BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 7 2012, pages 1028–1030
doi:10.1093/bioinformatics/bts062

Sequence analysis
Advance Access publication January 31, 2012

WegoLoc: accurate prediction of protein subcellular localization using weighted Gene Ontology terms

Sang-Mun Chi[1],* and Dougu Nam[2],*

[1]School of Computer Science and Engineering, Kyungsung University, Nam-gu, Suyoung-ro 309, Pusan and [2]School of Nano-Bioscience and Chemical Engineering, UNIST, UNIST-gil 50, Eonyang-eup, Ulsan, South Korea

Associate Editor: Martin Bishop

ABSTRACT

Summary: We present an accurate and fast web server, WegoLoc for predicting subcellular localization of proteins based on sequence similarity and weighted Gene Ontology (GO) information. A term weighting method in the text categorization process is applied to GO terms for a support vector machine classifier. As a result, WegoLoc surpasses the state-of-the-art methods for previously used test datasets. WegoLoc supports three eukaryotic kingdoms (animals, fungi and plants) and provides human-specific analysis, and covers several sets of cellular locations. In addition, WegoLoc provides (i) multiple possible localizations of input protein(s) as well as their corresponding probability scores, (ii) weights of GO terms representing the contribution of each GO term in the prediction, and (iii) a BLAST E-value for the best hit with GO terms. If the similarity score does not meet a given threshold, an amino acid composition-based prediction is applied as a backup method.

Availability: WegoLoc and User's guide are freely available at the website http://www.btool.org/WegoLoc

Contact: smchiks@ks.ac.kr; dougnam@unist.ac.kr

Supplementary information: Supplementary data is available at http://www.btool.org/WegoLoc.

Received on August 25, 2011; revised on January 19, 2012; accepted on January 27, 2012

The knowledge of protein subcellular localization (PSL) provides primary information for protein functions and many biotechnological applications (Glory and Murphy, 2007). As a massive number of novel proteins have been sequenced, accurate computational methods for PSL prediction have become essential in many bioinformatics analyses. Such prediction methods may be classified into different categories according to the features of the query sequences used. The most widely used methods are based on the amino acid compositions and sorting signals (Emanuelsson *et al.*, 2007; Höglund *et al.*, 2006; Horton *et al.*, 2007; Nair and Rost, 2005; Pierleoni *et al.*, 2006). Other powerful approaches use textual information (Brady and Shatkay, 2008; Fyshe *et al.*, 2008; Nair and Rost, 2002) or Gene Ontology (GO) annotations for the input (or the most similar) proteins (Blum *et al.*, 2009; Briesemeister *et al.*, 2009; Chi, 2010; Huang *et al.*, 2008; Lei and Dai, 2006; Mei *et al.*, 2011; Shen and Chou, 2009).

In particular, GO-based methods have been among the best predictors for PSL (Chi, 2010; Mei *et al.*, 2011). This may be due to the strong correlation between GO annotations and PSL, which implies that the GO annotations solely contain substantial information for predicting PSL. Current GO-based methods use all the GO terms as the information sources or select a small number of informative GO terms (Huang *et al.*, 2008; Lei and Dai, 2006; Lu and Hunter, 2005; Mei *et al.*, 2011; Shen and Chou, 2009). However, most of such methods assume that each GO term has an equal amount of contribution in predicting PSL, which limits the discriminating power of the methods.

