PROTEINS: Structure, Function, and Genetics 29:172–185 (1997)

# Understanding the Recognition of Protein Structural Classes by Amino Acid Composition

**Ivet Bahar,[1,2] Ali Rana Atilgan,[2] Robert L. Jernigan,[1]* and Burak Erman[2]**
[1]*Molecular Structure Section, Laboratory of Experimental and Computational Biology, Division of Basic Sciences, National Cancer Institute, National Institutes of Health, MSC 5677, Bethesda, Maryland*
[2]*Polymer Research Center, Bogazici University, and TUBITAK Advanced Polymeric Materials Research Center, Bebek 80815, Istanbul, Turkey*

***ABSTRACT*** **Knowledge of amino acid composition, alone, is verified here to be sufficient for recognizing the structural class, α, β, α+β, or α/β of a given protein with an accuracy of 81%. This is supported by results from exhaustive enumerations of all conformations for all sequences of simple, compact lattice models consisting of two types (hydrophobic and polar) of residues. Different compositions exhibit strong affinities for certain folds. Within the limits of validity of the lattice models, two factors appear to determine the choice of particular folds: 1) the coordination numbers of individual sites and 2) the size and geometry of non-bonded clusters. These two properties, collectively termed the *distribution of non-bonded contacts*, are quantitatively assessed by an eigenvalue analysis of the so-called *Kirchhoff* or *adjacency* matrices obtained by considering the non-bonded interactions on a lattice. The analysis permits the identification of conformations that possess the same distribution of non-bonded contacts. Furthermore, some distributions of non-bonded contacts are favored entropically, due to their high degeneracies. Thus, a competition between enthalpic and entropic effects is effective in determining the choice of a distribution for a given composition. Based on these findings, an analysis of non-bonded contacts in protein structures was made. The analysis shows that proteins belonging to the four distinct folding classes exhibit significant differences in their distributions of non-bonded contacts, which more directly explains the success in predicting structural class from amino acid composition. Proteins 29:172–185, 1997.** © 1997 Wiley-Liss, Inc.†

**Key words: non-bonded contacts; coordination of amino acids; Kirchhoff matrices; lattice models; singular value decomposition; secondary structure content prediction; contact patterns**



## INTRODUCTION

The existence of a correlation between amino acid composition and protein structural classes has been the object of a number of studies during the last decade,[1–7] after the original proposals of Nishikawa and Ooi[8] and of Nakashima et al.[9–11] and Chou.[12] Knowledge of the fractions of the 20 amino acids is now accepted to be sufficient, alone, for predicting the structural class of a given protein, α, β, α+β, or α/β.[12] The level of accuracy is, however, still variable, ranging from 60%[13] to near 100%[12] and clearly depends on the definitions of structural classes and the set of database structures considered for performing the analysis.

In the interest of gaining further insight into these observations, we performed a singular value decomposition (SVD) analysis of the amino acid compositions of the same sets of proteins as used by Chou.[12] Our use of the SVD technique for classifying proteins follows the recent analysis of Berry et al.[14] for classifying words with regard to their frequency of appearance in different texts. In their work, 16 words were classified according to their appearance in 17 different texts. The distances between different words were determined by SVD analysis, as well as the distance between different texts. Interestingly, two words may be near each other in the 16-dimensional space even if they never co-occur in the same text. Berry et al.[14] also discussed the technique of determining to which cluster of texts would a new word, not belonging to the original set of 16 words, be closest. If one replaces words with amino acids, and texts with proteins, the intelligent information retrieval method of Berry et al.[14] is exactly applicable to the classification of proteins on the basis of their amino acid compositions. The method simulta-

Contract grant sponsor: NATO Collaborative Research Grant Project; Contract grant number: CRG951240; Contract grant sponsor: Bogazici University Research Funds Project; Contract grant number: 96A0430.
*Correspondence to: Robert L. Jernigan, Molecular Structure Section, Laboratory of Experimental and Computational Biology, Division of Basic Sciences, National Cancer Institute, National Institutes of Health, MSC 5677, Room B-116, Bldg. 12B, Bethesda, MD 20892-5677.
Received 10 February 1997; Accepted 6 June 1997

neously groups proteins according to their amino acid compositions and amino acids according to their frequencies of occurrences in different classes. Application of this procedure in the present work indicates that compositions recognize the different structural classes with 81% accuracy. The procedure and results are briefly outlined in the next section. Our analysis is shown to be mathematically equivalent, but conceptually simpler, compared with the Mahalanobis distance approach of Chou.

In the remaining part of the study, we search for an answer to why composition should recognize structure. Our examination of α-, β-, α+β-, and α/β-proteins indicates that these classes do not exhibit significant differences in their degree of compactness, the average coordination number of residues being approximately constant in the different classes. Thus, the differences in the selection of certain classes by certain compositions are not correlated with the average density of non-bonded contacts. Instead, the distribution of non-bonded contacts emerges as an important factor distinguishing different classes. The distribution of non-bonded contacts is defined at two levels of approximation. On a coarse-grained level, the coordination numbers of individual residues characterize the distribution of non-bonded contacts. From a more detailed viewpoint, the size and geometry of clusters of non-bonded contacts, i.e., the spatial organization of groups of closely interacting residues, characterize the distribution of non-bonded contacts. The distribution of non-bonded contacts will be shown to be uniquely determined by the eigenvalues of the *Kirchhoff* or *adjacency* matrices [15,16] characteristic of a given tertiary structure.

To illustrate these issues, we analyze compact lattice models. We use simple two- and three-dimensional (3-D) compact lattice models of two types of residues, H (hydrophobic) and P (polar). Such simple models have proven useful in gaining insights into the general structural characteristics of proteins.[17] A recent study that motivated our approach is that of Li et al.[18] This study shows that only a few compact configurations are energetically selected by a large number of primary structures in a simple cubic $3 \times 3 \times 3$ lattice. These are named 'designable structures' because a large number of primary sequences fold into these structures. Exhaustive enumerations of all conformations for all possible sequences of H and P residues were made by Li et al.[18] to extract the lowest energy fold for each sequence and show that only a few structures were selected in general. In the $3 \times 3 \times 3$ lattice, there are 54 edges and 27 vertices on which residues may be placed. Thus, 28 edges are unoccupied by bonds and therefore form non-bonded H-H, H-P, or P-P contacts, irrespective of the overall 3-D fold. Because the total number of contacts is conserved, what could then be the factor that drives the recognition of a certain fold by a given sequence? Is it further possible to identify an energetically preferred fold by considering all amino acid sequences of fixed composition? Is it possible to assign a degeneracy number to particular folds, so that these will be favored from entropic effects? These issues will be addressed in Examination of Simple Lattice Models.

We will consider all possible conformations having a given compact shape, and for each conformation investigate the energetics of all permutations of H and P residues, i.e., generate and evaluate all sequences subject to a fixed 3-D fold and composition. Calculations will be repeated for different fractions of H and P residues. The aim is to identify the rules that govern the selection of certain structural classes by given amino acid compositions. We show that it is possible to cluster structures on the basis of their distribution of non-bonded contacts. This classification is uniquely determined by the eigenvalues of adjacency matrices. Sequences of a fixed composition cannot distinguish between distinct 3-D folds, unless these folds differ in their distribution of non-bonded contacts.

Finally, data bank structures[19,20] are revisited to explore the validity of the information inferred from the simple lattice simulations. Proteins belonging to different structural classes are verified to have significant differences in their distributions of non-bonded contacts, which may provide an explanation for the recognition of structural classes by amino acid composition.

The main body of the paper consists of three parts: In the first part, the determination of structural classes by the SVD method according to their amino acid compositions is presented. The question of why a structural class can be recognized by amino acid composition only, is addressed in the next section by using complete enumeration of all sequences or conformations of two-letter model chains on a lattice. Finally, in the following section, protein structures are revisited in the light of the indications given by lattice simulations, to verify that the distributions of non-bonded contacts do exhibit some net departures in the four structural classes. In the Conclusion, the major findings are summarized, with a discussion of additional factors that may affect the recognition of structural classes in proteins.

