

Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition

Yu-Dong Cai^{a,*} and Kuo-Chen Chou^b

^a Shanghai Research Centre of Biotechnology, Chinese Academy of Sciences, Shanghai 200233, China

^b Upjohn Laboratories, Pfizer, Kalamazoo, MI 49007-4940, USA

Received 21 April 2003

Abstract

In this paper, based on the approach by combining the “functional domain composition” [K.C. Chou, Y.D. Cai, *J. Biol. Chem.* 277 (2002) 45765] and the pseudo-amino acid composition [K.C. Chou, *Proteins Struct. Funct. Genet.* 43 (2001) 246; Correction *Proteins Struct. Funct. Genet.* 2044 (2001) 2060], the Nearest Neighbour Algorithm (NNA) was developed for predicting the protein subcellular location. Very high success rates were observed, suggesting that such a hybrid approach may become a useful high-throughput tool in the area of bioinformatics and proteomics.

© 2003 Elsevier Science (USA). All rights reserved.

Keywords: Hybrid algorithm; Proteins subcellular location; Functional domain composition; Pseudo-amino acid composition; Bioinformatics; Proteomics

Given the sequence of a protein, how to predict which subcellular location it belongs to? Owing to the fact that the localization of a protein in a cell is closely correlated with its biological function, and that the number of sequences entering into databanks has been rapidly increasing, the importance of the problem has become self-evident. Particularly, it is anticipated that many more new protein sequences will be derived soon because of the recent success of the human genome project, which has provided an enormous amount of genomic information in the form of three billion base pairs, assembled into tens of thousands of genes. Therefore, the challenge to address such a problem will become even more urgent and critical in the very near future. Actually, many efforts have been made trying to develop some computational methods for quickly predicting the subcellular locations of proteins [1–14]. Of these methods, some [1,2] are based on the N-terminal sorting

signals. Their merit is with a clear biological implication [15]. However, as pointed out by Reinhardt and Hubbard [5], “In large genome analysis projects gene are usually automatically assigned and these assignments are often unreliable for the 5'-regions” “This can lead to leader sequences being missing or only partially included, thereby causing problems for prediction algorithms depending on them.” Therefore, most of the existing algorithms were based on the information derived from entire protein sequences rather than their signal peptides alone. However, because of the difficulty due to the extreme variance in sequence order and length, the majority of these algorithms are actually based on the amino acid composition of an entire protein chain. According to the classical definition, the amino acid composition consists of 20 components, representing the occurrence frequency of each of the 20 native amino acids in a given protein, and hence a protein is represented by a 20D (dimensional) vector [16,17]. Obviously, if using the classical amino acid composition alone to represent a protein, all the sequence-order and sequence-length effects would be missed out and the prediction method underlain with

* Corresponding author. Present address: Biomolecular Sciences Department, UMIST, P.O. Box 88, Manchester, M60 1QD, UK. Fax: +44-161-236-0409.

E-mail address: y.cai@umist.ac.uk (Y.-D. Cai).

such a basis must bear a considerable intrinsic limitation. It is in only a few recent papers [10,11] that efforts to take into account the sequence-order effects were initiated through a novel concept, the so-called pseudo-amino acid composition [11]. It should be pointed out, however, that the pseudo-amino acid composition can make allowance for incorporating the partial or quasi-sequence order effects [10] only, but not the complete sequence-order effects. Therefore, the pseudo-amino acid composition may still miss some information which might be directly related to the function of a protein. Meanwhile, a completely different approach, the so-called functional domain composition [13], was proposed that incorporated the information of various functional types. The introduction of the functional domain composition represents an important progress in directly relating the localization of proteins with their function. However, owing to the fact that the current functional domain database [18] is far from complete yet, some proteins cannot be properly defined in terms of the functional domain composition, leading to some setback in practical application. In view of this, here a strategy is developed to represent a protein by combining the functional domain composition and pseudo-amino acid composition. The combination makes allowance for bringing out the best in each other. With that approach, the Nearest Neighbour Algorithm [19,20] was applied to predict the subcellular location of proteins and high success rates were observed.

