GOASVM: PROTEIN SUBCELLULAR LOCALIZATION PREDICTION BASED ON GENE ONTOLOGY ANNOTATION AND SVM

Shibiao Wan, Man-Wai Mak

Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University
Hung Hom, Hong Kong SAR, China
email: {shibiao.wan, enmwmak}@polyu.edu.hk

Sun-Yuan Kung

Dept. of Electrical Engineering
Princeton University
New Jersey, USA
email: kung@princeton.edu

ABSTRACT

Protein subcellular localization is an essential step to annotate proteins and to design drugs. This paper proposes a functional-domain based method—GOASVM—by making full use of Gene Ontology Annotation (GOA) database to predict the subcellular locations of proteins. GOASVM uses the accession number (AC) of a query protein and the accession numbers (ACs) of homologous proteins returned from PSI-BLAST as the query strings to search against the GOA database. The occurrences of a set of predefined GO terms are used to construct the GO vectors for classification by support vector machines (SVMs). The paper investigated two different approaches to constructing the GO vectors. Experimental results suggest that using the ACs of homologous proteins as the query strings can achieve an accuracy of 94.68%, which is significantly higher than all published results based on the same dataset. As a user-friendly web-server, GOASVM is freely accessible to the public at http://bioinfo.eie.polyu.edu.hk/mGoaSvmServer/GOASVM.html.

Index Terms— Protein subcellular localization; Gene Ontology Annotation; Gene Ontology; Support vector machines; GO terms.

1. INTRODUCTION

Determination of protein subcellular locations is indispensable for the annotation of protein functions, and it plays an important role in drug design. However, in the post-genomic era, the number of newly discovered proteins has been growing exponentially, making subcellular localization prediction by purely laboratory tests prohibitively expensive. Therefore, computational methods are developed to help biologists in selecting target proteins and designing related experiments. Recent years have witnessed impressive progress in dealing with this challenge. Localization methods can be generally divided into the following four categories.

(1) ***Sorting-signals based methods***. This group predicts the localization via the recognition of N-terminal sorting signals in amino acid sequences [1]. Nakai and Kanehisa in 1991 [2] proposed the earliest predictor—PSORT—using sorting signals. In 2006, they extended it to WoLF PSORT [3]. However, this group of methods could only deal with proteins that contain signal sequences. For example, the popular TargetP [4, 5] could only detect three locations: chloroplast, mitochondria and secretory pathway (extracellular).

(2) ***Composition-based methods***. This category focuses on the relationship between subcellular locations and the information embedded in the amino acid sequences such as amino-acid compositions (AA) [6], amino-acid pair compositions (PairAA) [6], gapped amino-acid pair compositions (GapAA) [7], and pseudo amino-acid composition (PseAA) [8]. This kind of methods is easy to implement, but usually have poorer performance than other methods.

(3) ***Homology-based methods***. These methods assume that homologous proteins are more likely to reside in the same subcellular location. Their performance is generally better than that of composition-based method as long as the homologs of the query sequences can be found in protein databases [9]. Recently, Mak et al. [10] proposed a homology-based predictor called PairProSVM. The predictor applies profile alignment to detect weak similarity between protein sequences. Homology based methods can detect as many locations as appeared in the dataset and can achieve comparatively high accuracy. But when the dataset contains sequences with low sequence similarity or the numbers of samples in different classes are imbalanced, the performance is still poor.

(4) ***Functional-domain based methods***. These approaches make use of the correlation between the function of a protein and its subcellular location. In [11], a sequence is mapped into the GO database so that a feature vector can be formed by determining which GO terms the sequence holds. Moreover, by exploiting the domain knowledge in subcellular localization, Huang et al. [12] proposed a searching algorithm called GOmining to discover the informative GO terms and classify them into instructive GO terms and essential GO terms for leveraging the information in the GO database. More recently, Xiao et al. [13] proposed a predictor called iLoc-Virus that uses the information in the GO annotation database to predict the subcellular localizations of virus proteins. Although the functional-domain based methods can often outperform sequence-based methods, they are only applicable to sequences that possess the required information, because so far not all sequences are functionally annotated.

This paper proposes a functional-domain based method called GOASVM, which is based on protein homology, gene ontology annotations, and support vector machines. Unlike our previous work [14] where homolog-based and functional-domain based predictors are fused, GOASVM uses PSI-BLAST [15] to find more relevant accession numbers (ACs)[1] to search against the GO annotation (GOA) database. This strategy leads to more GO terms and GO vectors for classification and hence improves classification performance. Sections 2 and 3 detail the GOA and the GO vector construction methods. Section 4 reports the performance of GOASVM on a popular benchmark dataset and show that it outperforms other functional-domain based predictors on the same dataset.

