UniLoc: A universal protein localization site predictor for eukaryotes and prokaryotes

Hsin-Nan Lin, Ching-Tai Chen, Ting-Yi Sung* and Wen-Lian Hsu*

Institute of Information Science, Academia Sinica, Taipei, Taiwan

* To whom correspondence should be addressed. Tel: +886-2-2788-3799; Fax: +886-2-2782-4814; Email: hsu@iis.sinica.edu.tw.

Correspondence may also be addressed to tsung@iis.sinica.edu.tw.

ABSTRACT

There is a growing gap between protein subcellular localization (PSL) data and protein sequence data, raising the need for computation methods to rapidly determine subcellular localizations for uncharacterized proteins. Currently, the most efficient computation method involves finding sequence-similar proteins (hereafter referred to as *similar proteins*) in the annotated database and transferring their annotations to the target protein. When a sequence-similarity search fails to find similar proteins, many PSL predictors adopt machine learning methods for the prediction of localization sites. We proposed a universal protein localization site predictor - UniLoc - to take advantage of implicit similarity among proteins through sequence analysis alone. The notion of related protein words is introduced to explore the localization site assignment of uncharacterized proteins. UniLoc is found to identify useful template proteins and produce reliable predictions when similar proteins were not available.

INTRODUCTION

The subcellular localization of a protein provides crucial information regarding the protein's biological function, and can be used to help to annotate newly sequenced genomes, to verify experimental results, and to identify potential drug targets [1]. Biologists world-wide have contributed localization data, but the gap between localization data and protein sequence data continues to increase, raising the need for computation methods to rapidly predict subcellular localization for uncharacterized proteins based on sequence information alone. The SwissProt database currently contains more than a half million manually annotated and reviewed protein

records. The annotations are taken from the literature and curator-evaluated computational analysis, and computational analysis has played an important role in PSL annotation. Around 90% of protein records in Swiss-Prot database are annotated through sequence similarity. For protein sequences without obvious sequence-similar proteins, a PSL predictor helps to provide plausible information quickly.

The most efficient approach to predict the subcellular localization sites for a target protein is to use similar proteins in the Swiss-Prot database as templates, and to transfer their annotations to the target sequence. When a similarity search fails to find such proteins, various protein subcellular localization (PSL) predictors use different features including amino acid composition[2-8], sequence patterns[9-12], physical-chemical features of protein sequences[11, 13-18], biological features from the literature and databases[4, 19-23], Gene Ontology annotations[24-31], and protein-protein interactions[32, 33]. Some approaches develop algorithms to combine the prediction results of existing PSL predictors and produce the consensus prediction, such as SUBAcon [34].

Many PSL predictors were built specifically for a target species due to the difference in subcellular locations of similar or homologous proteins. Users choose a target species to avoid noisy sequences. However, we found that proteins with the same localization sites in different species may share an implicit similarity, and a PSL predictor could take advantage of such similarities among proteins regardless of their species. It is also observed that sophisticated machine learning algorithms in state-of-the-art PSL predictors can make it difficult to trace the prediction back the most important features leading to a particular prediction [31]. Such information may be helpful for biologists to further determine what features are likely responsible for the localization.

In this study, we present UniLoc for PSL prediction. UniLoc is a successor of KnowPred_{site} [9]. The major improvements of UniLoc include: 1) UniLoc is able to predict single- and multi-localized proteins, whereas KnowPred_{site} could only estimate a probability of each localization site. So, it is difficult for KnowPred_{site} to evaluate its performance on the multi-localized protein prediction and to compare with existing methods. The performance of UniLoc is compared with state-of-the-art methods and a PSIBLAST-based predictor. 2) UniLoc identifies template proteins based on implicit local similarities and makes the prediction with the identified significant

templates, whereas KnowPred_{site} makes predictions with all possible proteins. 3) UniLoc is trained and evaluated with non-redundant proteins of eukaryotes and prokaryotes, whereas KnowPred_{site} was trained and evaluated with a set of redundant eukaryotic proteins. UniLoc was trained with a non-redundant data set consisting of 14754 eukaryotic proteins and 2585 prokaryotic proteins covering 1377 species. It not only predicts the localization sites for a protein sequence, but also provides potentially related proteins. In addition, UniLoc offers confidence scores to indicate the reliability of the prediction. The UniLoc web server is available at <http://bioapp.iis.sinica.edu.tw/UniLoc/>.

