

A Data Fusion Approach in Protein Homology Detection

Aydın Can Polatkan^{1,2}, Hasan Oğul² and Hayri Sever³

¹*Wilhelm-Schickard-Institute, Universität Tübingen*

²*Department of Computer Engineering, Başkent University*

³*Department of Computer Engineering, Çankaya University*

polatkan@gris.uni-tuebingen.de, hogul@baskent.edu.tr, sever@cankaya.edu.tr

Abstract

The discriminative framework for protein remote homology detection based on support vector machines (SVMs) is reconstructed by the fusion of sequence based features. In this respect, n-peptide compositions are partitioned and fed into separate SVMs. The SVM outputs are evaluated with different techniques and tested to discern their ability for SCOP protein super family classification on a common benchmarking set. It reveals that the fusion approach leads to an improvement in prediction accuracy with a remarkable gain on computer memory usage.

1. Introduction

High-throughput genome sequencing projects have been resulted in the accumulation of a large amount of raw sequence data in many databases. Owing to the experimental complications and obstacles in the structural and functional analysis of proteins, the amount of discrepancy between the number of known protein sequences and the number of experimentally determined structures has steadily increased in recent years. This situation has given occasion to the emergence of computational tools and methodologies that make automated annotations on structure and function of proteins using the sequence information available in public databases.

Two proteins are said to be homolog if they share a common evolutionary origin. Homology information is important in that it may imply a common structure and function between two proteins. Since we are often supplied, only with the sequence information, inferring homology using solely the sequence has been one of the central problems in computational biology. If we are able to find some number of similar proteins whose structural and functional analyses are already completed, the target protein can be easily annotated using the assumption that the sequence is the main determinant of the structure.

However, there are two main problems with this argument. First, the target protein may be entirely new and its structure is different from all of the proteins available in the databases. Second, in spite of the weak similarity between two protein sequences they may still have evolutionary relationships. The first problem is a bottleneck of computational biology and there is no method that works well at the moment. The second problem is known as *remote homology detection* problem, and various methods have been proposed in recent years. In spite of several successful attempts, they are either computationally inefficient or insufficient to work for all cases. The previous methods can be grouped into three stages; pair wise methods, generative methods and discriminative methods.

The early methods for homology detection were based on the pair wise sequence similarity inferred by dynamic programming based sequence alignment (Smith and Waterman, 1981). While the dynamic programming method finds an optimal score for similarity according to a predefined objective function, it suffers from long computation times for relatively long sequences. To speed up the alignment, some heuristic methods, such as BLAST (Altschul et al. 1990), have been developed to find a near-optimal alignment within a reasonable time. The general assumption is that two proteins are homolog if the sequence identity (the percentage of identical residues after the alignment of two sequences) is above 40%. The problem with pair wise sequence alignment is biological inaccuracy of evaluation with respect to only one known protein homolog. Considering only one protein to annotate a newly sequenced protein may lead to biologically uncertain results. This uncertainty comes from the fact that the *twilight zone* of sequence alignment sets a boundary for confidence levels for the detection of evolutionary relatedness of proteins (Rost, 1999). In most alignments this twilight zone falls between 20-40% sequences identities. Despite two proteins are not similar in terms of their sequences, i.e. the sequence identity is below 40%, they may still

share some important structural or functional features, which actually refers to remote or distant homology between proteins.

To take apart from the twilight zone of pair wise sequence alignment, family based comparisons have been proposed (Grundy, 1998). To improve the sensitivity of homology results, a representative set of sequences from the family is incorporated into the comparison, that is, the new protein is aligned concurrently with all (or some) of the protein sequences in a specified family. The utilization of multiple comparisons has increased three times the sensitivity of homology detection compared to pair wise comparisons (Park et al., 1998). Indeed, most of the proteins are classified within a protein family in available databases and the actual annotations are made over those families, thus, it seems more appropriate to use all family knowledge in homology modelling.

Additional accuracy can be gained by searching the available database for homology and refining the central profile model, iteratively. SAM-T98 method is an example of iterative family refinement methods (Karplus, Barrett and Hughey, 1998). In this method, sequences are first aligned to an initial model which is constructed based on some background distributions and then the model is improved iteratively by aligning the sequences to the current model to which the new statistical results are incorporated.

