Gene Ontology (GO) Molecular Function Prediction Based on Alignment Scores

Eser Aygun and Zehra Cataltepe

Istanbul Technical University, Computer Engineering Department
Ayazaga, Sariyer, TR-34469, Istanbul, Turkey
phone: + (90) 535-821-9803, fax: + (90) 212-285-3679
email: eser.aygun@gmail.com, cataltepe@itu.edu.tr
web: www2.itu.edu.tr/∼aygunes/eser, www3.itu.edu.tr/∼cataltepe

ABSTRACT

We investigate Gene Ontology (GO) Molecular Function prediction using alignment scores between protein sequences. We introduce a binary classification algorithm called Double Threshold Classifier (DTC). The parameters of the algorithm are two alignment score thresholds. In order to classify a sequence, DTC uses the alignment scores between the sequence and sequences in the training set. The algorithm decides on the class of the sequence based on the score distribution of positive and negative training examples with respect to the upper and lower thresholds.

We compare the DTC algorithm's performance to k-Nearest Neighbor Classifier (KNN) and Nearest Mean Classifier (NMC). As feature vectors, both KNN and NMC use the alignment scores between a sequence and the training sequences. On a GO Molecular Function dataset consisting of 1890 proteins and 7 classes, DTC outperforms both KNN and NMC. The 10-fold cross validation accuracy of DTC algorithm is around 83.40%, while the accuracies for KNN and NMC are 68.44% and 71.00% respectively. The total (training + test) running time of the DTC is also better than both KNN and NMC.

1. INTRODUCTION

The protein sequence databases have been growing with great speed, while the amount of proteins whose function are known does not increase as fast, because determining protein function experimentally is an expensive and time consuming process. On the other hand, determining functions of proteins is crucial in a number of fields, such as cancer detection and personalized drug design. Hence, automatic prediction of protein function becomes an important research area.

One of the most used protein function annotation databases is the Gene Ontology (GO) [2, 10] database. In this study we investigate prediction of GO Molecular Function of proteins, based on the sequence alignment scores of proteins. GO Molecular Function data have been used in a number of studies. For example, Protein Function Prediction (PFP) [13], ProtFun [15], Proteome Analyst (PA) [22].

In order to predict function, some studies have used the features extracted from the sequences, SVMProt [12] using PROFEAT features [17] is an example of this approach. Using sequence alignment scores as features is another approach followed in, for example, [18]. With these and other features, machine learning methods, such as Support Vector Machines (SVMs) [12, 23] and neural networks [16] have been used for function prediction.

There has been a link between sequence identity and function and this link has been interpreted in different ways in different studies. [8] mentions that sequence alignment methods, such as Needleman-Wunsch or Smith-Waterman, are known to be good at detecting homologs whose sequence identity is greater than 40%. According to [20] "for about 40 to 60% of all sequences from current genome projects, sequence homology suggests some aspects of function. However, a firm conclusion about function is not always clear". On the other hand [24] claims that "for pairs of domains that share the same fold, precise function appears to be conserved down to ∼40% sequence identity, whereas broad functional class is conserved to ∼25%. Interestingly, percent identity is more effective at quantifying functional conservation than the more modern scores (e.g. Pvalues)."

One of the goals of this paper is to shed some more light on the sequence alignment scores and thresholds on them within the framework of GO Molecular Function classes.

In addition to determining thresholds, we also introduce the Double Threshold Classifier (DTC) Algorithm, which uses these thresholds to predict whether a sequence belongs to a certain function class or not.

The rest of the paper is organized as follows: In section 2 we introduce the DTC algorithm. Section 3 reviews the other classification algorithms that we compare DTC against. The dataset we use is explained in Section 4. Section 5 goes through the alignment method that we use. Section 6 contains the results of our experiments and Section 7 concludes the paper.

2. DOUBLE THRESHOLD CLASSIFIER ALGORITHM

Based on the idea that, strong homology could point to function identity we wanted to base our classification algorithm on sequence alignment scores.

The classification algorithm that we introduce, Double Threshold Classifier (DTC), is a binary (one-against-all) classifier. In other words, it considers the members of a specific class as positive examples and the rest of the examples as negative examples. Therefore, for each molecular function class, there will be a separate classifier. It is also possible to combine the results of these individual classifiers to make an exact decision on the class, but we do not concentrate on this problem in our work.