## Determination of Structural Classes by SVD Technique

Let us consider a set of n proteins. We represent each protein i by a 19-dimensional array of fluctuations in fractions of residues of different types

$$\Delta \mathbf{r}_i = \begin{bmatrix} \mathbf{r}_{1i} - \bar{\mathbf{r}}_1 \\ \mathbf{r}_{2i} - \bar{\mathbf{r}}_2 \\ \cdots \\ \mathbf{r}_{19,i} - \bar{\mathbf{r}}_{19} \end{bmatrix} \quad (1)$$

Here the jth element of $\Delta\mathbf{r}_i$ is the difference between the fraction $r_{ji}$ of the amino acid of type j in the protein i, and the average fraction of amino acid j in the ensemble of n structures. We note only 19 of the 20 residue fractions constitute an independent basis set, because the fractions sum up to unity. We refer to $\Delta\mathbf{r}_i$ as the protein vector expressed in the original 19-dimensional space of the amino acid compositions. Protein vectors of Equation (1) are organized in a matrix $\mathbf{A}_{[19\times n]}$ of size 19×n as

$$\mathbf{A}_{[19\times n]} = = [\Delta\mathbf{r}_1\ \Delta\mathbf{r}_2\ \Delta\mathbf{r}_3 \ldots . \Delta\mathbf{r}_n] \qquad (2)$$

the subscript in brackets denoting the size of the matrix.

First, the **A** matrix of Equation (2) is constructed for a training set of 120 non-homologous proteins that were selected in previous works.[3,5,21] This set comprises 30 proteins from each structural class. The definitions given in previous work[3,12] are adopted here for the four folding classes: α-proteins have more than 40% of their residues participating in α-helices and less than 5% in β-sheets. β-Proteins have ≤5% of residues in α-helices and ≥40% in β-sheets. α+β-proteins consist of two separate domains, α and β; more than 15% of their residues are in α-helices, and more than 15% in β-sheets with 60% antiparallel β-sheets. Finally, α/β-proteins have ≥15% of residues in α-helices and ≥15% in β-sheets, with >60% parallel β-sheets.

A structural class is represented by the variable ξ. First, the average composition vector or the so-called norm is calculated for each ξ. Then, the SVD method outlined in the Appendix is applied to each subset of n = 30 proteins to determine the singular space representations of the protein vectors. The distances $d_i(\xi)$ of a protein from the four type ξ = α, β, α+β, and α/β of structural classes are found from [see Eq. (A3)]

$$d_i^2(\xi) = \Delta\hat{\mathbf{r}}_i(\xi)\cdot\Delta\hat{\mathbf{r}}_i(\xi) = \Delta\mathbf{r}_i^T(\xi)\,\mathbf{S}^{-1}(\xi)\,\Delta\mathbf{r}_i(\xi) \qquad (3)$$

where **S**(ξ) is the covariance matrix corresponding to structural class ξ, as defined in by Equations (A4) and (A5) in the Appendix. The smallest of the four $d_i(\xi)$ values obtained for each protein determines the structural class of that protein. Application of Equation (3) to all proteins in the training set verified that this criterion was satisfied with an accuracy level of 98%, in parallel with Chou's calculations.[3,5]

A summary of the results is given in Table I. Figure 1 illustrates the clustering of proteins of different classes. Here, the projections of 30 α-proteins and 30 β-proteins onto the plane spanned by the dominant singular directions $\mathbf{u}_1$ and $\mathbf{u}_2$ are displayed. Interestingly, these two dominant singular directions, alone, provide a sufficiently accurate clustering of the two types of structural classes. Classes α+β and α/β are not shown here for clarity, their loci being almost evenly distributed over both regions of the plane.

**TABLE I. Performance of SVD Analysis of Protein Structural Classes**

| Set | Structural class | Success rates (%)* p = 19 | p = 18 | p = 10 |
|---|---|---|---|---|
| Training | α | 100 | 100 | 90.0 |
| Training | β | 100 | 100 | 86.7 |
| Training | α + β | 96.7 | 86.7 | 63.3 |
| Training | α/β | 93.3 | 86.7 | 36.7 |
| | AVG | 97.5 | 93.3 | 69.2 |
| Prediction | α | 66.7 | 44.4 | 55.5 |
| Prediction | β | 90.1 | 95.5 | 81.8 |
| Prediction | α + β | 81.0 | 50.0 | 40.9 |
| Prediction | α/β | 66.7 | 66.7 | 22.2 |
| | AVG | 81.0 | 67.7 | 54.8 |

*p is the number of distinct types of amino acids.

In the second stage of calculations, the predictive power of the method is tested for 62 unknown proteins. The covariance matrices already determined for the four classes of protein structures, **S**(α), **S**(β), **S**(α+β), and **S**(α/β) are directly used. Application of Equation (3) to the set of unknown proteins is shown to predict correctly the structural class of 81% of the proteins, on average. The prediction rates for the α, β, α+β, and α/β classes are 67, 91, 81, and 67%, respectively (Table I). The list of the distances $d_i(\xi)$ for all test proteins i with respect to the four classes ξ = α, β, α+β, and α/β is available as Supplementary Material.

We note that a higher accuracy level (greater than 90%) was reported by Chou and collaborators,[3–5] although the method used in their calculations is mathematically identical to ours (see Appendix). The average compositions for the four structural classes were different in the two studies, despite the use of the same set of proteins. From our communication with this group, the origin of this puzzling difference between our results and theirs was found to be the use of different sets of data files, intact Brookhaven PDB in our case, and a form modified for use in DSSP in theirs. Their files generally contained fewer residues compared with intact PDB files, which led to differences in accuracies. Otherwise, the two methods were equivalent, as explained in the Appendix, and identical success rates were obtained upon the application of either method to the same set of input files.

Correct clustering of these classes necessitates the consideration of all of the 19 dimensions of the singular space. For example, if only identification of hydrophobic and polar residues were sufficient to classify proteins into their structural classes, then the minimal basis for SVD would be 2. Our analysis showed that the minimal basis is 19, i.e., the full set of 20 residues must be considered to achieve the highest accuracy level. Neglect of one or more types

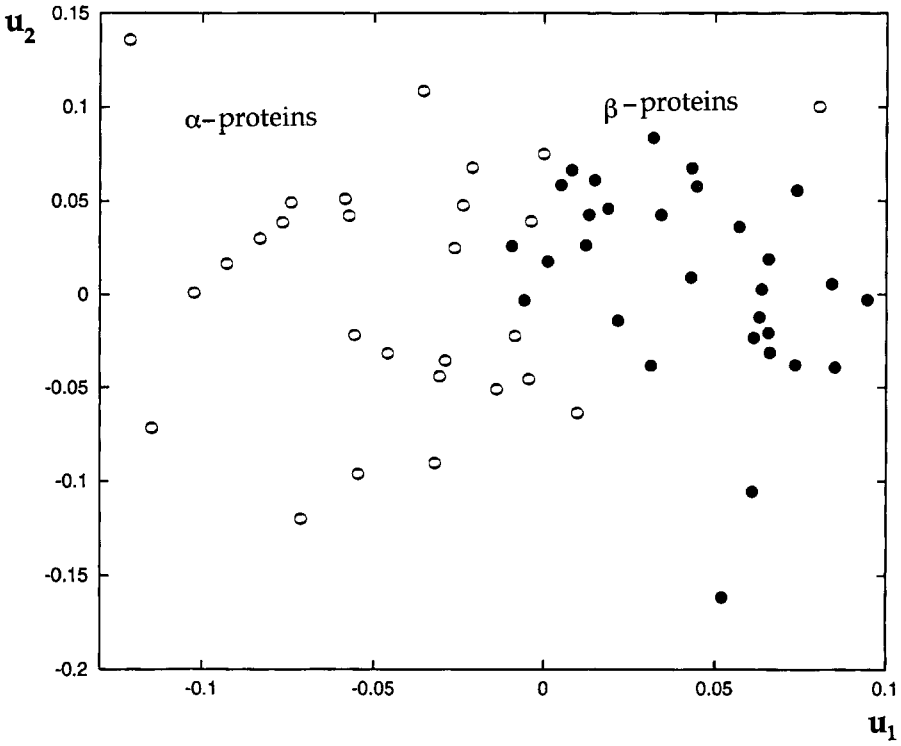


Fig. 1. Projection of proteins onto the frame spanned by the two principal axes $\mathbf{u}_1$ and $\mathbf{u}_2$. Sixty test structures are displayed, of which 30 are α-proteins (○) and 30 β-proteins (●).

