PROTEINS: Structure, Function, and Genetics 53:282–289 (2003)

Predicting Protein Quaternary Structure by Pseudo Amino Acid Composition

Kuo-Chen Chou[1,2*] **and Yu-Dong Cai**[3†]
[1]*Gordon Life Science Institute, Kalamazoo, Michigan*
[2]*Tianjin Institute of Bioinformatics and Drug Discovery (TIBDD), Tianjin, China*
[3]*Shanghai Research Center of Biotechnology, Chinese Academy of Sciences, Shanghai, China*

ABSTRACT **In the protein universe, many proteins are composed of two or more polypeptide chains, generally referred to as subunits, that associate through noncovalent interactions and, occasionally, disulfide bonds. With the number of protein sequences entering into data banks rapidly increasing, we are confronted with a challenge: how to develop an automated method to identify the quaternary attribute for a new polypeptide chain (i.e., whether it is formed just as a monomer, or as a dimer, trimer, or any other oligomer). This is important, because the functions of proteins are closely related to their quaternary attribute. For example, some critical ligands only bind to dimers but not to monomers; some marvelous allosteric transitions only occur in tetramers but not other oligomers; and some ion channels are formed by tetramers, whereas others are formed by pentamers. To explore this problem, we adopted the pseudo amino acid composition originally proposed for improving the prediction of protein subcellular location (Chou, Proteins, 2001; 43:246–255). The advantage of using the pseudo amino acid composition to represent a protein is that it has paved a way that can take into account a considerable amount of sequence-order effects to significantly improve prediction quality. Results obtained by resubstitution, jack-knife, and independent data set tests, have indicated that the current approach might be quite promising in dealing with such an extremely complicated and difficult problem.** **Proteins 2003;53:282–289.** **© 2003 Wiley-Liss, Inc.**

Key words: homo-oligomerization; quaternary attribute; quasi-sequence-order effects; bioinformatics; covariant-discriminant algorithm; proteomics

INTRODUCTION

Proteins are at the center of the action in biologic processes, and their function can only be understood based on the structure of their constituent polypeptide chains. The structural hierarchy in proteins is traditionally described in terms of four levels: primary, secondary, tertiary, and quaternary. Primary structure is defined by the amino acid sequence. Secondary structure is the local spatial arrangement of a polypeptide's backbone, without regard to the conformations of its sidechains. Tertiary structure refers to the three-dimensional (3D) structure of an entire polypeptide. Quaternary structure refers to the number of polypeptide chains (subunits) involved in forming a protein and the spatial arrangement of its subunits. The concept of quaternary structure is derived from the fact that many proteins are composed of two or more subunits that associate through noncovalent interactions and, in some cases, disulfide bonds. The fact that multisubunit proteins are so common can be elucidated by the following evolutional points of view.[1,2] (1) *Easier to repair.* For multisubunit proteins, defects can be repaired by simply replacing the flawed subunit. In this regard, the advantages of subunit construction over the synthesis of one huge polypeptide chain are analogous to the use of prefabricated components in constructing a building. The site of subunit manufacture can differ from the site of assembly into the final product, and the only genetic information necessary to specify the entire edifice is to identify its few different, self-assembling subunits. (2) *Indispensable in function.* The subunit construction of many enzymes provides the structural basis for the regulation of their activities, an indispensable function for many important biologic processes. Thus, in the protein universe, there are many different classes of subunit construction, such as monomer, dimer, trimer, tetramer, and so forth (Fig. 1). Note that oligomers may be homo-oligomers and hetero-oligomers; the former consist of identical polypeptide chains, whereas the latter are nonidentical. For example, the sodium channel is formed by a monomer,[3] whereas the potassium channel is formed by a homotetramer[4]; the acetylcholine-binding protein (AChBP) is formed by homopentamer[5], whereas the gamma-aminobutyric acid type A ($GABA_A$) receptor is formed by a heteropentamer.[6] The present study is limited to homo-oligomers.

In the vast majority of oligomeric proteins, the subunits are symmetrically arranged. However, proteins cannot have inverse or mirror symmetry, because such symmetric

†Current address: Biomolecular Sciences Department, University of Manchester Institute of Science and Technology, Manchester M60 1QD, UK.

*Correspondence to: Kuo-Chen Chou, Gordon Life Science Institute, Kalamazoo, MI 49009. E-mail: lifescience@chartermi.net

