

Sequence analysis

Predicting protein localization in budding yeast

Kuo-Chen Chou^{1,2,3,*} and Yu-Dong Cai^{1,4}¹Gordon Life Science Institute, San Diego, CA 92130, USA, ²Shanghai Jiaotong University, Biomedical Engineering, Shanghai 200030, China, ³Tianjin Institute of Bioinformatics and Drug Discovery (TIBDD), Tianjin, China and ⁴Biomolecular Sciences Department, UMIST, Manchester, M60 1QD, UKReceived on September 23, 2004; revised on October 13, 2004; accepted on October 18, 2004
Advance Access publication October 28, 2004

ABSTRACT

Motivation: Most of the existing methods in predicting protein subcellular location were used to deal with the cases limited within the scope from two to five localizations, and only a few of them can be effectively extended to cover the cases of 12–14 localizations. This is because the more the locations involved are, the poorer the success rate would be. Besides, some proteins may occur in several different subcellular locations, i.e. bear the feature of ‘multiplex locations’. So far there is no method that can be used to effectively treat the difficult multiplex location problem. The present study was initiated in an attempt to address (1) how to efficiently identify the localization of a query protein among many possible subcellular locations, and (2) how to deal with the case of multiplex locations.

Results: By hybridizing gene ontology, functional domain and pseudo amino acid composition approaches, a new method has been developed that can be used to predict subcellular localization of proteins with multiplex location feature. A global analysis of the proteins in budding yeast classified into 22 locations was performed by jack-knife cross-validation with the new method. The overall success identification rate thus obtained is 70%. In contrast to this, the corresponding rates obtained by some other existing methods were only 13–14%, indicating that the new method is very powerful and promising. Furthermore, predictions were made for the four proteins whose localizations could not be determined by experiments, as well as for the 236 proteins whose localizations in budding yeast were ambiguous according to experimental observations. However, according to our predicted results, many of these ‘ambiguous proteins’ were found to have the same score and ranking for several different subcellular locations, implying that they may simultaneously exist, or move around, in these locations. This finding is intriguing because it reflects the dynamic feature of these proteins in a cell that may be associated with some special biological functions.

Contact: kchou@san.rr.com**Supplementary information:** www.pami.sjtu.edu.cn/kcchou

1 INTRODUCTION

One of the fundamental goals in cell biology and proteomics is to identify the functions of proteins in the context of compartments that organize them in the cellular environment. To realize this, it is indispensable to first identify the subcellular locations of proteins.

However, it is time-consuming and costly to determine the localization of a newly found protein in a cell purely based on experiments. Particularly, we are facing the times the number of protein sequences is growing extremely fast. For instance, the total number of protein sequences entering into the SWISS-PROT databank was only 3939 in 1986, and now the number has jumped to 162,781 according to version 44.7 released on October 11, 2004. This is more than 41 times the size in 1986! With the explosion in the number of sequences, it is highly desirable to develop an automated method to quickly identify the subcellular location of a newly found protein. Actually, many efforts have been made (Cedano *et al.*, 1997; Chou, 2001; Chou and Cai, 2002, 2003a,b; Chou and Elrod, 1999b; Emanuelsson *et al.*, 2000; Feng, 2001; Hua and Sun, 2001; Nakai and Kanehisa, 1991, 1992; Nakashima and Nishikawa, 1994; Pan *et al.*, 2003; Park and Kanehisa, 2003; Reinhardt and Hubbard, 1998; Zhou and Doctor, 2003) during the last decade or so. The development in this area has generally followed two trends. One is to improve the prediction quality by extracting more and more useful information from a protein sequence, such as using the information from the amino acid composition (Cedano *et al.*, 1997; Reinhardt and Hubbard, 1998), to the amino acid pair composition (Park and Kanehisa, 2003), to the pseudo amino acid composition (Chou, 2001; Pan *et al.*, 2003), and to the functional domain composition (Cai *et al.*, 2003; Chou and Cai, 2002). The other trend is to enhance the practical application value by enlarging the coverage scope, such as from the scope of covering only two subcellular locations (Nakashima and Nishikawa, 1994) to five locations (Cedano *et al.*, 1997), to 12 locations (Chou and Elrod, 1999b; Park and Kanehisa, 2003), and to 14 locations (Chou and Cai, 2003b). Recently, using the GFP (green fluorescent protein) fluorescence technique, Huh *et al.* (2003) made a global analysis of protein localization in budding yeast, classifying the proteins into 22 distinct subcellular localization categories. Compared with the previous datasets, the dataset determined experimentally by these authors not only covers the largest scope so far, but also reflects the fact that some proteins may occur in several different subcellular locations; i.e. have the attribute with ‘multiplex locations’. Actually, all the previous methods were developed to deal with only the ‘mono-location’ case where a given protein is assumed to belong to one, and only one, subcellular location. Now we are facing a ‘multi-location’ problem. How to deal with the case of multiplex locations is a big challenge that was always artificially avoided in the previous treatments. The present study is devoted to addressing this problem.

