Prediction of Protein Signal Sequences

Article *in* Current Protein and Peptide Science · January 2003

DOI: 10.2174/1389203023380468 · Source: PubMed

1 author:

Kuo-Chen Chou
Gordon Life Science Institute
712 PUBLICATIONS **49,165** CITATIONS

SEE PROFILE

Prediction of Protein Signal Sequences

Kuo-Chen Chou*

Computer-Aided Drug Discovery, Pharmacia & Upjohn, Kalamazoo, MI 49007-4940, U S A

Abstract: Newly synthesized proteins have an intrinsic signal sequence, functioning as "address tags" or "zip codes", that is essential for guiding them wherever they are needed. Owing to such a unique function, protein signals have become a crucial tool in finding new drugs or reprogramming cells for gene therapy. However, to effectively use protein signals as a desirable vehicle in the field of proteomics, the first important thing is to find a fast and powerful method to identify the "address tag" or "zip code" entity. Although all signal sequences contain a hydrophobic core region, they show great variation in both overall length and amino acid sequence. It is this variation that makes it possible to deliver thousands of proteins to many different cellular locations by varieties of modes. It is also this variation that makes it very difficult to formulate a general algorithm to predict signal sequences. Nevertheless, various prediction models and algorithms have been developed during the past 17 years. This Review summarizes the development in this area, from the pioneering methods to neural network approaches, and to the sub-site coupling approaches. Meanwhile, the future challenges in this area, as well as some promising avenues for further improving the prediction quality, have been briefly addressed as well.

I. INTRODUCTION

A cell contains approximately 10^9 protein molecules. An adult human being is made up of approximately 10^{14} cells, or approximately 10^{23} protein molecules. It is interesting to see that the latter has the same order of magnitude as the Avogadro constant, isn't it? As shown in Fig. (**1**), a cell consists of many different compartments, or organelles, each surrounded by a membrane. The organelles are specialized to carry out different tasks. For example, the mitochondria function as the "power plants", producing energy needed by the cell. The cell nucleus contains the genetic material (DNA), governing all functions of the cell. And the endoplasmic reticulum is, together with the ribosomes, responsible for synthesizing proteins.

A large number of proteins with various essential functions are constantly being made within cells. These nascent proteins have to be transported either out of the cell, or to the different compartments - the organelles - within the cell. The number of amino acids - the building blocks making up all proteins - may in a single protein range from about 50 to several thousands, forming long, folded chains. Every chemical reaction essential to life depends on the services of proteins in one way or another. For example, proteins can serve as: the "beams and rafter" that frame the cell; the enzymes that catalyze thousands of specific chemical reactions; the "glue" that binds the body together; the "circuits" that power movement and thought; the hormones that course through our veins; the "guided missiles" that target infections; and much more. The proteins within a cell are constantly degraded and resynthesized.

*Address correspondence to this author at the Computer-Aided Drug Discovery, Pharmacia & Upjohn, Kalamazoo, MI 49007-4940, U S A; Tel: 269-833-8867; E-mail: kuo-chen.chou@am.pnu.com

How are newly made proteins transported across the membrane surrounding the organelles? And how are they directed to their correct location? To address these questions, Blobel and Sabatini [1] formulated a first version of "signal hypothesis". They postulated that proteins secreted out of the cell contain an intrinsic signal that governs them to and across membranes. Based on a series of elegant biochemical experiments, Blobel and his co-workers described in 1975 the various steps in these processes (see, e.g., [2-4]). It has become clear now that whether a protein will pass through a membrane into a particular organelle, become integrated into the membrane, or be exported out of the cell is determined by a specific amino acid sequence, the so-called topogenic signal. The signal consists of a peptide, i.e. a sequence of amino acids in a particular order that form an integral part of the protein. The signal peptides can be compared to "address tags" for a traveler's luggage to arrives at the expected destination, or compared to "zip codes" for a letter to reach its correct addressee.

The discovery of signal peptides has had an immense impact on modern cell biological research. Knowledge of signal peptides has helped explain the molecular mechanisms behind several genetic diseases. When a cell divides, large amounts of proteins are being made and new organelles are formed. If a sorting signal in a protein is changed, the protein could end up in a wrong cellular location and cause varieties of diseases. For example, in some forms of familial hypercholesterolemia, a very high level of cholesterol in the blood is due to deficient transport signals. Also, hereditary diseases, such as cystic fibrosis, are caused by the fact that proteins do not reach their proper destination. Knowledge of signal peptides will increase our understanding of processes leading to disease and hence can be used to develop new therapeutic strategies.

