FUSION OF CLEAVAGE SITE DETECTION AND PAIRWISE ALIGNMENT FOR FAST SUBCELLULAR LOCALIZATION

Man-Wai Mak

Dept. of Electronic and Information Engineering
The Hong Kong Polytechnic University,
Hong Kong SAR

Sun-Yuan Kung

Dept. of Electrical Engineering
Princeton University, USA

ABSTRACT

In recent years, homology-based and signal-based methods have been proposed for predicting the subcellular localization of proteins. While it has been known that homology-based methods can detect more subcellular locations than signal-based methods, the former generally requires a lot more computational resources during both training and prediction. The problem will become intractable for annotating large databases. One possible solution is to reduce the sequence length. This paper proposes to use the cleavage sites detected by signal-based methods (e.g., TargetP) to extract the sequence or profile segments that contain the most localization information for alignment. It was found that the method can reduce computation time of full-length alignment by 27-fold at a cost of only 8% reduction in prediction accuracy. Moreover, the method can increase the accuracy by 0.8% and at the same time reduce the computation time by 41%. Results also show that cutting the sequences at the cleavage sites detected by TargetP is better than cutting them at a fixed position.

Index Terms— Pairwise alignment; subcellular localization; cleavage sites; TargetP; profile; protein sequences.

1. INTRODUCTION

Determination of subcellular localization via experimental means is often time-consuming and laborious. As a result, the development of efficient and reliable computation techniques for annotating biological sequences has become increasingly important. Current approaches to subcellular localization either look for localization information (namely sorting signals) from short segments of amino acid sequences or extract relevant features from the whole sequences. The former is very fast, but it is limited to the detection of a few localizations only. The latter approach can theoretically detect as many localizations as in the training data, and its performance is usually better than just using the information of short segments. However, the approach requires a lot more computation resources for both training and prediction. A typical example is pairwise alignment in which an unknown sequence is aligned with each of the training sequences to form a score vector for classification. The idea is based on the notion that similarity (homology) in sequences, to a certain extent, also means closeness in function and structure.

This paper attempts to mitigate the computation burden of alignment-based approach by using the information provided by the signal-based approach. To this end, a cascaded fusion of the two approaches is proposed, where the segments that are rich in localization information are used for pairwise alignment. Experimental results on a large dataset show that the method can make use of the best property of both approaches and can reduce the computation time by 27 folds at just a 8% reduction in accuracy. We advocate that the method will be important for biologists to conduct large-scale protein annotation or for bioinformaticians to perform preliminary investigations on new algorithms that involve pairwise alignments.

This work was in part supported by The Research Grant Council of the Hong Kong SAR (Project Nos. PolyU 5241/07E and A-PH18. We thank Center for Biological Sequence Analysis, Technical University of Denmark for providing us the stand-alone version of TargetP.

2. SUBCELLULAR LOCALIZATION: BIOLOGICAL PERSPECTIVE

A cell contains internal membranes that enclose a number of compartments called organelles. Different organelles specialize in different functions. The majority of proteins are synthesized in the cytoplasm, a region between the nuclear and the plasma membrane. While many of them will remain in the cytoplasm, many others will need to be delivered to particular organelles within the cell or to the cell surface (secretion). For a cell to function properly, the proteins should be delivered to the correct compartments. The mechanism of delivering proteins to their respective organelles or outside the cell is called protein sorting or protein targeting, and the division of a cell into different organelles is called subcellular localization.

A protein can be represented by a sequence of amino acids arranged from the left to right. The left-end that contains a free amino is called the N-terminus and the right-end that contains a free carboxyl group is called the C-terminus. The amino acid sequence of a protein contains information about its organelle destination. Typically, the information can be found within a short segment of 20–100 amino acids. These short segments are generally known as sorting-signal sequences, targeting sequences, or signal peptides. For most proteins, the targeting sequences can be found in the N-terminus. But there are also exceptions. For instance, the ER retention signal that keeps the proteins in the endoplasmic reticulum can be found in the C-terminus [1]. For some proteins, the targeting sequences will be immediately cleaved from the polypeptide at the "cleavage site" once the translocation process has been completed [2].

3. SUBCELLULAR LOCALIZATION: COMPUTATIONAL PERSPECTIVE

Over the years, a number of *in-silico* subcellular localization methods have been proposed. These methods either look at the sorting signals of the amino acid sequences or extract localization information from the whole sequence.