Received 2 January 2003; Accepted 1 May 2003

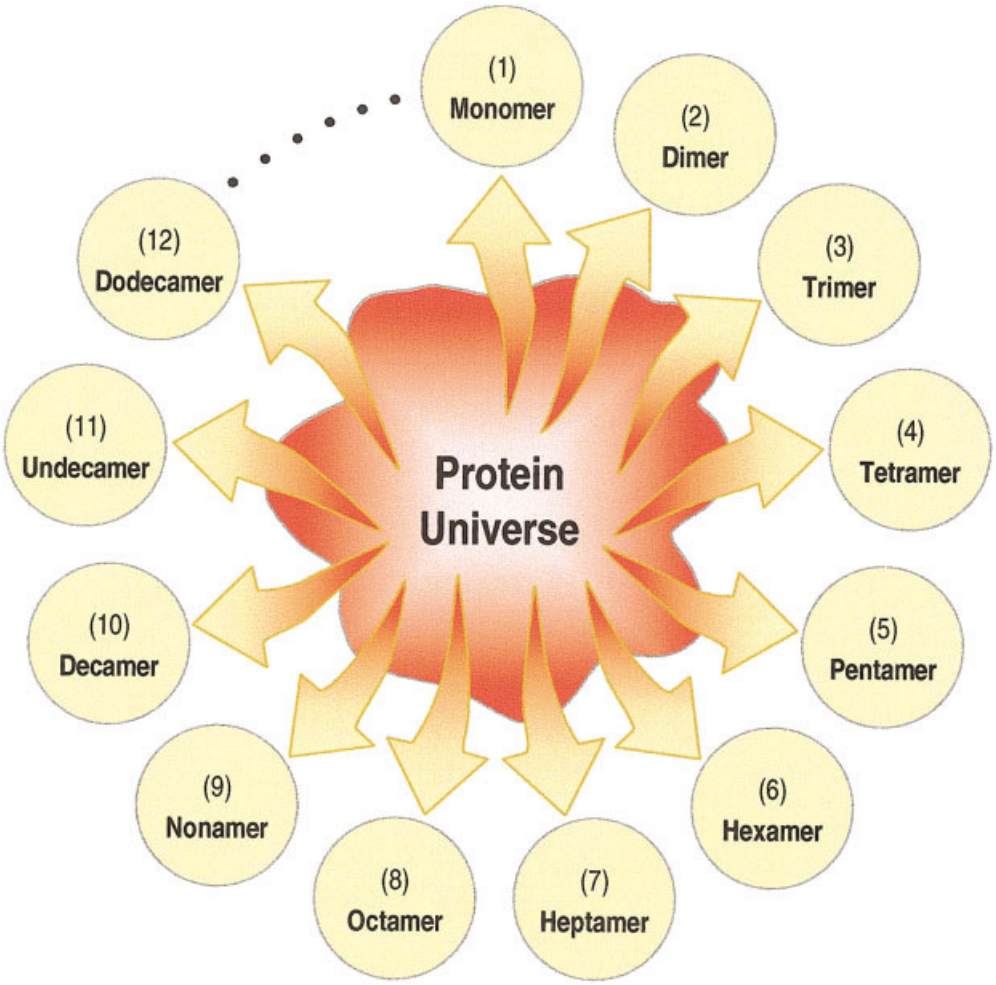

Fig. 1. A schematic drawing to illustrate that different polypeptide chains may form various oligomers.

operations would convert chiral L-residues to D-residues.[7] Accordingly, proteins can only have rotational symmetry. The following types of rotational symmetry occur in proteins: (1) Cyclic symmetry, C_n, where subunits are related (brought to coincidence) by a single axis of rotation of (360°/n) ($n = 2, 3, 4, \ldots$). For example, the object in Figure 2(a) has C_5 symmetry, with a 5-fold rotational axis. C_2 symmetry is most common in proteins; higher cyclic symmetries are relatively rare. (2) Dihedral symmetry, D_n, in which the symmetry is generated when an n-fold and a 2-fold rotation axis intersect at right angles. For example, the object in Figure 2(b) has the D_4 symmetry. (3) Other rotational symmetries, such as those that have the rotational symmetries of a tetrahedron [Figure 2(c)], a cube [Figure 2(d)], and an icosahedron [Figure 2(c)], with 12, 24, and 60 equivalent positions, respectively. For example, the spherical viruses have the subunit construction of icosahedron symmetry.[8,9]

Given a polypeptide chain, will it form a dimer, trimer, or any other oligomer, or exist only as a monomer? To the best of our knowledge, no report in literature has systematically addressed such a problem. It has been known for long that the functions of proteins are closely related to their quaternary structure, but so far, there is no automated, high-throughput tool to predict it. With the number of protein sequences entering into data banks rapidly increasing, the challenge to develop such an automated method has become even more critical and urgent. In view of this, our study was initiated in an attempt to explore the problem.

