Predicting Protein-Protein Interactions from Sequences in a Hybridization Space

ARTICLE *in* JOURNAL OF PROTEOME RESEARCH · FEBRUARY 2006

Impact Factor: 4.25 · DOI: 10.1021/pr050331g · Source: PubMed

2 AUTHORS:

Kuo-Chen Chou
Gordon Life Science Institute
526 PUBLICATIONS **35,465** CITATIONS
SEE PROFILE

Yu-Dong Cai
Shanghai University
210 PUBLICATIONS **7,206** CITATIONS
SEE PROFILE

research articles Journal of proteome research

Predicting Protein−Protein Interactions from Sequences in a Hybridization Space

Kuo-Chen Chou*,† and Yu-Dong Cai†,‡,§

Gordon Life Science Institute, 13784 Torrey Del Mar, San Diego, California 92130, Department of Chemistry, College of Sciences, Shanghai University, 99 Shang-Da Road, Shanghai 200436, China, and Department of Biomolecular Science, UMIST, P.O. Box 88, Manchester M60 1QD, United Kingdom

Received September 30, 2005

To understand the networks in living cells, it is indispensably important to identify protein−protein interactions on a genomic scale. Unfortunately, it is both time-consuming and expensive to do so solely based on experiments due to the nature of the problem whose complexity is obviously overwhelming, just like the fact that "life is complicated". Therefore, developing computational techniques for predicting protein−protein interactions would be of significant value in this regard. By fusing the approach based on the gene ontology and the approach of pseudo-amino acid composition, a predictor called "GO-PseAA" predictor was established to deal with this problem. As a showcase, prediction was performed on 6323 protein pairs from yeast. To avoid redundancy and homology bias, none of the protein pairs investigated has ≥40% sequence identity with any other. The overall success rate obtained by jackknife cross-validation was 81.6%, indicating the GO-PseAA predictor is very promising for predicting protein−protein interactions from protein sequences, and might become a useful vehicle for studying the network biology in the postgenomic era.

Keywords: Genomic scale • Gene ontology • Pseudo-amino acid composition • ISort predictor • GO-PseAA fusion classifier • Network biology • Yeast

1. Introduction

To understand the molecular underpinnings of life, it is indispensable to study protein−protein interactions. Proteins rarely function in isolation. Most of their functions essential to life are associated with protein−protein interactions. For instance, structural connections between cells are formed through protein−protein interactions; proteins are directed to the correct compartments of cells by binding to other proteins; some inhibitors of enzymes are proteins; proteins are modified and degraded by enzymes; protein messengers bind to protein receptors on the outer surface of cell membranes to send signals between cells; interactions between different protein subunits are the basis of allosteric changes in oligomers; protein−protein interactions underlie very large-scale movements in organisms, such as muscle contraction.

Actually, protein−protein interactions affect all processes in a cell. All cellular processes depend on precisely orchestrated interactions between proteins. Let us imagine a cell in which the specific interactions between proteins would suddenly disappear. The deprived cell would become "blind" and "deaf", completely paralytic, and would finally perish. Also, imagine a cell in which many abnormal interactions between proteins would suddenly occur, the unfortunate cell would completely lose control, leading to network confuse and a terrible disaster. This is because specific and normal protein−protein interactions are involved in almost all physiological processes. Therefore, characterizing protein−protein interactions, or understanding the interaction network, is vitally important.

The success of the human genome project has stimulated the emergence of a new and far more challenging area: protein network, a frontier to study the functional relationship of proteins in a cell. With the rapid increase of protein sequences in the postgenomic era, it is highly desired to develop an automated method for fast and accurate prediction of protein−protein interactions from protein sequences because knowledge thus obtained is very useful for the areas ranging from rational drug design to analysis of metabolic and signal transduction networks. The present study was initiated in an attempt to stimulate the development of this area.

2. Materials

For the protein−protein interactions of yeast, the original data was taken from STRING at http://string.embl.de.[1] Because a full description of a protein's function requires knowledge of all partner proteins with which it specifically associates, the word "interaction" or "association" in STRING can mean direct physical binding but can also mean indirect interaction such as participation in the same metabolic pathway or cellular process. The protein−protein interactions in STRING are classified into three categories: (1) high confidence, (2) medium

* To whom correspondence should be addressed. E-mail: kchou@san.rr.com.
† Gordon Life Science Institute.
‡ Shanghai University.
§ UMIST.

