Amino Acids (2007) 32: 493–496
DOI 10.1007/s00726-006-0466-z
Printed in The Netherlands

Amino Acids

Predicting secretory protein signal sequence cleavage sites by fusing the marks of global alignments

D.-Q. Liu[1], **H. Liu**[1], **H.-B. Shen**[1], **J. Yang**[1], and **K.-C. Chou**[1,2]

[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai, China
[2] Gordon Life Science Institute, San Diego, CA, USA

Received September 12, 2006
Accepted October 16, 2006
Published online November 15, 2006;

Summary. A newly synthesized secretory protein in cells bears a special sequence, called signal peptide or sequence, which plays the role of ''address tag'' in guiding the protein to wherever it is needed. Such a unique function of signal sequences has stimulated novel strategies for drug design or reprogramming cells for gene therapy. To realize these new ideas and plans, however, it is important to develop an automated method for fast and accurately identifying the signal sequences or their cleavage sites. In this paper, a new method is developed for predicting the signal sequence of a query secretory protein by fusing the results from a series of global alignments through a voting system. The very high success rates thus obtained suggest that the novel approach is very promising, and that the new method may become a useful vehicle in identifying signal sequence, or at least serve as a complementary tool to the existing algorithms of this field.

Keywords: Signal peptide – Cleavage site – Global alignment – Needleman–Wunsch algorithm – Secretory protein

1. Introduction

A signal sequence is a short peptide that functions as an ''address tag'' in directing a nascent protein to whereever it is supposed to be. If the signal sequence in a nacent protein was changed, the protein could end up in a wrong cellular location causing various weird diseases. Therefore, knowledge of signal sequences can be used to reprogram cells in a desired way for future cell and gene therapy. However, to realize this, the first important thing is to identify the signal sequence for a nacent protein. Because the number of nacent protein sequences entering into databanks has been increasing explosively in the post-genomic era, to timely use them for basic research and drug discovery (Chou, 2004; Lubec et al., 2005), it is highly desired to develop a computational method for fast and reliably identifying signal sequences. Actually, many efforts have been made in this regards (Arrigo et al., 1991; Bendtsen et al., 2004; Chou, 2001a, b, c; Emanuelsson et al., 1999; Folz and Gordon, 1987; Ladunga et al., 1991; Liu et al., 2005; McGeoch, 1985; Nielsen et al., 1997; Schneider et al., 1993; Schneider and Wrede, 1993; von Heijne, 1986; Wang et al., 2005a). A brief introduction for most of these methods can be found in some review papers (see, e.g., Chou, 2002 and Nakai, 2000).

Given a protein sequence, the first step in predicting its signal peptide is to identify whether it is a secretory protein or non-secretory protein. For the latter, no prediction is needed at all because it contains no signal peptide. Most of the existing methods are actually effective in classifying proteins as secretory or non-secretory, as pointed out in (Bendtsen et al., 2004). The present study was initiated in an attempt to develop a powerful method for predicting the signal peptide cleavage sites of secretory proteins.

2. Materials and methods

The benchmark dataset used in this paper is from SignalP (Nielsen et al., 1997), which can be downloaded from http://www.cbs.dtu.dk/ftp/signalp. It contains 1939 secretory proteins, of which 1011 belong to eukaryotes, 105 to E. coli, 416 to human, 266 to Gram-negative, and 141 to Gram-positive (Table 1). Because the signal sequences of secretory proteins are located at the N-terminal, for simplifying the problem, the dataset constructed by Nielsen et al. only contains the sequence of the signal peptide plus the first 30 amino acids of the mature protein for each of the secretory proteins included.

As shown in Fig. 1 (Chou, 2001b), to predict the signal sequence of a secretory protein, the key is to identify its cleavage site by the signal peptidase. Once the cleavage site is determined, the corresponding signal sequence is naturally clearly defined.

