PROTEIN FUNCTION PREDICTION WITH AMINO ACID SEQUENCE AND SECONDARY STRUCTURE ALIGNMENT SCORES

Eser Aygün[1], Caner Kömürlü[2], Zafer Aydin[3] and Zehra Çataltepe[1]

[1] Computer Engineering Department and [2]Informatics Institute, Istanbul Technical University
Maslak, 34469, Istanbul, Turkey
[3]School of Electrical and Computer Engineering, Georgia Institute of Technology,
Atlanta, GA 30332
eser.aygun@gmail.com, komurluc@itu.edu.tr, aydinz@ece.gatech.edu, cataltepe@itu.edu.tr

ABSTRACT

In this study, we use amino acid sequence and actual or predicted secondary structure for protein function prediction. We investigate both sequence-sequence and Hidden Markov Model (HMM) profile-profile similarity measures that use both sequence and secondary structure information. Our findings on a data set consisting of 5 Gene Ontology (GO) Molecular functions and 785 sequences show that actual secondary structure can help for protein function prediction when it is used with sequence-sequence alignment scores. We did not find a consistent improvement in function prediction when predicted or actual secondary structure is used with profile-profile alignment scores.

1. INTRODUCTION

Previous Work: Sequence-sequence amino acid alignment scores have already been used for protein function prediction (see for example GOtcha [17] and GoFigure [13]). For fold recognition problem, profile-profile alignment scores have been shown to perform better than sequence-sequence alignment scores [6]. Wallqvist et al. [21] used predicted and actual secondary structure in sequence alignment and showed that secondary structure information contributes to the accuracy of fold recognition. Aydin et al. [3] reports similar results. In [20], it has been shown that using predicted or actual secondary structure in profile-profile hidden Markov model alignments has also potential to improve fold recognition.

Contributions of our method: We evaluate the effect of using actual or predicted secondary structure for alignment scores for Gene Ontology (GO) [5] function prediction. We perform our experiments on a data set of 785 protein sequences from 5 first level GO Molecular Function classes. We compute the sequence-sequence alignment scores as well as profile-profile HMM alignment scores with varying contributions of actual or predicted secondary structure. We determine the best secondary structure contribution to the alignment scores based on our data set. We evaluate performance of different methods by means of AUC (area under the ROC curve) as in [9] .

Results: Using the actual secondary structure in sequence-sequence alignment helps significantly (95% confidence) for function prediction in 3 out of the 5 GO classes we considered. Using predicted secondary structure in sequence-sequence or profile-profile HMM alignment scores, or using actual secondary structure in profile-profile HMM alignment scores, does not result in an improvement in function prediction.

2. METHODS

2.1 Data

2.1.1 Labels from GOA

To obtain a list of annotated proteins, we referred to Gene Ontology Annotation (GOA) project [5]. GOA provides GO assignments for the proteins of human, mouse, rat, arabidopsis, zebra fish, chicken and cow. It also provides a Protein Data Bank (PDB) [22] association file, which contains only the assignments for the proteins present in the PDB database. To be able to fetch sequence and structure information from PDB, we used the PDB association file.

2.1.2 Ontology Structure from GO

We projected GO molecular function terms onto the first level of GO molecular function hierarchy over "is-a" relations. As "is-a" relations are allowed to be many-to-many in GO hierarchy, a term may lead to more than one first level term. Also, a protein may be associated with more than one term if it is known that it has multiple functions. We captured both cases during the labelling process and we generated five dimensional target vectors for five first level molecular function terms: GO:0005198 (structural molecule activity), GO:0005215 (transporter activity), GO:0030234 (enzyme regulator activity), GO:0030528 (transcription regulator activity), GO:0060089 (molecular transducer activity) (Please see Table 1).

2.1.3 Clusterings from PDB

We applied PDB's scheme to remove sequence homologs. PDB provides several clusterings of proteins generated with CD-HIT or BLASTClust algorithms for different sequence identities. According to the scheme, only the best representative of each cluster is kept for a given clustering. Thereby, potential homologs are removed and non-redundant datasets are obtained. In this work, we used clusterings generated by BLASTClust for 30% and 40% sequence identities, and we used clusterings generated with CD-HIT for 50% sequence identity.