1389-2037/02 $35.00+.00

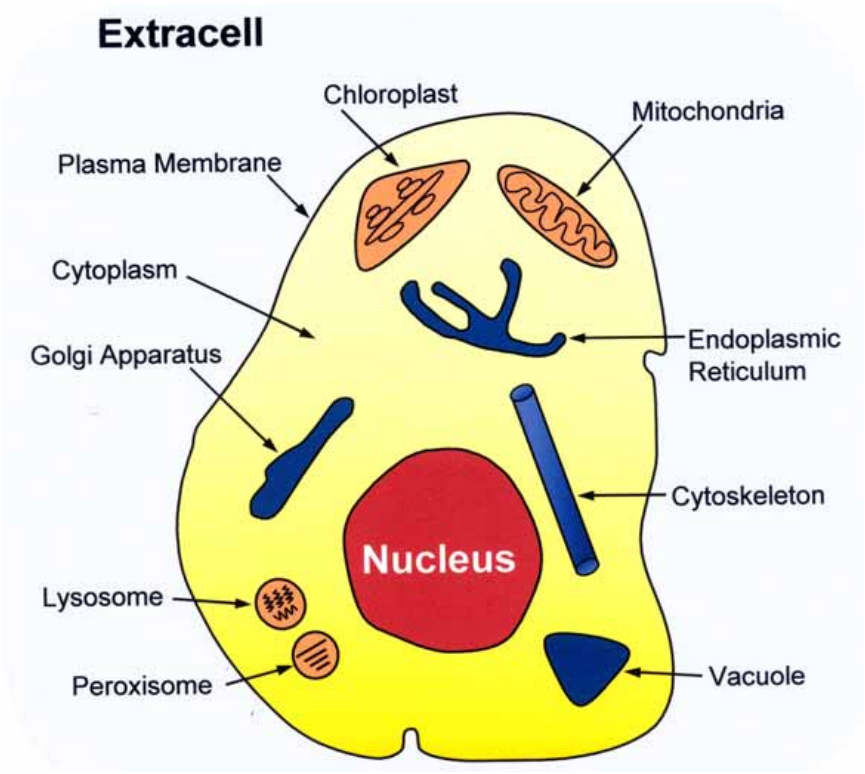

Fig. (1). A schematic drawing to show a cell consisting of many different compartments, or organelles, each surrounded by a membrane (reproduced from Chou and Elrod [44] with permission). Note that the vacuole and chloroplast organelles exist only in a plant cell. How are newly made proteins transported across the membrane surrounding the organelles? And how are they directed to their correct location? It is clear that all nascent proteins have an intrinsic signal sequence, functioning as "address tags", that is essential in controlling their pathway and guiding them wherever they are needed.

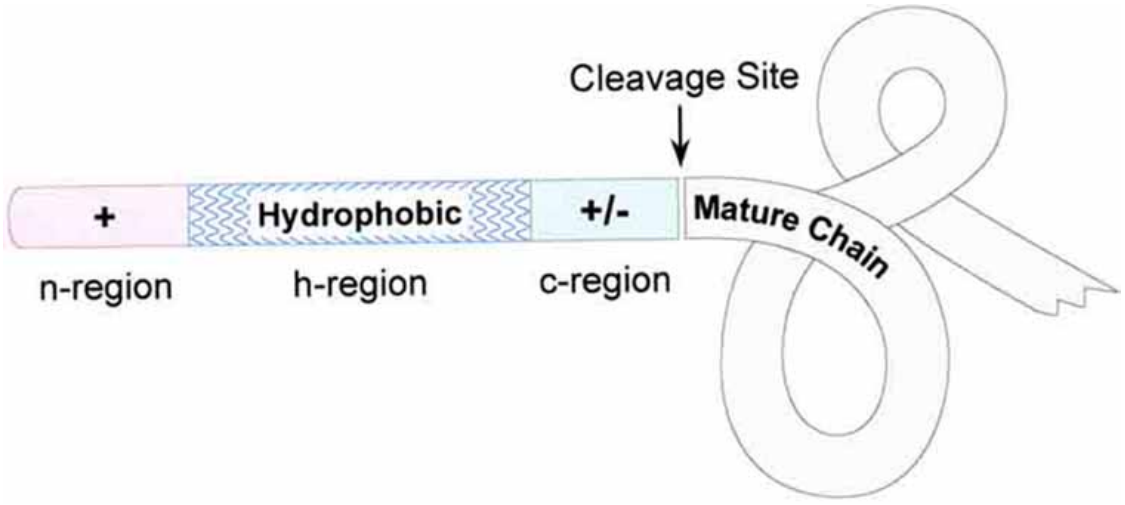

Fig. (2). A schematic drawing to show the three sub-regions of the tripartite structure of a signal peptide: (**i**) The n-region usually contains relatively hydrophilic (basic) residues with positive charges, such as Arg and Lys. (**ii**) The central h-region is dominated by hydrophobic residues and hence called "the hydrophobic core"; it comprises about 7 to 15 amino acid residues, and is the most essential part required for targeting and membrane insertion. (**iii**) The c-region contains more neutral but polar residues, often with Ala at the last and the third last sequence positions of the sub-region.

Today some drugs have already been produced in the form of proteins, e.g. growth hormone, insulin, and hemoglobin. Bacteria are usually used for the production of protein drugs. However, in order to be functionally proper, it is necessary to synthesize certain human proteins in more complex cells, such as yeast cells. The contemporary gene technology allows us to generate the genes of the desired proteins with sequences coding for transport signals. The cells with the modified genes can be efficiently used as "protein factories". Accordingly, knowledge of protein signals can then be used to reprogram cells in a specific way for future cell and gene therapy. Actually, protein signals have already become a crucial tool for researchers to construct new drugs that are targeted to a particular organelle to correct a specific defect. For example, by adding a specific tag to the desired proteins, one can, for instance, tag them for excretion, making them much easier to harvest [5].