*To whom correspondence should be addressed.

2 SYSTEMS AND METHODS

The experimental classification results by Huh *et al.* (2003) can be downloaded from the website <http://www.yeastgfp.ucsf.edu>. After excluding those whose sequences are not available, we have 4115 proteins, of which four proteins, i.e. YFL030W, YJL057C, YJL107C and YLR426W, do not have subcellular location, and 236 proteins whose locations are ambiguous. Thus, we have $4115 - 4 - 236 = 3875$ proteins left. The remaining 3875 proteins, which are clearly classified into 22 distinct subcellular locations, can serve as a solid basis for further development in predicting protein subcellular locations. Meanwhile, the 3875 proteins will also serve as a training dataset to predict the $4 + 236 = 240$ proteins whose subcellular locations are unknown or ambiguous. A breakdown of the 3875 proteins into 22 subcellular locations is given in Table 1, from which we can see that, owing to the fact that some proteins coexist in several different subcellular locations, the so-called ‘multiplex location’ feature as mentioned above, the total number of different proteins \tilde{N} is smaller than the total number of classified proteins \tilde{N} . The relationship between these two is given by

$$\tilde{N} = N + \sum_{\lambda=2}^{\mu} (\lambda - 1) \Phi_{\lambda}, \quad (1)$$

where μ is the number of the total subcellular locations investigated, and Φ_{λ} is the number of proteins that occur simultaneously in λ different subcellular locations. For instance, of the $N = 3875$ proteins provided by Huh *et al.* (2003), 2968 ($=\Phi_1$) occur in only one subcellular location, 1106 ($=\Phi_2$) in two different locations, 63 ($=\Phi_3$) in three different locations, 7 ($=\Phi_4$) in four different locations, 1 ($=\Phi_5$) in five different locations, and 0 in $\lambda(=6, 7, \dots, 22)$ different locations. Substituting these numbers into Equation (1), we have

$$\begin{aligned} \tilde{N} &= 3875 + (2 - 1) \times 1106 + (3 - 1) \times 63 + (4 - 1) \\ &\quad \times 7 + (5 - 1) \times 1 = 5132, \end{aligned} \quad (2)$$

which is fully consistent with the number of \tilde{N} derived from Table 1.

The key to improving the prediction quality of the protein subcellular location is to grasp the core features of a protein that are intimately related to the current theme, and then use these features to represent it. In this sense, we can use the source of Gene Ontology (GO) Consortium (Ashburner *et al.*, 2000) as a vehicle to formulate the prediction algorithm. The term ‘ontology’ was originally borrowed from philosophy, where an ontology is a systematic account of existence. In other words, an ontology is an explicit specification of a conceptualization. In the GO database, gene products are organized according to the following three principles in a species-independent manner: cellular components, molecular function and biological process.

The first principle is directly related to the subcellular localization, while the other two are associated with the molecular function of a protein and its acting object, and hence are also closely relevant to the subcellular location of a protein (Alberts *et al.*, 1994; Chou and Elrod, 1999a). Accordingly, it is anticipated that the prediction quality will be significantly improved if the GO database is used to define proteins according to the following steps.

Step 1 Mapping InterPro (Apweiler *et al.*, 2001) entries to GO, one can get a list of data called ‘InterProt2GO’ (<ftp://ftp.ebi.ac.uk/pub/databases/interpro/interpro2go/>), where each InterPro entry corresponds to a GO number. Since a protein may have one or more molecular functions, be used in one or more biological processes, and be associated with one or more cellular components, the relationships between InterPro and GO may be one-to-many. For instance, the InterPro entry ‘IPR_000003’ 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 entries (such as IPR_000001, IPR_000002 and IPR_000004) do not have the corresponding GO numbers in the InterProt2GO list.