Combination of functional domain composition and pseudo-amino acid composition

The original concept of the functional domain composition and the detailed procedure of how to use it to represent a protein were given in a pioneer paper [13], where the functional domain composition was defined in the SBASE-A database [18]. The SBASE-A database consists of 2005 functional domains and hence the functional domain composition thus defined corresponds to a 2005D (dimensional) vector. In this paper, the InterPro database, i.e., the integrated domain and motif database [21], was used to define the functional domain composition of a protein. InterPro release 5.2 (September 2002) contains 5875 entries. With each of the 5875 functional domains as a vector-base, a protein can be defined as a 5875D vector, as illustrated by the following procedures.

(1) Use the program IPRSCAN [21] to search InterPro database for a given protein, if there is a hit (e.g., IPR000307, meaning the protein contains a sequence very similar to that of the 307th domain of the InterPro database), then the 307th component of the protein in the 5875D functional domain space is assigned 1; otherwise, 0.

(2) The protein can thus be explicitly formulated as

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_{5875} \end{bmatrix}, \quad (1)$$

where

$$x_i = \begin{cases} 1 & \text{hit,} \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Thus, a protein is corresponding to a 5875D vector by using each of the 5875 functional domain sequences as a base. In other words, rather than the 20D space [17] in terms of the amino acid composition, or the $(20 + \lambda)$ D space of the pseudo-amino acid composition [11], or the 2002D space of the functional domain composition [13] based SBASE-A database [18], a protein is now defined in a 5875D space.

Because not all the proteins could get hits within InterPro database [21], for those proteins with which no hit was found, the pseudo-amino acid composition [11] was adopted to represent them, as given below:

$$\mathbf{X} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_{20} \\ p_{20+1} \\ \vdots \\ p_{20+\lambda} \end{bmatrix}, \quad (3)$$

where p_1, p_2, \dots, p_{20} represent the 20 components of the classical amino acid composition, while p_{20+1} is the first-tier sequence correlation factor, p_{20+2} the second-tier sequence correlation factor, and so forth. For the current study, we took $\lambda = 20$, i.e., the dimension of the pseudo-amino acid composition considered is 40. Given a protein, the 40 pseudo-amino acid components (cf. Eq. (3)) can be easily derived by following the procedures as described in a previous publication [11].

The Nearest Neighbour Algorithm

The Nearest Neighbour (NN) Algorithm [19,20] tries to classify the new patterns into their class membership by comparing the features of the unknown new patterns with the features of the patterns which have already been classified. It is particularly useful in the situations when the distributions of the patterns and the categories of the patterns are unknown. The approach will weigh heavily the evidence derived from the nearby patterns. It is attractive because it is simple to implement and has a low probability of error.

Suppose there are N proteins ($\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N$) which have been classified into categories $1, 2, \dots, \mu$. Now, for a query protein \mathbf{X} , how can we predict which category it belongs to? According to the nearest neighbour principle, the prediction can be formulated as follows. First, let us define a *generalized distance* between \mathbf{X} and \mathbf{X}_i ($i = 1, 2, \dots, N$) given by

$$D(\mathbf{X}, \mathbf{X}_i) = 1 - \frac{\mathbf{X} \cdot \mathbf{X}_i}{\|\mathbf{X}\| \|\mathbf{X}_i\|} \quad (i = 1, 2, \dots, N), \quad (4)$$

where $\mathbf{X} \cdot \mathbf{X}_i$ is the dot product of vectors \mathbf{X} and \mathbf{X}_i , and $\|\mathbf{X}\|$ and $\|\mathbf{X}_i\|$ their modulus, respectively. Obviously, when $\mathbf{X} \equiv \mathbf{X}_i$, we have $D(\mathbf{X}, \mathbf{X}_i) = 0$. Generally speaking, the generalized distance is within the range of 0 and 1; i.e., $0 \leq D(\mathbf{X}, \mathbf{X}_i) \leq 1$.