Actually, the identification of signal peptides has become a prerequisite in order to use such a technology effectively. However, since the number of protein sequences entering into data banks has been rapidly increasing, it is time-consuming and costly to identify the signal peptides solely based on experiments. For example, the number of protein sequence entries in SWISS-PROT [6] in 1986 was 3,939, and that in 1992 was 28,154, but that in 1999 was already 80,000. In view of this, we are facing a critical challenge: How to develop an automated algorithm to identify signal sequences of newly synthesized proteins. Particularly, many more new protein sequences will be derived soon owing to the recent success of human genome project, which has provided enormous amount of genomic information in the form of 3 billion base pairs, assembled into ten of thousands of genes. Accordingly, the challenge will become even more urgent and critical. This is the need of times. Actually, many efforts have been made trying to develop some computational methods for quickly predicting the signal peptides. Below, let us give a brief review in this area.

II. SIGNAL PEPTIDES

Signal sequences are usually N-terminal extensions, but they can also be located within a protein or at its C-terminal end (e.g., for "tail-anchored" membrane proteins [7]). All secreted proteins, as well as many transmembrane proteins, are synthesized with N-terminal signal peptides. They control the entry of virtually all secretory proteins to the pathway, both in eukaryotes and prokaryotes [8-10]. After their translocation, signal peptides of preproteins are cleaved off by membrane-bound enzymes, called signal peptidases [11]. As shown in Fig. (**2**), the N-terminal signal peptides (also called leader sequences) generally consist of the following three structurally, and, possibly, functionally distinct regions: (**i**) an N-terminal positively charged n-region, (**ii**) a central hydrophobic h-region, and (**iii**) a neutral but polar c-region [12]. The n-region usually contains relatively hydrophilic (basic) residues, such as arginine and lysine. The central h-region is the hydrophobic core comprising about 7 to 15 amino acid residues, and is the most essential part required for targeting and membrane insertion. The c-region contains more polar residues than the h-region, often with helix-breaking proline and glycine residues as well as small uncharged residues around the signal peptide cleavage site. Such a tripartite structural pattern might provide useful information for their identification. Unfortunately, signal peptides have little sequence similarity and their lengths are in extreme variation. As shown in Fig. (**3**), of the 1939 signal peptides the shortest one contains 8 amino acid residues, the longest one contains 90 residues, while the majority have a length within 18-25 residues. Comparative analysis of a large number of signal sequences reveals that the n-region contributes most to the variation in overall length [13]. The extreme variation in length and lack of a consensus sequence will certainly make the problem more difficult, especially in formulating a general algorithm for their prediction.

III. PREDICTION METHODS

Since very little work has been done in predicting the signal sequence within a protein or at the C-terminal end, the present review will be focused on the N-terminal signal peptide prediction. Because each method uses different training data and a different evaluation criteria, it is difficult to compare the performances of various methods from the literature. Also for the same reason, it might be misleading to simply list the success prediction rates claimed by the authors in introducing their methods. It happened quite often in statistical prediction that a method good for one dataset would not always be so while used on another dataset, particularly when the results were obtained without following an objective cross validation procedure. What is an objective cross validation procedure? In the literature of molecular biology, the single independent dataset test, sub-sampling test, and jackknife test are the three methods often used for cross-validation. Of the three cross validation methods, the jackknife test is the most objective one (see Mardia *et al*. [14] for the mathematical principle and Chou & Zhang [15] as well as some recent articles [16, 17] for a comprehensive discussion about this). Unfortunately, owing to various reasons, the single independent dataset test and sub-sampling test were used for cross validation in most of the papers in this area. Besides, any prediction method developed based on a training dataset usually has some limit or caveat. Over extension in applying a prediction method might often lead to an unreasonable result, in spite of how good the method is. This is just like the situation implied by the proverb that "even truth could become falsehood if overly extended" or "just one step across the frame of truth might lead to a fallacy", as often quoted in philosophical works. Therefore, in this review we would prefer to focus on the concepts and characteristics of various methods rather than the success rates claimed by their authors. We believe that each method has its advantage and disadvantage. A complement of one with the others is a wise way to apply the existing methods for predicting the signal peptides.