Formally, let Pos denote the set of all positive examples in the training set and $s(x, y)$ be the pairwise alignment score of two protein sequences x and y. We define four predicates for each test example x:

$$\begin{aligned} P_{strong}(x) &: \forall y \in Pos\ s(x,y) > t_{upper} \\ P_{weak}(x) &: \exists y \in Pos\ s(x,y) > t_{upper} \\ N_{strong}(x) &: \forall y \notin Pos\ s(x,y) > t_{lower} \\ N_{weak}(x) &: \exists y \notin Pos\ s(x,y) > t_{lower} \end{aligned} \quad (1)$$

where t_{upper} and t_{lower} are the score threshold parameters of the classifier. Accordingly, we decide that x is a positive example if:

$$P_{strong}(x) \vee \neg N_{weak}(x) \vee (P_{weak}(x) \wedge \neg N_{strong}(x)). \quad (2)$$

In equation (1), $P_{strong}(x)$ and $P_{weak}(x)$ corresponds to strong and weak belief that sequence x is a positive instance respectively. Similarly, $N_{strong}(x)$ and $N_{weak}(x)$ corresponds to strong and weak belief that sequence x is a negative instance. Equation (2) classifies x as a positive example if any of the following holds, x is strongly positive or x is not weakly negative or x is both weakly positive and not strongly negative.

We experimented to find the best values of t_{upper} and t_{lower}. Based on those experiments (Figure 1), the score thresholds t_{upper} and t_{lower} are optimized between 30% and 40%. When we use the whole data set to optimize these parameters, the accuracy is maximized when $t_{upper} = 39\%$ and $t_{lower} = 26\%$. However, to avoid using apriori knowledge about the data, the experiments are done for constant values of $t_{upper} = 40\%$ and $t_{lower} = 30\%$. We think that it is interesting that the threshold values pointed out by [24], 25 and 40%, are quite close to the ones we found for the GO Molecular Function classes.

3. OTHER CLASSIFICATION ALGORITHMS

Each feature vector we consider consists of more than 1000 dimensions (i.e. alignment scores of the sequence to all training sequences). Not all classification algorithms were suitable such a high dimensional input space. For example, according to our experiments, both Support Vector Classifier (SVC) and Linear Discriminant Classifier (LDC) failed to generate reasonable (i.e. significantly more than 50% accurate) results. We found out that k-Nearest Neighbor (KNN) and Nearest Mean Classification (NMC) algorithms produced better results and hence, in this paper, we chose to compare DTC to these two algorithms.

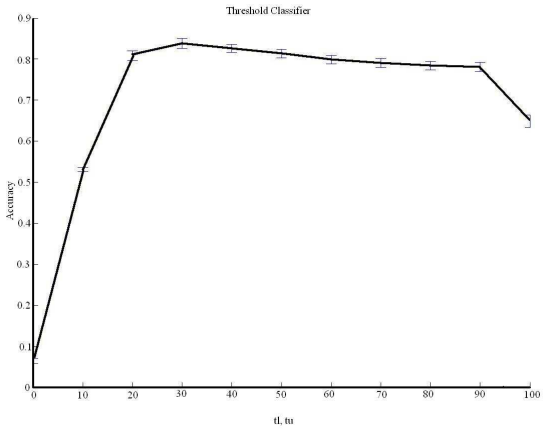

Figure 1: Prediction accuracy while t_{upper} and t_{lower} change at the same time.

3.1 KNN

k-Nearest Neighbor Classifiers (KNN) have been used in many pattern recognition applications and are very simple and intuitive. KNN needs very small training time, because it does not really need to learn any parameters, but only need to store the training data. In prediction (testing) mode, KNN decides on the class label based on the majority label of the closest training data. Closest is according to a distance or similarity metric. In this study, we use the Euclidean distance between the feature vectors as the distance metric. Since the distance between the test input and all training data needs to be computed, KNN suffers from long prediction times.

3.2 NMC

Nearest Mean Classifier (NMC) represents each class using the mean of the training data in that class. NMC algorithm is also known as centroid based classification and has been known to produce very good results for document categorization [11]. NMC labels a given example with the closest class label, hence it is very fast in recognition phase.

4. DATA SET

Protein function prediction differs from protein domain classification. Usually SCOP [7, 21] database is used for domain classification and chains (domains) instead of the whole protein needs to be classified as belonging to (possibly multiple) domains. For GO function prediction, which proteins exhibit which function is available through the Protein Data Bank PDB [4, 6] database. However, we do not know which chain(s) in the protein are responsible for the function.