The length of signal sequence is varied for different secretory proteins. A statistical analysis for the signal sequences in the aforementioned 1939 secretory proteins indicate that the shortest signal sequence contains eight

Table 1. The success rate of predicting the signal sequence cleavage sites by jackknife cross validation for the secretory proteins in each of the five different species

Species	Number of secretory proteins	Number of correct prediction[a]	Success rate (%)
E. coli	105	103	98.1
Gram-negative	266	263	98.9
Gram-positive	141	139	98.6
Human	416	413	99.3
Eukaryotic	1011	1003	99.2
Overall	1939	1921	99.1

[a] During the global sequence alignments, the parameters $d = 30$ and $e = 10$ were used for **NW**(d, e)

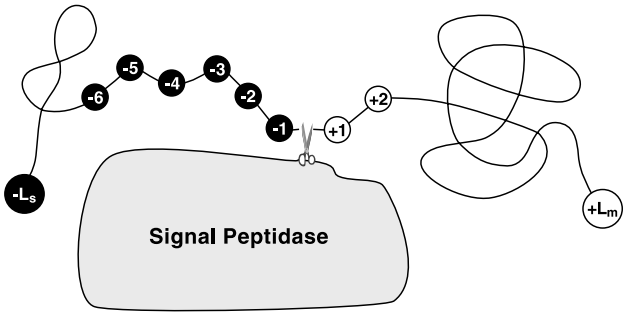

Fig. 1. A schematic drawing to show the signal sequence of a protein and how it is cleaved by the signal peptidase. An amino acid in the signal part is depicted as a black circle with a white number to indicate its sequential position, while that in the mature protein depicted as an open circle with a black number. The signal sequence contains L_S residues, and the mature protein L_m residues. The cleavage site is at the position (−1, +1), i.e. between the last residue of the signal sequence and the first residue of the mature protein. Reproduced from Chou (2001b) with permission

amino acid residues and the longest one contains 90 residues while the majority have a length within 18–25 residues. Facing such a problem with extreme variation in length and sequence, many investigators resorted to the "scaled window" approach (see, e.g. Chou (2001b)). However, a length-fixed window might not include sufficient information for an accurate prediction. Moreover, the "scaled window" approach could not avoid the imbalance situation with a small size of positive subset and very large size of negative subset in training the predictor.

The hidden Markov model (HMM) (Baldi and Brunak, 1998; Durbin et al., 1998) can be used to deal with this kind of the imbalance problem caused by the "scaled window" approach. The advantage of HMM is that it does not use windows of a fixed width, but threads an entire sequence through a trained model. However, the HMM approach was effective in discriminating signal peptides from signal anchors, but less effective for the signal cleavage site prediction (Nielsen and Krogh, 1998).

To incorporate more sequence information for identifying the signal cleavage site, we have developed a new global alignment algorithm based on Needleman–Wunsch algorithm (Needleman and Wunsch, 1970). The program can find the optimal match of any two sequences in spite of the difference of their lengths. In the global alignment approach, a query sequence is aligned with all the sequences in the training dataset one-by-one. Each of these alignments will leave a mark on the query sequence that is corresponding to the cleavage site of the known signal sequence. The mark is considered as a "conjectured cleavage site" for the query sequence. It is assumed that the more similar the two sequences are, the more credible the conjectured cleavage site will be.

According to Needleman–Wunsch algorithm (Needleman and Wunsch, 1970), finding the optimal alignment between two sequences is closely correlated with the following two basic elements: (1) substitution matrix, and (2) gap penalties.

The substitution matrices commonly used include BLOSUM, PAM, and JOHNSON. It is hard to tell which one is better than the other as they each has different suited occasion (Blake and Cohen, 2001). Among these three BLOSUM has a good reputation in sequence alignment and sequence query. BLOSUM50 means the alignment is generated using sequences sharing no more than 50% identity. BLOSUM50 is the most popular substitution matrix for pairwise alignment, providing the foundation for a number of database search techniques including BLAST and PSI-BLAST. In view of this, we choose BLOSUM50 in the current study.