10.1021/pr050331g CCC: $33.50
Published on Web 12/23/2005

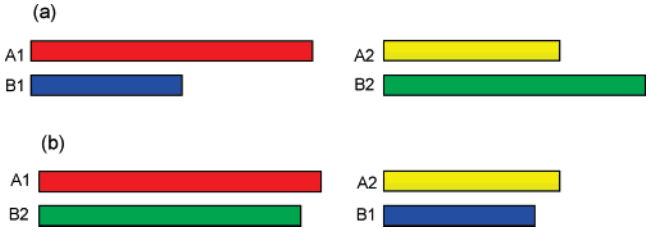

Figure 1. Illustration to show how to define the protein pairs which have ≥40% sequence identity to each other. Suppose protein pair A formed with proteins A1 (red) and A2 (yellow), and protein pair B with B1 (blue) and B2 (green). If (a) both the sequence identity between proteins A1 and B1 and that between A2 and B2 are ≥40%, or (b) both the sequence identity between proteins A1 and B2 and that between A2 and B1 are ≥40%, then the two protein pairs are defined as having ≥40% sequence identity to each other.

confidence, and (3) low confidence. These categories do not tell us anything about the "strength" of an interaction but, instead, simply refer to the reliability of having an interaction at all (i.e., the likelihood of not showing a "false-positive"). Furthermore, "interaction" here should not be interpreted as "physical binding" but, rather, "functional association". As an extreme case, two proteins could have a very high STRING score but not even touch each other in the cell. This could be the case, for example, if they are both part of a metabolic pathway and are simply associated with each other because they share an intermediate substrate. This kind of category in classifying protein−protein interactions is particularly useful when our focus is moving from a traditional investigation of proteins to a new frontier, the so-called "system biology" or "cellular networking".

Originally, we obtained 78 390 interaction couples between yeast proteins, of which 2455 belonged to the high confidence, 9400 the medium confidence, and 66 535 the low confidence. The data thus obtained are protein pairs. To reduce redundancy and homology bias for methodology development, all the protein pairs have been screened according to the following procedures. (1) For those pairs which had ≥40% sequence identity to one another, only one of them was kept. The ≥40% sequence identity between two protein pairs is defined as follows. Suppose protein pair A formed with proteins A1 and A2, and protein pair B with B1 and B2 (Figure 1). If both the sequence identity between proteins A1 and B1 and that between A2 and B2 are ≥40% or both the sequence identity between proteins A1 and B2 and that between A2 and B1 are ≥40%, then the two protein pairs are defined as having ≥40% sequence identity to each other. (2) Those protein pairs which contained a protein with less than 50 amino acids or longer than 5000 amino acids were removed. After these screening procedures, the total interaction pairs reduced to 59 248, of which 2111 belonged to the high confidence, 7430 the medium confidence, and 49 707 the low confidence. Furthermore, to make the three subsets have a similar size, we randomly picked 2109 pairs and 2103 pairs from the 7430 medium confidence pairs and the 49 707 low confidence pairs, respectively. Thus, the working dataset contained 6323 pairs, of which 2111 belonged to the high confidence, 2109 the medium confidence, and 2103 the low confidence. The accession numbers of the 6323 pairs are given in the Supporting Information A.

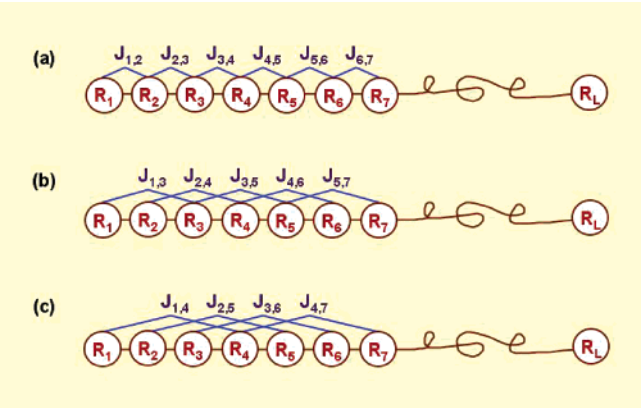

Figure 2. A schematic drawing to show (a) the first-rank, (b) the second-rank, and (3) the third-rank sequence-order correlation mode along a protein sequence. Panel a reflects the correlation mode between all the most contiguous residues, panel b that between all the second most contiguous residues, and panel c that between all the third most contiguous residues. Adapted from ref 2 with permission.

3. Method

The present classifier is established by fusing the classifier based on the pseudo-amino acid composition[2] and the classifier based on the Gene Ontology Consortium.[3] For easier understanding, let us first give a brief introduction of how it works for single individual proteins, and then extend the formulation for protein pairs.