MATERIALS AND METHODS

Data sets

A set of 538,010 proteins were obtained from UniProt/SwissProt database. Proteins with fewer than 50 amino acids or without experimentally determined PSL annotations were removed, leaving a subset of 65,506 proteins. Localization annotations were further categorized according to the taxonomy described in [35]. To construct a non-redundant benchmark data set, PSI-CD-HIT was used to exclude proteins that shared greater than 30% sequence identity with any other protein. Subsequently, localization sites were not considered if the number of proteins in a category was less than 30. Eventually, we have two benchmark data sets, one containing 14,754 eukaryotic proteins (denoted as *SP-Euk*), and the other containing 2,585 prokaryotic proteins (denoted as *SP-Prok*).

The numbers of localization sites of *SP-Euk* and *SP-Prok* are 16 and 5, respectively. For *SP-Euk*, 8798 proteins (~60%) are single-localized and 5956 proteins (~40%) are multi-localized. In contrast, most prokaryotic proteins (~92%) in *SP-Prok* are single-localized. *SP-Euk* and *SP-Prok* respectively embrace 836 and 541 species.

Similar words vs. related words

In the field of computational linguistics, n-gram models have been shown to achieve good results for document classification [36, 37]. Similar techniques have also been applied to classify and characterize protein sequences. N-grams in a protein sequence are a set of substrings of a fixed length n and are often referred to as protein words. Thus a protein sequence can be viewed as a

```

Query: 7  EWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDR 46
      ++ +VL  W  VEAD A HG  +L RLF  HPETL+ F +
Sbjct: 3  DFDMLKCVGPVEADHATHGSLVLRFLFTEHPETLKLFPK 42

```

Figure 1. An example of significant sequence alignment. The alignment shows that the sequence fragment from position 7 to position 46 of the query sequence is highly similar with that from position 3 to position 42 of the subject sequence.

text written in the protein language consisting of 20 amino acids. The mapping of protein words to protein functions is conceptually similar to the mapping of words to meaning.

BLAST's search algorithm [38] uses a sliding window to generate all possible words (referred to as the *initial word set*) that appear throughout the target sequence. Each word is then used as a seed to identify potentially similar proteins. To increase the sensitivity of protein words, BLAST defines similar words with respect to a word in the initial word set based on the alignment score of two words.

Combining the idea of similar words and the mapping of protein words to protein functions, we present the concept of related words to extend similar words. Given a protein sequence p , we perform a PSI-BLAST search against NCBI NR database. PSI-BLAST generally returns a number of sequence alignments (HSP with e-value < 0.001) between p and its similar proteins sp . All the words in p and sp are extracted by a sliding window of size n . Given a word w (without gaps) in p , its *related word* is defined as the word w' (without gaps) in sp that is aligned with w . It is noteworthy that the same word w can be related to a number of words w' if it is aligned with different protein sequences. Unlike similar words which were defined based on the alignment score of the word pair alone, related words are defined based on a significant sequence alignment of two proteins. In the example of Figure 1, EWQLVLH and DFDMLVL are related words, and so are HVWAKVE and KCWGPVE. We use a sliding window of size n to screen the alignment to extract all related word pairs. No gap is allowed in a related word pair, thus every word represents an n -gram. In the example of Figure 1, we can collect 34 related word pairs of length 7 from the alignment.

The choice of word length n is a trade-off between specificity and sensitivity, i.e., a larger n yields high specificity, whereas a smaller n produces high sensitivity. In sequence similarity

identification, a smaller n is preferred. The word length in this study was empirically set as 7 to achieve balance between specificity and sensitivity [9, 11, 39].

Given a protein sequence p , we apply PSI-BLAST[40] to search against NCBI nr database and obtain a group of high-scoring segment pairs (HSPs). Each HSP represents a pairwise sequence alignment between p and its similar protein sp . Every word w in p may obtain a number of distinct related words w' in sp . If the localization sites of protein p have been experimentally determined, the information can be hypothetically transferred to each of its related words. In this way, we can build a word dictionary, called *RelatedDict* given a group of proteins with subcellular localization information and their HSPs. An example of a related word entry in *RelatedDict* is shown in Table 1, where protein sources represent those proteins which contain an initial word related to the word entry (MSQTAPS) and has been annotated with localization information. In this example, there are four protein sources, three of which are single-localized into the cytoplasm, and one is multi-localized into the cytoplasm and nucleus.