Step 2 The GO numbers in the InterProt2GO database do not increase successively and orderly, and hence an operation to reorganize and compress the GO numbers obtained in Step 1 is needed. For example, after such an operation, the original GO numbers GO_0000012, GO_0000015,

Table 1. Breakdown of the 3875 proteins in budding yeast into 22 subcellular locations according to the experimental observations (Huh *et al.*, 2003)

Subcellular location	Number of proteins
1. Actin	32
2. Bud	25
3. Bud neck	61
4. Cell periphery	128
5. Cytoplasm	1767
6. Early Golgi	54
7. Endosome	45
8. ER	288
9. ER to Golgi	6
10. Golgi	41
11. Late Golgi	44
12. Lipid particle	22
13. Microtubule	20
14. Mitochondrion	516
15. Nuclear periphery	60
16. Nucleolus	162
17. Nucleus	1432
18. Peroxisome	21
19. Punctate composite	133
20. Spindle pole	59
21. Vacuolar membrane	57
22. Vacuole	159
Total number of classified proteins \tilde{N}	5132
Total number of different proteins \tilde{N}	3875

GO_0000030, ..., GO_0046413 would become GO-compress_0000001, GO-compress_0000002, GO-compress_0000003, ..., GO-compress_0001930, respectively. The database thus obtained is called GO-compress database or the 1930D GO database, whose dimensions have been reduced to 1930 from 46,413 of the original GO database.

Step 3 Each of the 1930 GO numbers will serve as a base to define a protein \mathbf{P} in terms of the following 1930D (dimensional) vector:

$$\mathbf{P} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_i \\ \vdots \\ a_{1930} \end{bmatrix}, \quad (3)$$

where $a_i = 1$ if there is a hit corresponding to the i th ($i = 1, 2, \dots, 1930$) GO number when using the program IPRSCAN (Apweiler *et al.*, 2001) to search InterPro functional domain database (release 6.1) (Apweiler *et al.*, 2001) for the protein \mathbf{P} ; otherwise, $a_i = 0$.

Step 4 If no hit (i.e. no corresponding GO number) is found in the entire 1930D GO-compress space, the protein \mathbf{P} formulated by Equation (3) will correspond to a naught vector. To cope with such a circumstance, the protein should be defined in the 7785D FunD (Functional Domain composition) space (Apweiler *et al.*, 2001), as given below:

$$\mathbf{P} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_j \\ \vdots \\ b_{7785} \end{bmatrix}, \quad (4)$$

where $b_i = 1$ if there is a hit corresponding to the i th ($i = 1, 2, \dots, 7785$) InterPro FunD database (Apweiler *et al.*, 2001) for the protein \mathbf{P} (Chou and Cai, 2002); otherwise, $b_i = 0$.

Step 5 If no hit is found even in the entire 7785D FunD space, the protein should be defined in the $(20 + \lambda)$ D PseAA (Pseudo Amino Acid composition) space, as given below:

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

where c_1, c_2, \dots, c_{20} represent the 20 components of the classical amino acid composition (Nakashima *et al.*, 1986; Chou and Zhang, 1993; Chou, 1995; Zhou, 1998), 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. Fig. 1 of Chou (2001)]. It is the additional λ components in Equation (5) that incorporate some sequence-order effects into the vector representation of a protein. Generally speaking, the larger the number of the λ components, the more the sequence-order effects incorporated. However, the number λ cannot exceed the length of a protein (i.e. the number of its total residues). Also, if the number of λ is too large, the overall success rate by jack-knife tests might be reduced (Chou, 2001). Therefore, for different training datasets, λ may have different optimal values. For the current study, the optimal value of λ is 37. Given a protein, the $(20 + 37) = 57$ pseudo amino acid components in Equation (5) can be easily derived by following the procedures as described in Chou (2001), the paper that introduced the concept of pseudo-amino acid composition. Thus, the protein that corresponds to a naught vector in both the 1930D GO space [Equation (3)] and the 7785D FunD space [Equation (4)] can always be explicitly defined in the 57D PseAA space [Equation (5)].