Another iterative method, called PSI-BLAST deploys BLAST search and refines the results iteratively (Altschul et al., 1997). Integrated sequence/structure alignments are iteratively performed for the search of homology in another method (Walqvist et al., 2000).

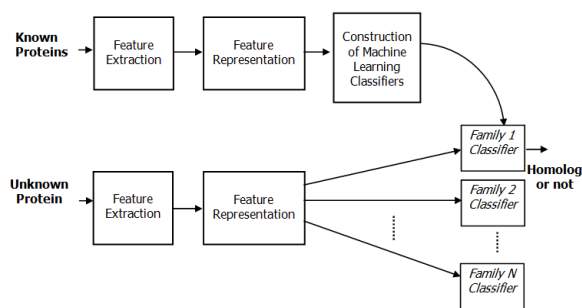


Figure 1. Discriminative homology detection model

The main problem with generative approaches is the fact that they produce so much false positives that they report a number of homolog though they are not. For this reason, the recent works on remote homology detection have begun to use a discriminative framework to make separation between homolog

(positive) and non-homolog (negative) classes. In contrast to generative methods, the discriminative methods focus on learning the combination of features that discriminate between the classes. These methods attempt to establish a model that differentiates between positive and negative examples. In other words, non-homolog is also taken into account.

The current methods using the discriminative approach differ in the feature extraction methods, the feature representation schemes and the type of the machine learning classifiers they have used. Among k-nearest neighbour method, neural networks and support vector machine, the last one has been reported as outperforming the others in many applications concerning with protein classification (Liao and Noble, 2003; Ding and Dubchak, 2001).

Discriminative methods are more successful than generative methods in terms of separation accuracy between true positives (homolog that are correctly predicted) and false positives (non-homolog that are incorrectly predicted as homolog). However, the training and testing phases require so much time with conventional workstations, which makes them inappropriate to use in practice. Thus, more efficient methods are required that preserve the classification accuracy.

First discriminative approach (SVM-Fisher) represents each protein by a vector of Fisher scores extracted from a profile Hidden Markov model constructed for a protein family and utilizes SVMs to classify the protein with those feature vectors (Jaakola, Diekhans and Haussler, 2000). A recent and more successful work, called SVM-Pair wise (Liao and Noble, 2003), combines the sequence similarity with the SVMs to discriminate between positive and negative examples. In SVM-Pair wise, both the training and test sets include positive and negative examples. This method was tested for dynamic-programming-based alignment scores and BLAST scores. Note that the latter one is referred as SVM-BLAST in the following sections. SVM-Pair wise approach is among the best methods in terms of accuracy, but it suffers from computational inefficiency since the alignment takes too much time for long sequences. Another drawback of this approach is that the alignment may force some residues to match even if they are evolutionary not related.

One of the effective approaches, developed to for these problems is protein n-peptide combinations. Approach frames each n-peptide long sub-sequences in the sequence as a percentage. To diminish the space complexity, for the increasing values of n, to amino acid alphabet is reduced regularly in order to resultant vector's adaptation to the recent memory resources (Ogul and Mumcuoglu, 2007).

In this solution, all feature inputs were given to a single classifier. In this paper, these feature inputs are classified into specific significant groups, according to the n-peptide compositions and reduced amino alphabets. These groups are given to several different classifiers to achieve a data fusion approach with a few techniques that are wandering in the narrowed search space by abstraction. The aim is to have better results with techniques that are converging in exact and leading to different regions of a solution. In that approach, to evaluate the output values of different classifiers, various cases like averaging, weighted averaging and choosing the most successful one in the training set are compared.

Each of these methods was tested on remote homology detection problem which is one of the major and actual problems of computational biology and results are presented relatively.

