Discrimination of GO term Annotated Proteins based on Amino Acid Occurrence and Composition

Y-h. Taguchi[1] and M. Michael Gromiha[2]

[1] Department of Physics, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan,
tag@granular.com,
[2] Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan
michael-gromiha@aist.go.jp

Abstract. In this paper, we have applied linear discriminant analysis and support vector machine for predicting GO term annotated proteins using amino acid occurrence/composition in uniref50 data set, i.e., uniprot with less than 50 % sequence identity. We found that our method could discriminate between proteins with at least one known GO term and those without any annotation at an AUC of 0.82 using three-fold cross validation test. Discrimination of the 38 most frequent GO terms is achieved with the maximum AUC of 0.91. Our method is solely based on amino acid sequence and hence it will be useful to predict GO term associations of newly obtained amino acid sequence without any annotated known homolog.

1 Introduction

Predicting molecular functions of proteins from their amino acid sequences is very important problem. Protein functions are mainly dictated by their structures [1] and the amino acid sequence of a protein contains all the necessary information encoded in its three-dimensional structure [2]. Hence, it is essential to develop algorithms for predicting protein functions from amino acid sequences. Although there are many investigations [3–14] to aim this purpose, almost all of them are based upon sequence similarity between targeted protein and other proteins with known functions. Recently, a sophisticated method, SIFTER [15] was proposed, which does not make use of sequence information directly but use the phylogenetic information derived from it for predicting the function. However, if there is very little sequence similarity between proteins, SIFTER cannot predict molecular functions since there will be no homolog, which is required to construct phylogenetic trees. More recently, Qiu *et al* [16] tried to predict GO terms with which each protein is annotated, without reference to sequence similarity. Although they have succeeded up to some extent, they had to make

M. Chetty, S. Ahmad, A. Ngom, S. W. Teng (Eds.): PRIB2008,
Supplementary Conf. Proc., pp.13-24, 2008.

use of information other than sequence similarity, e.g., protein structure similarity. This is hardly said to be useful, since there may be no known proteins annotated whose structure is similar with targeted newly found proteins. In this article, we propose a very new approach for predicting the molecular functions of proteins directly from amino acid sequence and without using the information from similar sequences. In our earlier work, we have showed that the amino acid occurrence is an important factor for protein fold recognition [17, 18]. In this work, we have utilized the information about amino acid occurrence for predicting the GO term associated proteins in uniref50[19], i.e., proteins with less than 50% sequence identity. The proposed method showed an AUC of 0.82 for discrimination between proteins associated with at least one GO term (in short, "GO protein") and those without any associated GO term (in short, "non-GO protein"). Discrimination of most frequent GO terms is achieved with the maximum AUC of 0.91.

2 Methods

2.1 Preparation of data set

We have prepared positive and negative sets of proteins as follows.

Discrimination between GO and non-GO proteins We have downloaded GO association file for uniprot from GOA[20] and uniref50[19] in fasta sequence format. Uniref50 is low redundancy set where no pairs of sequences have more than 50% sequence similarity. First we picked up sequences which are associated with any GO term. This has been done by choosing fasta sequences whose ID appear in "DB_Object_ID" columns in GOA "GO" associations. This results in 13329 sequences as GO proteins although 1686224 sequences are included into uniref50. This shows that less than 1% proteins have associated GO terms in uniref50 and more than 99 % proteins are non-GO proteins. We considered all the 13329 proteins as positive set. The negative set is the same number of proteins randomly chosen from non-GO proteins. The random selection for negative set is done only once since iterative re-sampling is time consuming because of large number of non-GO proteins.

Discrimination of specific GO terms We picked up proteins associated with specific GO terms in positive set of 13329 GO proteins. If the number of proteins is more than 30, we employ this set as the positive set for the GO term. Otherwise, we give up to discriminate this GO term, since discrimination for too few positive ones is erroneous. The negative set for this GO terms is the same number of proteins randomly selected from 13329 non-GO proteins as explained in the previous subsection. The random selection of negative set is repeated if the sampling error is too large to ignore.