This work was in part supported by The Hong Kong Research Grant Council, Grant No. PolyU5264/09E and HKPolyU Grant No. G-U877.

[1] Accession numbers are unique identifiers given to DNA or protein sequences. They are mainly used as foreign keys for referencing sequences in sequence databases.

2. GENE ONTOLOGY ANNOTATION DATABASE

2.1. Gene Ontology

Gene Ontology (GO)[2] is a set of standardized vocabularies that annotate the function of genes and gene products across different species. The term 'ontology' originally refers to a systematic account of existence. In the GO database, the annotations of gene products are organized in three related ontologies: cellular components, biological processes, and molecular functions. A cellular component is a component of a cell. It is a part of some larger objects such as an anatomical structure or a gene product group. A biological process is a sequence of events achieved by one or more ordered assemblies of molecular functions. A molecular function is achieved by activities that can be performed by individual or by assembled complexes of gene products at the molecular level.

2.2. Gene Ontology Annotation Database

As a result of the GO Consortium annotation effort, the Gene Ontology Annotation (GOA) database[3] has become a large and comprehensive resource for proteomics research [16]. The database provides structured annotations to non-redundant proteins from many species in UniProt Knowledgebase (UniProtKB) [17] using standardized GO vocabularies through a combination of electronic and manual techniques. It also includes a series of cross-references to other databases. Thus, the systematic integration of GO annotations and UniProtKB database can be exploited for subcellular localization. Specifically, given the accession number of a protein, a set of GO terms can be retrieved from the GOA database file.[4]

3. GOASVM METHOD

3.1. Retrieval of GO Terms

For proteins with known accession numbers (ACs), we directly retrieved the GO terms by using their ACs as the searching keys to search against the GOA database. For proteins without an AC, we used PSI-BLAST [15] to find their homologs and used their ACs as the searching keys. Specifically, given a query sequence, n homologs and n ACs will be found. This means that each sequence will produce either one set of GO terms from the true AC or n set of GO terms from the ACs of n homologs. In this work, we considered the top homolog (i.e., $n = 1$) only because it is the most relevant and its AC is more likely to bring us relevant information.

3.2. Construction of GO Vectors

Given a dataset, we used the procedure described in Section 3.1 to retrieve the GO terms of all of its proteins. Then, we determined the number of distinct GO terms corresponding to the dataset. Suppose T distinct GO terms were found; these GO terms form a GO Euclidean space with T dimensions. For each sequence in the dataset, we constructed a GO vector by matching its GO terms to all of the T GO terms. We have investigated two approaches to determine the elements of the GO vectors.

1. **1-0 value**. In this approach, each of the T GO terms represents one canonical basis of a Euclidean space, and a protein sequence is represented by a point with coordinates equal to either 0 or 1. Specifically, the GO vector of the i-th protein is denoted as:

$$\mathbf{p}_i = \begin{bmatrix} a_{i,1} \\ \vdots \\ a_{i,j} \\ \vdots \\ a_{i,T} \end{bmatrix} \text{ where } a_{i,j} = \begin{cases} 1 & \text{, GO hit} \\ 0 & \text{, otherwise} \end{cases} \quad (1)$$

where 'GO hit' means that the j-th GO term appears in the GOA search result using the AC of the i-th protein as the searching key.

2. **Term-Frequency (TF)**. This approach is similar to the 1-0 value approach in that a protein is represented by a point in a Euclidean space. However, unlike the 1-0 approach, it uses the number of occurrences of individual GO terms as the coordinates. Specifically, the GO vector $\mathbf{p}_i$ of the i-th protein is defined as:

$$\mathbf{p}_i = \begin{bmatrix} b_{i,1} \\ \vdots \\ b_{i,j} \\ \vdots \\ b_{i,T} \end{bmatrix} \text{ where } b_{i,j} = \begin{cases} f_{i,j} & \text{, GO hit} \\ 0 & \text{, otherwise} \end{cases} \quad (2)$$

where $f_{i,j}$ is the number of occurrences of the j-th GO term (term-frequency) in the i-th protein sequence. The rationale is that the term-frequencies may also contain important information for classification and therefore should not be quantized to either 0 or 1. Note that $b_{i,j}$'s are analogous to the term-frequencies commonly used in document retrieval. We have also tried forming the GO vectors by using inverse-sequence frequency (ISF) and TF-ISF [14], but they are inferior to TF.