For the gap penalties, the linear score and affine score are the two most popular methods for the global alignment. However, we have found that the affine score can improve the success rates for predicting the cleavage sites of signal sequences. The affine score, γ, can be formulated by the following equation:

$$\gamma(g) = -d - (g-1)e \tag{1}$$

where d is called the gap-open penalty, e called the gap-extension penalty, and g stands for the sequence length. Because using different parameters d and e for Needleman–Wunsch algorithm will result in different results, below we shall use **NW**(d, e) to denote the algorithm. Suppose

$$\mathbb{S} = \{S_1, S_2, \ldots, S_N\} \tag{2}$$

is a set of N secretory protein sequences each with known signal sequence. For a query secretory protein sequence, its global alignment with each of the sequences in Eq. (2) will generate N alignment pairs, as formulated below:

$$[[S, \mathbb{S}]] = \{[[S, S_1]], [[S, S_2]], \ldots, [[S, S_N]]\} \tag{3}$$

Suppose D_1 is the site in the query sequence S that corresponds to the known cleavage site in S_1 and regarded here as the deduced cleavage site of S from the alignment $[[S, S_1]]$, and D_2 is the deduced cleavage site of S from the alignment $[[S, S_2]]$, and so forth. Thus, we have N deduced cleavage sites according to Eq. (3); i.e.

$$\{D_1, D_2, \ldots, D_N\} \Rightarrow \{Y_1, Y_2, \ldots, Y_M\} \tag{4}$$

However, many of these deduced cleavage sites $D_j (j = 1, 2, \ldots, N)$ may be overlapped with each other, therefore the number of different deduced cleavage sites is much less than N, and can be expressed as $Y_i (i = 1, 2, \ldots, M < N)$, as shown on the right side of Eq. (4).

Now, let us define a score function given by

$$\mathcal{Q}_i = \sum_{j=1}^{N} \Delta(Y_i, D_j), \qquad (i = 1, 2, \ldots, M) \tag{5}$$

where the delta function is given by

$$\Delta(Y_i, D_j) = \begin{cases} 1, & \text{if } Y_i = D_j \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

The final decision is made by assigning Y_k of Eq. (4) as the signal sequence cleavage site for the query sequence S if

$$k = \mathbf{ArgMax}_i\{\mathcal{Q}_i\}, \qquad (i = 1, 2, \ldots, M) \tag{7}$$

where the operator $\mathbf{ArgMax}_i$ means taking the subscript with which the score function $\mathcal{Q}$ is the maximum. If there is a tie among two or more deduced cleavage sites, then the final predicted site will be randomly assigned to one of their corresponding sites, although this kind of tie case rarely happens and actually was not observed in the current study.

3. Results and discussion

As mentioned above, the alignment results by Needleman–Wunsch algorithm depend on the parameters d and e. The

former controls the number of gaps, while the latter, the gap-extension: the larger the d, the fewer the gaps are likely to occur; the larger the e, the odds in generating continuous gaps are lower. It is widely recognized that the gap-extension penalty e should be set to a value less than the gap-open penalty d, allowing long insertions and deletions to be penalized less than they would be by the linear gap cost (Durbin and Dear, 1998; Durbin et al., 1998). Most investigators used $d = 8$–12 and $e = 1$–2 by default. However, it was observed through this study that an optimal result could be obtained for predicting the cleavage sites of signal sequences by selecting $d = 30$ and $e = 10$. As an illustration, some examples are given in Appendix A to show how the difference of these parameters could affect the predicted results.

In statistical prediction, the following three cross-validation tests are often used to examine the power of a predictor: independent dataset test, sub-sampling test, and jackknife test. Of these three, the jackknife test is thought the most rigorous and objective one (see Chou and Zhang (1995) for a comprehensive review), and hence has been increasingly used by investigators (Chen et al., 2006; Chou and Shen, 2006; Feng, 2001, 2002; Gao et al., 2005; Guo et al., 2006; Liu et al., 2005; Luo et al., 2002; Niu et al., 2006; Sun and Huang, 2006; Wang et al., 2005b; Wen et al., 2006; Xiao et al., 2005, 2006a, b; Zhang et al., 2006; Zhou, 1998; Zhou and Doctor, 2003) in examining the power of various prediction methods. Here the power of the current method was therefore examined by the jackknife test as well. During the jackknifing process, each of the secretory protein sequences in the benchmark dataset is in turn taken out as a test sample and the prediction rule is trained based on the remaining sequences. The success rates thus obtained in predicting the cleavage sites of signal sequences for the 5 subsets are listed in Table 1, from which we can see that the success rates have been enhanced by 10–20% in comparison with rates by SignalP3.0 (Bendtsen et al., 2004), indicating that the fusing global alignment approach as proposed in this paper is indeed very powerful.