- **Sorting-Signal Based Methods**. This group of methods locates the proteins based on the existence of sorting signals [3]. PSORT [4] is one of the earliest predictor that uses sorting signals. Subsequent approaches use signal peptides, mitochondrial targeting peptides, and chloroplast transit peptides [5–7].
- **Whole-Sequence Based Methods**.[1] This type of method extracts localization information from the whole sequence. The methods are based on the notion that homologous sequences are also likely to have the same subcellular location [9]. Proteins are represented by sequences of alphabets. Because most classifiers work on numbers instead of strings, it is necessary to convert sequences of alphabets to fixed-length vectors for classification. One popular approach is to perform pairwise alignment between a query sequence with each of the training sequences, forming a score vector with dimensionality equal to the number of training sequences [10].

  The pairwise sequence alignment has also been extended to pairwise profile alignment to improve the sensitivity in detecting remote homolog and in classifying subcellular locations [11]. A profile is a matrix in which elements in a column specify the frequency of each amino acid appears in that sequence position. Given a sequence, a profile can be derived by aligning it with a set of similar sequences. The similarity score between a known and an unknown sequence can be computed by aligning the profile of the known sequence with that of the unknown sequence [11]. Because the comparison involves not only two sequences but also their closely related sequences, the score is more sensitive to detecting weak similarity between protein families. Research has also found that profile alignment can achieve better performance than sequence alignment in predicting subcellular locations [12].

Comparing the above two approaches, the sorting-signal based methods seem to be more direct, because they determine the localization from the sequence segments that contain the localization information. However, this type of method is typically limited to the prediction of a few subcellular locations only. For example, the popular TargetP [5] can only detect three localizations: chloroplast, mitochondria, and secretory pathway. The homology-based methods, on the other hands, can in theory predict as many localizations as available in the training data. The downside, however, is that the whole sequence is used for the homology search or pairwise alignment, without considering the fact that some segments of the sequence are more important or contain more information than the others. Moreover, the computation requirement will be excessive for long sequences. The problem will become intractable for database annotation where hundreds of thousands of proteins are involved.

4. COMBINING CLEAVAGE SITE DETECTION AND PAIRWISE ALIGNMENT

The computation burden of homology-based methods is mainly due to the alignment of the whole sequences. The question is: "Does every region of the sequence contain an equal amount of localization information?". The answer is a definitely 'No' because otherwise the signal-based methods will perform poorly. Then, the question becomes: "Which part(s) of the sequence should be used for alignment?" For this, the signal-based methods can provide a good solution because these methods scan the whole sequence to look for the signal peptide (i.e., informative region). In fact, the length of chloroplast transit peptide (cTP), mitochondrial targeting peptide (mTP), and secretory pathway signal peptide (SP) is under 100 amino acids only [1], as illustrated in Table 1. Given the fact that the majority of proteins in the Swissprot database have about a few hundred amino acids and that some proteins could have length up to 9000 amino acids,[2] tremendous computational saving can be obtained by aligning the pre-sequence region (from the N-terminus to the cleavage site) for those proteins containing a signal or targeting sequence.

Table 1. Length of chloroplast transit peptide (cTP), mitochondrial targeting peptide (mTP), and secretory pathway signal peptide (SP).

Peptide	Length (No. of Amino Acids)
cTP	20–100
mTP	6–85
SP	15–30

The above discussion suggests that the problem encountered by homology-based methods can be largely alleviated by a cascaded fusion of signal-based and homology-based methods. The fusion has three steps (see Fig. 1):

Step 1 **Cleavage site detection**. The cleavage site (if any) of a query sequence is determined by a signal-based method such as TargetP [5].

Step 2 **Pre-sequence selection**. The pre-sequence of the query is obtained by selecting from the N-terminus up to the cleavage site. If TargetP cannot find a site (a '-' character in the `TPlen` field) or the Reliability Class (RC) index is above a threshold η_{RC} (RC $= 1$ means most reliable and RC $= 5$ means least reliable), the sequence is cleaved-off at a default position P_{C} from the N-terminus.

Step 3 **Pairwise alignment**. The pre-sequence is aligned with each of the training pre-sequences to form a T-dim vector, which is fed to a one-versus-rest SVM classifier for prediction.