METHODS AND MATERIALS

Because any higher level structure is solely determined by its primary structure, in principle, the quaternary structure of a protein may also be predicted, based on its amino acid sequence. Unfortunately, if using the entire sequence of a protein as a sample to formulate a statistical prediction algorithm, one would face the difficulty of dealing with almost an infinite number of patterns, because of the extreme variation in both sequence order and length, as elaborated by Chou.[10] Accordingly, to formulate a feasible statistical prediction algorithm, one must express a protein in terms of a set of discrete numbers. The earlier approach was to represent proteins according their amino acid composition, which substantially reduced the number of samples and made the statistical treatment

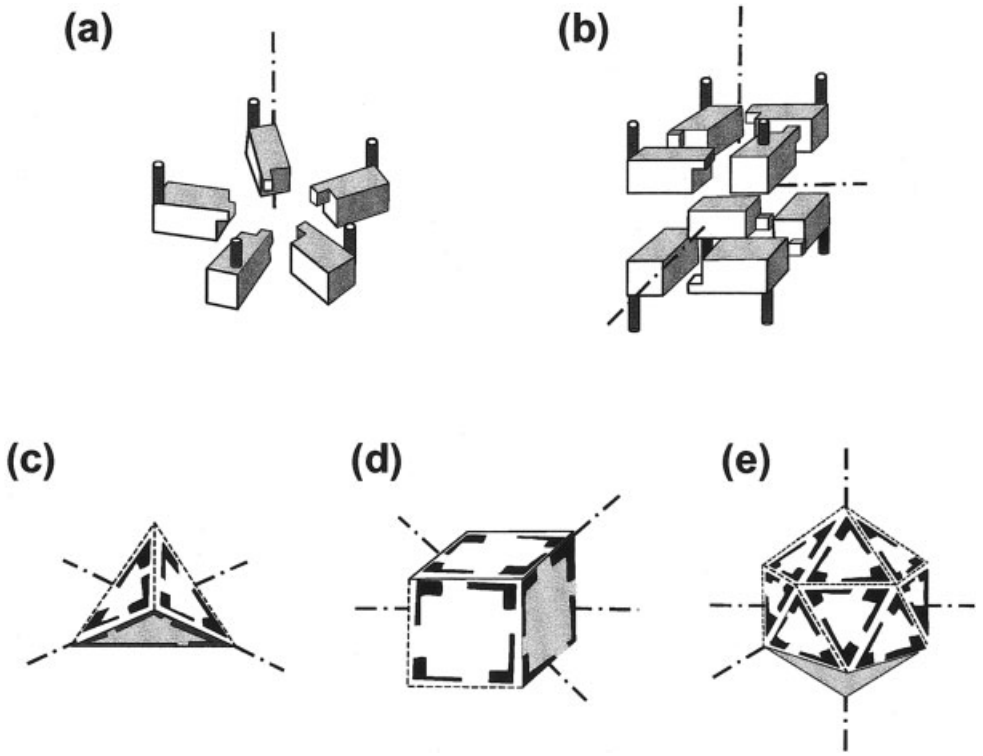

Fig. 2. A schematic drawing to illustrate protein oligomers with (**a**) C_5 symmetry, (**b**) D_4 symmetry, (**c**) tetrahedron symmetry, (**d**) cubic symmetry, and (**e**) icosahedron symmetry.

possible. Such an approach was widely used to predict protein structural class,[11–16] protein secondary structure content,[17–20] protein subcellular location,[21–25] and G protein–coupled receptor (GPCR) types.[26] According to the classical definition, amino acid composition consists of 20 components, representing the occurrence frequency of each of the 20 native amino acids in a given protein. Obviously, if, one uses classical amino acid composition to represent a protein, then all the sequence-order and -length effects are missed, and the prediction method based on such a representation must bear a considerable intrinsic limitation. Therefore, we are actually confronted with such a dilemma: if we wish to include the complete information of an entire protein chain, then the prediction becomes impracticable; if we wish to make the prediction feasible, then some of its information must be ignored. In view of this, can we find a compromise scenario in which a protein is still represented by a set of discrete numbers that, however, are also able to contain as much of the sequence-order effects as possible? The introduction of the pseudo amino acid composition,[10] a pioneer effort in this regard, has remarkably improved prediction of protein subcellular location. Unlike the classical amino acid composition that consists of only 20 components, the pseudo amino acid composition consists of 20 + λ discrete numbers, in which the first 20 numbers are the same as the 20 components in the classical amino acid composition, and the remainders represent λ sequence-order correlation factors with different ranks, as defined by the following equation[10]:

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

where L is the length of a given protein chain (Figure 3), τ_1 is the first-rank coupling factor that encodes the sequence-order correlation between all the most contiguous residues along a protein chain [Figure 3(a)], τ_2 is the second-rank coupling factor that encodes the sequence-order correlation between all the second most contiguous residues [Figure 3(b)], τ_3 is the third-rank coupling factor that encodes the sequence-order correlation between all the third most contiguous residues [Figure 3(c)], and so forth. In eq. (1), the coupling factor $J_{i,j}$ is given by

$$J_{ij} = \frac{1}{\Lambda}\sum_{g=1}^{\Lambda} [\Phi_g(\mathrm{R}_j) - \Phi_g(\mathrm{R}_i)]^2, \tag{2}$$

where $\Phi_g(\mathrm{R})$ is the gth function of the amino acids R, and Λ, the total number of the functions considered. In the