(**1**) Pioneer Methods

The first method for predicting the existence of a particular sorting signal in a protein sequence was developed by McGeoch [18], who built a discriminant function based on the net charge and length of the n-region, as well as the length and hydrophobicity of the h-region. The first method

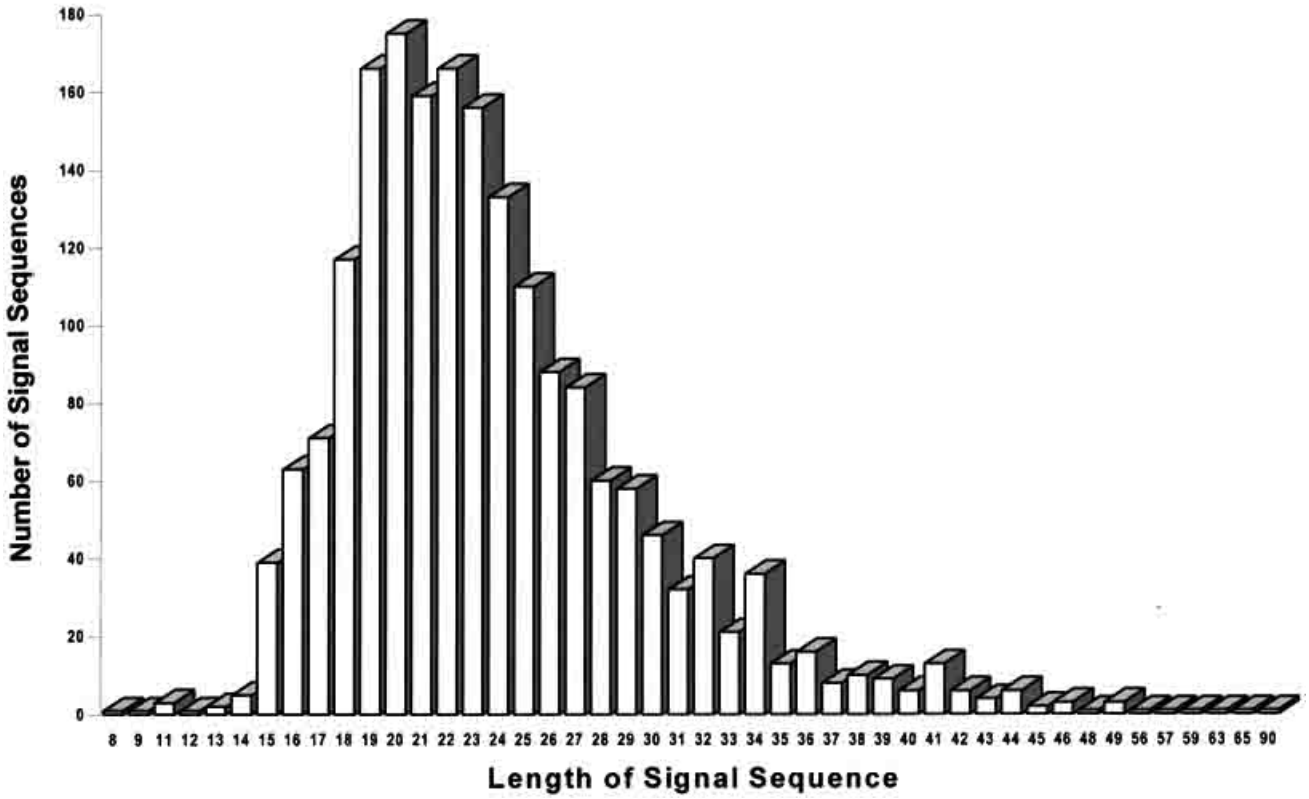

Fig. (3). A histogram to show the extreme variation of the 1939 signal peptides [26] in length. As shown from the figure, of the 1939 signal peptides the shortest one contains 8 amino acid residues, the longest one contains 90 residues, while the majority have a length within 18-25 residues. It was suggested by some comparative analysis that the n-region might contribute most to the variation in overall length [13].

for identifying signal peptide cleavage sites was developed by von Heijne [19], who introduced the weight-matrix approach to deal with the problem. The pioneer work by McGeoch and von Heijne was an important contribution that has greatly stimulated the development of this area.

The first approach by combining two different algorithms for predicting the signal peptide cleavage sites was formulated by Folz and Gordon [20]. One algorithm is to generate a probability score for each subsite based on a series of empirical rules, and the second algorithm is operated by using the statistical weight matrix approach. Since prediction of signal peptides is affected by many complicated factors, it is a savvy idea to introduce the approach of combining different algorithms. Actually, the idea and concept of the combination approach have also been used in various neural network methods developed later on.

(2) Neural Network Approaches

In 1991 Ladunga *et al*. [21] introduced a neural network approach, in which a multiplayer architecture that self-adjusts to the data was trained for predicting the N-terminal signal peptides. According to their report, if the network was applied to sequences first selected with the weight matrices, the success rate increased significantly, indicating that the combination approach was indeed a promising step for the improvement of prediction. Meanwhile, it was reported by Arrigo *et al*. [22] that the signal peptides could also be predicted by using Kohonen's self-organizing map that was originally used for identifying a new motif on nucleic acid sequence data. However, it is not clear how well this method performs on more general data sets.

A couple of years later, a series of progress were made by Schneider *et al*. [23-25] using a neural network trained by a genetic algorithm and representing each amino acid by 4-7 physicochemical properties.

In 1997 Nielsen *et al*. [26] developed a powerful algorithm that combines two networks, one to recognize the cleavage site and another to distinguish between signal peptides and non-signal peptides. The method by Nielsen *et al*. has been widely used not only because it is available on web site, but also because it contains very large sets of non-homologous prokaryotic and eukaryotic signal peptides. A similar neural network-based method was also trained to predict the chloroplast transit peptides and their cleavage sites [27]. The method has been constructed in much the same way as the one by Nielsen *et al*. [26] except the following two novel aspects: (i) the yes/no chloroplast transit peptide prediction is based on a neural network trained on the S-score outputs from the basic neural network, and (ii) the cleavage site prediction is not done using a neural network but by a simple weight matrix.