The prediction was performed with the ISort (Intimate Sorting) predictor, which can be briefed below. Suppose there are \mathbb{N} proteins ($\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_{\mathbb{N}}$) which have been classified into categories $1, 2, \dots, \mu$. Now, for a query protein \mathbf{P} , how can we predict which category it belongs to? To deal with this problem, let us define the following scale function to measure the similarity between \mathbf{P} and \mathbf{P}_i ($i = 1, 2, \dots, \mathbb{N}$):

$$\Lambda(\mathbf{P}, \mathbf{P}_i) = \frac{\mathbf{P} \cdot \mathbf{P}_i}{\|\mathbf{P}\| \|\mathbf{P}_i\|} \quad (i = 1, 2, \dots, \mathbb{N}) \quad (6)$$

where $\mathbf{P} \cdot \mathbf{P}_i$ is the dot product of vectors \mathbf{P} and \mathbf{P}_i , and $\|\mathbf{P}\|$ and $\|\mathbf{P}_i\|$ their modulus, respectively. Obviously, when $\mathbf{P} \equiv \mathbf{P}_i$, we have $\Lambda(\mathbf{P}, \mathbf{P}_i) = 1$, meaning they have perfect or 100% similarity. Generally speaking, the similarity is within the range of 0 and 1; i.e. $0 \leq \Lambda(\mathbf{P}, \mathbf{P}_i) \leq 1$. Accordingly, the ISort predictor can be formulated as follows. If the similarity between \mathbf{P} and \mathbf{P}_k ($k = 1, 2, \dots, \mathbb{N}$) is the highest, i.e.

$$\Lambda(\mathbf{P}, \mathbf{P}_k) = \mathbf{Max}\{\Lambda(\mathbf{P}, \mathbf{P}_1), \Lambda(\mathbf{P}, \mathbf{P}_2), \dots, \Lambda(\mathbf{P}, \mathbf{P}_{\mathbb{N}})\}, \quad (7)$$

where the operator **Max** means taking the maximum one among those in the brackets, then the query protein \mathbf{P} is predicted as belonging to the same category as of \mathbf{P}_k . The ISort predictor is particularly useful for the situation when the distributions of the samples are unknown.

During the course of prediction, the following self-consistency principle should be followed. If a query protein could be defined in the 1930D GO space [Equation (3)], then the prediction should be carried out based on those proteins in the training set that could also be defined in the same 1930D GO space. If all of the components for the query protein in the 1930D Go space are zero and hence it is defined by shifting to the 7785D functional domain space [Equation (4)], then the prediction should be conducted on the basis that all the rule parameters are derived from the same 7785D space. Finally, if all the components for the query protein in the 7785D functional domain space are also zero and its definition must be made by shifting to the $(20 + \lambda)$ D PseAA

space [Equation (5)], then the prediction should be carried out according to the principle that all the proteins in the training dataset be defined in the same PseAA space as well.

Accordingly, the current ISort predictor actually consists of three subpredictors: (1) the ISort-1930D GO predictor that operates in the compressed 1930D gene ontology space, (2) the ISort-7785D FunD predictor that operates in the 7785D functional domain composition space, and (3) the ISort-57D PseAA predictor that operates in the 57D pseudo-amino acid composition space with $\lambda = 37$. The entire process is called GO-FunD-PseAA hybridization approach.

3 SOME REMARKS ABOUT THE MONO-LOCATION AND MULTI-LOCATION PREDICTIONS

As mentioned at the beginning, all the previous studies (Cedano *et al.*, 1997; Chou, 2001; Chou and Cai, 2002, 2003a,b; Chou and Elrod, 1999b; Emanuelsson *et al.*, 2000; Feng, 2001; Hua and Sun, 2001; Nakai and Kanehisa, 1991, 1992; Nakashima and Nishikawa, 1994; Pan *et al.*, 2003; Park and Kanehisa, 2003; Reinhardt and Hubbard, 1998; Zhou and Doctor, 2003) were confined to within the scope of mono-location prediction. Here we are facing a multi-location problem, i.e. some proteins may coexist in several different subcellular locations. To deal with this kind of situation, it is instructive to highlight the difference between the mono-location and multi-location predictions according to the following points.

Training dataset For the mono-location case where a given protein belongs to one, and only one, subcellular location, the total number of samples in the training dataset can be expressed as

$$\mathbb{N} = \sum_{m=1}^{\mu} n_m, \quad (8)$$

where n_m is the number of proteins in the m th subcellular location. However, for the multi-location case, the total number of the samples in the training dataset should be instead expressed as [Equation (1)]

$$\tilde{\mathbb{N}} = \sum_{m=1}^{\mu} \tilde{n}_m, \quad (9)$$

where \tilde{n}_m has the same meaning as n_m of Equation (8) except that it is now associated with the multi-location case. Accordingly, we generally have $\tilde{n}_m \geq n_m$ because a protein may simultaneously occur in several different subsets. This implies that \mathbb{N} in Equations (6) and (7) should be replaced by $\tilde{\mathbb{N}}$ during the process of prediction.