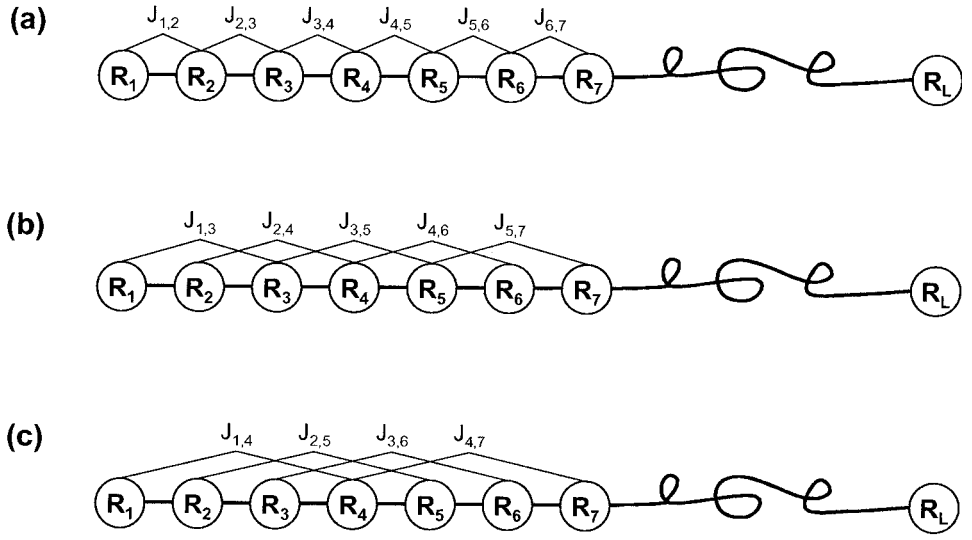

Fig. 3. A schematic drawing to illustrate (**a**) first tier, (**b**) second tier, and (**c**) third tier sequence-order correlation mode along a protein sequence. Panel (**a**) reflects the correlation mode between all the most contiguous residues; panel (**b**), that between all the second most contiguous residues, and panel (**c**), that between all the third most contiguous residues.

current study, three different functions ($\Lambda = 3$) are used to reflect the characters of an amino acid: $\Phi_1(\mathrm{R})$ refers to the hydrophobicity of amino acid R, $\Phi_2(\mathrm{R})$, its hydrophilicity, and $\Phi_3(\mathrm{R})$, its sidechain mass. Note that before substituting these functions into Eq. (2) we subjected them all to a standard conversion, as defined by the following operation:

$$\Phi_g(R_i) = \frac{\Phi_g^0(\mathrm{R}_j) - \sum_{j=1}^{20} \frac{\Phi_g^0(\mathrm{R}_j)}{20}}{\sqrt{\frac{\sum_{j=1}^{20}\left[\Phi_g^0(\mathrm{R}_j) - \sum_{j=1}^{20} \frac{\Phi_g^0(\mathrm{R}_j)}{20}\right]^2}{20}}} \qquad (g = 1, 2, 3), \tag{3}$$

where $\Phi_1^0(\mathrm{R}_i)$ is the original hydrophobicity value of the amino acid R_i, taken from Tanford,[27] $\Phi_2^0(\mathrm{R}_i)$ is the corresponding hydrophilicity value, taken from Hopp and Woods,[28] and $\Phi_3^0(\mathrm{R}_i)$ is the sidechain mass of R_i which can be obtained from any biochemistry textbook. Without loss of generality, we use the numerical indices 1, 2, 3, . . . ,20 to represent the 20 native amino acids, according to the alphabetical order of their single-letter codes: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y. The advantage of taking such a standard conversion [Eq. (3)] is that the data thus obtained have a zero mean value and will remain unchanged throughout the same conversion procedure again.

As seen in Figure 3, the sequence-order effect of a protein can to some extent be reflected through a set of discrete numbers, $\tau_1, \tau_2, \tau_3, \ldots, \tau_\lambda$, as defined by Eq. (2). Now, let us augment the formulation of amino acid composition to include such a set of sequence-coupling factors. To realize this, instead of using a 20D (dimensional) vector defined by 20 amino acid components,[11,14] let us use a $(20 + \lambda)$-D vector defined by $20 + \lambda$ components to represent a protein **X**; that is,

$$\mathbf{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_{20} \\ x_{20+1} \\ \vdots \\ x_{20+\lambda} \end{bmatrix}, \tag{4}$$

where

$$x_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \tau_j}, & (1 \le u \le 20) \\ \dfrac{w\tau_{u-20}}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \tau_j}, & (20 + 1 \le u \le 20 + \lambda) \end{cases} \tag{5}$$

where f_i is the normalized occurrence frequency of the 20 amino acids in the protein **X**, τ_j is the j-rank sequence-coupling factor computed according to Eqs. (1) and (3) for the protein **X**, and w is the weight factor for the sequence-order effect. Here, we chose $w = 0.05$. As we can see in Eqs. (4) and (5), the first 20 components reflect the effect of the amino acid composition, whereas the components from $20 + 1$ to $20 + \lambda$ reflect the effect of sequence order. A set of such $20 + \lambda$ components as formulated by Eqs. (4) and (5) is called the pseudo amino acid composition for protein **X**. Use of such a name is because it still keeps the main feature of amino acid composition; on the other hand, however, it contains the information beyond the classic amino acid composition. The pseudo amino acid composition thus defined has the following three advantages. (1)