In order to compare different function prediction methods,

abbr.	GO Molecular Function	# proteins
bin	binding (GOID:5488)	907
cat	catalytic act. (GOID:3824)	665
enz	enzyme regulator act. (GOID:30234)	54
sig	signal transducer act. (GOID:4871)	54
str	structural mol. act. (GOID:5198)	40
trc	transcription reg. act. (GOID:30528)	97
trp	transporter act. (GOID:5215)	73
	total	1890

Table 1: Number of proteins in each GO Molecular Function Class we considered.

we use the function categories under GO Molecular Function. We find out the PDB ID's of protein sequences that exhibit those functions through the PDB database. We only consider proteins whose chains contain between 50 to 150 amino acids.

Table 1 shows the molecular functions we consider and the number of proteins. We choose only proteins with a single chain because then we are sure that the function is exhibited by that specific chain. Note that we do not consider other categories since they contain less than 40 proteins.

5. SEQUENCE ALIGNMENT SCORES

We obtained the similarity scores between all sequences under the categories in Table 1 using ClustalW [3, 14] algorithm, which is a multiple alignment algorithm.

ClustalW is a well documented and open source program that can run on many platforms. It has been used in many studies and is very easy to use. We considered other alignment programs also. For example palign [9, 19] was reported to give more informative scores than ClustalW in [8]. However we could find very little documentation on the program. We also used blastpgp [1, 5], howevever we found out that we could not obtain any results at all for some of our sequences and we also found out that the order in which the sequences were given could make a big impact on the score obtained from blastpgp.

We used ClustalW in slow mode to get the exact pairwise alignment scores. Each score represents the number of identities in the best alignment divided by the number of residues compared (gap positions are excluded) times 100. The pairwise score is calculated independently of the substitution matrix (such as PAM, Dayhoff, Gonet) and gaps chosen.

GO Mol.Fn.	DTC	KNN	NMC	DTC % better
bin	68.29	57.03	58.20	19
cat	74.12	60.29	62.87	21
enz	93.12	80.31	76.19	16
sig	92.23	83.14	77.61	11
str	91.88	68.86	71.21	32
trc	78.78	64.97	73.05	19
trp	85.38	64.45	77.85	27
Avg	83.40	68.44	71.00	21
Std	3.12	3.11	2.42	

Table 2: Average 10-fold cross validation accuracies of DTC (Double Threshold Classifier), KNN (K-Nearest Neighbor) and NMC (Nearest Mean Classifier). The last two rows show the average and standard deviation of the accuracies for all the molecular function classes.

6. RESULTS

DTC can only be used for binary classification. Therefore, we train a classifier per GO Molecular Function in Table 1. We use the sequences that belong to the Molecular Function category as positive examples and all the other sequences as negative examples.

In order to evaluate the performance of the algorithms, we use 10-fold cross validation. We partition the training data into 10 different partitions. For i=1...10, we use the ith partition for validation and the rest of the data for training. We report the average of the validation accuracies.

Table 2 shows the average prediction accuracies for each Molecular Function and classification algorithm. The errorbars on these averages are on the order of 1 to at most 4. According to Table 2, DTC is always better than NMC and KNN. The last column shows the percentage difference between DTC accuracy and best of NMC and KNN for the Molecular Function for each row. DTC is 11 to 32 % better than DTC and NMC for different Molecular Function categories and it is 21 % better on the average.

In table 3 confusion matrix entries for DTC are shown. The number of TN and FP entries are large because there are more negative examples than positive examples for each of the binary classification problems.

Finally, Table 4 shows the time it takes to train and test each of the classifiers for all of the 10-fold cross validation runs. As expected, KNN is fast to train and and slow to test and NMC is slower to train and fast to test. DTC is faster than KNN to train and it is almost as fast as NMC to test. Hence when the sum of train and test times are considered, DTC is faster than both algorithms.

7. CONCLUSIONS AND FUTURE WORK

We have introduced the DTC (Double Threshold Algorithm) which is a fast and accurate method that can be used for GO Molecular Function prediction. We have also found out that 30% and 40% sequence alignment score thresholds are

GO Mol.Fn.	TP	TN	FP	FN
bin	835	437	546	72
cat	620	673	552	45
enz	53	1624	212	1
sig	52	1626	210	2
str	38	1640	210	2
trc	70	1535	258	27
trp	62	1568	249	11

Table 3: Confusion Matrix entries of DTC (Double Threshold Classifier). TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative.

