwise the results screen is displayed immediately after the analysis is finished.

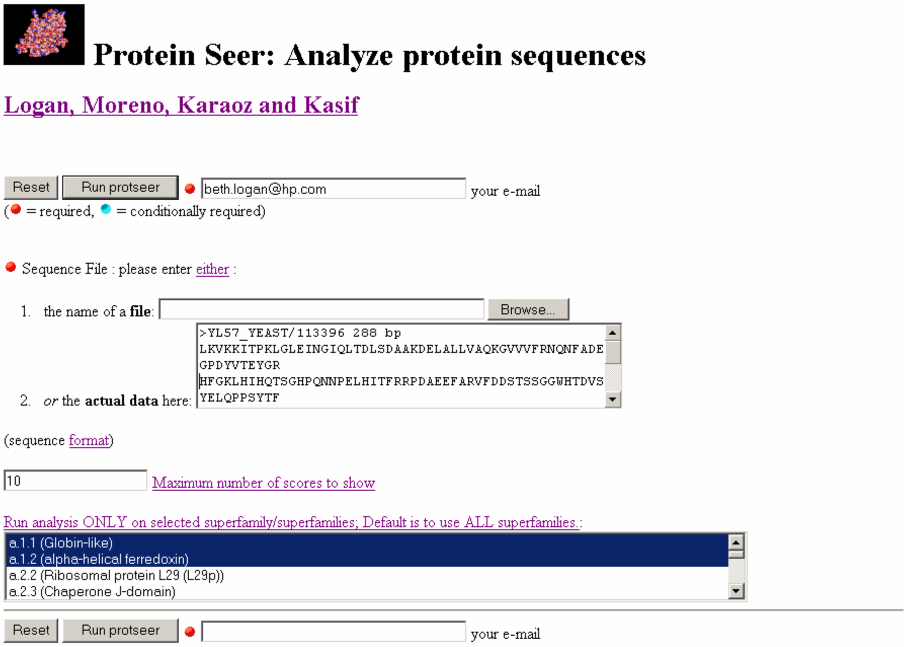
An example results page is shown in Figure 2. This page lists the top scoring superfamilies for BLAST, HMMER and SVM-BLOCKS with their scores. For BLAST and HMMER, the results are ranked by E-Value. For SVM-BLOCKS, the results are ranked by the score returned by the SVM classifier. The score represents the distance from the boundary between the positive and negative examples for each superfamily and is an indication of the confidence of the structural predictions. Strictly speaking, this score is not a probability and should be calibrated. We are currently investigating techniques to convert this score to a probability (an open problem for SVMs). Meanwhile, we have found that ranking by score produces good results.

The results page also displays graphically the BLOCK alignments with E-value greater than 1.0 for each query sequence. Additional links provide more information such as BLAST and HMMER alignments.

The web services interface is implemented in Java using the Apache Axis toolkit [14]. The interface implements a remote procedure call to return BLAST, SVM-BLOCKS and HMMER scores for a submitted query sequence and a given superfamily. This allows users to write their own client which queries the website and processes the results returned. Such an interface is much more convenient when a user has many sequences to process. The WSDL is at http://genomics15.bu.edu:8080/axis/services/RemHom?wsdl

III. Results

We conduct an information retrieval style study to assess how well a structurally unlabelled protein sequence can be classified into the correct superfamily and hence receive the correct structural assignment. We divide the data for the 738 SCOP superfamilies studied into disjoint training and testing sets. 70% of the data for each superfamily is used for training and the rest for testing. We use the data in the training set to train HMMER models and SVM classifiers for each superfamily. We construct a set of negative examples for each superfamily by sampling 1000 sequences from training data in other folds.



Protein Seer: Analyze protein sequences

Logan, Moreno, Karaoz and Kasif

● beth.logan@hp.com your e-mail
(● = required, ● = conditionally required)

● Sequence File : please enter either :

1. the name of a **file**:
2. *or* the **actual data** here:

```
>YL57_YEAST/113396 288 bp
LKVKKITPKLGLEINGIQLTDLSDAAKDELALLVAQKGVVVFRNQNFADE
GPDYVTEYGR
HFGKLHIHQTSGHPQNNPELHITFRRPDAEEFARVFDDSTSSGGWHTDVS
YELQPPSYTF
```

(sequence format)