4. Conclusions

Predicting the signal peptide or sequence of a secretory protein is very important to both basic research and drug design, but it is also extremely difficult. It has been found through this study that the approach by fusing the results from a series of global alignments is very promising for solving such a complicated problem.

Appendix A: sequence prediction

Here, let us give an example to show why selecting the current parameters can improve the prediction quality. Consider RNI_ECOLI, which can be expressed by the following scheme:

53 RNI_ECOLI 23 RIBONUCLEASE I PRECURSOR

```
M K A F W R N A A L L A V S L L P F S S A N A L A L Q A K Q A K Q...
S S S S S S S S S S S S S S S S S S S S S S S S C M M M M M M M M M...
```

In the above scheme, S and M indicate that the corresponding amino acids belong to the part of signal peptide and that of mature chain, respectively; while C indicates that the corresponding amino acid is located at the cleavage site. When $d = 10$, $e = 2$, the signal peptide cleavage site of RNI_ECOLI was predicted at the sequence position 25, which is false. But when $d = 18$, $e = 6$, the predicted cleavage site was at 23, which is correct.

To show how the two parameters affect the predicted result, let us consider the alignment of RNI_ECOLI with AMY1_ECOLO. The sequence scheme of the latter is given by:

47 AMY1_ECOLI 17 ALPHA-AMYLASE PRECURSOR

```
M K L A A C F L T L L P G F A V A A S W T S P G F P A F S E Q G T...
S S S S S S S S S S S S S S S S S S S S S C M M M M M M M M M M M M M M M M M M...
```

It was found that, when $d = 10$, $e = 2$, the alignment between RNI_ECOLI and AMY1_ECOLI is as follows:

```
M K A F W R N A A L L A V S L L P F S S A N A L A L Q A K Q A K Q...
M K L – – – – A A C F – L T L L P – – – G F A V A A S W T S P G –...
```

As we can see from above, the cleavage site of AMY1_ECOLI points to the 25th sequence position of RNI_ECOLI, and hence leading to a wrong prediction.

However, when $d = 18$, $e = 6$, the corresponding alignment is given by:

```
M K A F W R N A A L L A V S L L P F S S A N A L A L Q A K Q A K Q...
M K – – – – – L A A C F L T L L P – G F A V A A S W T S P G F P A...
```

It can be seen from the above alignment that the cleavage site of AMY1_ECOLI points to the 23rd sequence position of RNI_ECOLI, leading to a correct prediction.

The reason why the alignments between two same sequences with different parameters can lead to completely different results is due to the gaps inserted. Using different parameter values may generate different distribution of gaps. Therefore, selecting the optimal parameters can improve the prediction quality.

References

Arrigo P, Giuliano F, Scalia F, Rapallo A, Damiani G (1991) Identification of a new motif on nucleic acid sequence data using Kohonen's self-organizing map. Comput Appl Biosci 7: 353–357

Baldi P, Brunak S (1998) Bioinformatics: the machine learning approach. MIT Press, Cambridge/Mass

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340: 783–795

Blake JD, Cohen FE (2001) Pairwise sequence alignment below the twilight zone. J Mol Biol 307: 721–735

Chen C, Zhou X, Tian Y, Zou X, Cai P (2006) Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem 357: 116–121

Chou KC (2001a) Prediction of protein signal sequences and their cleavage sites. Proteins Struct Function Genet 42: 136–139

Chou KC (2001b) Prediction of signal peptides using scaled window. Peptides 22: 1973–1979

Chou KC (2001c) Using subsite coupling to predict signal peptides. Protein Eng 14: 75–79

Chou KC (2002) Review: Prediction of protein signal sequences. Curr Protein Pep Sci 3: 615–622

Chou KC (2004) Review: Structural bioinformatics and its impact to biomedical science. Curr Med Chem 11: 2105–2134

Chou KC, Shen HB (2006) Predicting protein subcellular location by fusing multiple classifiers. J Cell Biochem 99: 517–527