of residues invariably results in a decrease in the success rate. To check the suitability of simplified representations of amino acid types for recognizing structural classes, the success rates achieved by projecting the problem to a lower dimensional space than 19 have been explored. This is done in two ways: 1) Residues were combined into groups, depending on their chemical characteristics, or on their size and shape resemblances, and the SVD analyses were repeated with these reduced sets comprising $p < 20$ representative types of amino acids. 2) A subset of the small singular values $\lambda_m$ ($m > p$), which may represent the noise in the system,[14] was removed, and the matrices **A** and **S** were reconstructed with the remaining p dominant singular values by using Equation (A6). A systematic decrease in the success rate was observed with decreasing P, indicating that the detailed description of the protein composition in terms of the 20 different types of amino acids, or the 19-dimensional space of amino acid fractions, is required for achieving the best recognition of structural classes. For illustrative purposes, the fractions of correctly predicted proteins for different subsets of singular values, $p = 18$ and 10, are presented in the last column of Table I, showing the systematic decrease in accuracy levels with the use of $p < 19$.

An important merit of SVD analysis is the possibility of identifying clusters of amino acids, in parallel with the structural classification of proteins. This is similar to the clustering of words from the analysis of text or words by Berry et al.[14] The jth element of $\mathbf{u}_i$ may be interpreted, for example, as the projection of the jth residue to the ith basis vector of the singular space. Suppose one considers the subspace spanned by the three dominant singular directions $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$, i.e., the first three columns of $\mathbf{U}_{[19\times19]}$. One can readily locate the position of all types of residues in that subspace. Such a map reveals the distances between different types of residues, insofar as their fraction is effective in recognizing a structural class. Residues playing a comparable role in recognizing a structural class are closer to each other, whereas those distinguished by their unique identities are isolated.

Figure 2 illustrates the loci of amino acids obtained from the SVD of the complete set of 120 training proteins, projected into the space spanned by $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$. We note that the charged residues Lys, Glu, and Arg, the hydrophobic residue Leu, and the small residues Gly and Ala are distinguished by their distinct loci. The residues His, Tyr, Phe, Pro, and Trp aggregate into a cluster, in which Met and Ile, and on a broader scale Gln, Cys, Asn, and Asp participate. Ser is closest to Gly and Thr, whereas Ala is closest to Val, in good agreement with the known properties of these residues.

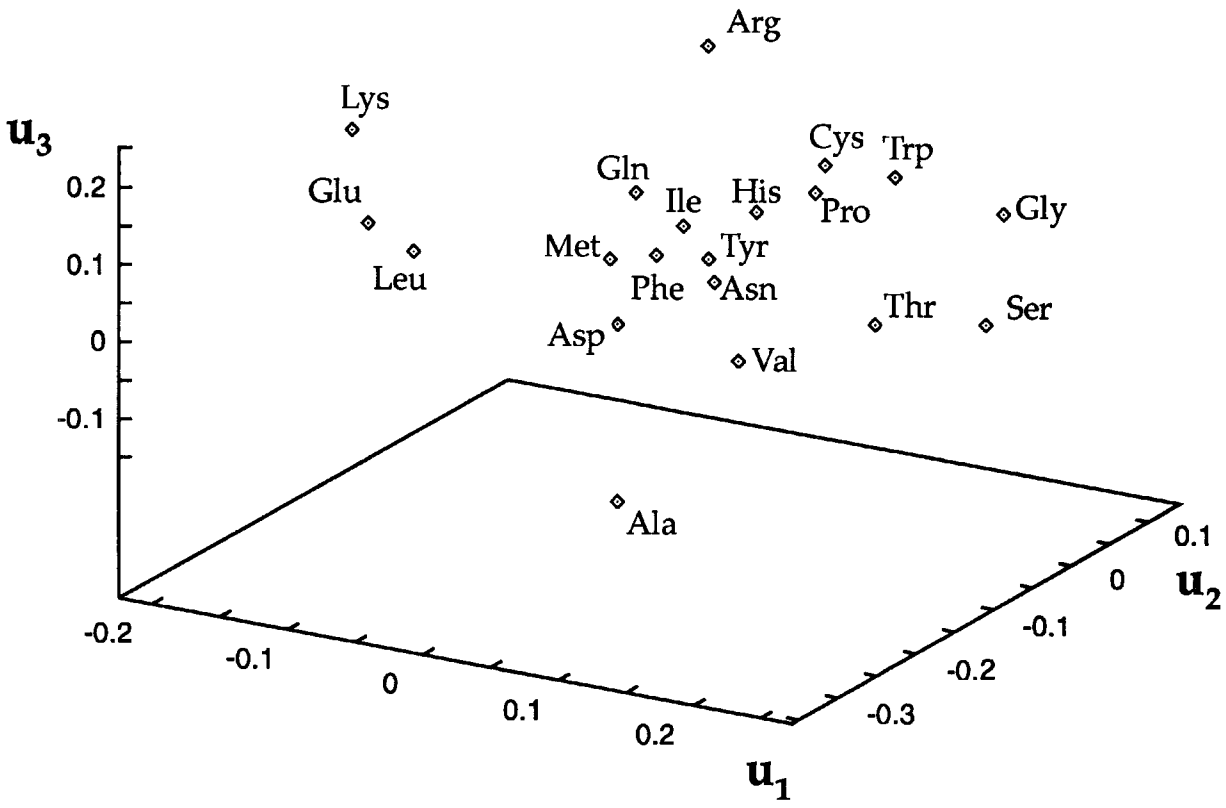


Fig. 2. Loci of amino acids in the space spanned by the three principal axes $\mathbf{u}_1$, $\mathbf{u}_2$, and $\mathbf{u}_3$ determined from the SVD analysis of the matrix **A**. See Equation (2).

## How Does Composition Recognize Structural Class? Examination of Simple Lattice Models

### *An illustrative example*

Let us first consider the three structures (a)–(c) displayed in Figure 3. These form the complete set of distinct conformations that may be assumed by a model protein of N = 9 residues confined to a 3 × 3 square lattice.[16] Suppose we want to identify the most stable structure among them. The drive for maximizing the total number of non-bonded contacts, which is a plausible criterion for selecting a given fold, is of no utility here, because all three structures are subject to the same number of non-bonded contacts (shown by the dashed lines). On the other hand, we note that residues in conformations (a) and (b) have the same distribution of coordination numbers: mainly, six vertices experience one non-bonded contact (z = 1) each, and one residue is subject to z = 2 contacts. The conformation (c), on the other hand, is distinguished by five vertices with coordination number z = 1 and one by z = 3. The corresponding distributions of non-bonded contacts may be designated as $\{2^1, 1^6\}$ and $\{3^1, 1^5\}$, respectively. Therefore, two distinct distributions of non-bonded contacts are discerned here, insofar as the coordination numbers of individual residues are concerned.

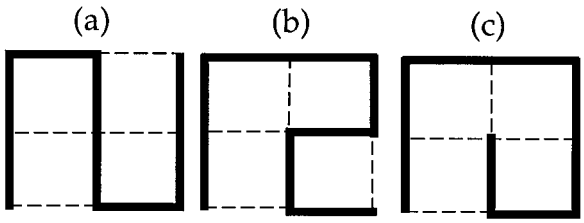


Fig. 3. Complete set of distinct conformations of nine-unit chains on a 3 × 3 square lattice (except mirror images). Non-bonded contacts are shown by the dashed lines. Conformations (**a**) and (**b**) have identical distributions of coordination numbers (distribution I): one unit has two non-bonded contacts in these conformations, and six units have a single non-bonded contact. Conformation (**c**), on the other hand, is distinguished by one unit with three non-bonded contacts, and five with a single contact, and thus represents another distribution (distribution II).

These will be referred to as distributions of non-bonded contacts I and II, respectively.