3.1. Protein Sample Representation by Pseudo-Amino Acid Composition (PseAA). According to the conventional amino acid composition, given a protein **P** with L amino acid residues,

$$R_1R_2R_3R_4R_5R_6R_7...R_L \tag{1}$$

where R_1 represents the residue at the sequence position 1, R_2 at position 2, and so forth, we can expressed it as a vector in a 20D (dimensional) space[4,5]; that is,

$$\mathbf{P} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_{20} \end{bmatrix} \tag{2}$$

where p_1 is the occurrence frequency of amino acid A in the protein, p_2 that of amino acid C, and so forth. Here, without loss of generality, the single codes of the 20 native amino acids are used according to their alphabetical order. If a protein was represented by a set of components (the so-called "discrete mode") as given by eq 2, all its sequence information would be lost. To keep its representation with a discrete mode but without completely losing its sequence-order information, we can use the PseAA[2] to represent the protein, as given by

$$\mathbf{P} = \begin{bmatrix} p_1 \\ \vdots \\ p_{20} \\ p_{20+1} \\ \vdots \\ p_{20+\lambda} \end{bmatrix} \tag{3}$$

where the first 20 elements are the same as in eq 2, and p_{20+1} is the 1st pseudo-amino acid component related to the 1st rank of sequence-order correlation (Figure 2), p_{20+2} is the 2nd

pseudo-amino acid component related to the 2nd rank of sequence-order correlation, and do forth. It is the additional λ components that incorporate some sequence-order effects into the representation of a protein (cf. Appendix A). For different datasets, λ usually has different optimal value.[2] For the current study, the optimal value of λ was 40. Given a protein, the (20 + 40) = 60 pseudo-amino acid components in eq 3 can be easily derived by following the procedures as given in Appendix A.

3.2. Protein Sample Representation by Gene Ontology Consortium (GO). What is ontology? In a brief way the word "ontology" is a specification of a conceptualization, although it has a long history in philosophy, where it refers to the subject of existence. GO is established according to the following three species-independent 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,[3,6,7] it is anticipated that GO consortium will be a very useful vehicle for investigating P−P interactions. The steps in using GO to represent protein samples are described as follows.

Step 1. Mapping of InterPro[8] entries to GO, one can get a list of data called "InterProt2GO", where each InterPro entry corresponds to a GO number.

Step 2. 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".[3] For example, "IPR000003" corresponds to "GO:0003677", "GO:0004879", "GO:0005496", "GO:0006355", and "GO:0005634". Also, because the current GO consortium is far from complete yet, some InterPro entries (such as IPR000001, IPR000002, and IPR000004) do not have the corresponding GO numbers in the InterProt2GO list. Furthermore, the GO numbers in InerProt2GO are not increasing successively and orderly, and hence, a reorganization/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 system thus obtained is called GO_compress, whose dimensions were reduced to 1930 from 46 413 in the original GO consortium.

Step 3. Each of the 1930 GO numbers in GO_compress will serve as a base to define a 1930D vector for a given protein **P**, as formulated below

$$\mathbf{P} = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_i \\ \vdots \\ g_{1930} \end{bmatrix} \tag{4}$$

where $g_i = 1$ if there is a hit corresponding to the ith (i = 1, 2, ..., 1930) GO number when using the program IPRSCAN[8] to search InterPro functional domain database (release 6.1)[8] for the protein **P**; otherwise, $g_i = 0$, as done in the case for defining the functional domain composition.[9]

Step 4. If no hit whatsoever is found for any of the 1930 GO numbers, the corresponding protein will be immediately considered as a failure for its prediction, and the naught vector thus obtained will be removed from the GO representation system.

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

$$\Lambda(\mathbf{P},\mathbf{P}_i) = \frac{\mathbf{P}\cdot\mathbf{P}_i}{||\mathbf{P}||\,||\mathbf{P}_i||}, \quad (i = 1, 2, ..., N) \tag{5}$$

where $\mathbf{P}\cdot\mathbf{P}_i$ is the dot product of vectors **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; that is, $0 \le \Lambda(\mathbf{P},\mathbf{P}_i) \le 1$. Accordingly, the ISort classifier can be formulated as follows. If the similarity between **P** and $\mathbf{P}_k$ (k = 1, 2, ..., or N) is the highest, that is,

$$\Lambda(\mathbf{P},\mathbf{P}_k) = \max\{\Lambda(\mathbf{P},\mathbf{P}_1), \Lambda(\mathbf{P},\mathbf{P}_2), ..., \Lambda(\mathbf{P},\mathbf{P}_N)\} \tag{6}$$

where the operator max means taking the maximum one among those in the brackets, then the query protein **P** is predicted belonging to the same category as $\mathbf{P}_k$. If there is a tie, the query protein may not be uniquely determined and will be randomly assigned among those with a tie, but cases such as that rarely occur. The ISort classifier is particularly useful for the situation when the distributions of the samples are unknown.