2. Methods

To discriminate between positive and negative examples SVMs are used. To train the SVMs, open-source software called SVM-Gist which is available at <http://www.cs.columbia.edu/compbio/svm>, is used. In the SVM-Gist software, a kernel function acts as the similarity score between pairs of input vectors. The base kernel is normalized in order to make that each vector has a length of 1, in the feature space, that is,

$$K(X, Y) = \frac{X.Y}{\sqrt{(X.X)(Y.Y)}}$$

where X and Y are the input vectors, K(.,.) is the kernel function, and “.” denotes the dot product. This kernel is then transformed into a radial basis kernel $K'(X, Y)$, as follows:

$$K'(X, Y) = e^{-\frac{K(X, X) - 2K(X, Y) + K(Y, Y)}{2\sigma^2}} + 1$$

where the width σ is the median Euclidean distance from any positive training example to the nearest negative example.

Each protein must be represented by a fixed-length feature vector to be fed into a machine learning classifier. We represent proteins by their n-peptide compositions with decremented amino acid alphabets. For each value of n, corresponding feature vector contains the fraction of each possible n-length substring in the sequence.

We use the reduced amino acid alphabets provided by (Murphy et al., 2000) in our method. These

alphabets have been produced using statistical techniques based on the information of certain BLOSUM matrices and justified by well-known biochemical amino acid classes.

In this approach, instead of sending the all input vectors which has all the features of n-peptide combinations to just one classifier, feature inputs are divided into significant groups. Afterwards these groups are sent to different classifiers and by this way a data fusion is tested.

1) Averaging (AVG): Output values of different SVM classifiers are averaged as if it's the output of only one SVM classifier. According to these values classification is actualized.

$$D = \frac{(dSVM_1 + dSVM_2 + \dots + dSVM_k)}{k}$$

2) Weighted Averaging (WAVG): By assigning a weight to each SVM classifier, output values of those SVM classifiers are collected by averaging by weight. These values are accepted as an output of just one SVM classifier and the classification process is actualized. The weights are assigned according to the classifiers success ratio on training sets. By this way most successful classifier on the training sets is ensured to be the most effective on the any of the test sets than the other classifiers.

$$D = \frac{(w_1 * dSVM_1 + w_2 * dSVM_2 + \dots + w_k * dSVM_k)}{(w_1 + w_2 + \dots + w_k)}$$

3) Picking the most successful on the training sets (MAX): After running each SVM classifier, before collecting the output values, the most successful classifier is picked and ran on the test sets. The result of the SVM was collected as true and the classification process is actualized.

$D = dSVM_x$, (x is the most successful classifier on the training set.)

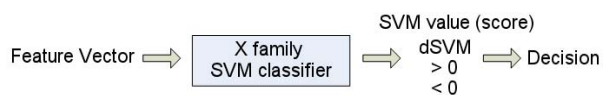


Figure 2. Classifier model for family X

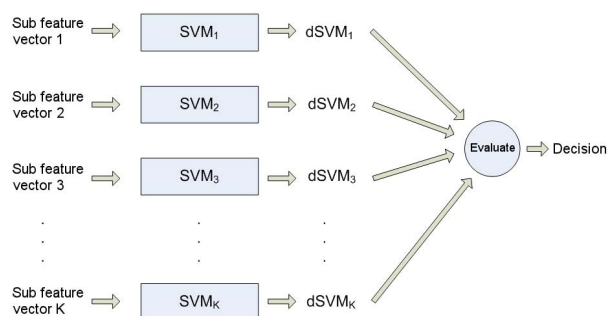


Figure 3. Proposed classifier model

The methods were tested to discern their ability to classify proteins into families on a subset of the SCOP family database (Murzin et al., 1995). Remote homology is simulated by excluding same family members from the training set and leaving proteins from same super family in the positive set. To achieve a fair comparison it is practiced on same experimental setup as which is practised for SVM-Pairwise and SVM-BLAST. The results are considered on account of certainty and efficiency.

The experiments are performed on a subset of the SCOP1.53 database including no protein pair with a pair wise similarity higher than an E-value (Altschul et al. 1990) of 10^{-25} . The training and test sets were separated as done in Liao and Noble's works resulting with 51 families to test.