Compared with the 400-D, first-order coupled amino acid composition[20] that contains the sequence-order effect only within a very short range of two adjacent amino acid residues, the pseudo amino acid composition incorporates much more sequence-order effects, namely those not only for the short range but for the medium and long range as well, as reflected by a series of sequence-coupling factors with different ranks of correlation (see Fig. 3 and Eq. (1). (2) The coupling factors are defined through a combination of several correlation functions [Eq. (2)] that make allowances for users to introduce any other biochemical quantities (in addition to the hydrophobicity, hydrophilicity, and sidechain mass, as adopted here) to obtain the optimal results for various cases. For example, the physicochemical distance[29] from amino acid R_i to amino acid R_j can also be used to define the coupling factor $J_{i,j}$.

As we can see in Eq. (4), the pseudo amino acid composition has the same formulation as the classical one, except that it contains more components. Therefore, all the existing prediction algorithms based on classical amino acid composition, such as the least Hamming distance algorithm,[12] the ProtLock algorithm[22] and the covariant discriminant algorithm,[24] can be applied to the pseudo amino acid composition by a straightforward augmentation procedure, as illustrated by Chou[10]; hence there is no need to repeat them here. It is instructive, however, to point out that because of the normalization condition imposed by Eq (5), the $20 + \lambda$ components of the pseudo amino acid composition are not independent. Therefore, a dimension-reduced operation achieved by leaving out one of the components and making the rest completely independent is needed when using the augmented covariant discriminant algorithm. In other words, a protein should be defined in a $(20 + \lambda - 1)$-D space instead of $(20 + \lambda)$-D space. Otherwise, a divergence difficulty occurs. However, which one of the $20 + \lambda$ components should be removed? Anyone of them. The reason is that according to the invariance Theorem given in Appendix A of Chou,[14] the values of the covariant discriminant function remain the same regardless of which $20 + \lambda$ components is left out. The theorem can also be used to address similar problems in other algorithms that involve a covariance matrix.[15,22,25]

To use an algorithm for statistical prediction, one has to first construct a training data set. We did this by using the SWISS-PROT databank,[30] released 8 March 2002. The data set construction was governed by the following criterias: (1) Clearness— collected were only those protein sequences with a clear annotation of their quaternary attribute. (2) Nonredundancy—for those proteins with high sequence similarity to each other, to avoid redundancy, only one of them was kept. (3) Statistical significance—those subsets were dropped from further consideration if they contained too few entries to be of statistical significance.

The training data set thus obtained consists of 3174 protein sequences, of which 382 are with annotation of monomer, 817 of dimer, 593 of trimer, 884 of tetramer, 54 of pentamer, 287 of hexamer, and 157 of octamer. They each contain more than 50 sequences. Moreover, according to the same criteria, an independent testing data set was also constructed that contains 332 protein sequences, of which 50 are with annotation of monomer, 102 of dimer, 56 of trimer, 80 of tetramer, 6 of pentamer, 28 of hexamer, and 10 of octamer. None of these protein sequences occur in the training data set; hence, they form a testing data set independent to the training data set. The training and testing data sets are given in the Online Supplemental Materials A and B, respectively, where the accession number rather than the SWISS-PROT name is used, because the accession number more stablly represents a unique protein sequence.

RESULTS AND DISCUSSION

To show the predicted results by means of the pseudo amino acid, we conducted the demonstration using the three most typical approaches in statistical prediction[31]: the resubstitution test, the jack-knife test, and the independent data set test. Also, as we see in Eq. (4) and Figure 3, the first 20 components in the $(20 + \lambda)$-D space represent the contribution from the amino acid composition, and the last λ components (from $20 + 1$ to $20 + \lambda$) represent the contribution from the sequence-order effect. The greater the number λ, the more the sequence-order effect is taken into account. To show how it affects the success rate, predictions were performed with different λ for each of the three test approaches, as reported below.

Resubstitution Test

The so-called resubstitution test is an examination for the self-consistency of a prediction method. When it was performed for this study, the quaternary attribute of each protein sequence in a data set was in turn identified by the rule parameters derived from the same data set, the so-called training data set. The success rates thus obtained for the 3174 protein sequences in the training data set (Online Supplemental Materials A) with different λ are summarized in Table I, where we can see, as expected, that the overall success rate is enhanced by increasing λ. When $\lambda = 0$ (i.e., no sequence-order effect is taken into account at all) the overall success rate is only 72.9%. When $\lambda = 25$ (i.e., the sequence-order effect is counted through 25 correlation factors), the rate is increased to 86.7%. When $\lambda = 50$ (i.e., the sequence-order effect is encoded into 50 correlation factors), the rate is further increased to 90.5%. Note that there is an upper limit for λ; that is, it must be smaller than the length of the shortest protein chain in the data set [see Fig. 1 and Eq. (1)]. For example, in this training data set, the protein sequence P28270 is only 51 amino acid residues long; hence, the maximum allowed for λ is 50, or the maximum dimension allowed here for the pseudo amino acid space is $(20 + 50)$-D = 70-D. Also, we should stress that, during the process of the resubstitution test, the rule parameters derived from the training data set include the query protein sequence information later plugged back in the test. This certainly underestimates the error and enhances the success rate, because the same proteins are used to derive the rule parameters and to test themselves. Accordingly, the success rate thus obtained