3.3. Multiclass SVM Classification

GO vectors are used for training one-vs-rest SVMs. Specifically, for an M-class problem (here M is the number of subcellular locations), M independent SVM are trained, one for each class. Denote the GO vector created by using the true AC of the i-th query protein as $\mathbf{q}_{i,0}$ and the GO vectors created by using the n homologous ACs as $\mathbf{q}_{i,j}$, $j = 1, \ldots, n$. Then, the score of the m-th SVM given the i-th query protein is

$$s_m(\mathbf{q}_i) = \sum_{j=0}^{n} w_j \left(\sum_{r \in \mathcal{S}_m} \alpha_{m,r} y_{m,r} K(\mathbf{p}_r, \mathbf{q}_{i,j}) + b_m \right) \quad (3)$$

where $\mathcal{S}_m$ is the set of support vector indexes corresponding to the m-th SVM, $y_{m,r} \in \{-1, +1\}$ are the class labels, $\alpha_{m,r}$ are the Lagrange multipliers, $K(\cdot, \cdot)$ is a kernel function, and w_j's are fusion weights such that $\sum_{j=0}^{n} w_j = 1$. The predicted class of the test sequence is given by

$$m^* = \arg\max_{m=1}^{M} s_m(\mathbf{q}_i). \quad (4)$$

Note that $\mathbf{p}_r$'s in Eq. 3 represents the GO training vectors, which may include the GO vectors created by using the true AC of the training sequences or their homologous ACs.

If the true ACs are not available, then only the ACs of the homologous sequences can be used for training the SVM and for scoring. In that case, $\mathbf{q}_{i,0}$ does not exist and $w_0 = 0$ in Eq. 3; moreover, $\mathbf{p}_r$ represents the GO training vectors created by using the homologous ACs only.

[2] http://www.geneontology.org
[3] http://www.ebi.ac.uk/GOA
[4] ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/

GO Vector Construction Method	OMCC	WAMCC	ACC
1-0 value	0.9252	0.9189	92.98%
TF	**0.9419**	**0.9379**	**94.55%**

Table 1. Performance of different GO-vector construction methods based on leave-one-out cross validation on the training set.

In this work, linear kernels and the top homolog were used, i.e., $K(\mathbf{p}_r, \mathbf{q}_{i,j}) = \langle \mathbf{p}_r, \mathbf{q}_{i,j} \rangle$ and $n = 1$ in Eq. 3. Therefore, when ACs are not available, $w_0 = 0$ and $w_1 = 1$. When ACs are available, $w_0 = 1$ and $w_1 = 0$.

4. EXPERIMENTS AND RESULTS

4.1. Dataset and Performance Measures

The performance of GOASVM was evaluated on Chou's dataset [18], which was created from Swiss-Prot 48.2. The dataset comprises 4150 proteins (2423 in the training set and 1727 in the independent test set) with 16 classes. The sequence similarity is cut off at 25%. We used the GOA database (released on 08-Mar-2011) as the retrieval database. In the experiments, 5450 and 5430 distinct GO terms were found, respectively, by using only ACs and only the ACs of homologous sequences as the searching keys.

Leave-one-out cross validation (LOOCV) and independent tests were used for performance evaluation. The performance measures include the overall accuracy (ACC), overall Mathew's correlation coefficient (OMCC) [10] and weighted average Mathew's correlation coefficient (WAMCC) [10]. The latter two have the advantage of avoiding the performance to be dominated by the majority classes.

4.2. Performance of GOASVM Predictor

Table 1 shows the performance of the GO-vector construction methods. Linear SVMs were used in both cases, and the penalty factor was set to 0.1. The results show that term-frequency (TF) performs almost 2% better than 1-0 value, which demonstrates that the frequencies of occurrences of GO terms could also provide information for subcellular locations. The results are biologically relevant because proteins of the same subcellular localization are expected to have a similar number of occurrences of the same GO term. In this regard, the 1-0 value approach is inferior because it quantizes the number of occurrences of a GO term to 0 or 1.

Table 2 shows the performance of different features and different SVM classifiers. The penalty factor for training the SVMs was set to 0.1 for both linear SVMs and RBF-SVMs. For RBF-SVMs, the kernel parameter was set to 1. The maximum gap length of GapAA [7] is 48 (the minimum length of all the sequences is 50). As AA, PairAA and PseAA produce low-dimensional feature vectors, the performance achieved by RBF-SVMs is better than that of the linear SVMs. So we just report the performance of RBF-SVMs here. As can be seen, amino-acid composition and its variant are not good features for subcellular localization. The highest accuracy is only 45.56%. Moreover, the performance of the homology-based method (last 2nd row) is also poor (only 45.23%). On the other hand, our proposed GOASVM can achieve a significantly better performance (94.68%), which is more than 40% (absolute) better than the composition-based and homology-based methods. This suggests that GO annotations can provide significantly richer information pertaining to protein subcellular localization than AA compositions and profile alignment.