Because there are two different representation systems for protein samples, we may generate three different classifiers: (1) ISort-60D PseAA classifier that operates in the 60D pseudo-amino acid composition space with $\lambda = 40$, (2) ISort-1930D GO classifier that operates in the 1930D GO space, and (3) GO-PseAA classifier that operates by fusing GO classifier and PseAA classifier. The rule by which to fuse the two is that the predicted result for a protein is always determined by the output from the ISort-1930D GO classifier if the query protein is not a naught vector in the 1930D GO space (cf. eq 4), but determined by that from the ISort-60D PseAA classifier when it is a naught vector, as can be formulated as follows:

$$\text{Result} = \begin{cases} \text{Output from GO,} & \text{if the query protein is not a naught vector} \\ \text{Output from PseAA,} & \text{otherwise} \end{cases} \tag{7}$$

All of the above equations were developed for single individual proteins. To study the interactions between two protein molecules, a formulation to cover protein pairs is needed. This can be done as follows. Suppose $\mathbf{P}^u$ and $\mathbf{P}^v$ are the uth protein and vth protein, respectively.

If expressed in the PseAA system (cf. eq 3), the u−v protein pair is expressed as

$$\mathbf{P}^u \oplus \mathbf{P}^v = \begin{bmatrix} p_1^u \\ p_2^u \\ \vdots \\ p_{20}^u \\ \vdots \\ p_{20+\lambda}^u \end{bmatrix} \oplus \begin{bmatrix} p_1^v \\ p_2^v \\ \vdots \\ p_{20}^v \\ \vdots \\ p_{20+\lambda}^v \end{bmatrix} = [p_1^u, p_2^u, ..., p_{20+\lambda}^u, p_1^v, p_2^v, ..., p_{20+\lambda}^v]^{\mathbf{T}} \quad (8a)$$

where p_1^u, p_2^u, ..., or p_1^v, p_2^v, ... have the same meaning as those in eq 3 except that they are now referred to a specified protein $\mathbf{P}^u$ or $\mathbf{P}^v$ instead of a general protein $\mathbf{P}$, the symbol $\oplus$ represents the sign of orthogonal sum, and $\mathbf{T}$ the transpose operator. Note that $\mathbf{P}^u \oplus \mathbf{P}^v$ is generally not equal to $\mathbf{P}^v \oplus \mathbf{P}^u$. Therefore, the protein pair should also include

$$\mathbf{P}^v \oplus \mathbf{P}^u = \begin{bmatrix} p_1^v \\ p_2^v \\ \vdots \\ p_{20}^v \\ \vdots \\ p_{20+\lambda}^v \end{bmatrix} \oplus \begin{bmatrix} p_1^u \\ p_2^u \\ \vdots \\ p_{20}^u \\ \vdots \\ p_{20+\lambda}^u \end{bmatrix} = [p_1^v, p_2^v, ..., p_{20+\lambda}^v, p_1^u, p_2^u, ..., p_{20+\lambda}^u]^{\mathbf{T}} \quad (8b)$$

If expressed in the GO system (cf. eq 4), such a $u-v$ protein pair is expressed by

$$\mathbf{P}^u \oplus \mathbf{P}^v = \begin{bmatrix} g_1^u \\ g_2^u \\ \vdots \\ g_i^u \\ \vdots \\ g_{1930}^u \end{bmatrix} \oplus \begin{bmatrix} g_1^v \\ g_2^v \\ \vdots \\ g_i^v \\ \vdots \\ g_{1930}^v \end{bmatrix} = [g_1^u, g_2^u, ..., g_{1930}^u, g_1^v, g_2^v, ..., g_{1930}^v]^{\mathbf{T}} \quad (9a)$$

and

$$\mathbf{P}^v \oplus \mathbf{P}^u = \begin{bmatrix} g_1^v \\ g_2^v \\ \vdots \\ g_i^v \\ \vdots \\ g_{1930}^v \end{bmatrix} \oplus \begin{bmatrix} g_1^u \\ g_2^u \\ \vdots \\ g_i^u \\ \vdots \\ g_{1930}^u \end{bmatrix} = [g_1^v, g_2^v, ..., g_{1930}^v, g_1^u, g_2^u, ..., g_{1930}^u]^{\mathbf{T}} \quad (9b)$$