As it is done in every classification process, methods of protein family classification are also struggle to balance between sensitivity (ignoring false positives) and specificity (selecting true positives). Before sensitivity and specificity evaluation, ROC score receiver operating characteristics score (Gribskov and Robinson, 1996) is used because of unequal distribution of positive and negative examples in the data set. This is a way to describe the whole set of scores with a single number. When the ROC score is 1 (one) it means that the positives are separated from negatives perfectly. When it is 0 (zero), it means that there are no positives that exist.

3. Results

Classification tests on the proteins taken from of SCOP family database are completed. After the tests of different methods applied, the ROC scores of the families are computed one by one. SVM performances of all methods on families are given in Table 1 and 2.

Feature vectors are extracted from those sequences by fronting on the n-peptide compositions within the sequence, in order to send to the classifiers. The SVM

performances of the method defined above are given in the Table 1.

Table 1. ROC Scores (n-peptide compositions)

ROC	N=1	N=2	N=3	N=1-3	AVG	WAVG	MAX
Average	0,8091	0,8464	0,7807	0,8741	0,8623	0,8672	0,882
Std.							
Dev.	0,1386	0,1328	0,182	0,1358	0,1318	0,1285	0,1103
Max	0,9963	0,998	0,9899	0,9989	0,999	0,999	0,998
Min	0,2958	0,3943	0,2957	0,2314	0,284	0,287	0,3943

As a different method for extracting the feature vectors, amino acids in the protein sequences are grouped all together. In this case, feature vectors are appeared after grouping the amino acids in the protein sequences which belongs to the families. In Table 2, SVM performances of such feature vectors sent the classifiers are given.

Table 2. ROC Scores (amino acid grouping)

ROC	$\Sigma 20$	$\Sigma 15$	$\Sigma 8$	$\Sigma 20-8$	ΣAVG	$\Sigma WAVG$	ΣMAX
Average	0,8091	0,7588	0,5645	0,8086	0,7968	0,8069	0,8339
Std. Dev.	0,1386	0,1592	0,19	0,1473	0,152	0,1476	0,1297
Max	0,9963	0,9967	0,9636	0,9996	0,9967	0,9963	0,9967
Min	0,2958	0,2147	0,2442	0,2273	0,2048	0,2069	0,2958

Second column in Table 2 which expresses the SVM performances of $\Sigma 20$ has equal values to Table 1's N=1 titled column, because both of the projections are synonymous. Here alphabet has a size of 20 and there isn't any grouping done because of all these reasons the feature vectors all has a size of 20.

Third column of Table 2 gives the SVM performances for $\Sigma 15$. This alphabet consists of 15 elements. As we can easily understand here amino acids are grouped in protein sequences. While extracting the feature vectors, similar amino acids are taken into account as they are the elements of the same dimension and computed in the same way. Four elements of the alphabet are united under one letter and all proteins are expressed like it. Fourth column also groups the letters.

Last which is the fifth column is a little bit different of the priors. Here alphabets with 20, 15 and 8 letters are grouped under the same idea but altogether. In this case distinct groups of those alphabets are united together to form a feature vector which has totally 28 dimensions. Protein sequences of the families are scaled in to one unit by different interpretation.

In last four methods given about extracting feature vectors; grouping amino acids are introduced by decrementing the letters in the alphabet, this decrease is provided by collecting more than one letters to symbolize just one in the alphabet.

Up to this point, how to extract the feature vectors for the protein sequences of families and SVM

performances of these feature vectors extracted from different methods are given. For all those eight methods given above, the approaches are all same, after the extraction of feature vectors, they are sent to one classifier and results are received by this way. In order to improve these results, instead of just using one classifier, n-peptide compositions are divided into significant groups to be sent more than one classifier.

In this direction, extracted multi dimensional feature vectors with the group numbers are sent to more than one classifier by this way a data fusion approach is achieved. This approach basically uses the model in Figure 2. Feature vectors are sent to classifiers in groups and by computing each classifier, classification score with evaluation methods, ROC scores for remote homology detection are computed.

