Predicting subcellular localization of proteins in a hybridization space

Yu-Dong Cai1,* and Kuo-Chen Chou2,3

1Biomolecular Sciences Department, UMIST, PO Box 88, Manchester M60 1QD, UK, 2Gordon Life Science Institute, San Diego, CA 92130, USA and 3Tianjin Research Institute of Bioinformatics and Drug Discovery (TRIBD), Tianjin, China

Received on July 14, 2003; revised on November 18, 2003; accepted on December 5, 2003
Advance Access publication February 5, 2004

ABSTRACT

Motivation: The localization of a protein in a cell is closely correlated with its biological function. With the number of sequences entering into databanks rapidly increasing, the importance of developing a powerful high-throughput tool to determine protein subcellular location has become self-evident. In view of this, the Nearest Neighbour Algorithm was developed for predicting the protein subcellular location using the strategy of hybridizing the information derived from the recent development in gene ontology with that from the functional domain composition as well as the pseudo amino acid composition.

Results: As a showcase, the same plant and non-plant protein datasets as investigated by the previous investigators were used for demonstration. The overall success rate of the jack-knife test for the plant protein dataset was 86%, and that for the non-plant protein dataset 91.2%. These are the highest success rates achieved so far for the two datasets by following a rigorous cross-validation test procedure, suggesting that such a hybrid approach (particularly by incorporating the knowledge of gene ontology) may become a very useful high-throughput tool in the area of bioinformatics, proteomics, as well as molecular cell biology.

Availability: The software would be made available on sending a request to the authors.

Contact: ycai@umist.ac.uk

1 INTRODUCTION

Given a protein sequence, how can one accurately predict which subcellular localization it belongs to? This is currently a very hot topic in molecular biology because it involves the three essential features of a protein; i.e. its biological objective, its biochemical activity, as well as its place in the cell where a gene product is active. Since the number of protein sequences entering into databanks has been rapidly increasing, the importance and urgency of the problem have become self-evident. Particularly, it is anticipated that many more new protein sequences will be derived soon owing to the recent success of the human genome project, which has provided an enormous amount of genomic information in the form of 3 billion base pairs, assembled into tens of thousands of genes. Therefore, the challenge of dealing with such a problem will become even more critical and urgent in the very near future. Actually, much efforts has been made trying to develop some computational methods for quickly predicting the subcellular locations of proteins (Cedano et al., 1997; Chou and Elrod, 1999; Claros et al., 1997; Emanuelsson et al., 2000; Nakai and Kanehisa, 1992; Nakashima and Nishikawa, 1994; Pan et al., 2003; Reinhardt and Hubbard, 1998; Zhou and Doctor, 2003). Of these methods, some (Claros et al., 1997; Nakai and Kanehisa, 1992) are based on the N-terminal sorting signals. Their merit is with a clear biological implication because newly synthesized proteins in vivo are governed by an intrinsic signal sequence to their destination, whether they are to pass through a membrane into a particular organelle, to become integrated into the membrane or to be exported out of the cell (Blobel, 1976). However, as pointed out by Reinhardt and Hubbard (1998), ‘In large genome analysis projects genes 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 amino acid composition derived from an entire protein sequence. However, the amino acid composition consists of only 20 components, representing the occurrence frequency of each of the 20 native amino acids in a given protein and corresponding to a 20-dimensional (20D) vector. Obviously, if it is used to represent a protein, all the sequence-order and sequence-length effects would be totally missed out and the prediction method underlain with such a basis must bear a considerable intrinsic limitation. To improve such a situation, a novel concept, the so-called pseudo-amino acid composition (Chou, 2001) was proposed and a remarkable improvement in prediction quality was observed. The pseudo-amino acid composition consists of 20 + λ components, of which the first 20 components are the same as those in the conventional amino acid composition, and the components...
from 20 + 1 to 20 + λ. represent λ sequence-order correlation factors of different ranks. Theoretically speaking, λ can be as large as the number of total amino acid residues in a protein. For example, if a protein consists of 50 amino acid residues, then its pseudo-amino acid composition can contain as large as 20 + 50 = 70 components, corresponding to a 70D vector. It is the additional λ components that make allowance for incorporating some sequence-order effects (Chou, 2001; Pan et al., 2003). However, the pseudo-amino acid composition only represents the partial (but not complete) sequence-order information, and hence may still miss some information that might be immediately related to the function of a protein. Subsequently, a completely different approach, the so-called functional domain composition (Chou and Cai, 2002) was proposed that incorporated the information of various functional domain 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 (Vlahovicek et al., 2002) 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 (Chou and Cai, 2002).