Table 1. Go classes and the number of sequences used for the experiments from each class.

GO term	No of Proteins
GO:0005198 (structural molecule activity)	171
GO:0005215 (transporter activity)	214
GO:0030234 (enzyme regulator activity)	127
GO:0030528 (transcription regulator activity)	208
GO:0060089 (molecular transducer activity)	119

2.1.4 Amino Acid Sequences and Secondary Structures from PDB

Last of all, we downloaded amino acid sequences and secondary structures of every protein via PDB web service. Eventually, we obtained three non-redundant datasets: NRB30, NRB40, and NRC50. We only report our results for NRB40 in this paper, however the results for NRB30 and NRC50 were similar.

The ratio of different secondary structure elements could affect function prediction success when secondary structure is predicted, because different types of secondary structure tend to be predicted with different accuracy [14]. In Figure 1, we show the ratios of secondary structure elements in our data set. Although the amount of beta sheets is less than the other types of secondary structure in general, there is no significant difference between the 5 function classes we work with.

2.2 Sequence Alignment

2.2.1 Parameters

In [21] secondary structure is used in the alignment score computation as follows:

$$w_{\alpha\beta} = \sum_{k}^{m_{ab}} (\alpha M^{aa}_{a_k,b_k} + \beta M^{ss}_{a_k,b_k}) + N_o g_o + N_e g_e \quad (1)$$

where w denotes the similarity score between two aligned sequences, M^{aa} denotes the amino acid similarity matrix and M^{ss} denotes the secondary structure similarity matrix, m_{ab} is the number of paired elements along the alignment, α and β denoted the weighted importance of amino acid and secondary structures. Finally, N_o denotes the number of gap openings and g_o denotes the gap opening penalty, N_e denotes the number of gap elongation and g_e denotes the gap extension penalty. Since Wallqvist et al. respect the relation, $\alpha = 1 - \beta$, in their experiments, we also kept this relation.

The following parameters were used for the alignment: Linear gap penalty; Gap extend: 8; AA scoring matrix: BLOSUM50; SS scoring matrix used in [21], show the full matrix; α (secondary structure contribution): 0.00, 0.25, 0.50, 0.75, 1.00.

2.2.2 Functional Identity Based on Sequence and Secondary Structure Alignment Scores

As shown in Figure 2, we computed the probability that two sequences with similarity above a threshold fall in the same function class. For our dataset, we see that if the secondary structure global alignment score is above 20%, then we can classify two sequences in the same function category. On the other hand for 40% sequence similarity, the same result holds. We could not use the global alignment scores for function prediction, because all sequences in our data set had sequence identities below 40%.

2.3 Hidden Markov Models

Hidden Markov Models are finite state machine based models and they are used to represent sequential data which are extracted from a probabilistic system. Hence, HMM's can be used in biological sequence analysis with such a point of view: HMM's can be used to represent matchings, insertions, deletions and unmatchings of sequences which are concepts of homology. HMM's can be used for both pairwise similarity and profile-profile similarity computation [7]. In profile-profile similarity computation, sequence profiles which measure how frequent corresponding residues are in a multiple sequence alignment, can be used. Since HMM's can model gaps and insertions, better than profiles, PFAM [18] database which includes HMMER profiles was established [8]. State-of-the-art tools for profile HMM extraction are HMMER [8], HMMSTR [4] and HHsearch [20]. Among these tools, HHsearch which is declared to be able to find distant homologies using profile-profile comparison, was preferred in this work. HHsearch has a good performance in homology detection [6, 20] and it allows use of actual or predicted secondary structure in the alignment.

2.3.1 HHsearch Method

HHsearch detects homology using pairwise alignment of profile-HMM's. A profile-HMM is a template that represents how much conserved the residues in a sequence are. For some proteins of similar function, some residues are more likely to happen at certain regions and some other residues are more likely to be deleted at some other regions. These patterns can be detected in a multiple sequence alignment which are used in profile-HMM's. A profile-HMM is built via modelling locations of some residues that are likely to be in, with M and I states, and modelling gaps/deletions with state G. Consequently, the aligned version of a sequence from a multiple sequence alignment block can be produced by a path on profile-HMM of that alignment.