Accordingly, the NN algorithm can be expressed as follows. If the generalized distance between \mathbf{X} and \mathbf{X}_k ($k = 1, 2, \dots$, or N) is the smallest; i.e.

$$D(\mathbf{X}, \mathbf{X}_k) = \text{Min}\{D(\mathbf{X}, \mathbf{X}_1), D(\mathbf{X}, \mathbf{X}_2), \dots, D(\mathbf{X}, \mathbf{X}_N)\}, \quad (5)$$

then the query protein \mathbf{X} is predicted as belonging to the same category as of \mathbf{X}_k . If there is a tie, the query protein is not uniquely determined, but cases like that rarely occur.

Since a query protein may or may not get a hit in searching the InterPro database [21], it is important to realize that if a query protein has no hit found during the prediction process (cf. Eq. (2)), then the query protein, as well as all the proteins in the training dataset, should be defined by the 40D pseudo-amino acid composition as given by Eq. (3); if a query protein can be defined in the 5875D functional domain composition, then prediction should be carried out based on those proteins in the training set that can be defined in the same 5875D space as well. Accordingly, the current NN predictor actually consists of two sub predictors: (1) the NN-5875D predictor that operates in the 5875D functional domain composition space and (2) the NN-40D predictor that operates in the $(20 + \lambda)$ D pseudo-amino acid composition with $\lambda = 20$.

Results and discussion

To show the power of the hybridization approach, the datasets constructed by Reinhardt and Hubbard [5] were used for demonstration. The reason we used the datasets constructed by other investigators rather than ours is to make the showcase more compelling and objective because they are available to public and have been used by a number of investigators [6,22–26] to test different prediction methods. The datasets consist of two parts: the prokaryotic set and the eukaryotic set. The former contains 997 prokaryotic protein sequences, of which 688 are cytoplasmic, 107 extracellular, and 202

periplasmic; the latter contains 2427 eukaryotic protein sequences, of which 684 are cytoplasmic, 325 extracellular, 321 mitochondrial, and 1097 nuclear. For a fast test of the power of a prediction algorithm, one can use such simple datasets including only three and four localizations, but for practical application, we would like to recommend using datasets covering more localization such as the one constructed by Chou and Elrod [7].

The computations were carried out on a Silicon Graphics IRIS Indigo workstation (Elan 4000). By searching the InterPro database for the 997 proteins in the prokaryotic set, 913 proteins got hits and 84 did not. And for the 2427 proteins in the eukaryotic set, 2239 got hits but 188 not. The results of these hits clearly indicate the need to combine the functional domain composition approach with the pseudo-amino acid composition approach. Otherwise, 84 proteins in the prokaryotic set and 188 proteins in the eukaryotic set would have no definition in the functional domain composition, leading to a failure of identifying their subcellular localization. On the other hand, in addition to the advantage of incorporating some sequence-order effects, the pseudo-amino acid composition can always be used to define a protein no matter whether it gets hits or not in searching the InterPro database. In view of this, a combination of the two approaches according to the following flowchart will optimize their synergy in operation: if a query protein got a hit by search InterPro database, then the NN-5875D predictor was used to predict its subcellular location; otherwise, the NN-40D predictor was used for the prediction.

Like most of previous investigators [6,22–26] in using the same datasets to demonstrate their methods, the examination was conducted by the re-substitution test and jackknife test, as reported below.

Re-substitution test

The so-called re-substitution test is an examination for the self-consistency of a prediction method. When the re-substitution test is performed for the current study, the subcellular location of each protein in the dataset is in turn identified using the rule parameters derived from the same dataset, the so-called training dataset. The results thus obtained are summarized in Table 1, from which we can see that the overall success rates for the 997 proteins in the prokaryotic set and the 2427 proteins in the eukaryotic set are both 100%, indicating a perfect self-consistency of the current predictor. However, during the process of the re-substitution test, the predictor derived from the training dataset includes the information of the query protein later plugged back in the test. This will certainly underestimate the error and enhance the success rate because the same proteins are used to train the predictor and to test themselves. Accordingly, the success rate thus obtained

Table 1

Overall success rates reported by investigators using different prediction methods for the prokaryotic and eukaryotic datasets,^a respectively