Success rate Suppose the proteins in budding yeast form a set S , which is the union of the 22 subsets; i.e.

$$S = S_1 \cup S_2 \cup S_3 \cup \dots \cup S_{21} \cup S_{22}, \quad (10)$$

where each subset corresponds to one of the 22 subcellular locations according to the order of Table 1. For the mono-location case, suppose the result operated by a predictor Ψ on \mathbf{P}_k^m , the k th protein in the m th subset, is the location belonging to the y_k^m th subset; i.e.

$$\Psi(\mathbf{P}_k^m) = y_k^m \quad (m = 1, 2, \dots, \mu; y_k^m = 1, 2, \dots, \mu), \quad (11)$$

then the overall success rate can be defined by

$$\frac{1}{\bar{N}} \sum_{m=1}^{\mu} \sum_{k=1}^{n_m} \delta[y_k^m, m], \quad (12)$$

where the delta function

$$\delta[y_k^m, m] = \begin{cases} 1, & \text{if } y_k^m = m, \\ 0, & \text{if } y_k^m \neq m. \end{cases} \quad (13)$$

However, for the multi-location case, the definition will be more complicated because the predicted result for a given protein may belong to one or more subcellular locations. Suppose $\tilde{\Psi}$ is a multi-location predictor, instead of Equation (11), we should have

$$\tilde{\Psi}(\mathbf{P}_k^m) = Y_k^m, \quad (14)$$

where Y_k^m is not a number but a set that is formed by one or more of the 22 subsets in Equation (10). Thus, the overall success rate is defined by

$$\frac{1}{\bar{N}} \sum_{m=1}^{\mu} \sum_{k=1}^{\tilde{n}_m} \Delta[Y_k^m, S_m], \quad (15)$$

where the Δ function is defined by

$$\Delta[Y_k^m, S_m] = \begin{cases} 1, & \text{if } S_m \in Y_k^m, \\ 0, & \text{if } S_m \notin Y_k^m. \end{cases} \quad (16)$$

Score of the scale function $\Lambda(\mathbf{P}, \mathbf{P}_i)$ The prediction is governed by the score of the similarity scale function according to Equation (6). Its interpretation is quite straightforward for the mono-location case; i.e. if $\Lambda(\mathbf{P}, \mathbf{P}_2)$ has the highest score, then the query protein \mathbf{P} is predicted to belong to the same location as \mathbf{P}_2 , the 2nd protein in the training dataset. For the multi-location case, however, the following two points should be realized. First, if \mathbf{P}_2 belongs to three different subcellular locations, then three identical highest scores are expected with each corresponding to one of the three locations. And the query protein \mathbf{P} is predicted to belong to these three locations as well. Secondly, as additional information, the results for the 2nd highest score and the 3rd highest score are also provided here.

4 RESULTS AND DISCUSSION

The computation was performed in a Silicon Graphics IRIS Indigo workstation (Elan 4000). According to steps 1–5 as described in Section 2, we obtained the following results (Table 2). For the 3875 different protein sequences in budding yeast, 2571 got hits in the GO database and hence were defined in the 1930D GO space, 539 of the remainder got hits in the FunD database and were hence defined in the 7785D FunD space, and finally the 765 proteins left were defined in the 57D PseAA space. For the 5132 classified proteins, the corresponding breakdown numbers are also given in Table 2. This means that if only the GO database was used, $3875 - 2571 = 1304$ proteins in budding yeast would have no definition, leading to a failure of identifying their localization. By incorporating the InterPro FunD database, we still have 765 proteins without definition (Table 2). That is why it is so important to hybridize with the pseudo-amino

Table 2. Breakdown of the 3875 different proteins and 5132 classified proteins, defined in the hybridization space of GO, FunD and PseAA composition

Dataset	1930D GO space	7785D FunD space	57D PseAA space	Total
Total number of different proteins \bar{N}	2571	539	765	3875
Total number of classified proteins \bar{N}	3425	725	982	5132