2.3.2 Similarity Measuring with HHsearch

HHsearch searches homology between a given protein and a protein database [20]. As the output, it produces a score list measuring the homology of the given protein to the dataset proteins above a threshold. The main idea behind this method is to use this homology score list as a similarity score list. Note that homology is a metric which measures, how many common residues do a couple of amino acid sequences have according to the evolution of the protein families. Since the functional sections are conserved during evolutionary process, the homology between two proteins points to the functional closeness of those proteins. From a coarse point of view, it can be said that, the more two proteins are similar, the more they are functionally related [19].

2.3.1 HHsearch Similarity with Amino Acid Only

HHsearch detects homology over a multiple sequence alignment. It uses PSI-BLAST [2] to obtain sequence alignments. We first produced a multiple sequence alignment database using PSI-BLAST. In this way, an alignment output per protein for whole data set was obtained. Next, an HMM per protein was produced. Then all these HMM's were collected and used to form the HMM database. HHsearch detects homology through this HMM database given an input protein's HMM. Hence for every protein, we obtained a score table. HHsearch, with default e-value, stores the scores of only 10 most similar proteins for the query protein. In this study, we use similarities between a protein and proteins in the training set as in SVM-PAIRWISE [16]. Hence we needed similarities for each protein pair. So we moved e-value to higher levels so that every score can exceed this threshold. Finally, we came up with score values for each protein. We extracted similarity matrix that can be used in conventional classifiers such as KNN or SVM [1], as told in section 3.1.

2.3.2 HHsearch Similarity with Amino Acid and Secondary Structure

HHsearch lets secondary structure information to be taken into account in homology detection. It uses secondary structure information for better HMM computation. Hence, HHsearch uses the secondary structure just after the alignment of each sequence and before the HMM extraction. HHsearch can be used directly with the PSIPRED tool [10]. By default, HHsearch uses predicted secondary structure. We implemented a converter tool which lets HHsearch use *actual* secondary structure, too. As a consequence, we made experiments with both predicted and actual secondary structure and compared these two cases.

In the alignment computation, HHsearch includes the secondary structure information according to:

$$S(\alpha) = S_{AA} + \alpha S_{SS} \qquad (2)$$

where S_{AA} the amino acid alignment score for two HMMs for two certain states, S_{SS} is the secondary structure alignment score for those states, α is the contribution of the secondary structure in the joint similarity and $S(\alpha)$ denotes the joint similarity. Soeding noted that $\alpha = 0.15$ gave good results for fold recognition. As a part of this work, the role of α was observed for 0, 0.25, 0.5, 0.75, 1. [20]

3. EXPERIMENTAL RESULTS

3.1. Classification

After the similarity scores are computed according to formulas (1) or (2), by using sequence-sequence or HMM profile-profile comparison and for different values of alpha, we used these similarities to produce classifiers for protein function prediction. As in [16], we used pairwise alignment scores between a protein and all proteins in the training data as the input vector.

The secondary structure which was in DSSP format was converted to HEL (H: α-helix, E: β-sheet, L: loop (coil)) format according to [11]. For sequence-sequence aligment, balign program, written by Eser Aygun of Istanbul Technical University, was used.

Since the dataset is multi-labeled, i.e. a protein can have multiple functions, one-against-all classification scheme is used. The function prediction must be performed for each function class individually. As a consequence, first of all for each function a different classifier is produced, resulting in 5 classifiers. For each classifier to be trained, all available, and at each step the data is separated into 2 parts, training set and test set. As the classifier, 1NN classifier is preferred since it has already produced good results [15] and since it runs much faster than other classifiers such as SVMs. The algorithm is tested with 10-fold cross-validation. The distribution of the samples to the classes at each fold is kept same as it is in the original dataset. To evaluate the results, AUC (Area under the ROC curve) is used of ROC's plot for each class are computed. Figure 3 shows the average AUC values for each value of alpha and for each of the five functions for sequence-sequence alignment scores.