2.2 Discriminant method and performance measure

We have applied linear discriminant analysis (LDA) and support vector machine (SVM) for the discrimination between positive and negative sets. Feature for LDA in this study is either amino acid occurrence or composition. For LDA, we have used lda module in MASS library in R[21]. The lda can give us probability of positive/negative by leave-one-out cross validation. For SVM, we have employed rbf kernel (ksvm module in kernlab packge in R). Parameter is tuned automatically as in the previous study[22]. Probabilities for positive set is computed by three fold cross validation as implemented in ksvm module. By changing the threshold probability between positive and negative classes, we can draw ROC curve for which we compute AUC, which is the performance measure in this study. AUC calculation and other ROC related analysis is performed by ROCR package[23] in R.

3 Results

3.1 Discrimination between GO and non-GO proteins

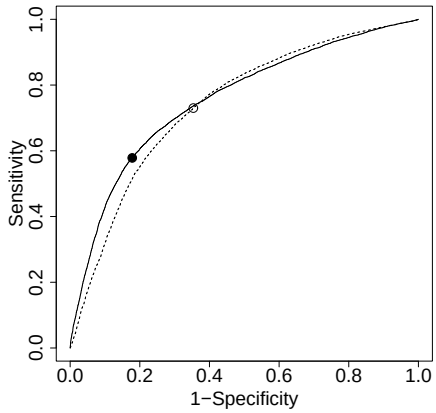


Fig. 1. ROC plot obtained for LDA using occurrence (solid curve) and composition (broken curve). Filled (open) circle indicates when threshold probability is equal to 0.5.

In Fig. 1, we have shown the ROC curve for discriminating GO and non-GO proteins using amino acid composition and occurrence. Interestingly these two methods achieve similar performance and the AUCs are 0.74 and 0.76, respectively. As we have not used any information other than sequence and the sequence

identity is less than 50 %, the performance of our method is remarkable. It is noteworthy that we use simple and fast linear discriminant analysis and hence our method can be applied to massive data set, too. When we employed SVM, AUCs rose to 0.80 (occurrence) and 0.82 (composition). Thus, SVM has better performance although it is more time consuming than LDA. CPU time is 720 sec. for SVM while that for LDA is only a few seconds when we use Intel(R) Core(TM)2, 1.06GHz.

3.2 Discrimination for specific GO terms

We have identified 38 GO terms (Table 1) that have more than 30 associated sequences in uniref50. We have computed AUC averaged over 10 independent selection of negative sets. Standard errors for AUC is less than 0.024 for all GO terms. As can be seen in Fig. 2(a), the performance of composition and occurrence is substantially different. Thus we employ either occurrence or composition so as to have better performance. Fig. 2(b) is the AUC ordered by performance, which is comparable to the method of Qiu *et al*[16]. in which several features other than sequence have been used for discrimination. Furthermore, for GO terms with relatively worse performance (i.e., from 20th to 30th best ranked GO terms), our results are significantly better than that of Qiu *et al*[16]. Although we did not try SVM because of requirement of long CPU time, it is expected that SVM would be better than LDA as seen in the previous section.

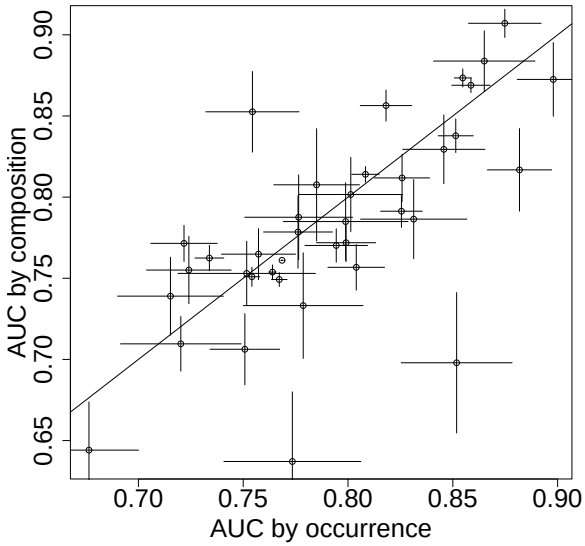
Table 1. GO terms used for discrimination. MF:Molecular Function, BP:Biological Process, CC:Cellular Components.