Table 3. Overall success rates of jack-knife cross-validation by the GO-FunD-PseAA predictor for the 5132 classified proteins in the budding yeast (see Table 1)

Counted scope ^a	Success rate
Ranking I	$\frac{3596}{5132} = 70.07\%$
Ranking I+II	$\frac{4328}{5132} = 84.33\%$
Ranking I+II+III	$\frac{4627}{5132} = 90.16\%$

^aThe counted scope is defined as follows. In ranking I, the predicted results with only the highest score are counted; in ranking I+II, those with both the highest and second highest scores are counted; in ranking I+II+III, those with all the first three highest scores are counted.

acid composition (PseAA), by which not only a protein can always be defined but also its sequence-order effects may considerably be reflected (Chou, 2001). Thus, the hybrid algorithm was operated according to the procedures: if a query protein was defined in the GO database, then the ISort-1930D GO predictor was used to predict its subcellular location; if the query protein could not be defined in the GO database but could be defined in the InterPro FunD database, then the ISort-7785D FunD predictor was used to predict its subcellular location; if the query protein could be defined neither in the GO database nor in the InterPro FunD database, then the ISort-57D PseAA predictor was used to predict its subcellular location.

As is well known, in statistical prediction the single independent dataset test, sub-sampling test and jack-knife test are the three methods often used for cross-validation. Of these three, the jack-knife test is deemed as the most rigorous and objective one [see the review by Chou and Zhang (1995) for a comprehensive discussion about this, and the monograph by Mardia *et al.* (1979) for the underlying mathematical principle]. Therefore, the jack-knife test has been used by more and more investigators (Feng, 2001; Hua and Sun, 2001; Pan *et al.*, 2003; Yuan, 1999; Zhou, 1998; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003) in examining the power of various prediction methods. With the current approach, the success rates by the jack-knife cross-validation for the 5132 classified proteins in budding yeast are given in Table 3, from which we can see that the overall success rate is 70.07%. It is instructive to mention the following

Table 4. Predicted results for the four proteins in budding yeast whose subcellular locations could not be observed by experiments^a

Subcellular location	Protein code			
	YFL030W	YJL057C	YJL107C	YLR426W
1. Actin				
2. Bud				
3. Bud neck				
4. Cell periphery	III			
5. Cytoplasm		III	I	
6. Early Golgi				
7. Endosome				
8. ER		I		
9. ER to Golgi				
10. Golgi				
11. Late Golgi				
12. Lipid particle		II		
13. Microtubule				
14. Mitochondrion	II			
15. Nuclear periphery	I			III
16. Nucleolus				
17. Nucleus			II	
18. Peroxisome				
19. Punctate composite			III	
20. Spindle pole				
21. Vacuolar membrane				I
22. Vacuole				II

^aThe roman numerals (I, II and III) reflect the ranking of hitting scores with I the highest, followed by II and III; e.g. for protein YJL057C: ER has the highest score (I) and hence the greatest likelihood where the protein will occur, followed by lipid particle (II) and cytoplasm (III).

two points. First, by following the procedures described in Equations (1), (9) and (14)–(16), those predictors which were established based on the amino acid composition, such as the Least Euclidean Distance algorithm (Nakashima and Nishikawa, 1994; Nakashima *et al.*, 1986), the Least Hamming Distance algorithm (Chou, 1989) and ProtLoc predictor (Cedano *et al.*, 1997), can be augmented to deal with the multi-localational case as well. However, the corresponding success rates obtained by those predictors were only 13.89, 14.03 and 13.95%, respectively. This implies that the success rate by the present approach is more than 56% higher. Secondly, as shown in Table 3, if ranking II (results with the 2nd highest score) and ranking III (results with the 3rd highest score) were also counted, the likelihood of hitting the localization of a protein in budding yeast could be as high as 90%.