3.2 Smith-Waterman Alignment Scores

Mean AUC values over 10 cross validation runs are computed for each function and for each value of α ($\alpha = 0.0, 0.25, 0.50, 0{,}75, 1$). As clearly seen in Table 2, contribution of secondary structure at level of 25% increases classification accuracy for each class (Please see Figure 3 and Table 2). We also performed a paired-t-test between the 10-fold-cross validation AUC values with $\alpha = 0.0$ and $\alpha = 0.25$. We found out that at 89% significance level $\alpha = 0.25$ gave larger AUC value for all 5

GO function categories (Please see Table 4). Predicted secondary structure does not cause an increase in accuracy except for “transcription regulator activity” (GO: 0030528) class (Please see Table 2).

In addition, it can be said that, AUC values obtained on the order of 90% points to the success of using sequence-sequence alignment scoring matrix with NN algorithm for function prediction.

3.3 HHsearch Scores

Similar to experiments with sequence-sequence alignment scores, for using profile-profile alignment scores also the mean AUC values over 10 cross validation runs are computed for each function and for each value of α ($\alpha = 0.0, 0.25, 0.50, 0{,}75, 1$). (Please see Table 3).

As seen in Table 3, using secondary structure (actual or predicted) did not result in a consistent improvement for function prediction. We also observed that profile-profile alignment scores performed worse than sequence-sequence alignment scores. This contradicts the results obtained for fold recognition by [21]. Through the use open source HMM software and bigger datasets, we are planning to work more on this issue in the near future.

4. CONCLUSIONS AND FUTURE WORK

In this work, we examined if secondary structure information can be used to help function prediction. We used Smith-Waterman sequence-sequence alignment scores and HHSearch profile-profile alignment scores together with nearest neighbour classifier. We found out that Smith-Waterman scores and actual secondary structure could improve function prediction, whereas predicted secondary structure does not. We also found out that HHSearch and secondary structure, whether actual or predicted, does not help with function prediction.

As future work, we are planning to work with other open source HMM tools such as HMMER [8] and SAM [12]. The dataset used in this study consisted of only 5 GO classes, we are planning to experiment on more GO classes.

We are also planning to work with other classifiers, including SVMs.

ACKOWLEDGEMENTS

Authors Aygün and Cataltepe were supported by Tubitak EEEAG Research Project 105E164. Author Kömürlü was supported by Tubitak BIDEB 2228 scholarship. We would like to thank Dr. Hakan Erdoğan of Sabanci Un. for useful discussions on HHSearch, and Dr. Ozlem Keskin of Koc University because of her help with the data set. Aslı Filiz of Istanbul Technical University prepared the figure showing the ratios of secondary structure elements in each class.

REFERENCES

[1] E. Alpaydin, *Introduction to Machine Learning*. The MIT Press, Oct. 2004.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman., “Basic Local alignment Search Tool,” *Journal of Molecular Biology,* (190)215,403-410, 1990.

[3] Z. Aydin, H. Erdogan and Y. Altunbasak, “Protein Fold Recognition using Residue-Based Alignments of Sequence and Secondary Structure,” *Acoustic, Speech and Signal Processing 2007*, Vol.1, 15-20 Apr. 2007, pp. I-349-I-352.

[4] C. Bystroff, V. Thorsson, and D. Baker “HMMSTR: a hidden markov model for local sequence structure correlations in proteins,” *Journal of Molecular Biology*, Vol. 301, issue 1, pp. 173–190, Aug. 2000.

[5] E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez and R. Apweiler, “The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology,” *Nucleic Acids Research*, vol. 32, no. 1, pp. D262-D266, 2004.

[6] J. Cheng and P. Baldi, “A machine learning information retrieval approach to protein fold recognition,” *Bioinformatics*, 22(12):1456–1463, 2006.

[7] R. Durbin, S. Eddy, A. Krogh and G. Mitchison, *Biological Sequence Analysis Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[8] S. R. Eddy, “Profile hidden Markov models,” *Bioinformatics*, vol. 14, pp. 755-763, 1998.

[9] L. J. Jensen, R. Gupta, H.-H. Stærfeldt and S. Brunak, “Prediction human protein function according to Gene Ontology categories,” *Bioinformatics,* Vol. 19 no. 5, pp. 635-642, 2003.

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

[11] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers,* Vol. 22, No. 12, pp. 2577-637, 1983.