MF	GO:0003674	GO:0003677	GO:0003700	GO:0003723
	GO:0005102	GO:0005515	GO:0005524	GO:0008022
	GO:0008134	GO:0008270	GO:0019899	GO:0019901
	GO:0042802	GO:0042803	GO:0046982	
BP	GO:0006355	GO:0006366	GO:0006412	GO:0006468
	GO:0007267	GO:0007275	GO:0008150	
CC	GO:0005575	GO:0005576	GO:0005615	GO:0005624
	GO:0005625	GO:0005634	GO:0005730	GO:0005737
	GO:0005739	GO:0005783	GO:0005794	GO:0005829
	GO:0005886	GO:0005887	GO:0016020	GO:0016021

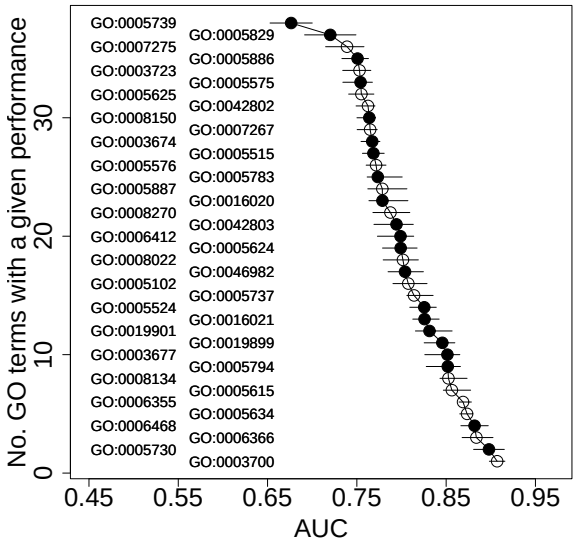
4 Discussion

4.1 Preparation of data set

One may argue that 50% is too large to regard that proteins do not have any sequence identity. To check this point, we have excluded any pairs of proteins



(a)



(b)

Fig. 2. (a) Comparison of AUC (open circles) by occurrence and composition. Horizontal(vertical) segments indicate 95 % confidence interval of AUC by occurrence(composition). Solid line corresponds to the cases when AUCs take same value for composition and occurrence. (b) AUC ordered by its performance. Filled (open) circle shows the results obtained with occurrence (composition) is for each GO term. GO terms are listed from bottom to top in the order of performance (Horizontal position does not have any meaning). Horizontal segments are 95 % confidence intervals.

between which sequence similarity is more than 20% by blastclust. Because of the limitation of CPU time, we picked up 1000 proteins from GO proteins and non-GO proteins and in total 2000 proteins are analyzed by blastclust. Then we have found that less than 10% proteins are removed and no pairs of proteins have more than 20% sequence identity if each of pair is taken from each of GO and non-GO proteins. It might be due to the fact that uniref50 has no pairs which share more than 50 % sequence identity for total more than one million proteins. If we consider very small part of this set, of course the largest sequence identity among this set is smaller than 50 %. Here we consider only 1% of uniref50 set, it is very small part of uniref50 and the largest sequence identity is much smaller than 50%. Thus, we do not refilter uniref50 to have smaller sequence similarity. Uniref50 is well defined and frequently used set. It is hopeful to use it as it is if possible.