Now let us use the 5132 classified proteins (Huh *et al.*, 2003) as the training dataset to predict the four proteins whose subcellular locations could not be determined by experiments and the 236 proteins whose subcellular locations were ambiguous (Huh *et al.*, 2003). The predicted results for the four location-unknown proteins are given in Table 4, where the roman numerals (I, II and III) reflect the ranking of likelihood. For example, nuclear periphery (I) has the highest likelihood for the subcellular location of protein YFL030W, and the next highest is mitochondrion (II), followed by cell periphery (III). The predicted results for the 236 ambiguous proteins are given in Online Supplementary Materials A. To help readers understand the data listed in the Online Supplementary

Materials A, the predicted results for the first five of the 236 proteins are summarized in Table 5 according to the format of Table 4. As we can see from Table 5, of the five proteins listed there, four have the same rankings for different subcellular locations, meaning that these proteins will coexist, or move around, in these locations. For example, protein YAR027W has ranking I for the following 20 locations: actin, bud, bud neck, cell periphery, cytoplasm, early Golgi, endosome, ER, ER to Golgi, Golgi, late Golgi, microtubule, mitochondrion, nuclear periphery, nucleolus, nucleus, punctate composite, spindle pole, vacuolar membrane and vacuole. This implies that YAR027W may coexist, or move around, in the 20 subcellular locations. It can be seen by looking at the data at the Online Supplementary Materials A that many of the proteins there have the same ranking for different subcellular locations. That is why the 236 proteins were attributed by Huh *et al.* (2003) as ‘ambiguous’ in subcellular location. According to our predicted results, these location-ambiguous proteins should be interpreted as those which coexist, or move around, in several different subcellular locations.

5 CONCLUSION

The key to enhancing the success rate of predicting protein subcellular location is to grasp the core features of proteins that are intimately related to their biological functions. This can be realized by defining a protein based on the GO (Ashburner *et al.*, 2000) and functional domain database (Apweiler *et al.*, 2001) developed recently. However, the current GO and functional domain database do not give a complete coverage so that some proteins cannot be meaningfully defined. Although the problem will be eventually solved as the GO and functional domain database increase in size, to deal with such a situation right now, a hybrid approach was introduced by combining them with the pseudo amino acid composition (Chou, 2001). With the latter, not only a protein can always be explicitly defined but also its sequence-order effects can be considerably incorporated. That is why a hybridization of these three approaches can yield the success rate that is far beyond the reach of the other existing methods, as demonstrated by a rigorous cross-validation test.

Particularly, the subcellular locations for the four proteins, whose localizations could not be determined by experiments (Huh *et al.*, 2003), have been explicitly predicted. Predictions were also made for the 236 proteins whose locations in budding yeast were ambiguous by experimental observations. According to our predicted results, however, it has been found that many of these proteins belong to several different subcellular locations, implying that they might simultaneously exist, or move around, in these locations. This finding is intriguing because it reflects the dynamic feature of these proteins in a cell that may have very special biological functions.

Just as the emergence of structural bioinformatics has greatly stimulated the process of both basic research and drug discovery (Chou, 2004), it is anticipated that the development of protein subcellular location prediction, particularly for cases with the multiplex location feature, will have important impacts on not only basic research but also on pharmaceutical industry and medical practice because proteins with such a dynamic feature are particularly interesting, and identifying differences in how proteins move within healthy and diseased cells is one critical way that doctors could diagnose disorders and gauge response to treatment.

Table 5. Predicted results for the first five of the 236 ambiguous proteins listed in the Online Supplementary Materials A^a

Subcellular location	Protein code YAL029C	YAL053W	YAR019C	YAR027W	YAR028W
1. Actin	II			I	
2. Bud	III			I	
3. Bud neck	I		II	I	I
4. Cell periphery	III	I	III	I	I
5. Cytoplasm	I		I	I	II
6. Early Golgi				I	
7. Endosome				I	
8. ER		II	III	I	
9. ER to Golgi				I	
10. Golgi				I	
11. Late Golgi	III			I	
12. Lipid particle				II	
13. Microtubule			II	I	
14. Mitochondrion			II	I	
15. Nuclear periphery				I	III
16. Nucleolus				I	
17. Nucleus	I		I	I	
18. Peroxisome				III	
19. Punctate composite				I	
20. Spindle pole				I	
21. Vacuolar membrane		III		I	
22. Vacuole	I		III	I	I

^aThe roman numerals (I, II and III) reflect the ranking of hitting scores. When a protein has the same ranking of hitting scores for different subcellular locations, it will have the same likelihood of occurring in these locations. For example, the protein YAR027W will have the highest likelihood of coexisting or moving around in the following 20 locations: actin, bud, bud neck, cell periphery, cytoplasm, early Golgi, endosome, ER, ER to Golgi, Golgi, late Golgi, microtubule, mitochondrion, nuclear periphery, nucleolus, nucleus, punctate composite, spindle pole, vacuolar membrane and vacuole.