[12] K. Karplus, et al., ”Hidden Markov Models for detecting remote protein homologies,” *Bioinformatics*, 14, : 846–856, 1998.

[13] S. Khan, G. Situ, K. Decker and C. J. Schmidt, “GoFigure: Automated Gene Ontology annotation,” *Bioinformatics*, Vol. 19 no. 18, pp 2484-2485, 2003.

[14] D. Kihara, "The effect of long-range interactions on the secondary structure formation of proteins," *Protein Science*, 14 (8), pp1955-1963, 2005.

[15] A. Kocsor, A. Kertesz-Farkas, L. Kajan, and S. Pongor, "Application of compression-based distance measures to protein sequence classification: a methodological study," *Bioinformatics*, 22(4):407–412, 2005.

[16] L. Liao and W. S. Noble, "Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships," *Journal of Computational Biology*, 10(6):857–868, 2003.

[17] D. M. A. Martin, M. Berriman and G. J. Barton, "GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes," *BMC Bioinformatics,* 2004, 5**:**178, 2004.

[18] The Pfam Protein Families Database, *Oxford Journals*, 2002.

[19] B. Rost, "Review: Protein Secondary Structure Prediction Continues to Rise," *Journal of Structural Biology,* Volume 134, Issues 2-3, pp. 204-218, May 2001.

[20] J. Soeding, "Protein homology detection by hmm-hmm comparison," *Bioinformatics*, 21:951–960, 2005.

[21] A. Wallqvist, Y. Fukunishi, L. R. Murphy, A. Fadel and R. M. Levy, "Iterative sequence/secondary structure search for protein homologs: comparison with amino acid sequence alignments and application to fold recognition in genome databases," *Bioinformatics*, 16(11):988-1002, Nov. 2000.

[22] J. Westbrook , Z. Feng , L. Chen, H. Yang and H. M. Berman, "The Protein Data Bank and structural genomics," *Nucleic Acids Research,* 31: 489-491, 2003.