TABLE I. Success Rates in Predicting Protein Quaternary Structures by Resubstitution, Jack-Knife, and Independent Data Set Tests

	Quaternary attribute							
	Monomer	Dimer	Trimer	Tetramer	Pentamer	Hexamer	Octamer	Overall
λ[a]	Success prediction rate by resubstitution test[b]							
0	281/382 = 73.6%	616/817 = 75.4%	363/593 = 61.2%	697/884 = 78.9%	51/54 = 94.4%	179/287 = 62.4%	127/157 = 80.9%	2314/3174 = 72.9%
25	325/382 = 85.1%	706/817 = 86.4%	474/593 = 79.9%	795/884 = 89.9%	54/54 = 100.0%	244/287 = 85.0%	153/157 = 97.5%	2751/3174 = 86.7%
50	335/382 = 87.7%	735/817 = 90.0%	543/593 = 91.6%	838/884 = 94.8%	5/54 = 9.3%	259/287 = 90.2%	156/157 = 99.4%	2871/3174 = **90.5%**
	Success prediction rate by jack-knife test[b]							
0	268/382 = 70.2%	589/817 = 72.1%	345/593 = 58.2%	666/884 = 75.3%	38/54 = 70.4%	154/287 = 53.7%	111/157 = 70.7%	2171/3174 = 68.4%
25	305/382 = 79.8%	676/817 = 82.7%	430/593 = 72.5%	746/884 = 84.4%	10/54 = 18.5%	177/287 = 61.7%	104/157 = 66.2%	2448/3174 = 77.1%
50	309/382 = 80.9%	700/817 = 85.7%	462/593 = 77.9%	755/884 = 85.4%	1/54 = 1.9%	180/287 = 62.7%	85/157 = 54.1%	2492/3174 = **78.5%**
	Success prediction rate by independent data set test[c]							
0	28/50 = 56.0%	72/102 = 70.6%	32/56 = 57.1%	60/80 = 75.0%	3/6 = 50.0%	14/28 = 50.0%	9/10 = 90.0%	218/332 = 65.7%
25	32/50 = 64.0%	78/102 = 76.5%	39/56 = 69.6%	68/80 = 85.0%	1/6 = 16.7%	19/28 = 67.9%	10/10 = 100.0%	247/332 = 74.4%
50	35/50 = 70.0%	85/102 = 83.3%	44/56 = 78.6%	72/80 = 90.0%	1/6 = 16.7%	21/28 = 75.0%	8/10 = 80.0%	266/332 = **80.1%**

[a]λ is the total number of different sequence-order correlation factors considered [see Eq. (1) and Fig. 3].
[b]Conducted for the 3174 protein sequences in the training data set given in Online Supplemental Materials A.
[c]Conducted for the 332 independent proteins given in Online Supplemental Materials B, based on the rule parameters derived from the 3174 protein sequences in the training data set.

represents some sort of optimistic estimation.[14,32] Nevertheless, the resubstitution test is absolutely necessary, because it reflects the self-consistency of a prediction method, especially for its algorithm part. A prediction algorithm certainly cannot be deemed a good one if its self-consistency is poor. In other words, the resubstitution test is necessary but not sufficient for evaluating a prediction method. As a complement, a cross-validation test for an independent testing data set is needed, because it can reflect the effectiveness of a prediction method in practical application. This is especially important for checking the validity of a training data set: whether it contains sufficient information to reflect all the important features to yield a high success rate in practical application.

Jack-knife Test

As is well known, the independent data set test, subsampling test, and jack-knife test are the three methods often used for cross-validation in statistical prediction. Among these three, however, the jack-knife test is deemed most effective and objective one; see, for example, a relevant review[31] for a comprehensive discussion about this, and a monograph[33] for the mathematical principle. During jack-knifing, each protein in the data set is in turn singled out as a tested protein, and all the rule-parameters are calculated based on the remaining proteins. In other words, the quaternary attribute of each protein is identified by the rule parameters derived with the use of all the other proteins except the one that is being identified. During the process of jack-knifing, both the training and testing data sets are actually open, and a protein will in turn move from one to the other. The results of a jack-knife test thus obtained for the 3174 proteins in the training data set are given in Table I. As expected, the success rates of jack-knife test are decreased compared with those of the resubstitution test. The decrement is more remarkable for small subsets, such as the pentamer and octamer, because the cluster-tolerant capacity[34] for small subsets is usually low. Hence, the information loss resulting from jack-knifing will have a greater impact on the small subsets than on the large ones. However, compared with the case of $\lambda = 0$, the overall success rates are significantly enhanced by incorporating sequence-order effect, such as in the case of $\lambda = 25$ or 50. Also, because of the information loss during jack-knifing, the overall success rate is not always monotonously increased with λ. Actually, different data sets may have different optimal values for λ to yield the highest overall jack-knife success rate, as discussed in protein subcellular location prediction.[10] It is expected that the overall jack-knife success rate can be further enhanced by improving the cluster-tolerant capacity of small subsets, by adding into them more new proteins that belong to these subsets.