4.2 Comparison between amino acid occurrence and composition in discrimination

Although in the discrimination of protein fold[17, 18] occurrence achieved significantly better performance than composition, in this study performance by these two did not differ much for discrimination between GO protein and non-GO protein. However, similar performance does not always mean same discrimination. In Fig. 3, we have shown scatter plot between occurrence and composition for discrimination and we noticed a substantial differences between them (Pearson's correlation coefficient is only 0.65). Table 2 shows the comparison when threshold probability is set to 0.5. In total, we have 13329 GO proteins and the same number of non-GO proteins. Among them, 407 proteins are predicted to be GO protein by occurrence while they are to be non-GO by composition. 4767 proteins are predicted to be non-GO by occurrence but to be GO by composition. Thus, it is clear that their discrimination differs from each other.

Table 2. Comparison of discrimination between composition and occurrence. Threshold probability is set to be 0.5.

	Negative	Positive [composition]
Negative	11807	4767
Positive	407	9677
[occurrence]		

Since we use LDA, it is possible to see what discriminates two classes. Figure 4 shows the comparison between basis vectors of discrimination. Positive(negative) values indicate GO (non-GO) proteins have more weights than non-GO (GO) proteins for the amino acids. Although weights are not always common, signs are often common. Hence, irrespective of the method weights of some specific amino acids are used for the discrmination. At present the reasons behind the

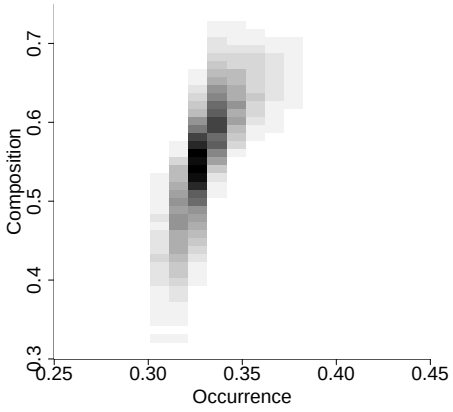


Fig. 3. Comparison between linear discriminant functions obtained by occurrence and composition. Both horizontal and vertical axes are normalized such that minimum (maximum) coordinate is 0(1). Only part of the plot is drawn. Darker means more density of points.

annotation of GO terms with respect to amino acid occurrence is not known, which might be revealed from the biological point of view. In addition to this, for occurrence, discrimination is mainly done by sequence length. If we try to discriminate GO and non-GO proteins only by length, we get AUC of 0.68 (GO proteins are, in average, longer than non-GO proteins) while AUC by occurrence is 0.76. Thus fairly large part of discrimination by occurrence is performed by sequence length. Longer protein has strong tendency to be annotated by GO terms. This result suggests that discrimination based on composition and occurrence is not similar as composition lacks the information of length.

4.3 Discrimination for specific GO terms

Since most of the negative sets consist of non-GO proteins, discrimination for each GO term is essentially discrimination between proteins associated with the GO term and proteins associated with non-GO terms. Here we have examined whether one can discriminate proteins annotated by specific GO term from those annotated by other GO terms. First, we noticed that selected 38 GO terms are very similar. Based upon "JiangConrath" option implemented in getTermSim function in GOSim package[24] in R, we have computed pairwise similarities. For all of Biological Process, Cellular Component and Molecular Funcion, mean similarity is more than 0.5. This means, remained GO terms are similar ones (Fig. 5 shows the hierarchy for remained 38 GO terms). This situation araised from the limitation that we consider GO term with more than 30 associated proteins. Since we consider uniref50, most of homolog is excluded from data set.

Thus, only GO terms with less specific functions left-out during this screenings. This results in survival of very similar GO terms. Since most of pairs of 38 GO terms even share at least one proteins, it is not reasonable to try to discriminate GO terms directly. This will be the future problem when more proteins are annotated with GO terms.