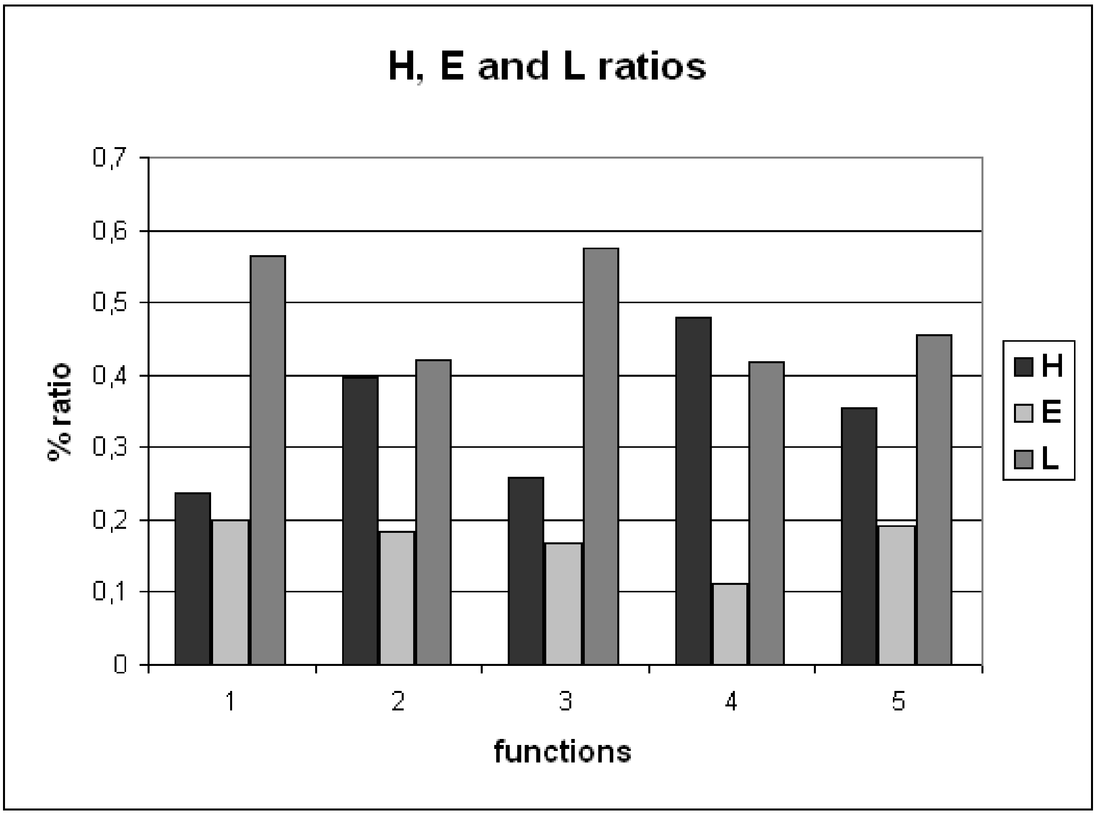

Figure 1. The ratios of each of the secondary structure elements in the protein sequences in the data set.

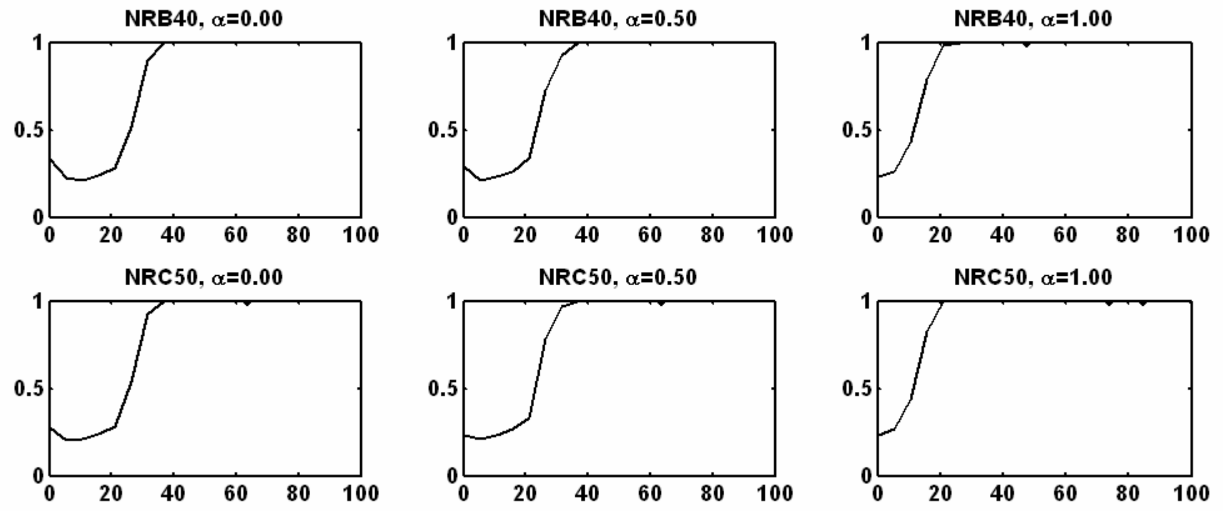

Figure 2. The probability that two sequences with amino acid sequence (alpha=0), secondary structure sequence (alpha=1) identity above a threshold belong to the same GO function category.

Table 2: Using Smith-Waterman and different contributions of secondary structure, the AUC values for each of the GO categories.

		α=0.00	α=0.25	α=0.50	α=0.75	α=1.00
Actual SS	GO:0005198	89.0%	90.8%	90.7%	84.2%	75.8%
	GO:0005215	90.2%	92.0%	90.7%	87.7%	82.1%
	GO:0030234	89.9%	91.4%	90.9%	86.3%	80.2%
	GO:0030528	89.5%	91.6%	91.4%	86.8%	80.4%
	GO:0060089	90.5%	91.4%	91.0%	88.2%	84.0%
Predicted SS	GO:0005198	89.0%	87.2%	86.0%	76.3%	71.3%
	GO:0005215	90.2%	88.6%	87.5%	84.7%	78.9%
	GO:0030234	89.9%	88.3%	87.0%	82.6%	81.1%
	GO:0030528	89.5%	90.0%	86.8%	82.5%	79.6%
	GO:0060089	90.5%	89.5%	87.5%	85.7%	82.2%