Let us next turn our attention to the selection of one of the two distributions of non-bonded contacts by a particular amino acid composition. For simplicity, two types of residues are considered: H and P. Their interaction potentials are denoted as $E_{HH}$, $E_{HP}$, and $E_{PP}$. Let us suppose the composition of the model protein is $N_H$:$N_P$ = 1:8, where $N_H$ and $N_P$ are the respective numbers of residues of type H and P.

Inasmuch as we are interested in the recognition of a structure by a composition, irrespective of the particular primary sequence, all permutations of H and P residues subject to the fixed composition will be taken into consideration. Precisely, an average over all $C(N_H, N_P) \equiv N!/N_H!N_P!$ combinations for fixed $N_H$ and $N_P$ will be examined. For the case $N_H = 1$, for example, nine sequences—or primary structures—are possible, some of which are energetically more favorable than others when subjected to the distribution of non-bonded contacts I or II. In particular, the sequence $P_8H$ will prefer the distribution II, provided that $E_{HP} < E_{PP}$. This preference will also be reflected in the weighted average over all nine primary structures threaded onto (II), resulting in a preference for the composition $N_H:N_P = 1:8$ for distribution of non-bonded contacts II. Clearly, the situation will be reversed, i.e., distribution I will be preferred, when $N_H:N_P = 8:1$, given that $E_{HH} < E_{HP}$.

In addition to this energetic effect, the number of conformations with a given distribution of non-bonded contacts, which we may simply term the degeneracy of a given distribution of non-bonded contacts, affects the choice of a distribution by a given composition. In the present simple example, the distribution of non-bonded contacts I comprises two conformations, (a) and (b), and in the absence of interaction energies, it is two times more probable than the distribution of non-bonded contacts II. This is the entropic contribution. It determines the preferred distribution of non-bonded contacts for homogeneous, all H, or all P, chains. In the case of heterogeneous chains, on the other hand, a competition between entropic and enthalpic effects is effective in setting the preference for a given distribution of non-bonded contacts. The most probable distributions resulting from the two contributions are shown in Figure 4 as a function of the composition of the chain. The curve displays the probability of occurrence of the most probable distribution of non-bonded contacts, which is either I or II depending on the composition, as indicated by the labels. It is interesting to observe the change in preference from one distribution of non-bonded contacts to another with changing composition. H-H, H-P, and P-P interaction energies are taken here as $E_{HH} = -3.0$ RT, $E_{HP} = -1.2$ RT, and $E_{PP} = 0$, in conformity with the structure-derived potentials for residues of type H and P.[22,23] This simple example shows that it is possible to group conformations into sets on the basis of their distribution of non-bonded contacts and identify the distributions of non-bonded contacts that are preferred for particular amino acid compositions. However, this analysis is not feasible with increased sizes of the chains, or in real proteins, unless a more systematic method is adopted. Such a method, applicable to longer chains, both on- or off-lattice, is presented next.

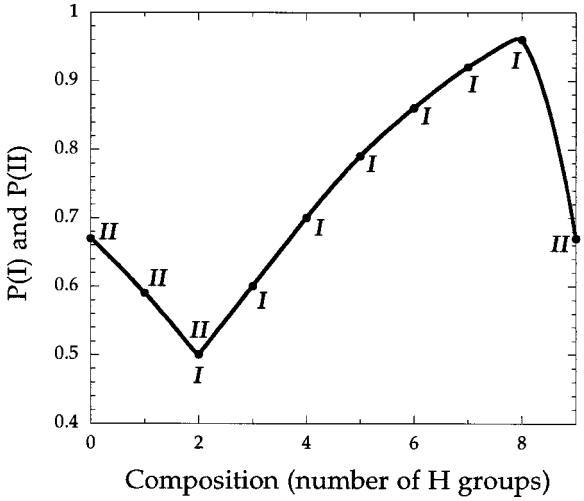


Fig. 4. Probability of occurrence of the two distributions I and II as a function of the composition of the nine-unit chain on the $3 \times 3$ square lattice. The abscissa represents the number $N_H$ of hydrophobic residues. Distribution I is preferred over II at higher concentrations $3 \leq N_H \leq 8$. At low concentrations of hydrophobic residues and in the extreme case of one component chains, distribution II is preferred. The two distributions are equally probable at $N_H = 2$.

### *General approach*

As a first step, an exhaustive enumeration of all conformations compatible with a given 3-D shape is made.[24] We note that in an $n \times n \times n$ cubic lattice all conformations have equal numbers of contacts, $n + n^{1/3} - 2n^{2/3}$, and the factor distinguishing the conformations is their distribution of non-bonded contacts, i.e., the distribution of coordination numbers, and the size and geometry of clusters of sites in close contact.

Next, those conformations having identical distributions of non-bonded contacts should be identified and assigned to different subsets. The characterization of the distributions of non-bonded contacts is conducted by the eigenvalue analysis of the corresponding Kirchhoff matrix **A**. The latter is a symmetric matrix of order m, for a chain comprising m interaction sites. The elements of **A** are defined as

$$\mathbf{A}_{ij} = \begin{cases} 1 & \text{if } i \neq j \text{ and } \mathbf{r}_{ij} \leq \mathbf{r}_c \\ 0 & \text{if } i \neq \text{ and } \mathbf{r}_{ij} > \mathbf{r}_c \\ -\sum \mathbf{A}_{ij} & \text{if } i = j \end{cases} \qquad (4)$$

The last summation in Equation (4) is done over all off-diagonal elements on a given column (or row). $r_{ij}$ is the distance between sites i and j, and $r_c$ is a cutoff separation defining the range of non-bonded contacts. Conformations having the same distribution of non-bonded contacts yield an identical set of eigenvalues upon transformation of their Kirchhoff matrices **A.** We note that the matrix **A** is similar in form to the

adjacency matrices of graph theory,[15,16] except for the definition of the diagonal elements as the negative sum of the other corresponding row (or column) elements. In this respect, **A** is equivalent in form to the rate matrices controlling the transitions between communicating states of multivariate stochastic processes (see for example Ref. 25).

Let us suppose that the eigenvalue analysis of the overall set of $\Omega$ conformations leads to W distinct distributions of non-bonded contacts. For convenience, these are designated as $\{\mathbf{\Phi}_k\}$, $1 \leq k \leq W$. Each of them comprises a total of $g_k$ conformations, such that $\Sigma_k g_k = \Omega$, i.e., $g_k$ is the degeneracy of the kth distribution of non-bonded contacts. To determine the distribution of non-bonded contacts preferred by a given composition ($N_H$, $N_P$), all combinations $C(N_H, N_P)$ of sequences are threaded onto one representative conformation chosen from each subset $\{\mathbf{\Phi}_k\}$. And an average energy is assigned to each distribution of non-bonded contacts on the basis of its total non-bonded energy averaged over all sequences as

$$\overline{\mathbf{E}}\{\mathbf{\Phi}_k;N_H\} = \frac{\sum_p E_p\{\mathbf{\Phi}_k\} \exp\left[-E_p\{\mathbf{\Phi}_k\}/RT\right]}{\sum_p \exp\left[-E_p\{\mathbf{\Phi}_k\}/RT\right]} \qquad (5)$$

where $1 \leq p \leq C(N_H, N_P)$. These are combined with the degeneracies $g_k$, to estimate the probability $P\{\mathbf{\Phi}_k; N_H\}$ of recognition of the kth distribution of non-bonded contacts by the composition ($N_H$, $N_P$),

$$\mathbf{P}\{\mathbf{\Phi}_k; N_H\} = \frac{g_k \exp\left[-\overline{\mathbf{E}}\{\mathbf{\Phi}_k, N_H\}/RT\right]}{\sum_K g_k \exp\left[-\overline{\mathbf{E}}\{\mathbf{\Phi}_k, N_H\}/RT\right]} \qquad (6)$$

We note that in the special case $E_{HH} = E_{HP} = E_{PP} = 0$, the subset with the highest degeneracy number will invariably be selected by all compositions.