In view of this, here, a strategy is developed to represent a protein by hybridizing the gene ontology composition, functional domain composition and pseudo-amino acid composition. The hybridization makes allowance for bringing out the best in one another. With that approach, the Intimate Sorting Algorithm was developed to predict the subcellular localization of proteins, and the high success rates yielded.

2 HYBRIDIZATION OF GENE ONTOLOGY, FUNCTIONAL DOMAIN AND PSEUDO-AMINO ACID COMPOSITIONS

To improve the quality of predicting protein subcellular location, a logic step is to catch the core features of a protein that are intimately related to its localization in a cell. According to the Gene Ontology (GO) Consortium (Ashburner et al., 2000), the GO database was established based on the following criteria: (a) biological process referring to a biological objective to which the gene or gene product contributes; (b) molecular function defined as the biochemical activity of a gene product and (c) cellular component referring to the place in the cell where a gene product is active. Since the above three criteria are not only the attributes of genes, gene products or gene-product groups, but also the core features reflecting the subcellular localization (Chou et al., 1998, 1999), it is anticipated that the prediction quality will be enhanced if the GO database is used to define proteins according to the following procedures.

By mapping of InterPro (Apweiler et al., 2001) to GO, one can get a list of data called ‘InterPro2GO’ (ftp://ftp.ebi.ac.uk/pub/databases/interpro/interpro2go/), where each InterPro entrance corresponds to a GO number. The relationships between InterPro and GO may be one-to-many, reflecting the biological reality that a particular protein may function in several processes, contain domains that carry out diverse molecular functions, and participate in multiple alternative interactions with other proteins, organelles or locations in the cell’ (Ashburner et al., 2000). For example, ‘IPR000003’ corresponds to ‘GO:0003677’, ‘GO:0004879’, ‘GO:0005496’, ‘GO:0006355’ and ‘GO:0005634’. Also, since the current GO database is far from complete yet, some InterPro entrances (such as IPR000001, IPR000002 and IPR000004) do not have the corresponding GO numbers in the InterPro2GO list.

Furthermore, the GO numbers in InterPro2GO are not increasing successively and orderly, and hence a reorganization and compression procedure was taken to renumber them. For example, after such a procedure, the original GO numbers GO:0000012, GO:0000015, GO:0000030, . . . , GO:0046413 would become GO_compress:0000001, GO_compress:0000002, GO_compress:0000003, . . . , GO_compress:0001930, respectively. The GO database thus obtained is called GO_compress database, whose dimensions were reduced to 1930 from 46 413 in the original GO database. Each of the 1930 entities in the GO_compress database will serve as a base to define a protein.

However, the current GO numbers do not give a complete coverage in the sense that some proteins might not belong to any of the GO numbers. Although the problem will eventually no longer exist as GO increases in size, to cope with such a situation right now, a hybrid approach was introduced by combining GO with the functional domain composition (Chou and Cai, 2002) and pseudo-amino acid composition (Chou, 2001), as described below.