Independent Data Set Test

Moreover, as a demonstration for practical application, predictions were also performed for the 332 independent

proteins (Online Supplemental Materials B), based on the rule-parameters derived from the training data set. The predicted results thus obtained are also summarized in Table I, in which we can see that when $\lambda = 50$, the overall success prediction rate is 80.1%.

Although some well-known algorithms developed by other investigators, such as the least Hamming distance algorithm[12] and the least Euclidean distance algorithm[11] for predicting protein structural class, and the ProtLock[22] for predicting protein cellular location, have never been used to predict protein quaternary attributes, the structural class and cellular location can generally be considered two components of the protein attribute. Therefore, it is instructive to compare those prediction algorithms with the current one, because all were developed for predicting the attribute of a protein based on its sequence. The results are as follows: If conducted on the same training and testing data sets, as provided in the Online Supplemental Materials A and B, the overall success rates obtained by the least Hamming distance algorithm for resubstitution, jack-knife, and independent data set tests are 41.3%, 40.6%, and 31.3%, respectively. Those by the least Euclidean distance algorithm are 43.3%, 42.8%, and 36.5%, respectively. And those by the ProtLock algorithm are 45.4%, 44.1%, and 39.2%, respectively. Accordingly, the augmented covariant discriminant algorithm incorporated with pseudo amino acid composition can yield success rates that are almost 40–50% higher than those obtained by the other prediction algorithms.

It is instructive to conduct a sequence-identity analysis for the proteins studied here. The sequence identity between two protein sequences is defined as follows: Suppose that the maximum number of residues matched by sliding one sequence along the other is *M,* and the alignment length is *L;* the sequence identity between the two sequences is defined as *M*/*L.* The treatment for gaps is according to CLUSTALW.[35] The average sequence identities thus obtained by the sequence match operation for the subsets of monomers, dimers, trimers, tetramers, pentamers, hexamers, and octamers were 0.092, 0.069, 0.083, 0.081, 0.132, 0.085, and 0.146, respectively. Accordingly, most sequences in a same subset have very low sequence identity, a clear indication of exclusion of redundant and homologous sequences. This is fully consistent with the above results. Otherwise, the success rates predicted by the methods based on classic amino acid composition alone, such as the least Hamming distance algorithm,[12] the least Euclidean distance algorithm,[11] and the ProtLock,[22] would not be that low. Therefore, introduction of the pseudo amino acid composition is a compelling approach to deal with nonhomologous sequences.

CONCLUSIONS

Let us imagine: If the protein samples are completely randomly distributed among the seven subsets considered, the rate of correct prediction by random assignment would generally be 1/7 = 14.3%; if the distribution is weighted according to the sizes of subsets, then the rate of correct identification by the weighted random assignment would be $(382/3174)^2 + (817/3174)^2 + (593/3174)^2 + (884/3174)^2 + (54/3174)^2 + (287/3174)^2 + (157/3174)^2 \cong 20.4\%$. Therefore, the rates of correct identification obtained by using the pseudo amino acid composition are much higher than the corresponding completely randomized and weighted randomized rate, suggesting that the individual primary sequences of an oligomeric protein do contain the information of its quaternary structure.

The average sequence identity obtained by a sequence-matching operation among all members in each of the subsets studied here is within the range of 0.06–0.15, suggesting that the high success rates reported here are due not to the trivial sequence similarity, but to the profound statistical essence that has been effectively grasped by the pseudo amino acid composition in discriminating the quaternary attributes of proteins according to their sequence characteristics.

We should pointed out that the current prediction of quaternary attribute is limited to the number of subunits, without reference to their special arrangement. For example, a prediction thus obtained may indicate that a given protein sequence will form a tetramer, but it cannot identify whether it has C_4 or D_4 symmetry. To further identify this, the training data set must be extended to cover the spatial arrangement as well. For the case of tetramers, the corresponding training data must be further classified into two subsets, one containing sequences for C_4 symmetry only, and the other for D_4 only. However, the existing protein data banks do not provide sufficient information for us to construct a training dataset with such a detailed distinction. It is expected that with the continuous increase of entries into data banks, a more complete training data set may be established to provide more distinguishable information. The current approach can be easily extended to cover the identification of the spatial arrangement as well.

ACKNOWLEDGMENTS

We are indebted to Raymond B. Moeller, Cynthia A. Brennan, Wendy Vanderheide, Katie Crawford, and Melissa Maneikis for their help in drawing the figures in this article.