Chou KC, Zhang CT (1995) Review: Prediction of protein structural classes. Crit Rev Biochem Mol Biol 30: 275–349

Durbin R, Dear S (1998) Base qualities help sequencing software. Genome Res 8: 161–162

Durbin RM, Eddy SR, Krogh A, Mitchison G (1998) Biological sequence analysis. Cambridge University Press, Cambridge

Emanuelsson O, Nielsen H, von Heijne G (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci 8: 978–984

Feng ZP (2001) Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. Biopolymers 58: 491–499

Feng ZP (2002) An overview on predicting the subcellular location of a protein. In Silico Biol 2: 291–303

Folz RJ, Gordon JI (1987) Computer-assisted predictions of signal peptidase processing sites. Biochem Biophys Res Commun 146: 870–877

Gao QB, Wang ZZ, Yan C, Du YH (2005) Prediction of protein subcellular location using a combined feature of sequence. FEBS Lett 579: 3444–3448

Guo YZ, Li M, Lu M, Wen Z, Wang K, Li G, Wu J (2006) Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform. Amino Acids 30: 397–402

Ladunga I, Czako F, Csabai I, Geszti T (1991) Improving signal peptide prediction accuracy by simulated neural network. Comput Appl Biosci 7: 485–487

Liu H, Yang J, Ling JG, Chou KC (2005) Prediction of protein signal sequences and their cleavage sites by statistical rulers. Biochem Biophys Res Commun 338: 1005–1011

Lubec G, Afjehi-Sadat L, Yang JW, John JP (2005) Searching for hypothetical proteins: theory and practice based upon original data and literature. Prog Neurobiol 77: 90–127

Luo RY, Feng ZP, Liu JK (2002) Prediction of protein structural class by amino acid and polypeptide composition. Eur J Biochem 269: 4219–4225

McGeoch DJ (1985) On the predictive recognition of signal peptide sequences. Virus Res 3: 271–286

Nakai K (2000) Protein sorting signals and prediction of subcellular localization. Adv Protein Chem 54: 277–344

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48: 443–453

Nielsen H, Engelbrecht J, Brunak S, von Heijne G (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 10: 1–6

Nielsen H, Krogh A (1998) Prediction of signal peptides and signal anchors by a hidden Markov model. Intell Syst Mol Biol 6: 122–130

Niu B, Cai YD, Lu WC, Zheng GY, Chou KC (2006) Predicting protein structural class with AdaBoost learner. Protein Peptide Lett 13: 489–492

Schneider G, Rohlk S, Wrede P (1993) Analysis of cleavage-site patterns in protein precusor sequences with a perceptron-type neural network. Biochem Biophys Res Commun 194: 951–959

Schneider G, Wrede P (1993) Signal analysis of protein targeting sequences. Protein Seq Data Anal 5: 227–236

Sun XD, Huang RB (2006) Prediction of protein structural classes using support vector machines. Amino Acids 30: 469–475

von Heijne G (1986) A new method for predicting signal sequence cleavage sites. Nucleic Acids Res 14: 4683–4690

Wang M, Yang J, Chou KC (2005a) Using string kernel to predict signal peptide cleavage site based on subsite coupling model. Amino Acids 28: 395–402 (Erratum, ibid. 2005, 29: 301)

Wang M, Yang J, Xu ZJ, Chou KC (2005b) SLLE for predicting membrane protein types. J Theor Biol 232: 7–15

Wen Z, Li M, Li Y, Guo Y, Wang K (2006) Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. Amino Acids (in press)

Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC (2005) Using complexity measure factor to predict protein subcellular location. Amino Acids 28: 57–61

Xiao X, Shao SH, Ding YS, Huang ZD, Chou KC (2006a) Using cellular automata images and pseudo amino acid composition to predict protein sub-cellular location. Amino Acids 30: 49–54

Xiao X, Shao SH, Huang ZD, Chou KC (2006b) Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J Comput Chem 27: 478–482

Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (2006) Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion. Amino Acids 30: 461–468

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17: 729–738

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins Struct Funct Genet 50: 44–48

Authors' address: Dan-Qing Liu, Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai 200240, China, Fax: +86-21-62-9334-2820, E-mail: danqingliu@gmail.com