(1) Use the program IPRSCAN (Apweiler et al., 2001) to search InterPro (release 6.1) database (Apweiler et al., 2001) for a given protein. If there is a hit corresponding to the i-th number of the GO_compress database, then the i-th component of the protein in the 1930D GO_compress space is assigned 1; otherwise 0. Thus, the protein can be formulated as

\[
P = \begin{bmatrix}
a_1 \\
a_2 \\
\vdots \\
a_i \\
\vdots \\
a_{1930}
\end{bmatrix},
\]

where

\[
a_i = \begin{cases} 
1, & \text{hit found in GO_compress} \\
0, & \text{otherwise.}
\end{cases}
\]
space (Apweiler et al., 2001), as given below

$$P = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_j \\ \vdots \\ b_{7785} \end{bmatrix},$$  \hspace{1cm} (3)

where

$$b_j = \begin{cases} 1, & \text{hit found in InterPro} \\ 0, & \text{otherwise}. \end{cases} \hspace{1cm} (4)$$

(3) If no hit was found at all even in the entire 7785D InterPro space, the protein should be defined in the \((20 + \lambda)D\) pseudo-amino acid space, as given below

$$P = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_{20} \\ c_{20+1} \\ \vdots \\ c_{20+\lambda} \end{bmatrix},$$  \hspace{1cm} (5)

where \(c_1, c_2, \ldots, c_{20}\) represent the 20 components of the classical amino acid composition, while \(c_{20+1}\) is the first-tier sequence order correlation factor, \(c_{20+2}\) the second-tier sequence order correlation factor, and so forth [cf. Figure 1 of Chou (2001)]. Generally speaking, the larger the number of these correlation factors, the more the sequence-order effects incorporated. However, the number \(\lambda\) cannot exceed the length of a protein (i.e. the number of its total residues).

For the current study, we took \(\lambda = 37\), i.e. the dimension of the pseudo-amino acid composition considered is \(20 + 37 = 57\). Given a protein, the 57 pseudo-amino acid components in Equation (5) can be easily derived by following the procedures as described in the original paper (Chou, 2001) which introduces the concept of pseudo-amino acid composition.

### 3 THE NEAREST NEIGHBOUR ALGORITHM

The Nearest Neighbour (NN) Algorithm (Cover and Hart, 1967; Friedman et al., 1975) 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 and 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 \((P_1, P_2, \ldots, P_N)\) which have been classified into categories 1, 2, \ldots, \(\mu\). Now, for a query protein \(P\), how can we predict which category it belongs to? According to the nearest neighbor principle, the prediction can be formulated as follows. First, let us define a generalized distance between \(P\) and \(P_k\) \((k = 1, 2, \ldots, N)\) given by

$$D(P, P_k) = 1 - \frac{P \cdot P_k}{\|P\| \cdot \|P_k\|}, \hspace{1cm} (k = 1, 2, \ldots, N),$$  \hspace{1cm} (6)

where \(P \cdot P_k\) is the dot product of vectors \(P\) and \(P_k\), and \(\|P\|\) and \(\|P_k\|\) their modulus, respectively. Obviously, when \(P = P_k\), we have \(D(P, P_k) = 0\). Generally speaking, \(D(P, P_k)\) is within the range of 0 and 1; i.e. \(0 \leq D(P, P_k) \leq 1\). According-ly, the NN algorithm can be expressed as follows. If the generalized distance between \(P\) and \(P_k\) \((k = 1, 2, \ldots, or N)\) is the smallest; i.e.

$$D(P, P_k) = \min\{D(P, P_1), D(P, P_2), \ldots, D(P, P_N)\},$$  \hspace{1cm} (7)

then the query protein \(P\) is predicted belonging to the same category as that of \(P_k\). If there is a tie, the query protein is not uniquely determined, but cases like that rarely occur.