Some membrane-bound proteins have sequences that play the same role as signal peptides but are not cleaved by signal peptidase. Instead, they are anchored to the membrane by the hydrophobic region. The uncleaved signal sequence is known as a signal anchor. It has proved to be very difficult for the neural network to discriminate a signal peptide with the signal anchor. To solve this problem, a different type of artificial neural network method was introduced [28, 29] that was based on the hidden Markov model [30, 31]. The term "hidden" refers to the invisibility of the underlying random walk between different states. The advantage of the hidden Markov model method is that it does not use windows of a fixed width, but threads an entire sequence through a trained model. A hidden Markov model is a chain of "states", each with a characteristic amino acid distribution. By using the hidden Markov model method, some improvement was observed, particularly in discriminating between signal peptides and signal anchors. However, it was less accurate when the method was used for cleavage site prediction.

A brief introduction of the above methods can also be found in four review papers by Claros *et al*. [12], Nielsen *et al*. [29], Nakai [32], and Ladunga [33], respectively. As pointed out by King [34], the advantages of neural network prediction methods are: (i) "readily available", and (ii) "often successful in practice"; the disadvantages are: (i) "very poor explanatory power", (ii) "little use of chemical or physical theory", and (iii) "statistically rather poorly characterized". To most protein chemists, the prediction principle and process by neural networks are just like a black box. Besides, probably because of the difficulty caused by the convergence rate and time-consuming problem, most neural network methods were tested by the sub-sampling analysis. Very few reports are seen to test neural work methods by means of jackknifing for cross validation. Furthermore, although the computational costs for training the networks was considerably higher, the prediction accuracy thus obtained was not always higher (and sometimes even lower) than the analytical methods. Below, we would like to introduce a different approach developed very recently.

(**3**) Sibsite Coupling Approach

The subsite coupling approach [35] was developed based on the sequence-encoded algorithm [36] and the scaled window approach [37]. As we mentioned at the beginning, the signal peptides generally have the so-called tripartite structure pattern as illustrated in Fig. (**2**). How can it help us reduce the scope in searching for possible signal peptides? To answer such a question, let us perform a statistical analysis based on a highly simplified model. Suppose a signal peptide consisting of 22 amino acid residues. The number of possible different combinations for a 22 residue sequence would be $20^{22} \sim 4.19 \times 10^{28}$. If the signal peptide is confined to such a tripartite structure pattern that its n-region contains 6 residues, h-region 11 residues, and c-region 5 residues. According to the respective features of the three sub-regions (see Fig. (**2**) and the corresponding text in Section II), the model is further simplified as follows. Suppose the n-region contains only Arg, Lys, and His (i.e. 3 basic residues with positive charge), the h-region only Ala, Cys, Gly, Ile, Leu, Met, Phe, Pro, Trp, Tyr, and Val (i.e. 11 hydrophobic residues), and c-region only Asp, Gln, Ser, and Thr (i.e. 4 polar and neutral residues) except the 1st and 3rd sequence positions from the end of c-region that are usually occupied by small and neutral residues such as Ala [35, 38]. Imposed with these constraints, the number of possible different combinations for the 22 residue sequence would be $3^6 \times 11^{11} \times 4^{5-3} \sim 3.33 \times 10^{15}$. Accordingly, the constraints of tripartite pattern do significantly reduce the search scope by a magnitude of 10^{13}. However, the number obtained after such a reduction is still an astronomic figure. Besides, the above constraints are based on an oversimplified case. Actually, neither the length of each of the 3 sub-regions nor the overall length of signal peptides is a constant but varies over a wide range (Fig. **3**). And in addition to the above special amino acids, the n-, h-, and c-regions may also contain some other residues. Therefore, the actual number obtained by taking into account the tripartite structure pattern would be much greater than 3.33×10^{15}.

To deal with such a grim situation, the subsite coupling approach [35] was emerging. The rationale of the subsite coupling approach is that, although the signal peptides are of extreme variation in both the sequence order and length, some intrinsic couplings might exist among their subsites. For instance, it has been observed from a statistical analysis for the 1939 secretory protein sequences [26] that the amino acid residues at the subsites -3, -1, and +1 are mostly occupied by Ala (Fig. **4**), while the occurrence frequencies of the other 19 amino acids at these subsites are relatively much lower. This suggests that a highly special match between the signal peptidase and the secretory protein at the subsites -3, -1 and +1 is required during the cleavage process, as illustrated by Fig. (**5**). Based on such a finding, some special terms that reflect the couplings among these subsites have been incorporated into the prediction algorithm. Some very encouraging results were observed by both self-consistency test and jackknife test [35].