Also, when using eqs 5 and 6 to predict P−P interactions, all the entries, whether they are query or training, should be in the form of protein pairs, as formulated in eqs 8 and 9. However, to avoid redundant calculations, if both $\mathbf{P}^u \oplus \mathbf{P}^v$ (cf. eqs 8a and 9a) and $\mathbf{P}^v \oplus \mathbf{P}^u$ (cf. eqs 8b and 9b) are used to express the input of a query protein pair, then only one of the two forms would suffice for all the protein pairs in the training dataset.

4. Results and Discussion

The computation was carried out in a Silicon Graphics IRIS Indigo workstation (Elan 4000). According to steps 1−4 as described in section 3.2, we obtained the following results. (a) For the 2111 protein pairs in the high confidence set, 1589 got hits in GO system and hence were meaningfully defined in the 1930D GO_compress space, and the remaining 522 protein pairs got no hits whatsoever and hence belong to a naught vector in the GO system. (b) For the 2109 protein pairs in the medium confidence set, 1596 were defined in the 1930D GO_compress space, and 513 belong to a naught vector. (c) For the 2103 protein pairs in the low confidence set, 1457 were defined in the 1930D GO_compress space, and 646 belong to a naught vector. This means that, if only the GO system was used, 2111 − 1589 = 522 protein pairs in the high confidence set, 2109 − 1596 = 513 protein pairs in the medium confidence set, and 2103 − 1457 = 646 protein pairs in the low confidence set would have no definition (Table 1), leading to a failure in identifying their P−P interaction feature. However, the naught vector problem would not occur in the PseAA representation system because a protein can always be meaningfully defined in the 60D PseAA space as long as it has a sequence of more than 40 amino acids. That is why, although the GO approach is extremely powerful, the overall success rate obtained by it might be lower than that obtained by the PseAA approach if the investigated dataset contains many proteins that could not meaningfully be defined by the GO representation. Accordingly, the GO-PseAA approach by fusing GO classifier and PseAA classifier as formulated by eq 7 will perform the best.

Table 1. Breakdown of the 6323 Protein−Protein Interaction Couples from STRING[1] into the Categories of Meaningful and Naught Vectors, Respectively, in the 1930D GO-Compress Space

rank of confidence	meaningful vector	naught vector	total
high	1589	522	2111
medium	1596	513	2109
low	1457	646	2103

Table 2. Jackknife Success Rates in Identifying the 6323 Protein−Protein Interactions Classified into High Confidence, Medium Confidence, and Low Confidence

rank of confidence	ISort-1930D GO classifier	ISort-60D PseAA classifier	GO-PseAA fusion classifier
high	$\frac{1387}{2111} = 65.7\%$	$\frac{1454}{2111} = 68.9\%$	$\frac{1753}{2111} = 83.0\%$
medium	$\frac{1451}{2109} = 68.8\%$	$\frac{1633}{2109} = 77.4\%$	$\frac{1792}{2109} = 85.0\%$
low	$\frac{1213}{2103} = 57.7\%$	$\frac{1349}{2103} = 64.2\%$	$\frac{1614}{2103} = 76.7\%$
total	$\frac{4051}{6323} = 64.1\%$	$\frac{4436}{6323} = 70.1\%$	$\mathbf{\frac{5159}{6323} = 81.6\%}$

As is well-known, in statistical prediction, the single independent dataset test, subsampling test, and jackknife test are the three methods often used for cross-validation. Of these three, the jackknife test is deemed as the most rigorous and objective one.[10] Therefore, the jackknife test has been used by more and more investigators[11−26] to examine the power of various prediction methods. For the aforementioned three classifiers, the success rates by the jackknife cross-validation for the 6323 protein pairs in the Supporting Information A are given in Table 2, from which we can see that the overall success rate by the GO classifier alone was 64.1% and that by the PseAA classifier alone was 70.2%, but that by the GO-PseAA classifier was 81.6%, indicating that the approach of fusing GO and PseAA classifiers is quite encouraging even for dealing with such a stringent dataset and complicated problem. Also, we observe from Table 1 that many failure predictions by the GO classifier are caused by the naught vector problem, implying that, with

the development of GO, more and more protein pairs can be covered by the GO representation and the overall success rate will be further enhanced.