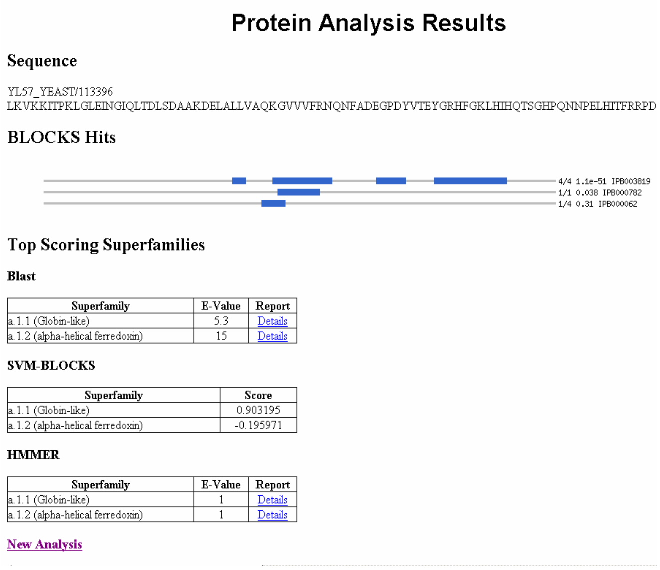
10 Maximum number of scores to show

Run analysis ONLY on selected superfamily/superfamilies; Default is to use ALL superfamilies.:

a.1.1 (Globin-like)
a.1.2 (alpha-helical ferredoxin)
a.2.2 (Ribosomal protein L29 (L29p))
a.2.3 (Chaperone J-domain)

● your e-mail

Fig. 1. Protein Seer Homepage



Protein Analysis Results

Sequence

YL57_YEAST/113396
LKVKKITPKLGLEINGIQLTDLSDAAKDELALLVAQKGVVVFRNQNFADEGPDYVTEYGRHFGKLHIHQTSGHPQNNPELHITFRRPD

BLOCKS Hits

Top Scoring Superfamilies

Blast

Superfamily	E-Value	Report
a.1.1 (Globin-like)	5.3	Details
a.1.2 (alpha-helical ferredoxin)	15	Details

SVM-BLOCKS

Superfamily	Score
a.1.1 (Globin-like)	0.903195
a.1.2 (alpha-helical ferredoxin)	-0.195971

HMMER

Superfamily	E-Value	Report
a.1.1 (Globin-like)	1	Details
a.1.2 (alpha-helical ferredoxin)	1	Details

New Analysis

Fig. 2 Protein Seer results page. Query sequence scored against the first two superfamilies.

We then form a test set by sampling 1000 sequences from the test data. Sampling is used to save processing time. For each sequence, we score it against each superfamily using the BLAST, HMMER and SVM-BLOCKS techniques. We then rank each superfamily by E-value for BLAST and HMMER or score for SVM-BLOCKS. The ranked superfamilies implicitly give the structural predications for each test sequence. Ideally, the superfamily of which the test sequence is a member of should be ranked 1.

Figure 3 shows the results from this study. The plot shows *log(Average Rank of the Correct Superfamily)* for each of the three techniques vs. *log(Number of Positive Training Examples)* with trend lines fitted. From this plot we can see that given sufficient training examples (around 50), the SVM-BLOCKS technique has better performance than BLAST or HMMER as it consistently assigns a higher rank the test sequence's true superfamily. When less than 50 training examples are available for a superfamily, BLAST has the best performance, although this is obviously dependent on how redundant the training sets are.

The results indicate that our website returns valid structural predictions, suggesting that it is a useful starting point for protein analysis. In cases where the three techniques give contradictory predictions, users can probe the supplied alignment and other information in order to evaluate the plausibility of the result.