IV. CONCLUSION AND PERSPECTIVE

Protein signal sequences play a central role in the targeting and translocation of nearly all secreted proteins and many integral membrane proteins in both prokaryotes and eukaryotes. The knowledge of signal sequences has become a crucial tool for pharmaceutical scientists who genetically modify bacteria, plants, and animals to produce effective drugs. To effectively use such a tool, the first important thing is to find a fast and effective method to identify the "zip-code" entity. Many different methods were proposed, and some encouraging results obtained, but they are far away from satisfactory yet.

To improve the prediction methods, further efforts can be made in the following several aspects. (**i**) Training dataset. The existing methods are all operated based on the statistical rules derived from a training dataset. Accordingly, it is vitally important to establish a training dataset that has extensive representativity. (**ii**) Subsite coupling. It has been proved through both the HIV protease cleavage prediction [39, 40] and protein tight turn prediction [41] that the subsite coupling effects are important for improving the prediction quality based on sequence-coded algorithms. As shown in

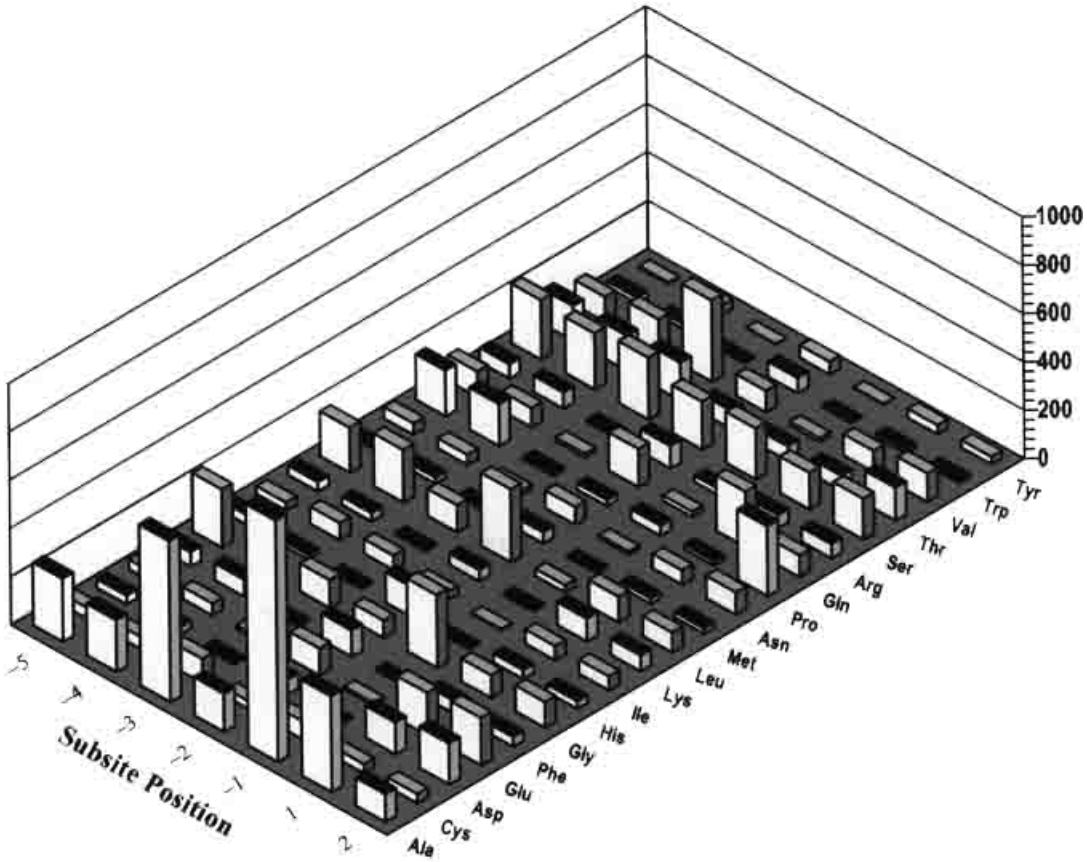

Fig. (4). A 3D histogram to show the frequencies of the 20 native amino acids that occur at the subsites proximal to the cleavage site, where the negative numbers represent the subsite positions for the signal sequence while positive numbers the subsite positions for the mature protein sequence, and the cleavage site is at (-1, +1). The frequencies were derived based on the 1939 secretory protein sequences [26]. As shown in the figure, the occurrence frequencies of Ala at the subsites -3, -1 and +1 are overwhelming in comparison with the other 19 amino acids, suggesting that a highly special match between the signal peptidase and the secretory protein at the subsites -3, -1 and +1 is required during the cleavage process, as illustrated by Fig.5.