Table 1. An example of a related word entry in *RelatedDict*. The word, MSQTAPS, is related to the words of the four protein sequences and it inherits their location information.

Word: MSQTAPS	
Protein Source	Localization sites
FNIP1_HUMAN	Cytoplasm
LYPA1_RAT	Cytoplasm
LYS2_SCHPO	Cytoplasm
LYS5_SCHPO	Cytoplasm; Nucleus

Identification of template proteins using related words and localization prediction

Given a target protein p and its related word set extracted from all the HSPs, each related word is used to match against the *RelatedDict*. If a word w matches a dictionary entry exactly, it means every protein source in this word entry shares a related word with p . UniLoc identifies template proteins based on the number of shared related words. Suppose there are N related words in the word set of p matched against *RelatedDict* and a protein q shares f related words with p , we define the *intersection percentage* of q over p as $(100 \times f / N)\%$. Protein q is considered to be a *template protein* if its intersection percentage is over 10%. In case a template protein is found, UniLoc

predicts the localization sites of p based on the top 5 template proteins. Assume $t_1, t_2, \dots, t_n$ ($n \leq 5$) are the top n template proteins, the confidence score $CS(s_i)$ for localization site s_i is defined as the ratio of its frequency to n . For example, if three out of five template proteins are localized into the nucleus, then $CS(nucleus)$ is $3/5 = 60\%$.

If none of template proteins can be found, UniLoc predicts the localization sites of p based on the individually matched related words. Given a matched word w , the confidence score $CSw(s_i)$ for localization site s_i is defined as the ratio of its frequency to the number of protein sources of the entry in *RelatedDict*. Take the related word in Table 1 as an example, $CSw(cytoplasm)$ is 100 ($= 100 \times 4 / 4$), whereas the $CSw(nucleus)$ is 25 ($= 100 \times 1 / 4$). If there are N matched word entries, the confidence score $CS(s_i)$ is the summation of all $CSw(s_i)$ divided by N , so that $CS(s_i)$ is normalized between 0 and 100.

The prediction result is derived from the confidence scores of all localization sites. For a single-localized protein, the prediction is simply the localization site with the highest confidence score; however, for a multi-localized protein, determining the prediction is more difficult because it is unclear how many localization sites should be reported. To handle this situation, we modify the multi-localized confidence score described in [11], denoted as $MLCS(s_i)$, to determine the prediction according to the confidence scores of all the sites. The $MLCS(s_i)$ for localization site s_i is defined as follows:

$$MLCS(s_i) = CS(s_1) + CS(s_i) - \frac{CS(s_1)^2 - CS(s_i)^2}{100},$$

where s_1 is the localization site with the highest score. UniLoc always reports s_1 (i.e., the top1) in the prediction and reports the remaining localization site s_i if the following two criteria are satisfied: 1) $CS(s_i) > CSthreshold$ and 2) $MLCS(s_i) > MLCSthreshold$. We will discuss the performance of different $CSthreshold$ and $MLCSthreshold$ in the results section.

A PSI-BLAST based prediction method

Since PSI-BLAST is a powerful tool for sequence similarity search, it has been widely used in the identification of protein templates for function and structure prediction. The most significant feature of PSI-BLAST is that it iteratively searches a protein database for similar sequences and uses position-specific scoring matrices to identify distantly related sequences. PSI-BLAST is more

sensitive than BLAST, therefore we implemented a predictor based on PSI-BLAST search results. Given a target protein p , we perform a PSI-BLAST search with iteration 3 against the benchmark data set (*SP-Euk* or *SP-Prok*) and use the best five proteins as templates for localization assignment. The assignment strategy is similar to that of UniLoc, except that UniLoc identifies templates by looking up related words in the *RelatedDict*. If PSI-BLAST fails to find any protein, no prediction will be made. The performance of this PSI-BLAST based predictor is used as the baseline for the benchmark data sets.