Using the proposal model in this paper, three different methods are applied on the compositions. These ones as given in the table form are, averaging (AVG), weighted averaging (WAVG) and maximising (MAX). According to the model, feature vectors belonging to the n-peptide compositions that have the number more than one, are sent to the classifiers at the same time, three different classifiers are used for the 1-peptide, 2-peptide and 3-peptide composition that used in the tests. At the end of the classification process, all classification scores are enrolled in the evaluation process and ROC scores are calculated for the selected method. After the tests made on SCOP dataset, ROC scores for the families are computed and these assets are given at the end of the table in order as averaging, weighted averaging and maximum. Proposed model, similarly applied on the 1-peptide compositions which are held by decrementing the amino acid alphabets and averaging, weighted averaging and maximising methods are tested. ROC score performance results are given in order in the last three columns.

In this paper, the results achieved by proposed, tested and applied model, methods and classifiers for homology detection are given in the tables. In Table 1 and Table 2, performance scores of the families and average, standard deviation, minimum and maximum assets are presented.

Table3. T-test scores of methods used for Remote Homology Detection (n-peptide compositions)

	N=1	N=2	N=3	N=1-3	ORT	AORT	MAX
N=1		1,83E-02	1,77E-01	7,38E-06	5,47E-07	1,23E-07	9,76E-08
N=2			4,01E-04	4,01E-04	6,46E-02	1,76E-02	2,24E-04
N=3				4,35E-06	2,59E-05	1,37E-05	5,14E-07
N=1-3					6,31E-02	2,82E-01	2,98E-01
ORT						9,14E-03	8,56E-03
AORT							2,05E-02
MAX							

To explore the statistical significance of differences between the results, paired-samples T-tests were carried out between remote homology detection methods used in this paper with a *p*-value threshold of 0.05 (Table 3). When the table is observed, it is clearly seen that MAX is statistically significant than all methods except N=1-3. Even if MAX and N=1-3 methods have different formats, the mathematical act in both methods are similar to each other and this is the reason that MAX and N=1-3 are not statistically significant against each other.

When standard deviation and the worst scores of ROC performances computed by the methods are evaluated for the families of SCOP dataset, ΣMAX method stands for the best proposed case for reduced amino acids alphabets approach. As seen in Table 2, method has the lowest standard deviation score, which points out the reliability of the method, when ΣMAX method is compared with the other in the Table 2. It proves that the method is the most durable against errors although changes in the family characteristics exist.

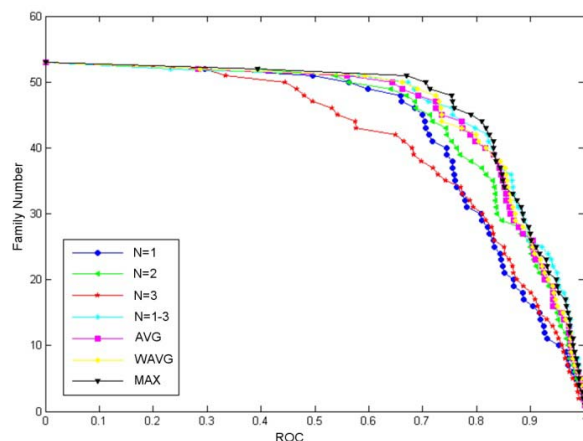


Figure 4. Relative Performances of Remote Homology Detection Methods (n-peptide compositions)

Seven methods for remote homology detection basically considering n-peptide compositions are performed on the families of the SCOP dataset. By means of the ROC – family number graph in figure 4, these methods are compared in respect of their relative performances. Each curve in the graph expresses a method. Each index on the graph represents the points where the methods ROC score exceeds the threshold value or in other words represents the families in the dataset. Presence of 51 families in the dataset causes 51 indexes on each curve. For each drawing, the highest curve means faultless homology detection performance. When we took a glance at the curves, it is

seen that the highest curve is MAX curve and the lowest one is N=3 curve. As a result, MAX method is easily be inferred to be better than the others.