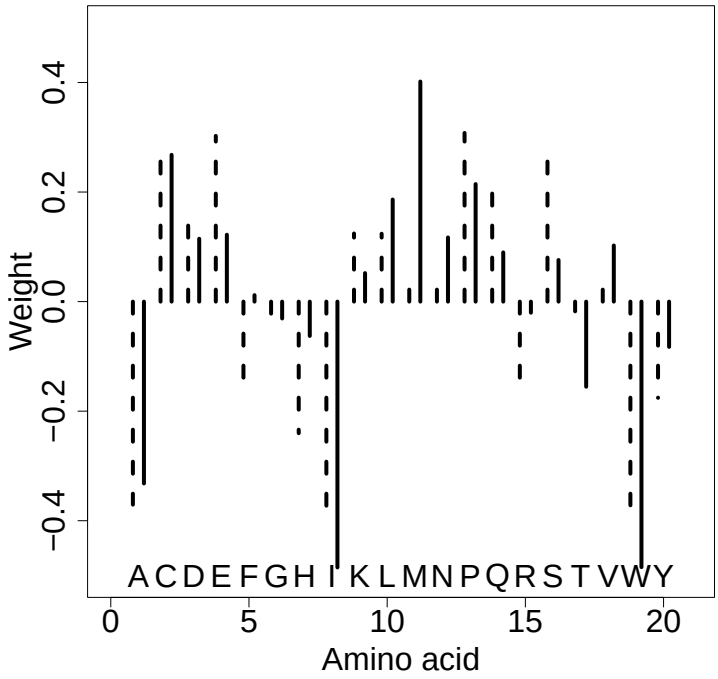
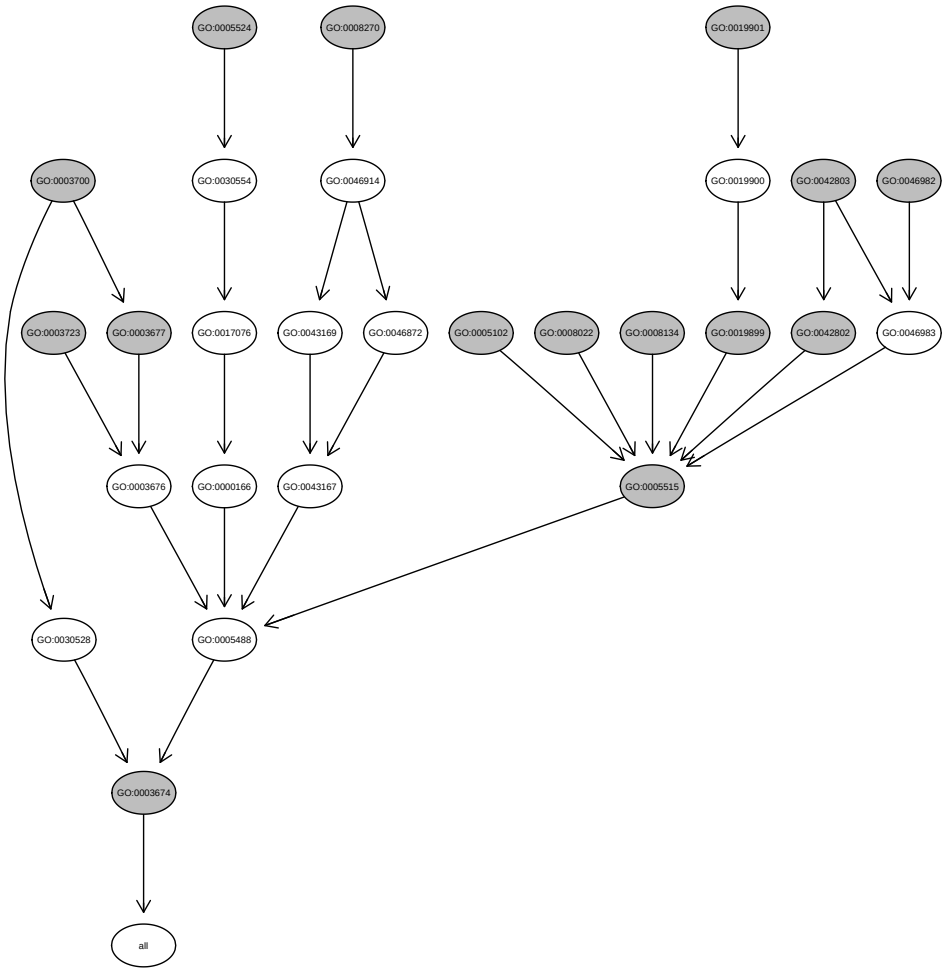


Fig. 4. Weight of amino acid in discrminant basis vector. Solid (broken) lines indicates discrimination by occurrence (composition). Positive (negative) weight means GO (no GO) proteins have more wieght. Weights are normalized such that squared sum is unity.