ACKNOWLEDGEMENTS

The authors wish to thank the anonymous reviewers whose constructive comments have greatly improved the presentation of this paper.

REFERENCES

- Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. and Watson, J.D. (1994) *Molecular Biology of the Cell*, Ch. 1, 3rd edn. Garland Publishing, New York, London.
- Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D.R., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, L., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M., Servant, F., Sigrist, C.J.A. and Zdobnov, E.M. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.
- Cai, Y.D., Zhou, G.P. and Chou, K.C. (2003) Support vector machines for predicting membrane protein types by using functional domain composition. *Biophys. J.*, **84**, 3257–3263.
- Cedano, J., Aloy, P., P'erez-Pons, J.A. and Querol, E. (1997) Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.*, **266**, 594–600.
- Chou, P.Y. (1989) Prediction of protein structural classes from amino acid composition. In Fasman, G.D. (ed.), *Prediction of Protein Structure and the Principles of Protein Conformation*. Plenum Press, New York, pp. 549–586.
- Chou, K.C. (1995) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. *Proteins*, **21**, 319–344.
- Chou, K.C. (2001) Prediction of protein cellular attributes using pseudo-amino-acid-composition. *Proteins*, **43**, 246–255 (Erratum, 2001, **44**, 60).
- Chou, K.C. (2004) Review: structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.*, **11**, 2105–2134.
- Chou, K.C. and Cai, Y.D. (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.*, **277**, 45765–45769.
- Chou, K.C. and Cai, Y.D. (2003a) A new hybrid approach to predict subcellular localization of proteins by incorporating Gene ontology. *Biochem. Biophys. Res. Commun.*, **311**, 743–747.
- Chou, K.C. and Cai, Y.D. (2003b) Prediction and classification of protein subcellular location: sequence-order effect and pseudo amino acid composition. *J. Cell. Biochem.*, **90**, 1250–1260 (Addendum, 2004, **91** (5) 1085).
- Chou, K.C. and Elrod, D.W. (1999a) Prediction of membrane protein types and subcellular locations. *Proteins*, **34**, 137–153.
- Chou, K.C. and Elrod, D.W. (1999b) Protein subcellular location prediction. *Protein Eng.*, **12**, 107–118.
- Chou, J.J. and Zhang, C.T. (1993) A joint prediction of the folding types of 1490 human proteins from their genetic codons. *J. Theor. Biol.*, **161**, 251–262.
- Chou, K.C. and Zhang, C.T. (1995) Review: prediction of protein structural classes. *Critical Rev. Biochem. Mol. Biol.*, **30**, 275–349.
- Emanuelsson, O., Nielsen, H., Brunak, S. and von Heijne, G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.
- Feng, Z.P. (2001) Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. *Biopolymers*, **58**, 491–499.
- Hua, S. and Sun, Z. (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721–728.
- Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S. and O'Shea, E.K. (2003) Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691.
- Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979) *Multivariate Analysis*, Chs. 11–13. Academic Press, London, pp. 322–381.

- Nakai,K. and Kanehisa,M. (1991) Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins*, **11**, 95–110.
- Nakai,K. and Kanehisa,M. (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, **14**, 897–911.
- Nakashima,H. and Nishikawa,K. (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol.*, **238**, 54–61.
- Nakashima,H., Nishikawa,K. and Ooi,T. (1986) The folding type of a protein is relevant to the amino acid composition. *J. Biochem.*, **99**, 152–162.
- Pan,Y.X., Zhang,Z.Z., Guo,Z.M., Feng,G.Y., Huang,Z.D. and He,L. (2003) Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. *J. Protein Chem.*, **22**, 395–402.
- Park,K.J. and Kanehisa,M. (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acid and amino acid pairs. *Bioinformatics*, **19**, 1656–1663.
- Reinhardt,A. and Hubbard,T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.*, **26**, 2230–2236.
- Yuan,Z. (1999) Prediction of protein subcellular locations using Markov chain models. *FEBS Lett.*, **451**, 23–26.
- Zhou,G.P. (1998) An intriguing controversy over protein structural class prediction. *J. Protein Chem.*, **17**, 729–738.
- Zhou,G.P. and Assa-Munt,N. (2001) Some insights into protein structural class prediction. *Proteins*, **44**, 57–59.
- Zhou,G.P. and Doctor,K. (2003) Subcellular location prediction of apoptosis proteins. *Proteins*, **50**, 44–48.