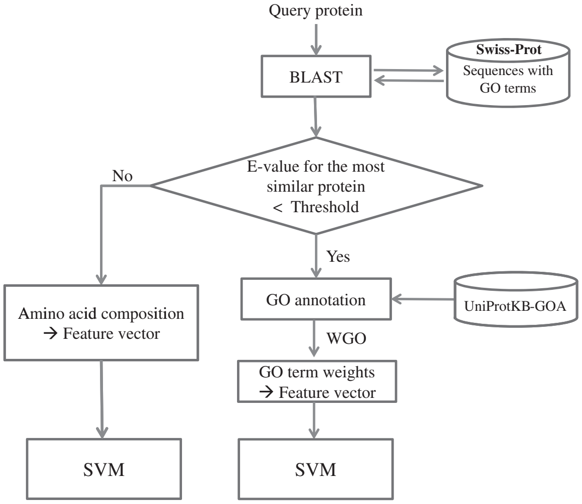
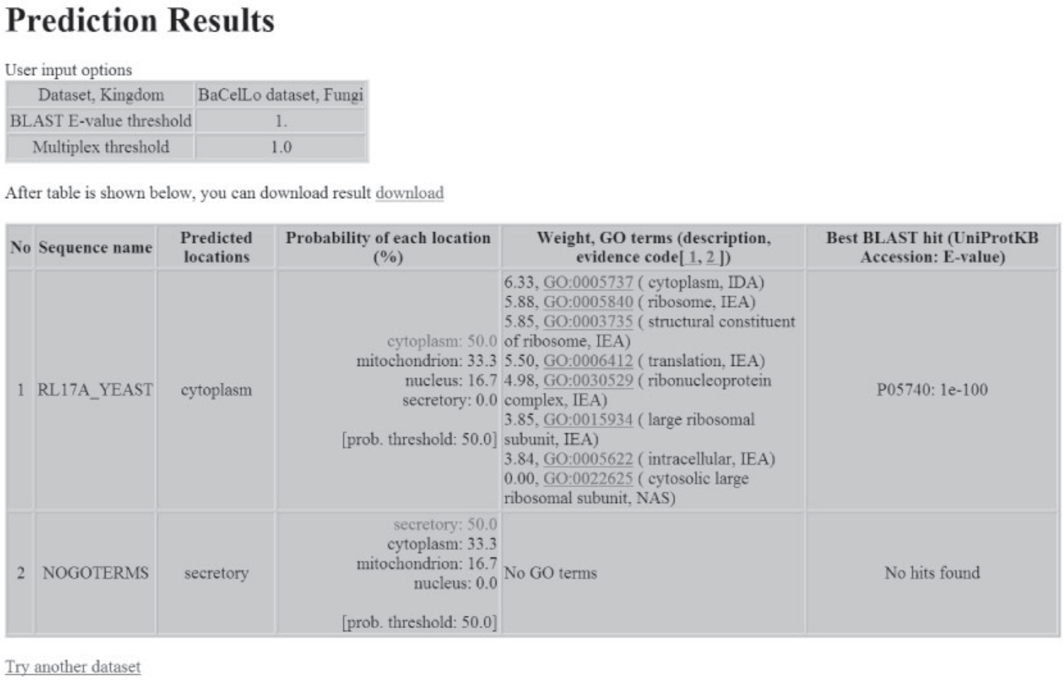

Query protein
BLAST
Swiss-Prot
Sequences with GO terms
E-value for the most similar protein < Threshold
No
Yes
Amino acid composition → Feature vector
GO annotation
UniProtKB-GOA
WGO
GO term weights → Feature vector
SVM
SVM

Fig. 1. The process of WegoLoc prediction. Either WGO terms or amino acid compositions are used for an SVM classifier.

To address the problem, a weighted GO algorithm (WGO) was introduced by one of the authors (Chi, 2010): for a given set of subcellular localizations, WGO calculates the log chi-square statistic for each GO term under the assumption of independence between GO terms and localizations. The chi-square statistic summarizes the differences between the observed and expected frequencies of proteins that share a GO term in each localization, and is used for weighting the GO term. A large chi-square value is assumed to indicate a high discriminating power of the GO term. The collection of the log chi-square values serves as the feature vector for the support vector machine (SVM) classifier, LIBSVM (Chang and Lin, 2001). See the paper by Chi (2010) for a detailed description of the algorithm.

WGO outperformed most state-of-the-art methods on previously used test datasets (Chi, 2010). In subsequent experiments, we identified several key options whose combination enabled the

*To whom correspondence should be addressed.



Prediction Results

User input options

Dataset, Kingdom	BaCelLo dataset, Fungi
BLAST E-value threshold	1.
Multiplex threshold	1.0

After table is shown below, you can download result download

No	Sequence name	Predicted locations	Probability of each location (%)	Weight, GO terms (description, evidence code[1, 2])	Best BLAST hit (UniProtKB Accession: E-value)
1	RL17A_YEAST	cytoplasm	cytoplasm: 50.0 mitochondrion: 33.3 nucleus: 16.7 secretory: 0.0 [prob. threshold: 50.0]	6.33, GO:0005737 (cytoplasm, IDA) 5.88, GO:0005840 (ribosome, IEA) 5.85, GO:0003735 (structural constituent of ribosome, IEA) 5.50, GO:0006412 (translation, IEA) 4.98, GO:0030529 (ribonucleoprotein complex, IEA) 3.85, GO:0015934 (large ribosomal subunit, IEA) 3.84, GO:0005622 (intracellular, IEA) 0.00, GO:0022625 (cytosolic large ribosomal subunit, NAS)	P05740: 1e-100
2	NOGOTERMS	secretory	secretory: 50.0 cytoplasm: 33.3 mitochondrion: 16.7 nucleus: 0.0 [prob. threshold: 50.0]	No GO terms	No hits found

Try another dataset

Fig. 2. A screenshot of an WegoLoc output.