Evaluation metrics

In this study, we performed 10-fold cross-validation on *SP-Euk* and *SP-Prok* to estimate the performance of UniLoc. In each experiment, a *RelatedDict* was compiled using one of the two data sets and PSL annotations are excluded if they belong to the proteins in the test set. To evaluate the prediction result, the *Top1 success rate*, *precision*, *recall*, and *F-score* were calculated. Top1 success rate is defined as the precision of the localization site with the highest confidence score in predictions; precision is defined as $TP / (TP + FP)$; recall as $TP / (TP + FN)$, where TP, FP, and FN are the numbers of true positives (correctly predicted localization sites), false positives (incorrectly predicted localization sites), and false negatives (missing localization sites), respectively. And F-score is defined as $(2 \times \text{precision} \times \text{recall}) / (\text{precision} + \text{recall})$.

RESULTS

Efficacy of protein related words

Protein related words are the basic units we used for localization site prediction. To demonstrate the efficacy of related words, we analyse the average number of common related words between two single-localized proteins. In this analysis, we randomly selected 10,000 protein pairs with the same localization site and another 10,000 protein pairs with different localization sites. Figure 2 shows the average number of common related words between these protein pairs (i.e., same localization site vs. different localization site). It is clear that protein pairs with the same localization share more related words than those with different localization. For example, the average number of common related words between two Golgi apparatus localized proteins is 1486; however, the average number of common related words between two proteins, one localized in

Golgi apparatus and the other localized somewhere else, is only 72. This suggests that Golgi apparatus localized protein sequences are more conserved. The average number of common related words between two extracell localized proteins is 59; the average number of common related words between two proteins, one localized into extracell and the other localized somewhere else, is 19. This implies that extracell localized protein sequences are more diverse.

CStreshold and MLCStreshold and overall performance on benchmark data sets

We use the two thresholds, CStreshold and MLCStreshold to determine what localization sites should be outputted. To identify a proper combination of the two parameters and avoid overfitting the benchmark data sets, we first evaluated the F-score of different CStreshold values with a fixed MLCStreshold of 0 to identify the best CStreshold, and then evaluated the F-score of different MLCStreshold values with the best CStreshold. Table 2 summarizes the performance evaluation on *SP-Euk*. It shows that a CStreshold of 30 and an MLCStreshold of 60 produce the best performance. We use the same thresholds to estimate the performance of UniLoc on *SP-Prok*. The detailed performance of UniLoc and the PSI-BLAST-based predictor on the two benchmark data sets is shown in Table 3. UniLoc significantly outperforms the PSI-BLAST-based predictor, suggesting that related words can help detect implicit similarities and identify more template proteins with the same localization sites.

Figure 2. The average number of common related words analysis on protein pairs with the same or different localization site.

Table 2. Performance of UniLoc with different combinations of CSthreshold and MLCSthreshold

CSthreshold + MLCSthreshold (0)	10	20	30	40	50	60	70	80	90
F-score	68.3	67.5	68.6	67.8	67.4	66.9	66.5	66.4	66.3
MLCSthreshold + CSthreshold (30)	10	20	30	40	50	60	70	80	90
F-score	68.6	68.6	68.6	68.6	68.8	68.9	68.7	67.9	67.3

Table 3. The overall performance on the benchmark data sets

	<i>SP-Euk</i>		<i>SP-Prok</i>	
	UniLoc	PSI-BLAST predictor	UniLoc	PSI-BLAST predictor
Top1 success rate	71.9	57.7	66.2	50.7
Precision	65.0	51.4	71.2	49.7
Recall	73.2	58.1	81.5	51.8
F-score	68.9	54.5	76.0	50.7

Similarity among proteins from different species

We assume proteins with the same localization sites may share implicit similarity. To verify this hypothesis, we estimated UniLoc's performance with the constraint that prediction is made only using the annotation from proteins within the same species and compared to the performance without this constraint. From Table 4, it is clear that UniLoc's performance is much better without the species constraints. Thus we can expect that the more proteins involved in the related word dictionary, the better UniLoc can predict for a target protein.

Table 4. Performance comparison on *SP-Euk* with/without the species constraints.

<i>SP-Euk</i>	<i>No species constraints</i>	<i>With species constraints</i>	<i>Difference</i>
Top1 success rate	71.9	59.7	+12.2
Precision	65.0	54.5	+10.5
Recall	73.2	64.7	+8.5
F-score	68.9	59.2	+9.7

The reliability of confidence scores

Confidence scores play an important role in computational analysis and are indicators of prediction reliability. To estimate the efficacy of UniLoc's confidence scores, we analyse the relationship between confidence scores and the corresponding precisions. From Figure 3, it can be observed that there is a strong positive correlation between the two factors. The correlation coefficient is 0.95 suggesting that UniLoc's confidence scores can faithfully reflect prediction reliability.