Table 3 compares the performance of GOASVM against three state-of-the-art GO-based methods. As Euk-OET-PLoc [18] could not produce valid GO vectors for some proteins in Chou's dataset, it uses PseAA as a backup. ProLoc-GO [12] uses either the ACs of proteins as searching keys or uses the ACs of homologs returned from BLAST [19] as searching keys. Our proposed method can also use either ACs or sequences as inputs. Unlike Euk-OET-PLoc and ProLoc-GO, GOASVM uses PSI-BLAST to find the top-ranked homolog. Table 3 shows that for ProLoc-GO, using ACs as input performs better than using sequences (ACs of homologs) as input. However, the results for GOASVM are not conclusive in this regard because under LOOCV, using ACs as input performs better than using sequences, but the situation is opposite under independent tests. Table 3 also shows that no matter using ACs as input or sequences as input, GOASVM performs better than Euk-OET-PLoc and ProLoc-GO.

4.3. Performance of GOASVM Using Old GOA Database

The newer the version of GOA database, the more annotation information it contains. To investigate how the updated information affects the performance of GOASVM, we performed experiments using an earlier version (published in Oct. 2005) of the GOA database and compared the results with Euk-OET-PLoc [18]. Comparison between the last and second last rows of Table 4 reveals that using newer versions of the GOA database can achieve better performance than using older versions. This suggests that annotation information is very important to the prediction. The results also show that GOASVM significantly outperforms Euk-OET-PLoc, suggesting that the GO vector construction method and classifier (term-frequency and SVM) in GOASVM are superior to the those used in Euk-OET-PLoc (1-0 value and K-NN).

5. CONCLUSIONS

This paper proposes a functional-domain based method – GOASVM – to predict subcellular locations of proteins. The accession numbers (ACs) of query proteins are used as keys to search against the GOA database to find the GO terms. For proteins without an AC, PSI-BLAST is used to find their homologs and the ACs of these homologs are used as the searching keys. Then, GO terms are used to construct the GO vectors, which are subsequently classified by SVMs. Results on a recent dataset demonstrate that GOASVM outperforms the state-of-the-art GO-based methods, homology-based method, and methods based on amino acid compositions. It was also found that the frequency of occurrences of GO terms provides useful information for classification. For readers' convenience, a user-friendly web-server for GOASVM was designed and it is freely accessible to the public at http://bioinfo.eie.polyu.edu.hk/mGoaSvmServer/GOASVM.html.

6. REFERENCES

[1] K. Nakai, "Protein sorting signals and prediction of subcellular localization," *Advances in Protein Chemistry*, vol. 54, no. 1, pp. 277–344, 2000.

[2] K. Nakai and M. Kanehisa, "Expert system for predicting protein localization sites in gram-negative bacteria," *Proteins: Structure, Function, and Genetics*, vol. 11, no. 2, pp. 95–110, 1991.

[3] P. Horton, K. J. Park, T. Obayashi, and K. Nakai, "Protein subcellular localization prediction with WOLF PSORT," in *Proc. 4th Annual Asia Pacific Bioinformatics Conference (APBC06)*, 2006, pp. 39–48.

Classifier	Feature	OMCC	WAMCC	ACC
RBF SVM	AA [6]	0.3489	0.2873	38.96%
RBF SVM	AA+PairAA [6]	0.3872	0.3148	42.55%
Linear SVM	AA+PairAA+GapAA(48) [7]	0.4193	0.3497	45.56%
RBF SVM	PseAA [8]	0.3700	0.2868	40.94%
Linear SVM	Profile-alignment scores [10]	0.4158	0.3827	45.23%
Linear SVM	GO (Using PSI-BLAST)	**0.9432**	**0.9388**	**94.68%**

Table 2. Performance of different features and different SVM classifiers based on leave-one-out cross validation on the training set. The input data to the feature extraction stage are amino-acid sequences. *OMCC*: Overall MCC; *WAMCC*: Weighted average MCC; *ACC*: Overall accuracy. See [10] for the definition of these performance measures.