### *Results for H-P model chains on 3 × 2 × 2 lattice*

In parallel with the procedure described above, all compact conformations are first generated on the 3 × 2 × 2 lattice. The following efficient method is adopted here for the exhaustive enumeration of all conformations. There are $n_v = 12$ vertices and $n_e = 20$ edges, and consequently $n_c = n_e - n_v + 1 = 9$ non-bonded contacts in the 3 × 2 × 2 lattice. In the absence of constraints, the number N of ways of distributing the $n_c$ contacts over $n_e$ edges is $N = n_e!/(n_e - n_v + 1)!(n_v - 1)! = 167{,}960$. We note that there are two different types of points in the 3 × 2 × 2 lattice: 1) lattice points at a corner and 2) lattice points along an external edge, but not at a corner. The coordination numbers of these two types cannot be more than 2 and 3, or less than 1 and 2, respectively. One has to reject those configurations that violate the stated maximum and minimum contact number conditions. In addition to these single site constraints, those conformations that lead to the isolation of one edge or face must be eliminated. Considering these restrictions, the total number of conformations generated on a 3 × 2 × 2 lattice reduces to 680.

In the second stage, Kirchhoff matrices are determined for each conformation. $r_c$ is taken as the length of a lattice edge. The eigenvalue analysis of these matrices lead to W = 21 distinct distributions of non-bonded contacts, each of them characterized by a unique set of eigenvalues. These are presented in Table II. The second column gives the degeneracy $g_k$ of each distribution of non-bonded contacts, and the succeeding nine columns are the corresponding non-zero eigenvalues $\lambda_i$, $1 \leq i \leq 9$. The probabilities of selection of these distributions of non-bonded contacts by the model chains of different compositions are listed in Table III. It is interesting to note from the two tables that the distributions of non-bonded contacts that are distinguished by their larger $\lambda_1$ values also exhibit the largest probabilities of being selected by a particular composition. This is consistent with the fact that the departure of the eigenvalues from a uniform distribution reflects the singularity of the distribution of non-bonded contacts, and those distributions of non-bonded contacts exhibiting more singular, unique distributions are more readily recognized by a given amino acid composition.

We note that three distributions of coordination numbers are accessible on a 3 × 2 × 2 lattice: A: $\{3^2, 2^2, 1^8\}$; B: $\{3^1, 2^4, 1^7\}$; and C: $\{2^6, 1^6\}$, where the exponents indicate repeating coordination numbers. However, the coordination numbers are not the only quantity distinguishing a given distribution of non-bonded contacts; the size and geometry of non-bonded clusters also contribute to the definition of a given distribution of non-bonded contacts and leads to 21 distinct subsets, identified by the eigenvalue analysis. This feature is illustrated in Figure 5. Here two conformations with identical distributions of coordination numbers, $\{3^1, 2^4, 1^7\}$, but different geometries are displayed. The types of coordination numbers corresponding to each distribution of non-bonded contacts are indicated in parentheses in the first column of Table III. The fact that these assignments closely conform with the hierarchy of eigenvalue distributions also indicates that the coordination numbers play a dominant role in defining the eigenvalue distributions (or the distributions of non-bonded contacts), but a finer level characterization of the structures in each group (A, B, or C) is achieved by the eigenvalue analysis of the Kirchhoff matrices.

Finally, the probabilities of selection of the coordination number distributions A, B, and C as a function of composition are shown in Figure 6. In parallel with Figure 4, the preferred distributions are con-

**TABLE II. Degeneracies ($g_k$) and Eigenvalues ($\lambda_i$) for Subsets of 3 × 2 × 2 Lattice Conformations Having Identical Distributions of Non-Bonded Contacts**

| k | $g_k$ | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\lambda_5$ | $\lambda_6$ | $\lambda_7$ | $\lambda_8$ | $\lambda_9$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 8 | 4.81 | 3.00 | 2.53 | 2.00 | 2.00 | 2.00 | 1.00 | 0.66 | 0.00 |
| 2 | 16 | 4.68 | 3.41 | 2.33 | 2.00 | 2.00 | 2.00 | 0.73 | 0.58 | 0.25 |
| 3 | 16 | 4.64 | 3.41 | 2.72 | 2.00 | 2.00 | 1.41 | 1.00 | 0.58 | 0.22 |
| 4 | 16 | 4.56 | 3.00 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 0.44 | 0.00 |
| 5 | 16 | 4.48 | 3.00 | 2.69 | 2.00 | 2.00 | 2.00 | 1.00 | 0.83 | 0.00 |
| 6 | 32 | 4.44 | 3.14 | 2.62 | 2.00 | 2.00 | 2.00 | 1.18 | 0.38 | 0.24 |
| 7 | 16 | 4.41 | 3.00 | 2.62 | 2.62 | 2.00 | 1.59 | 1.00 | 0.38 | 0.38 |
| 8 | 64 | 4.34 | 3.41 | 2.47 | 2.00 | 2.00 | 2.00 | 1.00 | 0.58 | 0.19 |
| 9 | 32 | 4.33 | 3.10 | 3.00 | 2.27 | 2.00 | 1.40 | 1.00 | 0.63 | 0.26 |
| 10 | 48 | 4.30 | 3.41 | 2.62 | 2.00 | 2.00 | 2.00 | 0.70 | 0.58 | 0.38 |
| 11 | 32 | 4.23 | 3.36 | 3.00 | 2.18 | 2.00 | 1.00 | 1.00 | 1.00 | 0.22 |
| 12 | 32 | 4.17 | 3.62 | 2.62 | 2.31 | 2.00 | 1.38 | 1.00 | 0.52 | 0.38 |
| 13 | 16 | 4.00 | 3.41 | 2.00 | 2.00 | 2.00 | 2.00 | 2.00 | 0.58 | 0.00 |
| 14 | 8 | 4.00 | 3.00 | 3.00 | 2.00 | 2.00 | 2.00 | 1.00 | 1.00 | 0.00 |
| 15 | 64 | 3.85 | 3.41 | 2.77 | 2.00 | 2.00 | 2.00 | 1.23 | 0.58 | 0.15 |
| 16 | 48 | 3.80 | 3.25 | 3.00 | 2.44 | 2.00 | 1.55 | 1.00 | 0.75 | 0.20 |
| 17 | 32 | 3.73 | 3.00 | 3.00 | 3.00 | 2.00 | 1.00 | 1.00 | 1.00 | 0.27 |
| 18 | 96 | 3.73 | 3.41 | 3.00 | 2.00 | 2.00 | 2.00 | 1.00 | 0.58 | 0.27 |
| 19 | 32 | 3.62 | 3.62 | 2.62 | 2.62 | 2.00 | 1.38 | 1.38 | 0.38 | 0.38 |
| 20 | 48 | 3.62 | 3.41 | 3.00 | 2.62 | 2.00 | 1.38 | 1.00 | 0.58 | 0.38 |
| 21 | 8 | 3.41 | 3.41 | 3.41 | 2.00 | 2.00 | 2.00 | 0.58 | 0.58 | 0.58 |

**TABLE III. Probabilities of Distinct Distributions of Non-Bonded Contacts for H-P Model Chains on a 3 × 2 × 2 Lattice, as a Function of Composition***