Figure 3. Relationship between confidence scores and precision.

Match rate and prediction performance

Figure 4. Relationship between match rate and prediction precision. The two factors show a strong positive correlation with a correlation coefficient of 0.99.

A word related to a target protein contributes to the prediction only if it can be exactly matched to a dictionary entry. Intuitively, the more related words that are matched, the more information UniLoc obtains for the prediction. We define the *match rate* as the ratio of the number of matched entries to the total number of related words of the target protein. Figure 4 shows the relationship between match rate and prediction accuracy on *SP-Euk* and *SP-Prok*. It can be observed that proteins with a higher match rate had better prediction results, suggesting that the prediction performance of UniLoc can be further improved if *RelatedDict* is enriched by adding more annotated proteins and related words.

The relationship between intersection percentages and precisions

UniLoc identifies a template protein q according to its intersection percentage to the target protein. In general, the higher intersection percentage it has, the more similar it is to the target protein. Figure 5 shows the strong positive relationship between intersection percentage and prediction efficiency (correlation coefficient = 0.93). The average precision is above 60% when the

intersection percentage is above 10%. This shows that the intersection percentage can fairly reflect the PSL similarity to the target protein regardless of their sequence identity.

Comparison with existing predictors

We compare the prediction performance of UniLoc with several selected state-of-the-art methods. A PSL predictor was selected if it was built for general eukaryotic or prokaryotic proteins and is available online or as a standalone program. Among the selected methods, some adopt similarity search and some do not. To conduct fair comparisons, all methods are separated by category (eukaryota or prokaryota, denoted as euk/prok) and method (with or without similarity search, denoted as wts/wos). The first group includes YLoc [31] and Euk-mPLoc [28] (euk and wts); the second group includes PSortb [41] (prok and wts); the third includes CELLO[14], Wolf PSORT[18], and SherLoc2 [42] (euk and wos); and the fourth includes CELLO (prok and wos). CELLO provides different prediction modes so it is included in two groups.

Not every selected PSL predictor can handle the whole protein sequences in the benchmark data sets, and some only cover a subset of subcellular locations. Thus each method was only tested on

Figure 5. The relationship between intersection percentage and prediction accuracy. The correlation coefficient is 0.93.

a specialized subset of *SP-Euk/SP-Prok* with the locations covered by the method. When UniLoc is compared to PSL predictors with similarity search, a BLAST search for similar proteins is first performed in the SwissProt database. The annotation of the target protein is ignored and its prediction is made based on the top five similar proteins as we do in the PSI-BLAST based

prediction method. Table 5 summarizes the comparison between UniLoc and the selected methods in the first and second groups. UniLoc outperforms the selected PSL predictors on both the *SP-Euk* and *SP-Prok* data sets. The average F-scores of UniLoc, Euk-mPLoc2, and YLoc on *SP-Euk* are 84.3, 74.6 and 65.9, respectively; the average F-scores of UniLoc and PSortb on *SP-Prok* are 88.4 and 79.4, respectively. Table 6 summarizes the comparison between UniLoc and the selected methods in the third and fourth groups. Likewise, UniLoc compares favorably to the selected PSL predictors both on *SP-Euk* and *SP-Prok* for all measures, with the exception that WolfPSORT yields the highest recall as it sacrifices precision. As all selected methods were trained separately based on their own training sets, it is likely that some of the test proteins in the benchmark data sets were already involved in their training process. Thus their evaluation could be overestimated.

Table 5. Performance comparison among UniLoc and the selected methods with a homology search step.

Data set	Predictor	Top1 success rate	Precision	Recall	F-score
<i>SP-Euk</i>	UniLoc	86.8	82.9	85.8	84.3
	Euk-mPLoc2	80.0	79.8	70.0	74.6
	YLoc	65.4	63.4	68.5	65.9
<i>SP-Prok</i>	UniLoc	87.0	87.2	89.6	88.4
	PSortb	80.8	80.8	78.1	79.4

Table 6. . Performance comparison among UniLoc and the selected methods which do not use homology search.