The following self-consistency principle should be followed in using the hybridization approach practically. If a query protein was defined in the 1930D GO_compress space [Equation (1)], then the prediction should be carried out based on those proteins in the training set that could be defined in the same 1930D space. If all the components for the query protein in the 1930D GO_compress space were zero and hence it was defined by shifting to the 7785D functional domain space [see Equation (3)], then the prediction should be conducted on the basis that all the rule parameters were derived from the same 7785D space. Finally, if all the components for the query protein in the 7785D functional domain space were also zero and its definition must be made by shifting to the \((20 + \lambda)D\) pseudo-amino acid composition space [Equation (5)], then the prediction should be carried out according to the principle that all the proteins in the training set be defined in the same pseudo-amino acid composition space as well. Accordingly, the current NN predictor actually consists of three sub predictors: (a) the NN-1930D predictor that operates in the 1930D GO_compress space, (b) the NN-7785D predictor that operates in the 7785D functional domain composition space and (c) the NN-57D predictor that operates in the 57D pseudo-amino acid composition space with \(\lambda = 37\) (Chou, 2001).

### 4 RESULTS AND DISCUSSION

To benchmark the prediction quality of the current method against others and to make the comparison more objectively, the datasets constructed by other investigators were used for demonstration. The datasets constructed by Emanuelsson et al. (2000) contain two redundancy-reduced sets. One is the plant set that consists of 940 proteins, of which 141 are destined for the chloroplast, 368 for the mitochondrion, 269 for the secretory pathway and 162 for the other localizations.
such as nuclear and cytosolic. The other is the non-plant set that consists of 2738 proteins, of which 371 are destined for the mitochondrion, 715 for the secretory pathway and 1652 for the other localizations such as nuclear and cytosolic. According to the report by Emanuelsson et al. (2000), the overall success rate predicted by the TargetP predictor for the 940 plant proteins classified into four different categories was 85%, and that for the 2738 non-plant proteins classified into three categories was 90%. These were the highest success rates so far reported in literature for the aforementioned plant and non-plant datasets. Now for the exactly same datasets, we used the hybrid approach to perform prediction.

The computation was carried out in a Silicon Graphics IRIS Indigo workstation (Elan 4000). According to the search procedures as described in Section 2, we obtained the following results. (a) For the 940 protein sequences in the plant set, 748 got hits in the GO database and hence were defined in the 1930D GO_compress space; 92 of the remainder got hits in the InterPro database (release 6.1) and hence were defined in the 7785D InterPro space, and the 100 proteins that were finally left were defined in the 57D pseudo amino acid space. (b) For the 2738 protein sequences in the non-plant set, 1989 were defined in the 1930D GO_compress space; 458 defined in the 7785D InterPro space and 291 defined in the 57D pseudo amino acid space. This means that, if only the GO database was used, 192 proteins in the plant set and 749 in the non-plant set would have no definition, leading to a failure in identifying their localization. By incorporating the InterPro database, we still have 100 proteins in the plant set and 291 proteins in the non-plant set without definition (Table 1). That is why it is so important to hybridize with the pseudo-amino acid composition, by which not only a protein can always be defined but its sequence-order effects may also be considerably reflected (Chou, 2001). Thus, the hybrid algorithm was operated according to the following steps: if a query protein was defined in the GO_compress database, then the NN-1930D predictor was used to predict its subcellular location; if the query protein could not be defined in the GO_compress database but defined in the InterPro database, then the NN-7785D predictor was used to predict its subcellular location; if the query protein could neither be defined in the GO_compress database nor in the InterPro database, then the NN-57D predictor was used to predict its subcellular location.

The prediction quality was examined by the jackknife test. Compared with the independent dataset test and sub-sampling test often adopted in biology, the jackknife test is thought the most objective and effective method for cross-validation in statistics (Mardia et al., 1979) and widely used recently (Feng, 2001; FENG and Zhang, 2001; HUA and Sun, 2001; Pan et al., 2003; YUAN, 1999; ZHOU, 1998; ZHOU and Assa-Munt, 2001). This is because in the independent dataset test, the selection of a testing dataset is quite arbitrary, and the accuracy thus obtained lacks an objective criterion unless the testing dataset is sufficiently large (Chou, 1995; CHOU and Zhang, 1995). As for the sub-sampling test in which a given dataset is divided into several subsets, the problem is that the number of possible divisions might be too large to be handled. For example, in the treatment by Emanuelsson et al. (2000), the number of proteins in each of the subgroups of the 940 plant sequences and 2738 non-plant sequences was ‘truncated by a number divisible by five’ and then ‘divided into five equally sized parts for cross-validation’. Four of them were used as the training data and one as the testing data. Thus, the number of possible divisions for the plant set of 940 proteins classified into four subsets would be \( \Pi = \Pi_1 \times \Pi_2 \times \Pi_3 \times \Pi_4 \), where \( \Pi_1 \) is the number of the possible divisions for the 141 chloroplast proteins and can be formulated (after the aforementioned truncation procedure) as