Chou [35], an encouraging result was observed after taking into account the [-3, -1, +1] subsite coupling effect derived from the 1939 secretory protein sequences (Fig. **5**). It is anticipated that with more and more training data accumulated, some other subsite couplings might be gradually revealed. Incorporation of these subsite coupling effects might further improve the prediction quality. (**iii**) Entire sequence approach. As mentioned above, to overcome the limitation imposed by the width-fixed windows as formulated in most of the existing methods, the hidden Markov model method was introduced that threads an entire sequence through a trained model. The entire sequence effect can also be taken into account via the quasi-sequence-order approach [42] and the pseudo-amino-acid-composition approach [43]. The essence of these two approaches is to introduce a set of discrete numbers, the so-called sequence-order-coupling numbers, generated by going through an entire protein sequence according to different correlation ranks. Remarkable improvements have been observed in predicting both protein subcellular locations and membrane protein attributes [42, 43]. Since these approaches are particularly suitable to characterize the tripartite structures of signal peptides with different lengths, it is anticipated that a similar improvement might also be obtained accordingly, at least in predicting whether a query sequence contains a sorting signal, as well as in discrimination between signal peptides and signal anchors. (**iv**) Combined approach. Owing to the extreme variation of signal peptides in both overall length and amino acid sequence, it might be effective to resort to the strategy of combining various different approaches. Developed by focusing at different aspects of the problem, each of the existing approaches might have remarkable advantages for some cases and disadvantages for the other. A complimentary combination of these methods might be a promising avenue. For instance, the entire sequence approach is more powerful in identifying the existence of a signal peptide for a given protein sequence but less powerful in predicting the cleavage site of a signal peptide, but the sub-site coupling approach is just opposite. Thus, the overall prediction quality might be improved by combining the subsite coupling operation with the entire sequence operation. In the combined approach a query protein sequence is screened by the entire sequence operation first: if the result is negative, no further operation will be needed; if the result is positive, the protein sequence will undergo the subsite coupling operation for identifying the cleavage site. Since neural works are particularly capable in performing multi-layer operation, inclusion of the neural network approaches might further strengthen the combined approach.

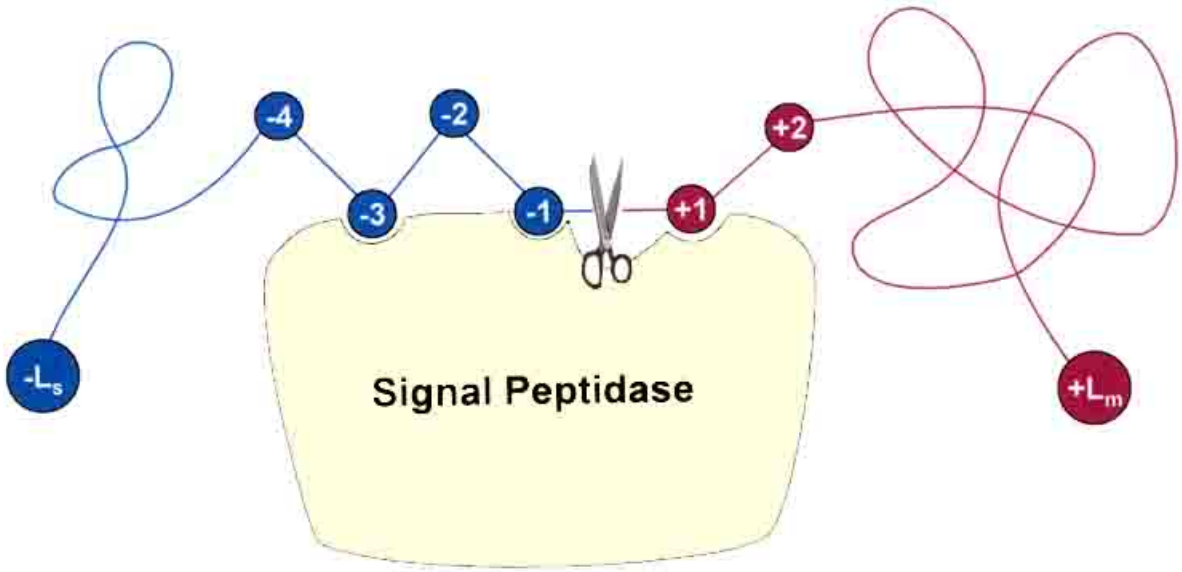

Fig. (5). A schematic drawing to show the [-3, -1, +1] subsite coupling mechanism. An amino acid in the signal sequence is colored blue, while that in the mature protein is magenta. The cleavage site is between the subsites -1 and +1. During the cleaving process, a highly special match is required between the residues at subsites -3, -1, and +1 of the secretory protein and their counterparts in the signal peptidase. Based on such a model, the [-3, -1, +1] subsite coupling algorithm was proposed [35].

ACKNOWLEDGEMENTS

The authors would like to thank Raymond B. Moeller, Cynthia A. Ludlow, Wendy Vanderheide, and Katie Crawford of Pharmacia's Graphic Service Group for their help of drawing the figures in this paper.

REFERENCES

[1] Blobel, G. & Sabatini, D. D. (**1971**) Ribosome-membrane interactiion in eukaryotic cells in *Biomembranes* (Manson, L. A., ed) pp. 193-195, Plenum Publishing Corporation, New York.

[2] Blobel, G. & Dobberstein, B. (**1975**) *Journal of Cell Biology, 67*, 852-862.

[3] Blobel, G. & Dobberstein, B. (**1975**) *Journal of Cell Biology, 67*, 835-851.

[4] Blobel, G. (**1976**) *Biochemical and Biophysical Research Communications, 68*, 1-7.