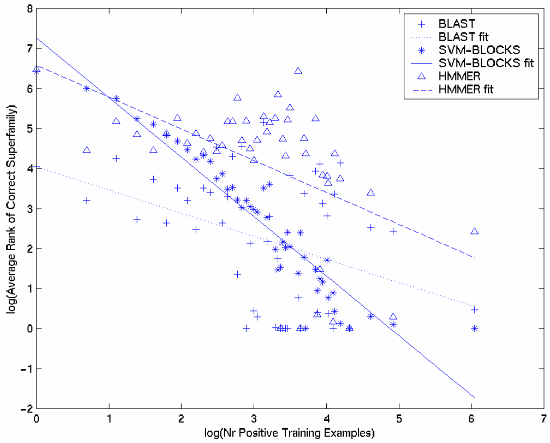


Fig. 3: log(Average Rank of the Correct Superfamily) for each of the three techniques vs. log(Number of Positive Training Examples) with trend lines fitted for 1000 sequences in SCOP 1.63

V. Conclusion

We have presented a publicly available web server which compares submitted protein sequences to SCOP 1.63 PDB95 structural superfamilies. Structural predictions are produced using three techniques: BLAST, HMMER and SVM-BLOCKS. The website returns ranked lists of likely superfamilies for the three techniques supplemented with additional alignment information. A web services interface is also available. We have additionally presented the results of investigations suggesting that the SVM-BLOCKS technique outperforms BLAST and HMMER when sufficient examples are available to learn the classifiers.

Since we introduced the first version of the SVM-BLOCKS technique in 2001 [7,8], a number of methods that provide a structural classification of proteins have been developed based on different SVM kernels [12]. In particular the technique reported in [15] is the most similar to our original proposal, although it uses emotifs instead of the BLOCKS database. The distinguishing characteristics of our approach are its simplicity and a relatively natural ability to interpret the results using block hits.

We are currently working with biologists to obtain feedback on the performance of the site. We anticipate refining the results returned as a result of this study. Additionally, we are working on improving the performance of the SVM-BLOCKS and HMMER techniques on those superfamilies which have few examples by expanding the training set using PSI-BLAST. We also intend to add PSI-BLAST to the website as another structural prediction technique.

Acknowledgment

We would like to thank Dave Goddeau for his help setting up the web services and Yu Zheng for providing code to visualize the BLOCK hits.

References

[1] S. K. Burley, S. C. Almo,, J. B. Bonanno, M. Capel , M. R. Chance, , et al., “Structural genomics: beyond the human genome project.” *Nat Genet*, vol. 23, no. 2, pp. 151-157, 1999.

[2] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, Z., et al., “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.” *Nucleic Acids Res* vol. 25, no. 17, pp. 3389-3402, 1997.

[3] S. K. Bateman, L. Coin, et al, “The Pfam protein families database”, Nucleic Acids Research 32 Database Issue: D138-41, 1999.

[4] S. Eddy, 1998. http://hmmer.wustl.edu.

[5] C. Burges, “A tutorial on support vector machines for pattern recognition”, *Data mining and knowledge discovery journal,* 1998.

[6] A. G. Murzin, S. E. Brenner, T. Hubbard and C. Chothia, “SCOP: a structural classification of proteins database for the investigation of sequences and structures”, *J. Mol. Biol.* 247, pp. 536-540, 1995.

[7] B. Logan, P. Moreno, B. Suzek, Z. Weng and S. Kasif, “Remote homology detection using feature vectors formed using alignments of small motifs”, 2002, 6th Annual International Conference on Computational Molecular Biology (RECOMB).

[8] B. Logan, P. Moreno, B. Suzek, Z. Weng and S. Kasif, “A study of remote homology detection”, Hewlett-Packard Labs Technical Report CRL-2001-5, 2001.

[9] J. M. Chandonia, et al., “The ASTRAL Compendium in 2004”, *Nucleic Acids Research*, vol. 32, pp. D189-D192, 2004.

[10] J. D. Thompson, D. G. Higgins, T. J. Gibson, “CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice”, *Nucleic Acids Res.* Vol. 22, pp. 4673-4680, 1994.

[11] S. S. Henikoff, and J. G. Henikoff, “Protein family classification based on searching a database of blocks”, *Genomics* vol. 19, pp. 97-107, 1994.

[12] W. S. Noble, “Support vector machine applications in computational biology”, in *Kernel Methods in Computational Biology*, B. Scheolkopf, K. Tsuda and J.-P. Vert, ed, MIT Press, 2004,

[13] C. Letondal, “A Web interface generator for molecular biology programs in Unix”, *Bioinformatics*, vol. 17, no. 1, pp. 73-82, 2001.

[14] http://ws.apache.org/axis

[15] A. Ben-Hur and D. Brutlag, “Remote homology detection: a motif based approach”, *Bioinformatics*, vol. 19, suppl. 1., pp. 26-33, 2003.