| k† | $P\{\Phi_k, 1\}$ | $P\{\Phi_k, 2\}$ | $P\{\Phi_k, 3\}$ | $P\{\Phi_k, 4\}$ | $P\{\Phi_k, 5\}$ | $P\{\Phi_k, 6\}$ | $P\{\Phi_k, 7\}$ | $P\{\Phi_k, 8\}$ | $P\{\Phi_k, 9\}$ | $P\{\Phi_k, 10\}$ | $P\{\Phi_k, 11\}$ | $P\{\Phi_k, 12\}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 (A) | 0.029 | 0.093 | 0.125 | 0.293 | 0.149 | 0.071 | 0.040 | 0.026 | 0.019 | 0.015 | 0.013 | 0.012 |
| 2 (A) | 0.058 | 0.188 | 0.250 | 0.152 | 0.103 | 0.069 | 0.050 | 0.039 | 0.032 | 0.028 | 0.025 | 0.024 |
| 3 (A) | 0.058 | 0.186 | 0.166 | 0.142 | 0.106 | 0.072 | 0.052 | 0.040 | 0.033 | 0.028 | 0.025 | 0.024 |
| 4 (B) | 0.035 | 0.032 | 0.039 | 0.062 | 0.143 | 0.107 | 0.069 | 0.049 | 0.036 | 0.029 | 0.024 | 0.024 |
| 5 (B) | 0.035 | 0.025 | 0.027 | 0.044 | 0.047 | 0.046 | 0.041 | 0.035 | 0.030 | 0.027 | 0.024 | 0.024 |
| 6 (B) | 0.071 | 0.065 | 0.068 | 0.057 | 0.079 | 0.085 | 0.078 | 0.068 | 0.059 | 0.053 | 0.048 | 0.047 |
| 7 (B) | 0.035 | 0.032 | 0.029 | 0.017 | 0.016 | 0.017 | 0.019 | 0.021 | 0.023 | 0.024 | 0.024 | 0.024 |
| 8 (B) | 0.142 | 0.101 | 0.092 | 0.088 | 0.154 | 0.175 | 0.162 | 0.138 | 0.120 | 0.107 | 0.097 | 0.094 |
| 9 (B) | 0.071 | 0.050 | 0.044 | 0.030 | 0.031 | 0.034 | 0.039 | 0.044 | 0.046 | 0.047 | 0.048 | 0.047 |
| 10 (B) | 0.106 | 0.075 | 0.053 | 0.030 | 0.037 | 0.047 | 0.057 | 0.065 | 0.069 | 0.071 | 0.072 | 0.071 |
| 11 (B) | 0.071 | 0.036 | 0.030 | 0.026 | 0.030 | 0.035 | 0.040 | 0.044 | 0.047 | 0.047 | 0.048 | 0.047 |
| 12 (B) | 0.071 | 0.036 | 0.022 | 0.015 | 0.023 | 0.031 | 0.039 | 0.044 | 0.047 | 0.047 | 0.048 | 0.047 |
| 13 (C) | 0.010 | 0.004 | 0.003 | 0.004 | 0.007 | 0.018 | 0.028 | 0.029 | 0.027 | 0.025 | 0.023 | 0.024 |
| 14 (C) | 0.005 | 0.002 | 0.002 | 0.002 | 0.007 | 0.038 | 0.029 | 0.023 | 0.017 | 0.014 | 0.011 | 0.012 |
| 15 (C) | 0.039 | 0.016 | 0.013 | 0.012 | 0.027 | 0.074 | 0.112 | 0.117 | 0.109 | 0.101 | 0.091 | 0.094 |
| 16 (C) | 0.030 | 0.011 | 0.008 | 0.006 | 0.012 | 0.021 | 0.034 | 0.048 | 0.059 | 0.065 | 0.068 | 0.071 |
| 17 (C) | 0.020 | 0.007 | 0.004 | 0.003 | 0.004 | 0.006 | 0.010 | 0.016 | 0.025 | 0.036 | 0.046 | 0.047 |
| 18 (C) | 0.059 | 0.022 | 0.015 | 0.011 | 0.017 | 0.034 | 0.064 | 0.096 | 0.118 | 0.129 | 0.137 | 0.141 |
| 19 (C) | 0.020 | 0.007 | 0.005 | 0.003 | 0.005 | 0.011 | 0.021 | 0.032 | 0.039 | 0.043 | 0.046 | 0.047 |
| 20 (C) | 0.030 | 0.010 | 0.006 | 0.003 | 0.004 | 0.007 | 0.013 | 0.023 | 0.037 | 0.054 | 0.068 | 0.071 |
| 21 (C) | 0.005 | 0.002 | 0.001 | 0.000 | 0.001 | 0.001 | 0.002 | 0.004 | 0.006 | 0.009 | 0.011 | 0.012 |

*Compositions are indicated by the arguments of the probabilities; $P\{\Phi_k, n\}$ represents the probability of selection of the kth by the chain containing n residues of type H, $1 \leq n \leq 12$.
†Letters in parentheses refer to the coordination number distributions A: $\{3^2, 2^2, 1^8\}$; B: $\{3^1, 2^4, 1^7\}$; and C: $\{2^6, 1^6\}$.

nected by a boldface curve. It is interesting to observe the strong dependence of the coordination type preferences on composition: B is selected when $N_H = 1$ and $5 \leq N_H \leq 9$; the range $2 \leq N_H \leq 4$ exhibits a preference for A in spite of the low degeneracy number (40) of this class, and finally, the presence of a large proportion of hydrophobic residues $10 \leq N_H \leq 12$ leads to a preference for C.

### Revisiting PDB Structures

Having extracted this information from lattice simulations, it is interesting to go back to real

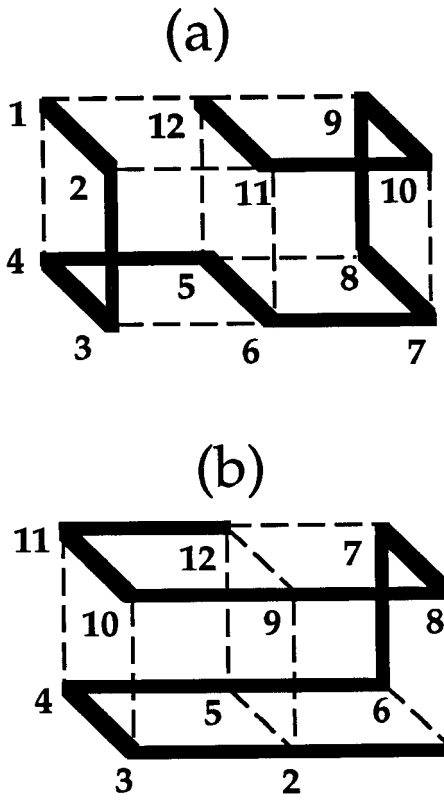


Fig. 5. Two distinct distributions of non-bonded contacts of 12-unit chains on the $3 \times 2 \times 2$ lattice. Units are indexed from 1 to 12, for clarity. Non-bonded contacts are shown by the dashed lines. Both conformations exhibit the same distribution of coordination numbers, one unit has three non-bonded contacts, four units have two non-bonded contacts, and seven experience a single contact. However, their coordination geometries differ.

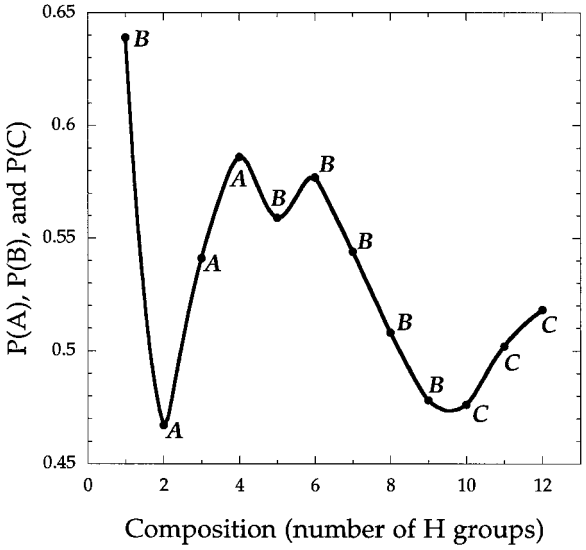


Fig. 6. Probabilities of the three coordination number distributions A, B, and C on the $3 \times 2 \times 2$ lattice, as a function of the composition of H-P model chains. The abscissa represents the number $N_H$ of hydrophobic residues. The preferred coordination types are connected by a boldface curve.

protein structures and check whether the distribution of coordination numbers, for example, differs in the different structural classes. Precisely, do the four structural classes exhibit differences in their coordination number distributions? Can we also trace some differences in the distribution of non-bonded contacts by eigenvalue analysis of Kirchhoff matrices?