5. Conclusion

The knowledge of protein networks is indispensable for understanding the molecular underpinnings of life. To expedite the study of protein networks, it is important to develop automated methods for predicting protein−protein interactions according to their sequences.

Introduction of the gene ontology approach (GO), hybridized with the pseudo-amino acid composition (PseAA) approach,[2] is a promising approach for predicting protein−protein interactions. This is fully consistent with the scientific logic because the current hybrid approach has combined the gene product and quasi-sequence-order effects. The gene product is closely correlated with the biological process, molecular function, and cellular components. Therefore, the approach via GO can grasp the core feature of protein−protein interaction in cellular networks. And the PseAA approach can play a complementary role by incorporating a considerable amount of sequence-order effects for those protein pairs that could not be covered by GO.

Appendix A

For reader's convenience, we give here a brief description of computing the pseudo-amino acid components for a given protein sequence. For details, the readers are referred to ref 2, where the original concept of PseAA was introduced.

Owing to the huge number of possible sequence-order patterns, it is hard to directly incorporate the sequence-order information into a classifier of discrete model. Nevertheless, we can indirectly and partially incorporate its effects through the steps given below. For a protein chain with L amino acid residues as expressed by eq 1, we can derive λ discrete numbers according to the following equations:

$$\begin{cases} \tau_1 = \dfrac{1}{L-1}\sum_{i=1}^{L-1} J_{i,i+1} \\ \tau_2 = \dfrac{1}{L-2}\sum_{i=1}^{L-2} J_{i,i+2} \\ \tau_3 = \dfrac{1}{L-3}\sum_{i=1}^{L-3} J_{i,i+3} \\ \cdots\cdots\cdots\cdots \\ \tau_\lambda = \dfrac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} J_{i,i+\lambda} \end{cases}, \qquad (\lambda < L) \qquad \text{(A1)}$$

where τ_1 is called the 1st-tier correlation factor that reflects the sequence-order correlation between all the most contiguous residues along a protein chain (Figure 2a), τ_2 the second-tier correlation factor that reflects the sequence-order correlation between all the 2nd most contiguous residues (Figure 2b), τ_3 the third-tier correlation factor that reflects the sequence-order correlation between all the third most contiguous residues (Figure 2c), and so forth. In eq A1, the coupling factor $J_{i,j}$ is defined by

$$J_{i,j} = \frac{1}{3}\{[h_1(\mathrm{R}_i) - h_1(\mathrm{R}_j)]^2 + [h_2(\mathrm{R}_i) - h_2(\mathrm{R}_j)]^2 + [M(\mathrm{R}_i) - M(\mathrm{R}_j)]^2\} \quad \text{(A2)}$$

where $h_1(\mathrm{R}_i)$, $h_2(\mathrm{R}_i)$, and $M(\mathrm{R}_i)$ are associated with the hydrophobicity value, hydrophilicity value, and side-chain mass of the amino acid R_i, respectively; and $h_1(\mathrm{R}_j)$, $h_2(\mathrm{R}_j)$, and $M(\mathrm{R}_j)$ are the corresponding values for the amino acid R_j. They are generated from the original values of hydrophobicity, hydrophilicity, and side-chain mass through a *standard conversion* as given below:

$$\begin{cases} h_1(\mathrm{R}_i) = \dfrac{h_1^0(\mathrm{R}_i) - < h_1^0 >}{\mathrm{SD}(h_1^0)} \\ h_2(\mathrm{R}_i) = \dfrac{h_2^0(\mathrm{R}_i) - < h_2^0 >}{\mathrm{SD}(h_2^0)} \\ M(\mathrm{R}_i) = \dfrac{M^0(\mathrm{R}_i) - < M^0 >}{\mathrm{SD}(M^0)} \end{cases} \qquad \text{(A3)}$$

where the symbols $h_1^0(\mathrm{R}_i)$, $h_2^0(\mathrm{R}_i)$, and $M^0(\mathrm{R}_i)$ represent the original hydrophobicity value,[27] hydrophilicity value,[28] and the side-chain mass for amino acid R_i, respectively (Table A1), $\langle h_1^0 \rangle$ is the mean of the original hydrophobicity values over 20 native amino acids, and $\mathrm{SD}(h_1^0)$ is the corresponding standard deviation and so forth. The data obtained through such a standard conversion (eq A3) will have a 0 mean value and will remain unchanged if going through the same conversion procedure again. As we can see from eqs A1−A3 as well as Figure 2, a considerable amount of sequence-order information has been incorporated into the λ correlation factors through the hydro-