Table 1. The sensitivity and accuracy of PSL predictors on several datasets (%)

Predictors	Ba (Ani.)	Ba (Fun.)	Ba (Pl.)	Ho (Ani.)	Ho (Fun.)	Ho (Pl.)	Sh (Hu.)
MultiLoc2	80/73	66/60	72/76	38/57	30/31	30/30	–
BaCelLo	69/64	71/57	61/69	–	–	–	–
WoLF PSORT	69/71	62/51	46/57	24/58	17/22	17/20	–
SherLoc2	76/71	61/59	69/69	39/54	–	–	–
GO-TLM	78/79	91/80	89/81	–	–	–	–
Hum-mPLoc 2.0	–	–	–	–	–	–	– / 63
WegoLoc	**98/98**	**99/99**	**95/98**	**94/95**	**92/92**	**94/90**	**80/72**

The numbers represent AVG/ACC and are given in percentages. The top-scoring AVG and ACC are highlighted in bold. Since most methods are applied for their own application domains, not all the test results are available. Ba, Ho and Sh stand for BaCelLo IDS, Höglund IDS and Shen human datasets, respectively. Ani., Fun., Pl. and Hu. stand for animals, fungi, plants and human, respectively.

high performance of WGO method: (i) classification algorithms (SVM versus KNN); (ii) annotation retrieving tools (GOA versus InterProScan); and (iii) weighting GO terms (weighted versus uniform weight). Each former option showed better accuracy. Further explanations and detailed test results are available in the Supplementary Material (Section 5). Selecting the former options indicated above, we constructed a web server called *WegoLoc* that implements the WGO algorithm (Chi, 2010).

Figure 1 describes the processing of an input sequence by WegoLoc. When a query sequence is entered, it searches GO-annotated proteins for the one with the highest sequence similarity over a given threshold, and uses all the corresponding GO terms and their weights to classify the input sequence. If no such sequence is found, the amino acid composition of the query sequence instead of the GO weights is used as a feature vector for the SVM classifier. In such cases, the GO terms and BLAST *E*-value are not available as illustrated in the second result (No. 2) of the table in Fig. 2. WegoLoc supports three eukaryotic kingdoms (animals, fungi and plants) and in particular human for several sets of cellular locations.

Table 1 shows the performance of PSL predictors on several test datasets. The first one, the BaCelLo independent dataset (IDS) (Casadio *et al.*, 2008) consists of proteins with <30% sequence identity with those in the BaCelLo training dataset (Pierleoni *et al.*, 2006). This dataset covers four localizations for animal and fungal proteins and five localizations for plant proteins. The second, the Höglund IDS (Höglund *et al.*, 2006) is collected in a way similar to the BaCelLo's. The training data in the Höglund dataset covers 9 localizations for animal and fungal proteins and 10 localizations for plant proteins. The last, the Human dataset (Shen and Chou, 2009) covers 14 localizations. The performance of each PSL predictor is measured by the average sensitivity (AVG) and the overall accuracy (ACC). Let $SE_i = T_i/n_i$, where n_i is the number of proteins and T_i is the number of correctly predicted proteins at the localization i, then AVG$=\sum_{i=1}^{C} SE_i/C$ and ACC$=\sum_{i=1}^{C} T_i/N$, where C is the number of subcellular locations used and N is the total number of proteins in the test set.

As shown in Table 1, WegoLoc outperformed other methods [MultiLoc2 (Blum *et al*., 2009), BaCelLo (Pierleoni *et al*., 2006), WoLF PSORT (Horton *et al*., 2007), SherLoc2 (Briesemeister *et al*., 2009), GO-TLM (Mei *et al*., 2011) and Hum-mPLoc 2.0 (Shen and Chou, 2009)] on each test dataset. Further explanations on the datasets, performance criteria and prediction accuracies of WegoLoc on each localization are available in the Supplementary Material (Sections 1–4).