Data set	Predictor	Top1 success rate	Precision	Recall	F-score
<i>SP-Euk</i>	UniLoc	71.9	65.0	73.2	68.9
	Wolf PSORT	64.6	49.0	81.5	61.2
	CELLO	63.5	62.9	56.6	59.6
	SherLoc2	64.0	64.0	51.7	57.2

<i>SP-Prok</i>	UniLoc	66.2	71.2	81.5	76.0
	CELLO	53.9	52.2	56.3	54.2

CONCLUSIONS

In this study, we demonstrated that UniLoc produced the best performance on the two benchmark data sets compared with several selected PSL methods. We present the concept of related protein words and explore subcellular localization site assignment of an uncharacterized protein by means of identifying its templates. UniLoc identifies template proteins based on the intersection of related words with target proteins.

In summary, we have demonstrated that related words can help to detect the implicit similarity between proteins with the same localization sites, and template proteins with high intersection percentages show a high positive correlation with target proteins. A limitation of UniLoc is the difficulty in producing reliable predictions for proteins which share fewer words with the dictionary. However, this issue will gradually be addressed as more annotated proteins are included each year.

REFERENCES

1. Chou, K.C. and H.B. Shen, *Cell-PLOC: a package of Web servers for predicting subcellular localization of proteins in various organisms*. Nature Protocols, 2008. **3**(2): p. 153-162.
2. Pierleoni, A., et al., *BaCelLo: a balanced subcellular localization predictor*. Bioinformatics, 2006. **22**(14): p. E408-E416.
3. Xie, D., et al., *LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST*. Nucleic Acids Research, 2005. **33**: p. W105-W110.
4. Nair, R. and B. Rost, *Mimicking cellular sorting improves prediction of subcellular localization*. Journal of Molecular Biology, 2005. **348**(1): p. 85-100.
5. Chou, K.C. and Y.D. Cai, *Prediction and classification of protein subcellular location - Sequence-order effect and pseudo amino acid composition*. Journal of Cellular Biochemistry, 2003. **90**(6): p. 1250-1260.
6. Mooney, C., Y.H. Wang, and G. Pollastri, *SCLpred: protein subcellular localization prediction by N-to-1 neural networks*. Bioinformatics, 2011. **27**(20): p. 2812-2819.
7. Hua, S.J. and Z.R. Sun, *Support vector machine approach for protein subcellular localization prediction*. Bioinformatics, 2001. **17**(8): p. 721-728.