### *Coordination numbers*

The following analysis is done for 40 proteins from each structural class. First, the distributions of coordination numbers, $P\{\mathbf{z}; \Lambda\}$, are evaluated for all classes, considering all residues, by using a set of 40 proteins from each class. The method conducted by Miyazawa and Jernigan[26] for different residues by using 1100 data bank structures is followed. Mainly, the numbers z of $C^\alpha$ atoms in the neighborhood ($r \leq 7.0$ Å) of a central $C^\alpha$ are analyzed. The frequencies of the different coordination numbers are examined. Here all $C^\alpha$ atoms are included, the degree of solvent exposure or burial of residues being also a characteristic of structural classes that needs to be taken into consideration. The results are displayed in Figure 7 as a function of the coordination number z, for the four classes. $\alpha$ and $\beta$ classes exhibit the most distinct behavior. The results for $\alpha+\beta$ and $\alpha/\beta$ lie between the curves for $\alpha$ and $\beta$ proteins.

The results show that the distribution of coordination numbers of the two classes exhibit significant departures from each other. $\alpha$-proteins are distinguished by their large proportion of $z = 6$ contacts; $\beta$-proteins are distinguished by the most frequent occurrence of coordination number $z = 4$. These most probable coordination numbers may be attributed to residue pairs $(i, i \pm 1)$, $(i, i \pm 3)$, $(i, i \pm 4)$ in $\alpha$-helices, and pairs $(i, i \pm 1)$, $(i, i \pm 2)$ in $\beta$-strands. It is to be noted that the individual proteins yield similar curves to the two mean curves shown in Figure 7, thus verifying the reproducibility of this analysis.

The examination of coordination number is not sufficient alone for the discrimination of all structural classes, particularly for distinguishing the classes $\alpha/\beta$ and $\alpha+\beta$. In fact, a further property, the geometry of non-bonded contacts, in addition to coordination numbers, was pointed out in the above arguments to characterize the distribution of non-bonded contacts of distinct structural classes. Distinct geometries of non-bonded contacts may be identified by the eigenvalue analysis of Kirchhoff matrices, which will be considered in the next subsection, indicating simply the differences in contact map patterns.

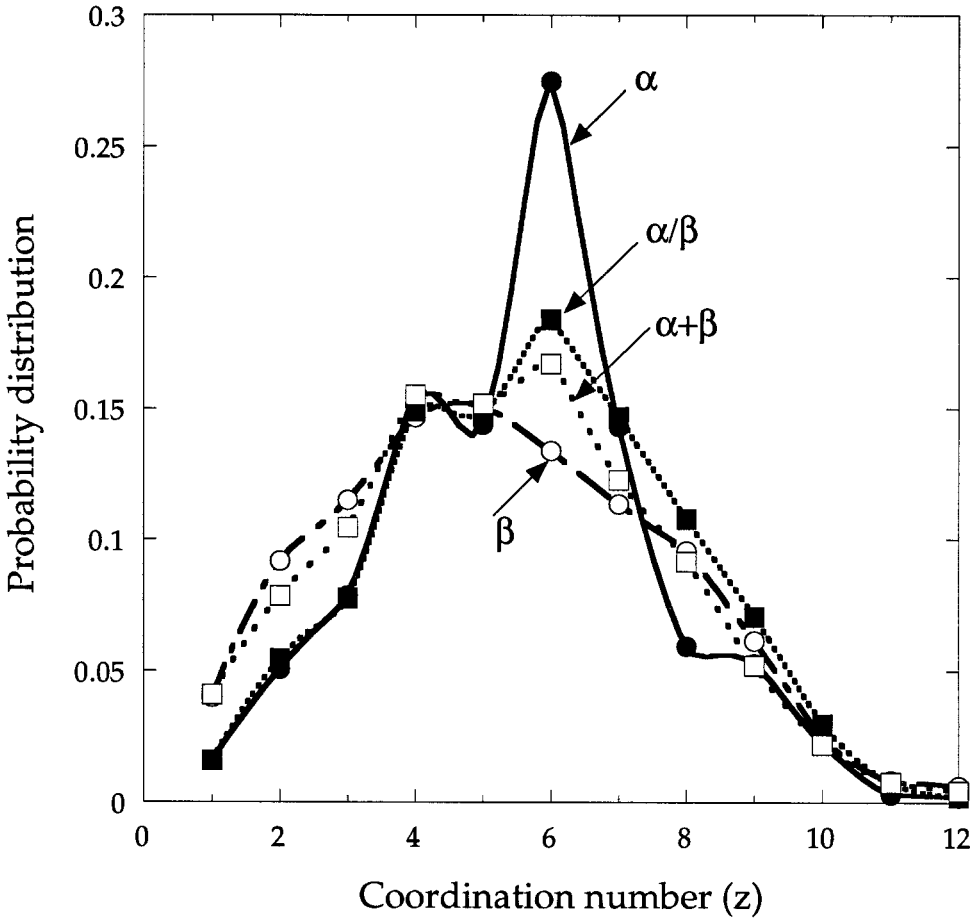


Fig. 7. Coordination number distributions for different structural classes. The curves represent the averages over all $C^{\alpha}$'s belonging to the indicated structural class. The distribution curves are normalized over the range $1 \leq z \leq 12$.

### *Eigenvalue distributions*

In lattice simulations, we have characterized the distribution of non-bonded contacts in a specific conformation by the eigenvalues of the corresponding Kirchhoff matrix **A** [Eq. (4)]. We now repeat the same analysis for proteins extracted from the PDB. $C^{\alpha}$ atoms located in the neighborhood ($r \leq 7.0$ Å) of a central $C^{\alpha}$ atom are considered. The percent contribution of the dominant 12 eigenvalues for each structural class are shown in Figure 8 as a function of eigenvalue index. The curves are obtained by averaging over 40 proteins for each class. The results for the α/β- and α+β-proteins exhibit strong departures from each other and therefore provide a useful means of identifying the two classes. The curves for the α-and β-proteins, on the other hand, are approximately superimposable.

The departure of the curve for α/β-proteins from that of α+β-proteins may be explained as follows. The Kirchhoff matrix for a given protein is analogous in form to a contact map, assuming all residue pairs whose α-carbons are separated by 7 Å or less to be in contact. Helices lead to entries parallel to the diagonal, shifted by three or four rows or columns with respect to the main diagonal (−45°). β-strands, on the other hand, are represented by arrays parallel or perpendicular to the main diagonal, depending on whether they are aligned parallel or antiparallel, respectively, to each other. As a result, the contact map, and/or the Kirchhoff matrix, for an α/β-protein consists predominantly of arrays parallel to the main diagonal, whereas that of α+β-proteins comprise segments both parallel and perpendicular to the main diagonal. Because of their more similar geometries, the eigenvalues of the former structure exhibit higher degeneracy numbers, compared with those of the latter. Repeated eigenvalues lead to horizontal segments in the eigenvalue distributions curves, hence the lower slope of the α/β-curve in Figure 8, compared with that of the curve for α+β-proteins.

## DISCUSSION AND CONCLUSION

There are well-established methods[27] for predicting secondary structures in proteins. It is clear that