Table A1. The Amino Acid Parameters Used for Deriving the Pseudo-Amino Acid Components (cf. eqs A2−A3)

code	hydrophobicity[a] (h_1^0)	hydrophilicity[b] (h_2^0)	side-chain mass[c] (M)
A	0.62	−0.5	15
C	0.29	−1.0	47
D	−0.90	3.0	59
E	−0.74	3.0	73
F	1.19	−2.5	91
G	0.48	0.0	1
H	−0.40	−0.5	82
I	1.38	−1.8	57
K	−1.50	3.0	73
L	1.06	−1.8	57
M	0.64	−1.3	75
N	−0.78	2.0	58
P	0.12	0.0	42
Q	−0.85	0.2	72
R	−2.53	3.0	101
S	−0.18	0.3	31
T	−0.05	−0.4	45
V	1.08	−1.5	43
W	0.81	−3.4	130
Y	0.26	−2.3	107

[a] The hydrophobicity values were taken from ref 27. [b] The hydrophilicity values were taken from ref 28. [c] In the unit of Da (dalton), the amino acid side-chain mass can be derived from any biochemistry text book.

phobic and hydrophilic values as well as the side-chain masses of the amino acid residues along a protein chain.

Suppose f_i ($i = 1, 2, ..., 20$) are the occurrence frequencies of 20 native amino acids in a protein. Fusing these frequencies with the correlation factors τ_k ($k = 1, 2, ..., \lambda$) of eq A1, we obtain an augmented discrete form; that is,

$$\begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{20} \end{bmatrix} \oplus \begin{bmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_\lambda \end{bmatrix} = \begin{bmatrix} f_1 \\ \vdots \\ f_{20} \\ \tau_1 \\ \vdots \\ \tau_\lambda \end{bmatrix} \quad \text{(A4)}$$

Thus, the PseAA components $p_1, p_2, ..., p_{20+\lambda}$ in eq 3 can be uniquely derived by normalizing the $20 + \lambda$ elements in eq A4 according to the following equations:

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \tau_j}, & (1 \le u \le 20) \\ \dfrac{w\tau_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \tau_j}, & (20+1 \le u \le 20+\lambda) \end{cases} \quad \text{(A5)}$$

where w is the weight factor. In the current study, we chose $w = 0.05$ to make the results of eq A5 within the range easier to be handled (w can be, of course, assigned with other values, but this would not have a significant impact on the final results).

As we can see from the above derivation, the first 20 components in eq 3 represent the classic amino acid composition, while the components from $20 + 1$ to $20 + \lambda$ are λ correlation factors along a protein chain reflecting the effect of sequence order. A set of such $20 + \lambda$ components is called the pseudo-amino acid composition or abbreviated as PseAA. We use such a name because it still has the main feature of amino acid composition, but on the other hand, it contains information beyond the conventional amino acid composition.

The pseudo-amino acid composition thus defined has the following three advantages. (1) It contains more sequence-order effects not only than the 20D conventional amino acid composition[4,29] but also than the 210D pair-coupled amino acid composition[30] and the 400D first-order-coupled amino acid composition,[31] as reflected by a series of sequence-coupling factors with different tiers of correlation (see Figure 2 and eq A1). (2) The coupling factors are defined by a combination of correlation functions that allows users to introduce any other biochemical quantities (in addition to the hydrophobicity, hydrophilicity, and side-chain mass as explicitly expressed in eq A2) to obtain the optimal results for various cases concerned. (3) The pseudo-amino acid composition has the same formulation as the conventional one, except that it contains more components (eq A5 or eq 3); accordingly, all the existing prediction algorithms based on the conventional amino acid composition can be straightforwardly extended to cover the pseudo-amino acid composition as well.

Supporting Information Available: List of accesion numbers of the 6323 pairs of yeast proteins. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) von Mering, C.; Jensen, L. J.; Snel, B.; Hooper, S. D.; Krupp, M.; Foglierini, M.; Jouffre, N.; Huynen, M. A.; Bork, P. STRING: known and predicted protein−protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **2005**, *33*, D433−437.

(2) Chou, K. C. Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins* (Erratum: Chou, K. C. *Proteins* **2001**, *44*, 60) **2001**, *43*, 246−255.

(3) 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.; Sherlock, G. Gene ontology: tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25−29.