The training of the SVM classifier follows a similar procedure. More specifically, the cleavage sites of T training sequences are firstly detected by TargetP or set to the default position P_{C} if TargetP cannot find any cleavage sites. Then, T training pre-sequences are obtained by cleaving off at the corresponding cleavage sites. Pairwise alignments are then performed to create a $T \times T$ symmetric score matrix whose column vectors are used to train a one-vs-rest SVM classifier. See [13] for the details of pairwise alignment and the training of SVM classifiers.

Note that the same training and prediction procedures are also applicable to the fusion of cleavage site detection and profile alignment. The only modification is that the pre-sequences in Steps 2 and 3 are replaced by pre-profiles, i.e., profiles starting from the left-end to the position corresponding to the cleavage site.

5. EXPERIMENTS AND RESULTS

5.1. Experiments

The dataset introduced by Haung and Li [14] was used in the experiments. This dataset was created by selecting all eukaryotic proteins with annotated subcellular locations from SWISSPROT 41.0 and by setting the identity cut-off to 50%. The dataset comprises 3572 proteins (622 cytoplasm, 1188 nuclear, 424 mitochondria, 915 extracellular, 26 golgi apparatus, 225 chloroplast, 45 endoplasmic reticulum, 7 cytoskeleton, 29 vacuole, 47 peroxisome, and 44 lysosome).

[1] This paper focuses on the homology-based method in this category. For a review of other methods, see [8].

[2] The dataset used in this work contains sequences with length between 61 and 5702 amino acids.

Table 2. Computation time and accuracy of profile alignment with or without cleavage site detection (CSD). For the rows with CSD, the cleavage sites (if any) were obtained by selecting the "Non-plant" option in TargetP; the overall accuracies are slightly lower if the "Plant" option is selected. *Default Pos.* P_C: Default cleaved-off position, i.e. the length of profiles used for alignment when TargetP cannot find a cleavage site. When this entry is an 'L', no cleaving will be applied to the profiles whose sequences do not have a cleavage site. *Time*: time taken on a 3.2GHz Pentium IV CPU for creating the whole pairwise scoring matrix (3572×3572) using profile alignment. *Speedup factor*: Speedup factor with respect to full-length alignment (198.9 hr.). *ProAlign*: pairwise profile alignment. *Cyt*: Cytoplasm; *Nuc*: Nuclear; *Mit*: Mitochondria; *Ext*: Extracellular; *Gol*: Golgi Apparatus; *Chl*: Chloroplast; *ER*: End. Reticulum; *Cto*: Cytoskeleton; *Vac*: Vacuole; *Per*: Peroxisome; *Lys*: Lysosome.

Default Pos. P_C	Method	Time (hr.)	Speedup Factor	Overall Acc. (%)	Accuracy of Individual Subcellular Locations (%)										
					Cyt	Nuc	Mit	Ext	Gol	Chl	ER	Cto	Vac	Per	Lys
L	TargetP(non-plant)	0.08	–	–	–	–	71.9	89.2	–	–	–	–	–	–	–
	TargetP(plant)	0.08	–	–	–	–	70.1	72.4	–	58.2	–	–	–	–	–
	ProAlign	198.9	–	75.3	51.1	90.3	66.5	89.7	15.4	59.6	44.4	0.0	0.0	46.8	36.4
	ProAlign + CSD	117.2	1.7	75.9	50.3	93.9	73.1	92.7	0.0	43.6	4.4	0.0	10.3	42.6	2.3
100	ProAlign	12.6	15.8	68.7	45.2	75.9	67.7	87.7	7.7	56.9	37.8	0.0	10.3	42.6	29.6
	ProAlign + CSD	8.6	23.1	70.2	47.3	82.1	65.8	88.1	7.7	53.8	0.0	0.0	13.8	44.7	11.4
80	ProAlign	9.9	20.1	68.0	43.7	75.3	67.0	87.4	3.8	58.2	37.8	28.6	10.3	31.9	22.7
	ProAlign + CSD	7.3	27.2	69.1	39.7	82.6	65.1	91.3	3.8	48.4	0.0	0.0	20.7	10.6	18.2
50	ProAlign	6.4	31.1	65.3	42.1	72.0	63.7	85.6	0.0	55.1	26.7	0.0	10.3	27.7	20.5
	ProAlign + CSD	5.4	36.8	65.3	42.8	73.8	62.3	87.3	0.0	50.2	2.2	0.0	10.3	23.4	6.8