5 Acknowledgement

This work has been partly supported by the Grant-in-Aid for Creative Scientific Research No.19500254 of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) from 2007 to 2008. We are grateful for their support.



(a)MF

Fig. 5. GO hierarchy used in this study. Shaded parts are used for GO terms. The lower most GO term is "all", i. e. root. The second lowest GO terms are roots of (a) Molecular Function (MF), (b), Biological Process (BP) and (c) Cellular Component (CC)

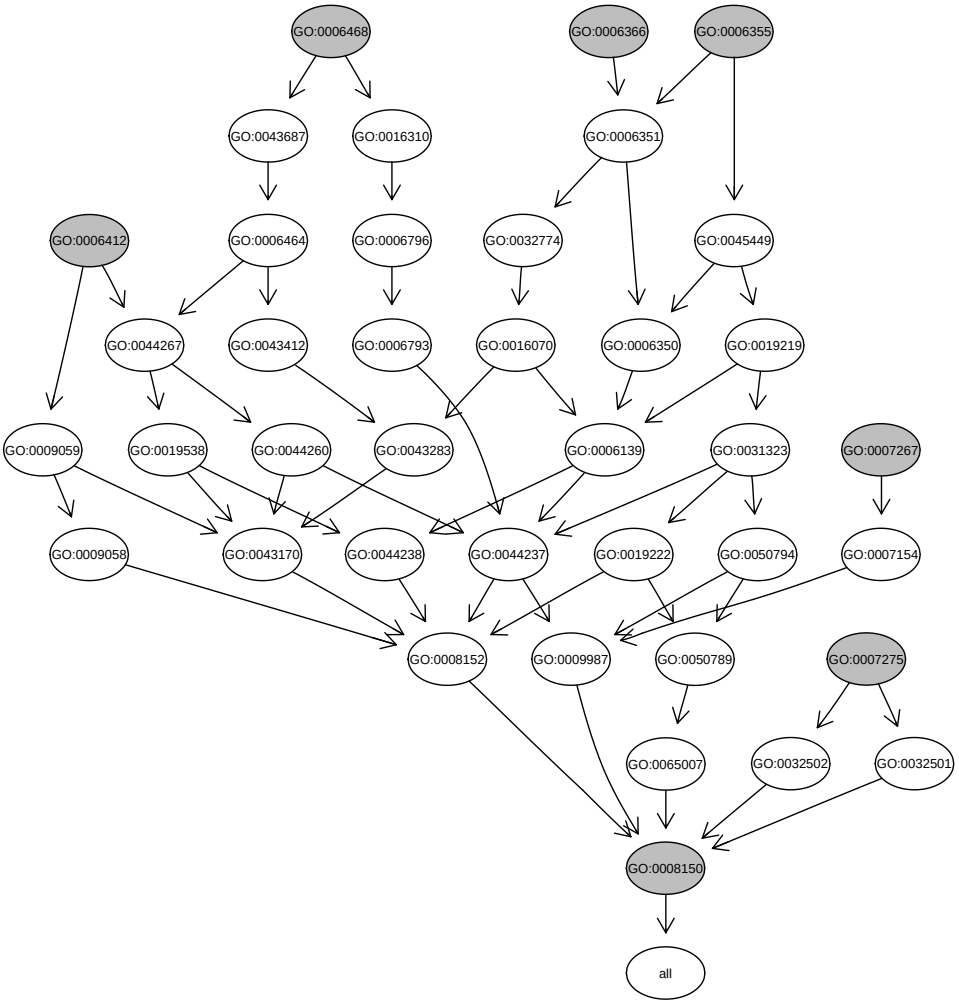

GO:0006468
GO:0006366
GO:0006355
GO:0043687
GO:0016310
GO:0006351
GO:0006412
GO:0006464
GO:0006796
GO:0032774
GO:0045449
GO:0044267
GO:0043412
GO:0006793