4. Conclusion

Protein classification and one of its subsequent problems, remote homology detection, have been extensively studied over the last decade. It is inferred from the recent works that the discriminative methods outperform the others in terms of prediction accuracy with some concession in computational complexity. Two main issues have discussed in those studies. First, SVM is agreed to be most accurate classifier among all alternatives. Second, three of the feature representation schemes (pair wise similarity scores, motif content and n-peptide compositions) are considered to be the most proper choices with no significant superiority to each other. In the meanwhile, recent studies on pattern classification and data mining systems have shown that the fusion of input features is resulting with improvement in both accuracy and computational complexity of classification systems.

This study is an implementation of data fusion approach for remote homology detection based on SVMs with n-peptide compositions. It is shown that the fusion of compositional features can be resulted with an improvement in prediction accuracy. Another advantage of using this approach is considerable gain in memory space complexity. It can easily be argued that a proportional amount of gain is attained to the number of classifiers used for each of the partitioned input vector. Therefore, the use of data fusion is noticed to provide an effective solution for practical implementation and wide usage of discriminative systems for protein classification.

Among three evaluation techniques for separated classifier outputs, the one that picks up the result of most reliable classifier, for which the reliability is resolved during the training stage, performs best on the given set. It also outperforms the predictions based on single classifiers. Since each SVM trained from a separate input vector produces a discriminant score on a distinct scale, averaging techniques have always some component than can not be computationally proven. This is the possible reason for that they perform worse than the solution that intuitively selects the most reliable classifier.

Apart from its results on remote homology detection, this work also tells that data fusion method promises better results in other SVM based classification problems. Testing of the methods proposed in this paper on other protein recognition problems such as sub cellular localization prediction,

operand prediction and functional annotation would also be applicable as well.

5. References

- [1] Altschul, S., Gish, W., Miller, W., Myers, E. W., Lipman, D., (1990). A basic local alignment search tool, *J.Mol.Biol.*, 251, 403-410.
- [2] Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D., (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, 25, 3389-3402.
- [3] Ding, C. ve Dubchack, I., (2001). Multi-class protein fold recognition using support vector machines and neural networks, *Bioinformatics*, 17, 349-358.
- [4] Gribskov, M. ve Robinson, N. L., (1996), "Use of receiver operating characteristic analysis to evaluate sequence matching", *Comput. Chem.*, 20, 25-33.
- [5] Grundy, W. N., (1998), "Homology Detection via Family Pairwise Search", *J.Comp.Biol.*, 5, 479-492.
- [6] Jaakola, T., Diekhans, M., Haussler, D., (2000) "A discriminative framework for detecting remote homologies", *J.Comput.Biol.*, 7, 95-114.
- [7] Karplus, K., Barrett, C., Hughey, R., (1998). Hidden Markov Models for detecting remote protein homologies, *Bioinformatics*, 14, 846-856.
- [8] Liao, L. ve Noble, W. S., (2003), "Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships", *J.Comput.Biol.*, 10, 857-868.
- [9] Murphy, L. R., Wallqvist, A., Levy, R. M., (2000) "Simplified amino acid alphabets for protein fold recognition and implications for folding", *Protein Eng.*, 13, 149-152.
- [10] Murzin, A. G., Brenner, S. E., Hubbard, T., Chothia, C., (1995), "SCOP: A structural classification of proteins database for the investigation of sequences and structures", *J.Mol.Biol.*, 247, 536-40.
- [11] Oğul, H., Mumcuoğlu, E., (2007), "A discriminative method for remote homology detection based on n-peptide compositions with reduced amino acid alphabets", *Biosystems*, 87, 75-81.
- [12] Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T., Chothia, C., (1998) "Sequence comparisons using multiple sequences detect tree times as many remote homologues as pairwise methods", *J.Mol.Biol.*, 284, 1201-1210.
- [13] Rost, B., (1999), "Twilight zone of protein sequence alignments", *Protein Eng.*, 12, 85-94.
- [14] Smith T. F. ve Waterman, M. S., (1981), "Identification of common molecular subsequences", *J.Mol.Biol.*, 147, 195-197.
- [15] Wallqvist, A., Fukunishi, Y., Murphy, L. R., Fadel, A., Levy, R. M., (2000), "Iterative sequence/structure search for protein homologs: comparison with amino acid sequence alignments and application to fold recognition in genome databases", *Bioinformatics*, 16, 988-1002.