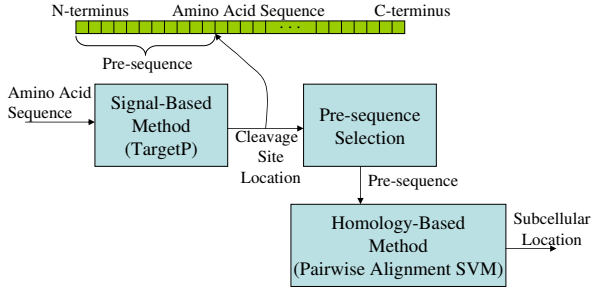

Fig. 1. Cascaded fusion of signal-based and homology-based methods for speeding up the prediction process. The signal-based method, such as TargetP, is used as a pre-processor that reduces the sequence length for the computationally expensive homology-based method. For profile alignment, the pre-sequence is replaced by a pre-profile and pre-sequence selection becomes pre-profile selection.

We used TargetP for cleavage site detection and PairProSVM [12] for classification of TargetP-cleaved profiles. The Reliability Class threshold η_{RC} mentioned in Section 4 was set to 3, and the default cleaved-off position P_C was set to 50, 80, 100, or L, where L is the sequence length. Note that when $P_C = L$, only the sequences with a cleavage site will have length shortened, and no cleaving will be applied to those without a cleavage site. We measured the time taken to create a 3572×3572 alignment-score matrix on a 3.2GHz Pentium IV CPU for different P_C. And, for each P_C, the alignment time with or without cleavage site detection was recorded. The alignment time for full-length profiles was used as a reference against which the computation time of aligning TargetP-cleaved profiles is compared.

The results for profile alignment are shown in Table 2.[3] The first column specifies the default cleaved-off positions P_C for the profiles. The length of the profiles for alignment depends on whether cleavage site detection (CSD) was applied or not. For the rows without CSD, all profiles were cleaved off at P_C or at the C-terminus end, whichever is smaller. On the other hand, rows with CSD means that sequences with detectable cleavage sites were cleaved off at the sites, and sequences that do not have a detectable site were cleaved off at the default positions P_C. When $P_C = L$ and no CSD was applied, full-length profiles were used for alignment.

[3] A similar trend was also observed in the sequence alignment case.

5.2. Performance of TargetP

The first two rows of Table 2 show the performance of TargetP on the dataset. Note that TargetP can only detect mitochondria targeting peptide (mTP) and secretory pathway signal peptide (SP)—extracellular—when users select the option "Non-plant", and it can detect mTP, SP, and chloroplast transit peptide (cTP) when the option "Plant" is selected. Because neither the protein IDs nor accession numbers are known, we have tried both options. The results show that TargetP's performance on non-plant is slightly better, because the dataset contains mainly non-plant proteins. The results also show that TargetP is very fast. Bear in mind, however, that TargetP can only detect at most three types of proteins, and it requires users to select plant/non-plant before performing prediction. This restriction significantly limits the applicability of TargetP.

5.3. Performance of Cascaded Fusion

Table 2 shows that the computation time for full-length profile alignment is a striking 199 hours, which suggests that full-length alignment is computationally prohibitive for most practical applications. Therefore, it is imperative to limit the length of the sequences or profiles before alignment. The advantage of limiting the length will become evident when we compare the alignment time between the full-length "ProAlign" (3rd row) and "ProAlign + CSD" (last row). The reduction in computation time is 36 folds (from 199 hours to 5.4 hours). This dramatic reduction comes with a price of 13% reduction in accuracy (from 75.3% to 65.3%). This result suggests that the cascaded fusion of TargetP and pairwise alignment allows biologists to trade speed with accuracy.

Because it is impractical to perform full-length alignments and shortening the sequences or profiles is a viable solution, it is important to compare the computation time (3rd column) and overall ac-

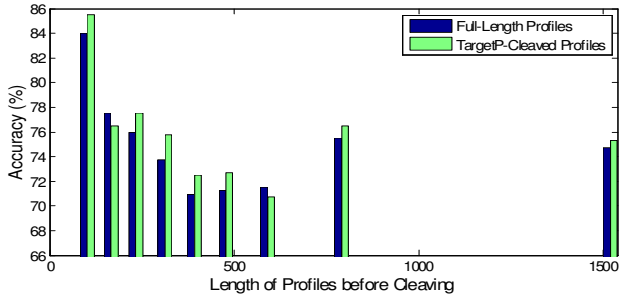

Fig. 2. . Subcellular localization accuracies of profiles with different length before (blue bars) and after (green bars) applying TargetP for pre-profile selection. For the green bars, full-length profiles were used for alignment if TargetP cannot find any cleavage sites. Each bar specifies the average accuracy of a group of 400 profiles whose mean full-length before pre-profile selection is shown on the horizontal axis.