\[
\Pi_1 = \frac{(141 - 1)!}{(28!28!28!28!15!)},
\]

where 1 in the numerator is the truncated number. Likewise, for the 368 mitochondrion, 269 secretory and 162 other proteins, we have

\[
\Pi_2 = \frac{(368 - 3)!}{73!73!73!73!5!},
\]

\[
\Pi_3 = \frac{(269 - 4)!}{53!53!53!53!5!},
\]

and

\[
\Pi_4 = \frac{(162 - 2)!}{32!32!32!32!5!},
\]

respectively. Of \( \Pi_1, \Pi_2, \Pi_3 \) and \( \Pi_4 \), the smallest is \( \Pi_1 \approx 4 \times 10^{91} \), indicating that the number of total possible divisions would be \( \Pi \gg 10^{95} \). This is an astronomical figure, which is too large to be handled by any existing computers. Furthermore, it is conceivable that, for the non-plant set of 2738 proteins, the corresponding number of total possible divisions would be even greater. Hence, in any practical sub-sampling test as conducted by Emanuelsson et al. (2000), only a very small fraction of the possible divisions was investigated, and the results thus obtained could hardly avoid arbitrariness and might be overestimated, as will be further illustrated later. Here it is instructive to point out that, although the 5-fold sub-sampling cross-validation is also often used in biological

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GO_compress space</th>
<th>InterPro space</th>
<th>57D pseudo-amino acid composition space</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plant</td>
<td>748</td>
<td>92</td>
<td>100</td>
<td>940</td>
</tr>
<tr>
<td>Non-plant</td>
<td>1989</td>
<td>458</td>
<td>291</td>
<td>2738</td>
</tr>
</tbody>
</table>

Table 1. Breakdown of the proteins defined in the hybridization space of gene ontology, functional domain and pseudo-amino acid composition
literature for neural network predictors when the dataset is small, it is not a rigorous test but just a compromise. For the mathematical principle and a comprehensive discussion in this regard, refer to a monograph (Mardia et al., 1979) and a review paper (Chou and Zhang, 1995). Why are the sub-sampling cross-validations still seen in literature for testing neural network predictors? This is because (1) sub-sampling is much easier than jackknifing in programming and (2) sub-sampling tests take much shorter computation time than jackknifing, especially for neural network-based algorithms since the process of jackknifing must in turn train the neural network parameters one by one for each of the individual predictions. Therefore, very few reports in literature have been seen for the jackknife rates by using neural network-based predictors because it would otherwise take an extremely long computation time. That is why the data for the jackknife rates by TargetP and PSort are not available. But there is no such problem for the current algorithm. The current algorithm took longer time to perform prediction than the algorithms based on the pure analytical mathematics, such as the covariant discriminant algorithm (Chou and Elrod, 1999) and the augmented covariant discriminant algorithm (Chou, 2000). But it was computationally much more efficient than the neural networks (Cai et al., 2002) and support vector machines (Cai and Chou, 2000) because no convergence requirement whatsoever was involved during computation. For example, it took only a few hours of CPU times by a Silicon Graphics IRIS Indigo workstation (Elan 4000) to complete the jackknife test for the 940 proteins of the plant set. The overall success rates thus obtained are given in Table 2. For facilitating comparison, the rates obtained by the other predictors, such as TargetP (Emanuelsson et al., 2000) and PSort (Nakai and Horton, 1999; Nakai and Kanehisa, 1992), are also listed in the same table. From Table 2 we can see the following: (1) The overall success rates obtained by the current approach, which has combined the gene product, functional domain and sequence-order effects, are the highest, indicating that the subcellular localization of a protein is closely related to its gene product and function in both the plant and non-plant cases. (2) At a first glance, the success rates by the current approach seem only slightly higher than those by TargetP. Nevertheless, the rates for TargetP were obtained by conducting the sub-sampling test rather than the rigorous jackknife test. As mentioned above, the rates thus obtained could not avoid arbitrariness and were likely overestimated. For example, if one uses the same sub-sampling test procedure as used by Emanuelsson et al. (2000), the overall success rates by the current approach could reach 95.7 and 99.0% for the plant set and non-plant set, respectively, which is 9–10% higher than the corresponding rates by TargetP.