[5] Hagmann, M. (**1999**) *Science, 286*, 666-666.

[6] Bairoch, A. & Apweiler, R. (**1997**) *Nucleic Acids Research, 25*, 31-36.

[7] Kutay, U., Ahnert-Hilger, G., Hartmann, E., Wiedenmann, B. & Rapoport, T. A. (**1995**) *EMBO J., 14*, 217-223.

[8] Gierasch, L. M. (**1989**) *Biochemistry, 28*, 923-930.

[9] Rapoport, T. A. (**1992**) *Science, 258*, 931-936.

[10] Zheng, N. & Gierasch, L. M. (**1996**) *Cell, 86*, 849-852.

[11] Pugsley, A. (**1993**) *Microbiol. Rev., 57*, 50-108.

[12] Claros, M. G., Brunak, S. & von Heijne, G. (**1997**) *Curr. Opin. Struct. Biol., 7*, 394-398.

[13] Martoglio, B. & Dobberstein, B. (**1998**) *Trends in Cell Biology, 8*, 410-415.

[14] Mardia, K. V., Kent, J. T. & Bibby, J. M. (**1979**) In *Multivariate Analysis* pp. 322 and 381, Academic Press, London.

[15] Chou, K. C. & Zhang, C. T. (**1995**) *Critical Reviews in Biochemistry and Molecular Biology, 30*, 275-349.

[16] Cai, Y. D. (**2001**) *PROTEINS: Structure, Function, and Genetics, 43*, 336-338.

[17] Zhou, G. P. & Assa-Munt, N. (**2001**) *PROTEINS: Structure, Function, and Genetics, 44*, 57-59.

[18] McGeoch, D. J. (**1985**) *Virus Res., 3*, 271-286.

[19] von Heijne, G. (**1986**) *Nucleic Research, 14*, 4683-4690.

[20] Folz, R. J. & Gordon, J. I. (**1987**) *Biochem. Biophys. Res. Comm., 146*, 870-877.

[21] Ladunga, I., Czako, F., Csabai, I. & Geszti, T. (**1991**) *Comput. Appl. Biosci., 7*, 485-487.

[22] Arrigo, P., Giuliano, F., Scalia, F., Rapallo, A. & Damiani, G. (**1991**) *Comput. Appl. Biosci., 7*, 353-357.

[23] Schneider, G., Rohlk, S. & Wrede, P. (**1993**) *Biochem. Biophys. Res. Comm., 194*, 951-959.

[24] Schneider, G. & Wrede, P. (**1993**) *J. Mol. Evol., 36*, 586-595.

[25] Schneider, G. & Wrede, P. (**1993**) *Protein Seq. Data Anal., 5*, 227-236.

[26] Nielsen, H., Engelbrecht, J., Brunak, S. & von Heijne, G. (**1997**) *Protein Engineering, 10*, 1-6.

[27] Emanuelsson, O., Nielsen, H. & von Heijne, G. (**1999**) *Protein Science, 8*, 978-984.

[28] Nielsen, H. & Krogh, A. (**1998**) *Intell. Syst. Mol. Biol., 6*, 122-130.

[29] Nielsen, H., Brunak, S. & von Heijne, G. (**1999**) *Protein Engineering, 12*, 3-9.

[30] Durbin, R. M., Eddy, S. R., Krogh, A. & Mitchison, G. (**1998**) *Biological Sequence Analysis*, Cambridge University Press, Cambridge.

[31] Baldi, P. & Brunak, S. (**1998**) *Bioinformatics: The Machine Learning Approach*, MIT Press, Cambridge.

[32] Nakai, K. (**2000**) *Advances in Protein Chemistry, 54*, 277-344.

[33] Ladunga, I. (**2000**) *Current Opinion in Biotecnilogy, 11*, 13-18.

[34] King, R. D. (**1996**) Prediction of secondary structure in *Protein Structure Prediction: A Practical Approach* (Sternberg, M. J. E., ed) pp. 79-97, IRL Press, Oxford.

[35] Chou, K. C. (**2001**) *Protein Engineering, 14*, 75-79.

[36] Chou, K. C. (**2001**) *PROTEINS: Structure, Function, and Genetics. 42*, 136-139.

[37] Chou, K. C. (**2001**) *Peptides, 22*, 1973-1979.

[38] von Heijne, G. (**1984**) *Journal of Molecular Biology, 173*, 243-251.

[39] Chou, K. C. (**1993**) *Journal of Biological Chemistry, 268*, 16938-16948.

[40] Chou, K. C. (**1996**) *Analytical Biochemistry, 233*, 1-14.

[41] Chou, K. C. (**2000**) *Analytical Biochemistry, 286*, 1-16.

[42] Chou, K. C. (**2000**) *Biochemical & Biophysical Research Communications, 278*, 477-483.

[43] Chou, K. C. (**2001**) *PROTEINS: Structure, Function, and Genetics. 43*, 246-255 (Erratum: *Proteins: Struct. Funct. Genet.*, **2001**, Vol. *44*, 60).

[44] Chou, K. C. & Elrod, D. W. (**1999**) *Protein Engineering, 12*, 107-118.