Comparing it with other tools, WegoLoc provides a wide coverage on eukaryotic species, locations and test datasets (Table 1). Moreover, it provides fast computation for predicting PSL. It takes 2 s per sequence on average, though the time may vary depending on the network condition and the number of users, and will further be speeded up by adopting the accelerated versions of the BLASTP and GPUs (graphics processing units) (Vouzis and Sahinidis, 2011): see Supplementary Fig. S1 for the current processing time of WegoLoc. The input format for processing multiple sequences is explained in the User's guide from our web page. It allows up to 3000 sequences at a time in principle, but we recommend using $\leqslant$500 sequences per each execution. Lastly, it provides detailed information as shown in Fig. 2: (i) multiple possible localizations as well as the corresponding probability scores if the multiplex threshold is set below 1.0: localizations with a probability score higher than [multiplex threshold × highest probability score] will also be assigned to the query protein; (ii) description and links for GO terms and their weights; and (iii) the BLAST E-value for the best hit with GO terms. In addition to this web output, all the results can be downloaded as a text file with a well-defined format for easy parsing. See Supplementary Material (Section 6) for detailed explanations on the prediction results.

Overall, we constructed a highly accurate and fast predictor, WegoLoc, for PSL by combining several key components for the system including the WGO method (Chi, 2010). WegoLoc provides a wide coverage on eukaryotic species and datasets as well as multiple localizations of proteins for the query sequences. Currently, we only provide related GO terms for describing the input protein, but plan to extend the interpretation of our predictions e.g. implications for disease (Park *et al*., 2011).

Funding: Kyungsung University Research grants in 2011.

Conflict of Interest: none declared.

REFERENCES

Blum,T. *et al*. (2009) MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics*, **10**, 274.

Brady,S. and Shatkay,H. (2008) EpiLoc: a (working) text-based system for predicting protein subcellular location. *Pac. Symp. Biocomput.*, **13**, 604–615.

Briesemeister,S. *et al.* (2009) SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins. *J. Proteome Res.*, **8**, 5363–5366.

Casadio,R. *et al*. (2008) The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Brief. Funct. Genomic Proteomic*, **7**, 63–73.

Chang,C.-C. and Lin,C.-J. (2001) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 27:1–27:27.

Chi,S.M. (2010) Prediction of protein subcellular localization by weighted gene ontology terms. *Biochem. Biophys. Res. Commun.*, **399**, 402–405.

Emanuelsson,O. *et al.* (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.*, **2**, 953–971.

Fyshe,A. *et al.* (2008) Improving subcellular localization prediction using text classification and the gene ontology. *Bioinformatics*, **24**, 2512–2517.

Glory,E. and Murphy,R.F. (2007) Automated subcellular location determination and high-throughput microscopy. *Dev. Cell*, **12**, 7–16.

Höglund,A. *et al.* (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics*, **22**, 1158–1165.

Horton,P. *et al.* (2007) WoLF PSORT: protein localization predictor. *Nucleic Acids Res.*, **35**, W585–W587.

Huang,W.L. *et al.* (2008) ProLoc-GO: utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization. *BMC Bioinformatics*, **9**, 80.

Lei,Z. and Dai,Y. (2006) Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. *BMC Bioinformatics*, **7**, 491.

Lu,Z. and Hunter,L. (2005) Go molecular function terms are predictive of subcellular localization. *Pac. Symp. Biocomput.*, **10**, 151–161.

Mei,S. *et al*. (2011) Gene ontology based transfer learning for protein subcellular localization. *BMC Bioinformatics*, **12**, 44.

Nair,R. and Rost,B. (2002) Inferring sub-cellular localization through automated lexical analysis. *Bioinformatics*, **18** (Suppl. 1) S78–S86.

Nair,R. and Rost,B. (2005) Mimicking cellular sorting improves prediction of subcellular localization. *J. Mol. Biol.*, **348**, 85–100.

Park,S. *et al.* (2011) Protein localization as a principal feature of the etiology and comorbidity of genetic diseases. *Mol. Syst. Biol.*, **7**, 494.

Pierleoni,A. *et al.* (2006) BaCelLo: a balanced subcellular localization predictor. *Bioinformatics*, **22**, e408–e416.

Shen,H.B. and Chou,K.C. (2009) A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0. *Anal. Biochem.*, **394**, 269–274.

Vouzis,P.D. and Sahinidis,N.V. (2011) GPU-BLAST: using graphics processors to accelerate protein sequence alignment. *Bioinformatics*, **27**, 182–188.