GO:0016070
GO:0006350
GO:0019219
GO:0009059
GO:0019538
GO:0044260
GO:0043283
GO:0006139
GO:0031323
GO:0007267
GO:0009058
GO:0043170
GO:0044238
GO:0044237
GO:0019222
GO:0050794
GO:0007154
GO:0008152
GO:0009987
GO:0050789
GO:0007275
GO:0065007
GO:0032502
GO:0032501
GO:0008150
all

Fig.5 (b)BP

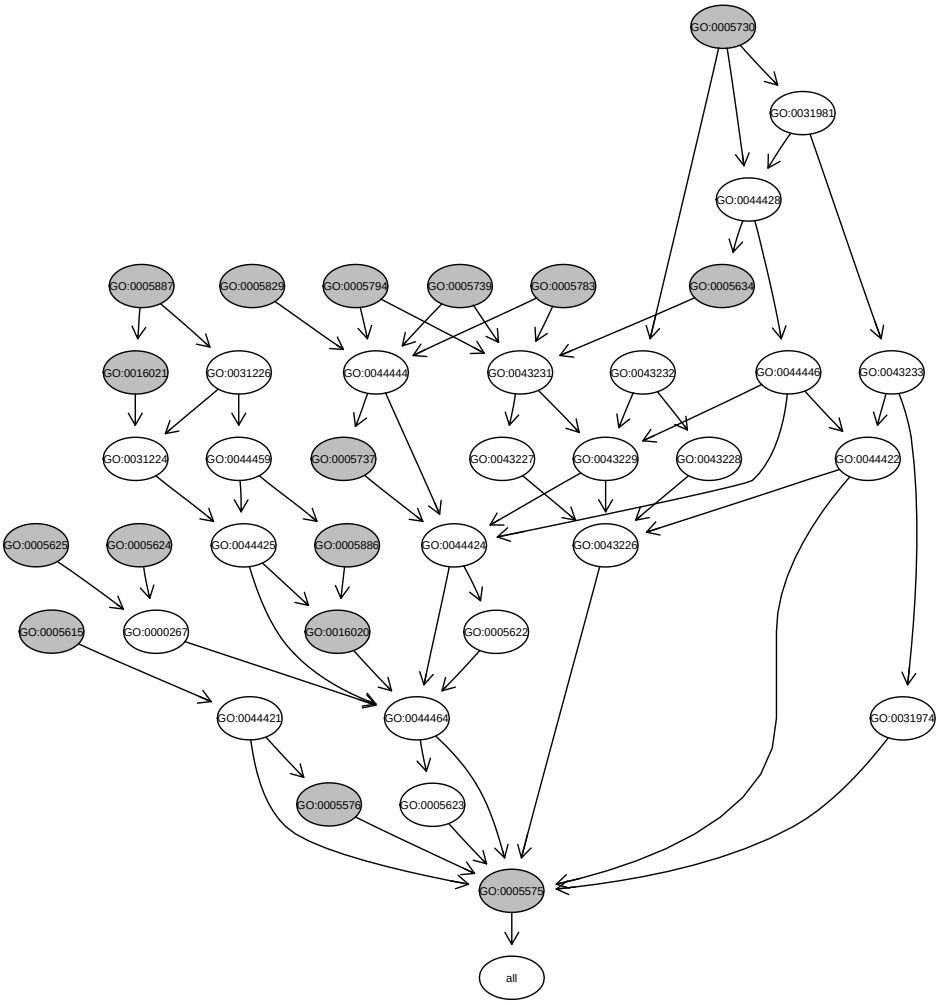


Fig.5(c)CC

References

1. Branden, C. Tooze, C., Introduction to protein structure. Garland Publishing Inc, New York, 1999.
2. Anfinsen, C.B., Principles that govern the folding of protein chains. Science. **181** (1973) 223–230.
3. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., Basic local alignment search tool. J. Mol. Biol. **215** (1990) 403–410