Method	Input Data	Feature	Accuracy	
			LOOCV on Training Set	Independent Test Set
ProLoc-GO [12]	S	GO (using BLAST)	86.6%	83.3%
ProLoc-GO [12]	AC	GO (No BLAST)	89.0%	85.7%
Euk-OET-PLoc [18]	S + AC	GO + PseAA	81.6%	83.7%
GOASVM	S	GO (usig PSI-BLAST)	**94.68%**	93.86%
GOASVM	AC	GO (No PSI-BLAST)	94.55%	**94.61%**

Table 3. Comparing GOASVM with state-of-the-art GO-based methods based on the training and testing datasets. *LOOCV*: leave-one-out cross validation; *S*: Sequences; *AC*: Accession Number.

Method	Feature		Accuracy	
	Main	Backup	LOOCV on training set	Independent test set
Euk-OET-PLoc [18]	GO (GOA2005)	PseAA	81.6%	83.7%
GOASVM	GO (GOA2005)	PseAA	86.42%	89.11%
GOASVM	GO (GOA2011)	–	**94.55%**	**94.61%**

Table 4. Performance of GOASVM based on different versions of the GOA database. The 2nd column specifies the publication year of the GOA database being used for constructing the GO vectors. For proteins without a GO term in the GOA database, pseudo amino-acid composition (PseAA) was used as the backup feature. When the latest GOA database is used (last row), only one protein in the dataset does not have a GO term. Therefore, we assigned '0' to all of the elements in the GO vector of this protein instead of using PseAA. *LOOCV*: leave-one-out cross validation.

[4] O. Emanuelsson, H. Nielsen, S. Brunak, and G. von Heijne, "Predicting subcellular localization of proteins based on their N-terminal amino acid sequence," *J. Mol. Biol.*, vol. 300, no. 4, pp. 1005–1016, 2000.

[5] O. Emanuelsson, S. Brunak, G. von Heijne, and H. Nielsen, "Locating proteins in the cell using TargetP, SignalP, and related tools," *Nature Protocols*, vol. 2, no. 4, pp. 953–971, 2007.

[6] H. Nakashima and K. Nishikawa, "Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies," *J. Mol. Biol.*, vol. 238, pp. 54–61, 1994.

[7] K. J. Park and M. Kanehisa, "Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs," *Bioinformatics*, vol. 19, no. 13, pp. 1656–1663, 2003.

[8] K. C. Chou, "Prediction of protein cellular attributes using pseudo amino acid composition," *Proteins: Structure, Function, and Genetics*, vol. 43, pp. 246–255, 2001.

[9] R. Nair and B. Rost, "Sequence conserved for subcellular localization," *Protein Science*, vol. 11, pp. 2836–2847, 2002.

[10] M. W. Mak, J. Guo, and S. Y. Kung, "PairProSVM: Protein subcellular localization based on local pairwise profile alignment and SVM," *IEEE/ACM Trans. on Computational Biology and Bioinformatics*, vol. 5, no. 3, pp. 416 – 422, 2008.

[11] K. C. Chou and H. B. Shen, "Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization," *Biochemical and Biophysical Resarch Communication*, vol. 347, pp. 150–157, 2006.

[12] W. L. Huang, C. W. Tung, S. W. Ho, S. F. Hwang, and S. Y. Ho, "ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization," *BMC Bioinformatics*, vol. 9, no. 80, 2008.

[13] X. Xiao, Z.C. Wu, and K.C. Chou, "iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites," *Journal of Theoretical Biology*, vol. 284, pp. 42–51, 2011.

[14] S. Wan, M. W. Mak, and S. Y. Kung, "Protein subcellular localization prediction based on profile alignment and Gene Ontology," in *2011 IEEE International Workshop on Machine Learning for Signal Processing (MLSP'11)*, 2011.

[15] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, pp. 3389–3402, 1997.

[16] E. Camon, M. Magrane, D. Barrel, D. Binns, W. Fleischnann, P. Kersey, N. Mulder, T. Oinn, J. Maslen, and A. Cox, "The gene ontology annotation (GOA) project: Implementation of GO in SWISS-PROT, TrEMBL and InterPro," *Genome Res.*, vol. 13, pp. 662–672, 2003.

[17] D. Butler, "NIH pledges cash for global protein database," *Nature*, vol. 419, no. 101, 2002.

[18] K. C. Chou and H. B. Shen, "Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers," *J. of Proteome Research*, vol. 5, pp. 1888–1897, 2006.

[19] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *J. Mol. Biol.*, vol. 215, pp. 403–410, 1990.