	DTC	KNN	NMC
Training time (sec)	0.02	2.05	624.08
Testing time (sec)	1.72	103.24	1.32
Total time (sec)	1.73	105.28	625.40

Table 4: The training, test and total time (in seconds) of DTC, KNN and NMC algorithms to process all 1890 proteins for 10 cross validation runs.

important for functional identity in GO Molecular Function classes.

We are planning to continue our work using a bigger data set than the one we used here. We also intend to extend DTC so it can be used not only for binary but also for multiway classification. Finally we are planning to compare the performance of DTC to other function prediction methods, such as using PROFEAT features and SVM for example.

ACKNOWLEDGEMENTS

Both authors were supported by means of Tubitak Research Project EEEAG 105E164. Author Cataltepe thanks her family for their support while she tries hard to be a mom and researcher.

REFERENCES

[1] S. F. Altschul, W. Gish, W. Miller, E. Y. Myers, and D. J. Lipman. A basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

[2] M. Ashburner, C C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, and J. Eppig et al. Gene ontology: tool for the unification of biology. *Nat. Genet.*, 25:25–29, 2000.

[3] ClustalW WWW Service at the European Bioinformatics Institute. http://www.ebi.ac.uk/clustalw.

[4] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000. http://www.rcsb.org/pdb.

[5] BLAST. Basic local alignment search tool. http://www.ncbi.nlm.nih.gov/blast/.

[6] P.E. Bourne and et.al. The distribution and query systems of the rcsb protein data bank. *Nucleic Acids Research*, 32:D223, 2004. http://www.rcsb.org/pdb.

[7] M. A. G. Brenner, S. E. T. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

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

[9] A. Elofsson. A study on how to best align protein sequences. *Proteins*, 15:330–339, 2002.

[10] GO. http://www.geneontology.org/index.shtml.

[11] E.H. Han and G. Karypis. Centroid-based document classification: Analysis & experimental results. In *4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, pages 424–431, 2000.

[12] L. Y. Han, C. Z. Cai, Z.L.Ji, Z.W.Cao, J. Cui, and Y. Z. Chen. Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Research*, 32(21):6437–6444, 2004.

[13] T. Hawkins, S. Luban, and D. Kihara. Enhanced automated function prediction using distantly related sequences and contextual association by pfp. *Protein Science*, 15:1550–1556, 2006.

[14] J.D.Thompson, D. G.Higgins, and T.J.Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680, 1994.

[15] L. J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames, C. Kesmir, H. Nielsen, H. H. Sterfeldt, K. Rapacki, C. Workman, C. A. F. Andersen, S. Knudsen, A. Krogh, A. Valencia, and S. Brunak. Prediction of human protein function from post-translational modifications and localization features. *J. Mol. Biol.*, 319:1257–1265, 2002.

[16] L.J. Jensen, R. Gupta, H.H. Starfeldt, and S. Brunak. Prediction of human protein function according to gene ontology categories. *Bioinformatics*, 19(5):635–642, 2003.

[17] Z.R. Li, H.H. Lin, L.Y. Han, L. Jiang, X. Chen, and Y.Z. Chen. Profeat: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Research,*, 34:W32–W37, 2006.

[18] 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.

[19] PALIGN. http://www.bioinfo.se/palign/.

[20] B. Rost. *Rising accuracy of protein secondary structure prediction. in: 'Protein structure determination, analy-*

sis, and modeling for drug discovery' (ed. D Chasman), 207-249. New York: Dekker, 2003.

[21] SCOP. http://scop.mrc-lmb.cam.ac.uk/scop/.

[22] D. Szafron, P. Lu, R. Greiner, D. S. Wishart, B. Poulin, R. Eisner, Z. Lu, J. Anvik, C. Macdonell, A. Fyshe, and D. Meeuwis. Proteome analyst: Custom predictions with explanations in a web-based tool for high-throughput proteome annotations. *Nucleic Acids Research,*, 32:W365–W371, July 2004.

[23] A. Vinayagam, R. Konig, J. Moormann, F. Schubert, R. Eils, K. Glatting, and S. Suhai. Applying support vector machines for gene ontology based gene function prediction. *BMC Bioinformatics*, 5(1), 2004.

[24] C.A. Wilson, J. Kreychman, and M. Gerstein. Assessing annotation transfer for genomics: Quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.*, 297:233–249, December 2000.