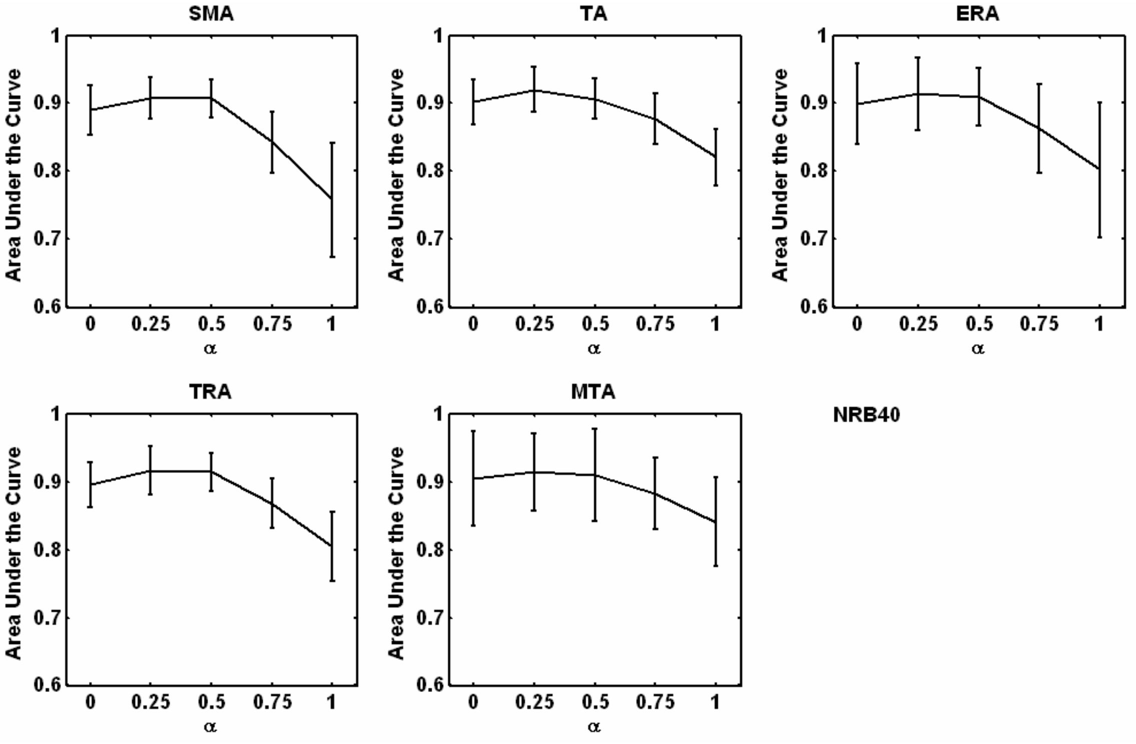

Figure 3. ROC curves for actual secondary structure and Smith-Waterman alignment scores.

Table 3: Using HHSearch and different contributions of secondary structure, the AUC values for each of the GO categories.

		α=0.00	α=0.25	α=0.50	α=0.75	α =1.00
Actual SS.	Class 1	80.0%	75.0%	70.0%	76.0%	80.0%
	Class 2	84.0%	83.0%	79.0%	77.0%	80.0%
	Class 3	84.0%	84.0%	83.0%	87.0%	83.0%
	Class 4	85.0%	83.0%	80.0%	81.0%	84.0%
	Class 5	83.0%	82.0%	82.0%	81.0%	82.0%
Predicted SS	Class 1	76.0%	78.0%	78.0%	74.0%	71.3%
	Class 2	81.0%	82.0%	81.0%	82.0%	78.9%
	Class 3	82.0%	83.0%	83.0%	83.0%	81.1%
	Class 4	83.0%	82.0%	82.0%	82.0%	79.6%
	Class 5	84.0%	82.0%	82.0%	83.0%	82.2%

Table 4: t-test results for the hypothesis that AUC values of α=0.25 are better than α=0.00.

	p-value	Conf. Intv.
Class 1	3%	-0,0031
Class 2	1%	-0,0059
Class 3	11%	0,0058
Class 4	0%	-0,0102
Class 5	7%	0,001