curacy (5th column) between "ProAlign" and "ProAlign + CSD" for different default cleaved-off positions (50, 80 and 100). Evidently, aligning TargetP-cleaved profiles can reduce the computation time by up to 31.7% (from 12.6 hr to 8.6 hr) and increase the overall accuracy by 2.2% (from 68.7% to 70.2%). This suggests that it is possible to achieve the best of both worlds: reduction in computation time and increase in prediction accuracy.

We also observe that CSD degrades the prediction performance of chloroplast (Chl) slightly because TargetP is not able to predict cTP when the non-plant option is selected. In such case, the cleavage site locations for chloroplast predicted by TargetP may be inaccurate. CSD can, however, increase the prediction accuracy of Extracellular. This is because the sTP can be found in the N-terminus, and removing the amino acids beyond the cleavage site helps the alignment focuses on the relevant features in the sequences and disregard noise. Comparing the rows with CSD against the ones without CSD suggests that CSD hurts the prediction performance of endoplasmic reticulum (ER) and lysosomes (Lys) significantly. This is primarily because of the removal of the C-terminal retention signal [1].

Fig. 2 plots the prediction accuracy against profile length with CSD (green bars) or without CSD (blue bars). Each bar represents the mean accuracy of a group of 400 profiles whose mean length (before shortening) is shown on the horizontal axis. The figure shows two interesting phenomena: (1) the accuracy is generally higher for short profiles and (2) shortening the profiles whose corresponding sequences have a detectable cleavage site helps improve the performance. These phenomena arise mainly because the key localization information is located at the N-terminus and long profiles (or sequences) have more irrelevant regions that interfere the alignment.

6. CONCLUSIONS

In this paper, a fusion method that combines the advantages of TargetP and pairwise alignment has been proposed for speeding up subcellular localization prediction. There are four key findings: (1) aligning TargetP-cleaved profiles is significantly faster than aligning full-length profiles, with just a small reduction in prediction accuracy; (2) cutting the profiles at the cleavage sites detected by TargetP can achieve a higher accuracy and prediction speed than cutting them at a fixed position; (3) the non-plant option in TargetP leads to incorrect detection of chloroplasts' cleavage sites, which in turn has a significant impact on the prediction accuracy of chloroplasts; and (4) ignoring the C-terminus part of the sequences or profiles in pairwise alignment degrades the prediction accuracy of endoplasmic reticulum and lysosomes significantly.

7. REFERENCES

[1] 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.

[2] H. Lodish, et al., *Molecular cell biology*, New York: W.H. Freeman, 6th edition, 2008.

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

[4] 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.

[5] 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.

[6] H. Nielsen, J. Engelbrecht, S. Brunak, and G. von Heijne, "Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites," *Protein Engineering*, vol. 10, pp. 1–6, 1997.

[7] O. Emanuelsson, H. Nielsen, and G. von Heijne, "Chlorop, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites," *Protein Science*, vol. 8, pp. 978–984, 1999.

[8] J. Gardy and F. Brinkman, "Methods for predicting bacterial protein subcellular localization," *Nature Reviews Microbiology*, vol. 4, no. 10, pp. 741–751, 2006.

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

[10] L. Liao and W. S. Noble, "Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships," *J. Comput. Biol.*, vol. 10, no. 6, pp. 857–868, 2003.

[11] H. Rangwala and G. Karypis, "Profile-based direct kernels for remote homology detection and fold recognition," *Bioinformatics*, vol. 21, no. 23, pp. 4239–4247, 2005.

[12] 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*, to appear.

[13] J. Guo, M. W. Mak, and S. Y. Kung, "Eukaryotic protein subcellular localization based on local pairwise profile alignment SVM," in *2006 IEEE International Workshop on Machine Learning for Signal Processing (MLSP'06)*, 2006, pp. 391–396.

[14] Y. Huang and Y. D. Li, "Prediction of protein subcellular locations using fuzzy K-NN method," *Bioinformatics*, vol. 20, no. 1, pp. 21–28, 2004.