Investigators	Prokaryotic set ^b		Eukaryotic set ^c	
	Re-substitution (%)	Jackknife (%)	Re-substitution (%)	Jackknife (%)
Chou and Elrod [6]	90.4	86.5	N/A	N/A
Yuan [22]	N/A	89.1	N/A	73.0
Cai and Chou [23]	96.1	84.4	95.6	70.6
Feng [24]	93.5	89.2	N/A	N/A
Feng and Zhang [25]	97.7	90.4	N/A	N/A
Hua and Sun [26]	N/A	91.4	N/A	79.4
Authors of this paper	100	89.3	100	90.4

^a The datasets used here were constructed by Reinhardt and Hubbard [5].^b Contains 997 protein sequences, of which 688 are cytoplasmic, 107 extracellular, and 202 periplasmic.^c Contains 2427 protein sequences, of which 684 are cytoplasmic, 325 extracellular, 321 mitochondrial, and 1097 nuclear.

represents an optimistic estimation [7,17,27,28]. Nevertheless, the re-substitution test is absolutely necessary because it reflects the self-consistency of an identification method, especially for its algorithm part. An identification algorithm certainly cannot be deemed as a good one if its self-consistency is poor. In other words, the re-substitution test is necessary but not sufficient for evaluating an identification method. As a complement, a cross-validation examination is needed.

Jackknife test

As is well known, in statistical prediction the single independent dataset test, sub-sampling test, and jackknife test are the three methods often used for cross-validation. Of these three, the jackknife test is deemed as the most effective and objective one; see, e.g., Chou and Zhang [29] for a comprehensive discussion about this, and Mardia et al. [30] for the mathematical principle. During jackknifing, each protein in the dataset is in turn singled out as a tested protein and all the rule-parameters are calculated based on the remaining proteins. In other words, the subcellular location of each protein is identified by the rule parameters derived using all the other proteins except the one which is being identified. During the process of jackknifing both the training dataset and testing dataset are actually open, and a protein will in turn move from one to the other. The overall success rates thus obtained for the prokaryotic set and those for the eukaryotic set are also given in Table 1. Meanwhile, for facilitating comparison, the corresponding rates reported by the previous investigators are listed in the same table as well. As shown in Table 1, the overall success rates obtained by the current methods have been remarkably improved. For example, none of the jackknife rates reported by the previous investigators for the eukaryotic set have ever exceeded 80%, but the rate by the current method has exceeded 90%, a significant enhancement.

Moreover, although the current nearest neighbour algorithm took longer time to perform prediction than the algorithms based on pure analytical mathematics,

such as the covariant discriminant algorithm [7] and the augmented covariant discriminant algorithm [10], it was computationally much more efficient than the neural networks [12] and support vector machines [23] because no convergence requirement was involved during computation. For example, it took about 5 h CPU time by a Silicon Graphics IRIS Indigo workstation (Elan 4000) to complete the jackknife test for the 2427 proteins of the eukaryotic set. Therefore, the computation speed would not be an issue for the nearest neighbour algorithm if the computer can be used to handle the algorithms of neural networks or support vector machines.

Conclusion

The pseudo-amino acid composition approach [11] and the functional domain composition approach [31] are two completely different approaches developed for improving the prediction quality of protein subcellular location. They are both quite powerful, but each has its own limitation. The present study has demonstrated that a combination of the two different approaches can make them complement each other, and that the introduction of the nearest neighbour algorithm can make allowance for bringing out the best in each other and making each shining more brilliantly in the other's company. This is the essence why the success rates predicted by the current method are superior to those by many other methods for the same datasets as reported by previous investigators. It has not escaped our notice that such a hybrid approach can also be used to improve the prediction quality for other protein attributes [32], such as G-protein-coupled receptor types [33,34] and enzyme family classes [35].

Acknowledgments

The authors would like to express their gratitude to the anonymous reviewer whose comments were very helpful in improving the presentation of this communication.