4. Khan, S., Situ, G., Decker, K., Schmidt, C.J., GoFigure: Automated gene ontology annotation. Bioinformatics **18** (2003) 2484–2485
5. Martin, D.M.A., Berriman, M., Barton, G.J., GOtcha: A new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinformatics **5** (2004) 178–195
6. Groth, D., Lehrach, H., Henning, S., GOblet: A platform for Gene Ontology annotation of anonymous sequence data. Nucleic. Acids Res. **32** (2004) 313–317
7. Zehetner, G., Ontoblast function: From sequences similarities directly to potential function annotations by ontology terms. Nucleic. Acids Res. **31** (2003) 3799–3803
8. Lee C.I., Irizarry, K. The Genemine system for genome/proteome annotation and collaborative data mining. IBM Syst J. **40** (2001) 592–603
9. Pazos F., Sternberg, M. J. E., Automated prediction of protein function and detection of functional sites from structure. Proc. Natl. Acad. Sci. U S A **101** (2004) 14754–14759
10. Frishman, D., Mokrejs, M., Kosykh, D., Kastenmuller, G., Kolesov, G., et al. The PEDANT genome database. Nucleic Acids Res. **31** (2003) 207–211
11. Gaasterland T., Sensen, C.W., Magpie: Automated genome interpretation. Trends Genet. **12** (1996) 76–78
12. Andrade, M.A., Brown, N.P., Leroy, C., Hoersch, S., deDaruvar, A., et al. Automated genome sequence analysis and annotation. Bioinformatics **15** (1999) 391–412
13. Tatusov, R.L., Galperin, M.Y., Natale, D.A., Koonin, E.V., The COG database: A tool for genome-scale analysis of protein functions and evolution. Nucleic Acids. Res. **28** (2000) 33–36
14. Perrierem, G., Duret, L., Gouy, M., HOBACGEN: Database system for comparative genomics in bacteria. Genome Res. **10** (2000) 379–385
15. Engelhardt, B.E., Jordan, M.I., Murator, K.E., Brenner, S.E., Protein Molecular Function Prediction by Bayesian Phylogenomics. PLoS Comput. Biol. **1** (2005) e45, Oct. 2005.
16. Qiu, J., Hue M., Ben-Hur, A., Vert, J-P., Noble, W.S., A structural alignment kernel for protein structures. Bioinformatics **23** (2007) 1090–1098
17. Taguchi, Y-h., Gromiha, M.M., Protein fold Recognition based upon the amino acid occurrence. in Pattern Recognition in Bioinformatics, (2007, Springer, Berlin) LNBI 4774 120–131
18. Taguchi, Y-h., Gromiha, M.M., Application of amino acid occurrence for discriminating different folding types of globular protein. BMC Bioinformatics **8** (2007) 404
19. ftp://ftp.uniprot.org/pub/databases/uniprot/uniref/uniref50/
20. ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/
21. R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, (Vienna, Austria 2007) ISBN 3-900051-07-0, URL http://www.R-project.org.
22. Taguchi, Y-h., Gromiha, M.M., Gene Ontology Term Prediction Based upon Amino Acid Occurrence, 2008 International Joint Conference on Neural Networks, IEEE,(2008) 616–621.
23. Sing, T., Sander, O., Beerenwinkel, N., Lengauer, T., ROCR: visualizing classifier performance in R. Bioinformatics **21**(2005) 3940–3941.
24. Froehlich, H., Speer, N., Poustka, A., Beissbarth, T., GOSim - An R-Package for Computation of Information Theoretic GO Similarities Between Terms and Gene Products. BMC Bioinformatics, **8** (2007) 166.