(4) Chou, K. C. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. *Proteins* **1995**, *21*, 319−344.

(5) Chou, K. C.; Zhang, C. T. Predicting protein folding types by distance functions that make allowances for amino acid interactions. *J. Biol. Chem.* **1994**, *269*, 22014−22020.

(6) Chou, J. J.; Li, H.; Salvessen, G. S.; Yuan, J.; Wagner, G. Solution structure of BID, an intracellular amplifier of apoptotic signalling. *Cell* **1999**, *96*, 615−624.

(7) Oxenoid, K.; Chou, J. J. The structure of phospholamban pentamer reveals a channel-like architecture in membranes. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 10870−10875.

(8) 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.; Zdobnov, E. M. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **2001**, *29*, 37−40.

(9) Chou, K. C.; Cai, Y. D. Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.* **2002**, *277*, 45765−45769.

(10) Chou, K. C.; Zhang, C. T. Review: prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275−349.

(11) Zhou, G. P. An intriguing controversy over protein structural class prediction. *J. Protein Chem.* **1998**, *17*, 729−738.

(12) Yuan, Z. Prediction of protein subcellular locations using Markov chain models. *FEBS Lett.* **1999**, *451*, 23−26.

(13) Feng, Z. P. Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. *Biopolymers* **2001**, *58*, 491−499.

(14) Hua, S.; Sun, Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* **2001**, *17*, 721−728.

(15) Zhou, G. P.; Assa-Munt, N. Some insights into protein structural class prediction. *Proteins* **2001**, *44*, 57−59.

(16) Luo, R. Y.; Feng, Z. P.; Liu, J. K. Prediction of protein strctural class by amino acid and polypeptide composition. *Eur. J. Biochem.* **2002**, *269*, 4219−4225.

(17) Pan, Y. X.; Zhang, Z. Z.; Guo, Z. M.; Feng, G. Y.; Huang, Z. D.; He, L. Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. *J. Protein Chem.* **2003**, *22*, 395−402.

(18) Zhou, G. P.; Doctor, K. Subcellular location prediction of apoptosis proteins. *Proteins* **2003**, *50*, 44−48.

(19) Xiao, X.; Shao, S. H.; Ding, Y. S.; Huang, Z. D.; Chou, K. C. Using cellular automata images and pseudo amino acid composition to predict protein sub-cellular location. *Amino Acids*, published online July 28, 2005, DOI 10.1007/s00726-005-0225-6.

(20) Xiao, X.; Shao, S.; Ding, Y.; Huang, Z.; Huang, Y.; Chou, K. C. Using complexity measure factor to predict protein subcellular location. *Amino Acids* **2005**, *28*, 57−61.

(21) Wang, M.; Yang, J.; Xu, Z. J.; Chou, K. C. SLLE for predicting membrane protein types. *J. Theor. Biol.* **2005**, *232*, 7−15.

(22) Feng, K. Y.; Cai, Y. D.; Chou, K. C. Boosting classifier for predicting protein domain structural class. *Biochem. Biophys. Res. Commun.* **2005**, *334*, 213-217.

(23) Shen, H.; Chou, K. C. Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. *Biochem. Biophys. Res. Commun.* **2005**, *334*, 288−292.
(24) Shen, H. P.; Yang, J.; Liu, X. J.; Chou, K. C. Using supervised fuzzy clustering to predict protein structural classes. *Biochem. Biophys. Res. Commun.* **2005**, *334*, 577−581.
(25) Liu, H.; Wang, M.; Chou, K. C. Low-frequency Fourier spectrum for predicting membrane protein types. *Biochem. Biophys. Res. Commun.* **2005**, *336*, 737−739.
(26) Gao, Y.; Shao, S. H.; Xiao, X.; Ding, Y. S.; Huang, Y. S.; Huang, Z. D.; Chou, K. C. Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. *Amino Acids* **2005**, *28*, 373−376.
(27) Tanford, C. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *J. Am. Chem. Soc.* **1962**, *84*, 4240−4274.
(28) Hopp, T. P.; Woods, K. R. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 3824−3828.
(29) Chou, K. C.; Elrod, D. W. Protein subcellular location prediction. *Protein Eng.* **1999**, *12*, 107−118.
(30) Chou, K. C. Using pair-coupled amino acid composition to predict protein secondary structure content. *J. Protein Chem.* **1999**, *18*, 473−480.
(31) Liu, W.; Chou, K. C. Protein secondary structural content prediction. *Protein Eng.* **1999**, *12*, 1041−1050.

PR050331G