References

- [1] K. Nakai, M. Kanehisa, *Genomics* 14 (1992) 897–911.
- [2] M.G. Claros, S. Brunak, G. von Heijne, *Curr. Opin. Struct. Biol.* 7 (1997) 394–398.
- [3] H. Nakashima, K. Nishikawa, *J. Mol. Biol.* 238 (1994) 54–61.
- [4] J. Cedano, P. Aloy, J.A. P'erez-pons, E. Querol, *J. Mol. Biol.* 266 (1997) 594–600.
- [5] A. Reinhardt, T. Hubbard, *Nucleic Acids Res.* 26 (1998) 2230–2236.
- [6] K.C. Chou, D.W. Elrod, *Biochem. Biophys. Res. Commun.* 252 (1998) 63–68.
- [7] K.C. Chou, D.W. Elrod, *Protein Eng.* 12 (1999) 107–118.
- [8] K.C. Chou, D.W. Elrod, *Proteins Struct. Funct. Genet.* 34 (1999) 137–153.
- [9] K.C. Chou, *Curr. Protein Peptide Sci.* 1 (2000) 171–208.
- [10] K.C. Chou, *Biochem. Biophys. Res. Commun.* 278 (2000) 477–483.
- [11] K.C. Chou, *Proteins Struct. Funct. Genet.* 43 (2001) 246–255 (Erratum: *Proteins Struct. Funct. Genet.* 2044 (2001) 2060).
- [12] Y.D. Cai, X.J. Liu, X.B. Xu, K.C. Chou, *J. Cell. Biochem.* 84 (2002) 343–348.
- [13] K.C. Chou, Y.D. Cai, *J. Biol. Chem.* 277 (2002) 45765–45769.
- [14] G.P. Zhou, K. Doctor, *Proteins Struct. Funct. Genet.* 50 (2003) 44–48.
- [15] K.C. Chou, *Curr. Protein Peptide Sci.* 3 (2002) 615–622.
- [16] K.C. Chou, C.T. Zhang, *J. Biol. Chem.* 269 (1994) 22014–22020.
- [17] K.C. Chou, *Proteins Struct. Funct. Genet.* 21 (1995) 319–344.
- [18] J. Murvai, K. Vlahovicek, E. Barta, S. Pongor, *Nucleic Acids Res.* 29 (2001) 58–60.
- [19] T.M. Cover, P.E. Hart, *IEEE Trans. Inform. Theory* IT-13 (1967) 21–27.
- [20] J.H. Friedman, F. Baskett, L.J. Shustek, *IEEE Trans. Inform. Theory* C-24 (1975) 1000–1006.
- [21] R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M.D.R. Croning, R. Durbin, L. Falquet, W. Fleischmann, L. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N.J. Mulder, T.M. Oinn, M. Pagni, F. Servant, C.J.A. Sigrist, E.M. Zdobnov, *Nucleic Acids Res.* 29 (2001) 37–40.
- [22] Z. Yuan, *FEBS Lett.* 451 (1999) 23–26.
- [23] Y.D. Cai, K.C. Chou, *Mol. Cell Biol. Res. Commun.* 4 (2000) 172–173.
- [24] Z.P. Feng, *Biopolymers* 58 (2001) 491–499.
- [25] Z.P. Feng, C.T. Zhang, *Int. J. Biol. Macromol.* 28 (2001) 255–261.
- [26] S. Hua, Z. Sun, *Bioinformatics* 17 (2001) 721–728.
- [27] Y.D. Cai, *Proteins Struct. Funct. Genet.* 43 (2001) 336–338.
- [28] G.P. Zhou, N. Assa-Munt, *Proteins Struct. Funct. Genet.* 44 (2001) 57–59.
- [29] K.C. Chou, C.T. Zhang, *Crit. Rev. Biochem. Mol. Biol.* 30 (1995) 275–349.
- [30] K.V. Mardia, J.T. Kent, J.M. Bibby, *Multivariate Analysis*, Academic Press, London, 1979, p. 322, 381.
- [31] K.C. Chou, Y.D. Cai, *Proteins Struct. Funct. Genet.*, 2003 (in press).
- [32] K.C. Chou, in: P.W. Weinrer, Q. Lu (Eds.), *Gene Cloning & Expression Technologies*, Eaton Publishing, Westborough, MA, 2002, pp. 57–70 (Chapter 54).
- [33] K.C. Chou, D.W. Elrod, *J. Proteome Res.* 1 (2002) 429–433.
- [34] D.W. Elrod, K.C. Chou, *Protein Eng.* 15 (2002) 713–715.
- [35] K.C. Chou, D.W. Elrod, *J. Proteome Res.* 2 (2003) 183–190.