8. Guo, J. and Y.L. Lin, *TSSub: eukaryotic protein subcellular localization by extracting features from profiles*. *Bioinformatics*, 2006. **22**(14): p. 1784-1785.
9. Lin, H.N., et al., *Protein subcellular localization prediction of eukaryotes using a knowledge-based approach*. *Bmc Bioinformatics*, 2009. **10**.
10. Blum, T., S. Briesemeister, and O. Kohlbacher, *MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction*. *Bmc Bioinformatics*, 2009. **10**.
11. King, B.R. and C. Guda, *ngLOC: an n-gram-based Bayesian method for estimating the subcellular proteomes of eukaryotes*. *Genome Biology*, 2007. **8**(5).
12. Zybailov, B., et al., *Sorting Signals, N-Terminal Modifications and Abundance of the Chloroplast Proteome*. *Plos One*, 2008. **3**(4).
13. Hoglund, A., et al., *MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition*. *Bioinformatics*, 2006. **22**(10): p. 1158-1165.
14. Yu, C.S., et al., *Prediction of protein subcellular localization*. *Proteins-Structure Function and Bioinformatics*, 2006. **64**(3): p. 643-651.
15. Su, E.C.Y., et al., *Protein subcellular localization prediction based on compartment-specific features and structure conservation*. *Bmc Bioinformatics*, 2007. **8**.
16. Wang, J.R., et al., *Protein subcellular localization prediction for Gram-negative bacteria using amino acid subalphabets and a combination of multiple support vector machines*. *Bmc Bioinformatics*, 2005. **6**.
17. Chang, J.M., et al., *PSLOoc: Protein subcellular localization prediction based on gapped-dipeptides and probabilistic latent semantic analysis*. *Proteins-Structure Function and Bioinformatics*, 2008. **72**(2): p. 693-710.
18. Horton, P., et al., *WoLF PSORT: protein localization predictor*. *Nucleic Acids Research*, 2007. **35**: p. W585-W587.
19. Lee, K., et al., *PLPD: reliable protein localization prediction from imbalanced and overlapped datasets*. *Nucleic Acids Research*, 2006. **34**(17): p. 4655-4666.
20. Chou, K.C. and Y.D. Cai, *Predicting protein localization in budding yeast*. *Bioinformatics*, 2005. **21**(7): p. 944-950.
21. Bhasin, M., A. Garg, and G.P.S. Raghava, *PSLPred: prediction of subcellular localization of bacterial proteins*. *Bioinformatics*, 2005. **21**(10): p. 2522-2524.
22. Gardy, J.L., et al., *PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis*. *Bioinformatics*, 2005. **21**(5): p. 617-623.
23. Gardy, J.L., et al., *PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria*. *Nucleic Acids Research*, 2003. **31**(13): p. 3613-3617.
24. Li, L.Q., et al., *An Ensemble Classifier for Eukaryotic Protein Subcellular Location Prediction Using Gene Ontology Categories and Amino Acid Hydrophobicity*. *Plos One*, 2012. **7**(1).
25. Shen, H.B., J. Yang, and K.C. Chou, *Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction*. *Amino Acids*, 2007. **33**(1): p. 57-67.
26. Fyshe, A., et al., *Improving subcellular localization prediction using text classification and the gene ontology*. *Bioinformatics*, 2008. **24**(21): p. 2512-2517.
27. Wan, S.B., M.W. Mak, and S.Y. Kung, *mGOASVM: Multi-label protein subcellular localization based on gene ontology and support vector machines*. *Bmc Bioinformatics*, 2012. **13**.
28. Chou, K.C. and H.B. Shen, *A New Method for Predicting the Subcellular Localization of Eukaryotic Proteins with Both Single and Multiple Sites: Euk-mPLoc 2.0*. *Plos One*, 2010. **5**(3).
29. Huang, W.L., et al., *ProLoc-GO: Utilizing informative Gene Ontology terms for sequence-based prediction of protein subcellular localization*. *Bmc Bioinformatics*, 2008. **9**.

30. Chi, S.M. and D. Nam, *WegoLoc: accurate prediction of protein subcellular localization using weighted Gene Ontology terms*. Bioinformatics, 2012. **28**(7): p. 1028-1030.
31. Briesemeister, S., J. Rahnenfuhrer, and O. Kohlbacher, *YLoc-an interpretable web server for predicting subcellular localization*. Nucleic Acids Research, 2010. **38**: p. W497-W502.
32. Lee, K., et al., *Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species*. Nucleic Acids Research, 2008. **36**(20).
33. Shin, C.J., et al., *Protein-protein interaction as a predictor of subcellular location*. BMC Systems Biology, 2009. **3**.
34. Hooper, C.M., et al., *SUBAcon: a consensus algorithm for unifying the subcellular localization data of the Arabidopsis proteome*. Bioinformatics, 2014. **30**(23): p. 3356-3364.
35. Chou, K.C. and H.B. Shen, *Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization*. Biochemical and Biophysical Research Communications, 2006. **347**(1): p. 150-157.
36. Peng, F.C., D. Schuurmans, and S.J. Wang, *Augmenting naive Bayes classifiers with statistical language models*. Information Retrieval, 2004. **7**(3-4): p. 317-345.
37. Boyce, B.R., *Automatic Text-Processing - the Transformation Analysis, and Retrieval of Information by Computer - Salton, G.* Journal of the American Society for Information Science, 1990. **41**(2): p. 150-151.
38. Altschul, S.F., et al., *Basic Local Alignment Search Tool*. Journal of Molecular Biology, 1990. **215**(3): p. 403-410.
39. Lin, H.N., et al., *Improving protein secondary structure prediction based on short subsequences with local structure similarity*. BMC Genomics, 2010. **11**.
40. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. Nucleic Acids Research, 1997. **25**(17): p. 3389-3402.
41. Yu, N.Y., et al., *PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes*. Bioinformatics, 2010. **26**(13): p. 1608-1615.
42. Briesemeister, S., et al., *SherLoc2: A High-Accuracy Hybrid Method for Predicting Subcellular Localization of Proteins*. Journal of Proteome Research, 2009. **8**(11): p. 5363-5366.