5 CONCLUSION

From both the rationality of testing procedure and the success rates of test results, hybridization of the gene ontology approach, functional domain approach (Chou and Cai, 2002), and the pseudo-amino acid composition approach (Chou, 2001) can significantly improve the prediction quality of protein subcellular location in both the plant and non-plant cases. This is fully consistent with the scientific logic because the current hybrid approach has combined the gene product, functional domain and quasi-sequence-order effects. The gene product is closely correlated with the biological process, molecular function and cellular component; while the functional domain (Chou and Cai, 2002) and quasi-sequence-order (Chou, 2001) have each proved to play an important role in determining the subcellular localization of a protein. The introduction of the nearest neighboring algorithm, i.e. NN predictor, can make allowance for bringing out the best in each and making one shine more brilliantly in the others’ company. It is anticipated that the power of the hybridization approach will be further enhanced with the continuous development in the relevant databases, such as incorporating the new version of functional domain database (Vlahovichek et al., 2002) and SMART database (Letunic et al., 2002) where extensive annotation for each domain family is available, providing information relating to function, subcellular localization, phyletic distribution and tertiary structure.

It has not escaped our notice that the hybridization approach can also be used to improve the prediction quality for other protein attributes (Chou, 2002; Jensen et al., 2002), such as protein functional classes (Jensen et al., 2002), enzyme categories (Chou and Elrod, 2003; Jensen et al., 2002) and protein quaternary structure attributes (Chou and Cai, 2003).

---

**Table 2. Comparison of localization predictor performances on the redundancy-reduced 940 plant proteins and 2738 non-plant proteins**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Test method</th>
<th>Overall success rate (%)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearest Neighbor (NN)</td>
<td>Jackknife</td>
<td>86.1</td>
<td>This work Emanuelsson et al. (2000)</td>
</tr>
<tr>
<td>TargetP</td>
<td>Sub-sampling</td>
<td>85.0</td>
<td>Nakai and Horton (1999); Nakai and Kanehisa (1992)</td>
</tr>
<tr>
<td>PSort</td>
<td>Sub-sampling</td>
<td>69.8</td>
<td>Nakai and Horton (1999); Nakai and Kanehisa (1992)</td>
</tr>
</tbody>
</table>

---

(1) The breakdown of the rate into those obtained from the three spaces (cf. Table 1) is: 662/748 = 88.5% for the 748 proteins defined in the 193D GO compres space; 90/92 = 97.8% for the 92 proteins in the 7785D InterPro space; and 57/100 = 57.0% for the remaining 100 proteins in the 57D pseudo-amino acid composition space.

(2) The breakdown of the rate into those obtained from the three spaces (cf. Table 1) is: 866/1989 = 93.8% for the 1989 proteins defined in the 193D GO compres space; 450/458 = 98.3% for the 458 proteins in the 7785D InterPro space; and 182/291 = 62.5% for the remaining 291 proteins in the 57D pseudo-amino acid composition space.
REFERENCES