REFERENCES

1. Klotz IM, Darnell DW, Langerman NR. Quaternary structure of proteins. In: Neurath H, Hill RL, editors. The proteins. 3rd edition, Vol. 1. New York: Academic Press; 1975. p 226–241.
2. Einstein E, Schachman HK. Determing the roles of subunits in protein function. In: Creighton TE, editor. Protein function: A practical approach. London: IRL Press; 1989. p 135–176.
3. Chen Z, Alcayaga C, Suarez-Isla BA, O'Rourke B, Tomaselli G, Marban E. A "minimal" sodium channel construct consisting of ligated S5-P-S6 segments forms a toxin-activatable ionophore. J Biol Chem 2002;277:24653–24658.
4. Doyle DA, Morais CJ, Pfuetzner RA, Kuo A, Gulbis JM, Cohen SL, Chait BT, MacKinnon R. The structure of the potassium channel: Molecular basis of K plus conduction and selectivity. Science 1998;280:69–77.
5. Brejc K, van Dijk WJ, Klaassen RV, Schuurmans M, van der Oost J, Smit AB, Sixma TK. Crystal structure of an ACh-binding protein reveals the ligand-binding domain of nicotinic receptors. Nature 2001;411:269–276.
6. Tretter V, Ehya N, Fuchs K, Sieghart W. Stoichiometry and

assembly of a recombinant GABAA receptor subtype. J Neurosci 1997;17:2728–2737.
7. Matthews BW, Bernhard SA. Structure and symmetry in oligomeric proteins. Annu Rev Biophys Bioeng 1973;6:257–317.
8. Caspar DLD, Klug A. Physical principles in the construction of regular viruses. Cold Spring Harb Sym 1962;27:1–24.
9. Harrison SC, Jack A. Structure of tomato bushy stunt virus: Three-dimensional X-ray diffraction analysis at 16 Å resolution. J Mol Biol 1975;97:173–191.
10. Chou KC. Prediction of protein cellular attributes using pseudo-amino-acid-composition. Proteins 2001;43:246–255. (Erratum: Proteins 2001;44:60)
11. Nakashima H, Nishikawa K, Ooi T. The folding type of a protein is relevant to the amino acid composition. J Biochem 1986;99:152–162.
12. Chou PY. Prediction of protein structural classes from amino acid composition. In: Fasman GD, editor. Prediction of protein structure and the principles of protein conformation. New York: Plenum Press; 1989. p 549–586.
13. Chou KC, Zhang CT. Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem 1994;269:22014–22020.
14. Chou KC. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins 1995;21:319–344.
15. Zhou GP. An intriguing controversy over protein structural class prediction. J Protein Chem 1998;17:729–738.
16. Bahar I, Atilgan AR, Jernigan RL, Erman B. Understanding the recognition of protein structural classes by amino acid composition. Proteins 1997;29:172–185.
17. Krigbaum WR, Knutton SP. Prediction of the amount of secondary structure in a globular protein from its amino acid composition. Proc Natl Acad Sci USA 1973;70:2809–2813.
18. Muskal SM, Kim S.-H. Predicting protein secondary structure content: A tandem neural network approach. J Mol Biol 1992;225: 713–727.
19. Zhang CT, Zhang Z, He Z. Prediction of the secondary structure content of globular proteins based on structural classes. J Protein Chem 1996;15:775–86.
20. Liu W, Chou KC. Protein secondary structural content prediction. Protein Eng 1999;12:1041–1050.
21. Nakashima H, Nishikawa K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J Mol Biol 1994;238:54–61.
22. Cedano J, Aloy P, Pérez-Pons JA, Querol E. Relation between amino acid composition and cellular location of proteins. J Mol Biol 1997;266:594–600.
23. Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 1998;26: 2230–2236.
24. Chou KC, Elrod DW. Protein subcellular location prediction. Protein Eng 1999;12:107–118.
25. Zhou GP, Doctor K. Subcellular location prediction of apoptosis proteins. Proteins 2003;50:44–48.
26. Chou KC, Elrod DW. Bioinformatical analysis of G-protein-coupled receptors. J Proteome Res 2002;1:429–433.
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 TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci USA 1981;78:3824–3828.
29. Schneider G, Wrede P. The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: De novo design of an idealized leader peptidase cleavage site. Biophys J 1994;66:335–344.
30. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Res 2000;25:31–36.
31. Chou KC, Zhang CT. Review: Prediction of protein structural classes. Crit Rev Biochem Mol Biol 1995;30:275–349.
32. Zhou GP, Assa-Munt N. Some insights into protein structural class prediction. Proteins 2001;44:57–59.
33. Mardia KV, Kent JT, Bibby JM. Multivariate analysis. London: Academic Press; 1979. p 322, 381.
34. Chou KC. A key driving force in determination of protein structural classes. Biochem Biophys Res Com 1999;264:216–224.
35. Thompson JD, Higgins DG, Gibson TJ. CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.