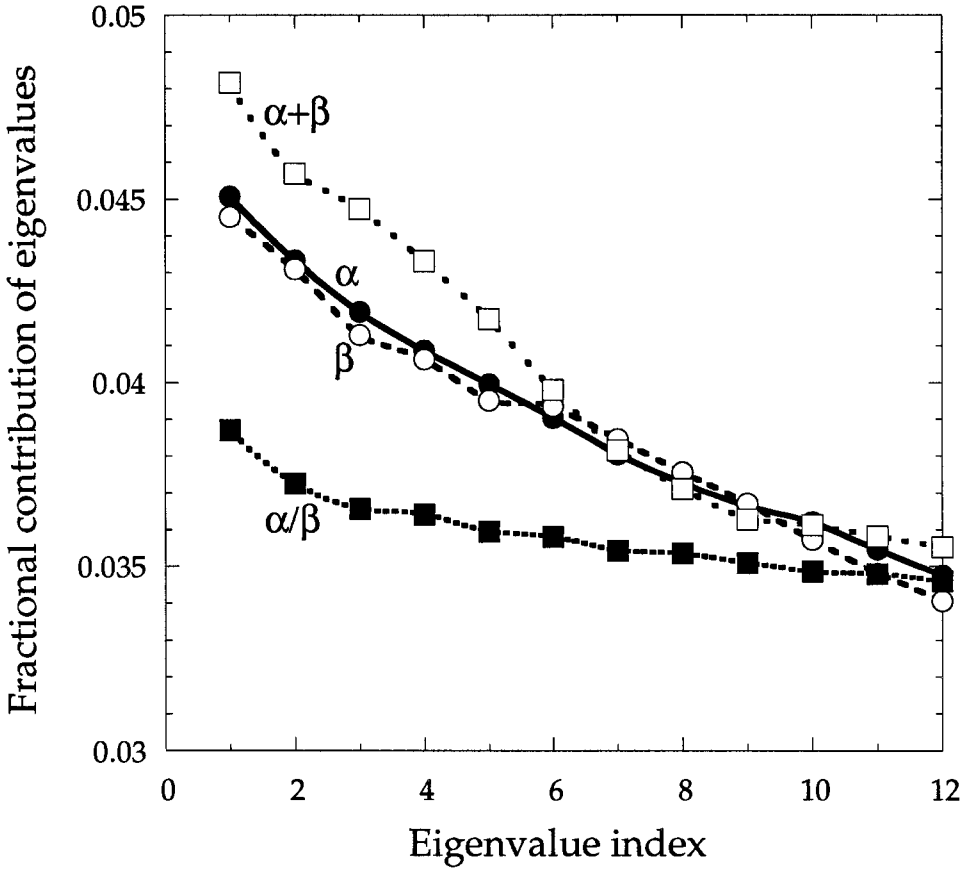


Fig. 8. Distribution of eigenvalues of Kirchhoff (contact) matrices for each structural class. The dominant 12 eigenvalues $\lambda_i$, $1 \leq i \leq 12$, are displayed. Curves represent the average behaviors obtained from the transformation of the Kirchhoff matrices of non-bonded contacts for 40 proteins from each class.

knowledge of structural class is an important step in protein folding problem in that it can improve the accuracy of secondary structure prediction[28] or reduce the conformational space during the search of the tertiary structure in globular proteins. The accurate prediction of structural classes, or secondary structural content, from amino acid composition alone is, in this respect, an important issue, which has been the object of a number of recent studies.[3,5,12,13,29] In the present study, our objectives have been to validate the relationship between amino acid composition and structural class by using a SVD method and to gain insights into the physical origin of the recognition of protein structural classes by amino acid composition.

Within the scope of these objectives, a straightforward method for the identification of structural classes is presented. This is based on the SVD of 19-dimensional protein vectors, the elements of which are the fractions of the amino acids present in the protein. The method automatically clusters proteins of different structural classes into different regions of a 19-dimensional space. The axes of the space as found by SVD maximize the distance between the proteins. One advantage of the method is that it also permits grouping residues in relation to their participation in different structural classes.

Neglect of one or more singular directions leads to a systematic decrease in the accuracy (Table I). A minimal basis set of all the 19-residue fractions is required for achieving the highest accuracy level when the four classes are considered together. The observation from Table I that the prediction for α-helices is better when 10 parameters are used is contrary to this conclusion and is thought to be an insignificant coincidence. We have also tried to improve the accuracy level by introducing more information about the amino acid composition of proteins. For example, the residue pairs located at positions i and i+3 were examined, and the proportions of H-H, H-P, and P-P pairs were added to protein vectors as additional information on the sequence composition. However, such considerations lead to a marginal increase (up to 82%) in the prediction level, which does not justify further elaboration in that direction. A new analytic vector decomposition technique ex-

ploring the limits of secondary structure content predictions relying solely on amino acid composition of the query protein, also indicates that couplings between amino acids do not significantly improve the success rates.[29]

One possible way of increasing the prediction rates of the present approach could be to construct more detailed protein vectors, in which data from Dayhoff substitution matrices are included in addition to amino acid fractions. Substitution matrices recently obtained[30] for amino acids at spatially conserved locations may be used, for example, for this purpose. It is interesting to notice that the charged residues are clustered therein in the same group, regardless of the type of their charge, in contrast to other substitution matrices. The location on the protein and consequently the possible types of non-bonded contacts with the environment appear to dominate in this behavior, which suggests that the combination of these substitution matrices with amino acid fractions may be useful in structural class predictions.

It should be noted that the apparent relatively high accuracy level (81%) attained in the present study, which exceeds the success rates (75%) of structural class predictions using traditional secondary structure prediction techniques (including those[6,31] combining evolutionary information and neural networks) may be due to some biases, as pointed out by Eisenhaber et al.[13] These are for example, the preselection of test sets, which may not be adequately representative of all unrelated proteins, the adoption of structural class definitions with extreme secondary structure contents, which thereby remove about 35% of the PDB structures without any class assignment, the ambiguities in the definitions of $\alpha+\beta$ and $\alpha/\beta$ classes having some $\beta$-sheets, including both parallel and antiparallel strands. In this respect, the structural class definitions proposed by Nakashima et al.[11] are pointed[13] out to be more appropriate. An extensive analysis of different methods led Eisenhaber and collaborators[13] to the conclusion that knowledge of amino acid composition alone cannot lead to a success rate higher than 60%. Nevertheless, these analyses show that amino acid composition does recognize structural classes to an accuracy level comparable with that of the much more complex secondary structure prediction methods. And in the second part of our study, we sought an explanation for the 'composition-recognizes-class property' by a thorough examination of simple model chains on a lattice.

In the interest of gaining an understanding into the origins of the selectivity of structural classes by amino acid composition, exhaustive enumerations of all conformations and all primary structures for simple H-P model chains were performed. These calculations revealed that the distribution of non-bonded contacts in a given 3-D fold is the important parameter controlling its recognition by an amino acid composition. For example, structures that permit the burial of all hydrophobic residues at a given composition, will be selected by those compositions of residues.

For compact structures permitting the same number of non-bonded contacts, the distribution of non-bonded contacts physically refers to the coordination number distributions and the size and geometry of clusters of non-bonded contacts. Mathematically, the distribution of non-bonded contacts is uniquely obtainable from the eigenvalue distribution of the Kirchhoff connectivity matrix for a given protein. In the absence of energetic effects, the selection of a particular distribution of non-bonded contacts scales with its degeneracy, i.e., the number of conformations exhibiting the same eigenvalue distribution, whereas for heterogeneous systems, enthalpic effects will also come into play, in addition to this entropic effect.

In view of the complexity and heterogeneity of protein structures, a one-to-one identification of a structural class with a given well-defined distribution of non-bonded contacts is not possible. Instead, each class exhibits a broad range of distributions of non-bonded contacts. However, examination of PDB structures do confirm that proteins belonging to different structural classes differ in the coordination number distributions of their residues, on average, and in the eigenvalue distributions of their adjacency matrices, as illustrated in Figures 7 and 8.

Li et al.[18] pointed out that certain 3-D structures that are 'protein-like' with secondary structures and symmetries are thermodynamically more stable than ordinary structures, as evidenced by their selection as the lowest energy state by a significantly large number of primary structures. These structures are thus easy to design and are stable against mutations. Here we go one step further and show that certain distributions of non-bonded contacts, rather than particular detailed 3-D structures, may be selected. And the amino acid composition, rather than the detailed primary structure, may be sufficient for the selection.

## ACKNOWLEDGMENTS

Support of the NATO Collaborative Research Grant Project CRG951240 and Bogazici University Research Funds Project 96A0430 is gratefully acknowledged.

## REFERENCES

1. Zhang, C.T., Chou, K.C. An optimization approach to predicting protein structural class from amino acid composition. Protein Sci. 1:401–408, 1992.
2. Dubchak, I., Holbrook, S.R., Kim, S.H. Prediction of protein folding class from amino acid composition. Proteins 16:79–91, 1993.
3. Chou, K.C. Does the folding type of a protein depend on its amino acid composition? FEBS Lett. 363:127–131, 1995.

4. Zhang, C.T., Chou, K.C., Maggiora, G.M. Predicting protein structural classes from amino acid composition: Application of fuzzy clustering. Protein Eng. 8:425–435, 1995.
5. Chou, K.C. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins 21:319–344, 1995.
6. Rost, B., Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19:55–72, 1994.
7. Reczko, M., Bohr, H. The DEF data base of sequence based protein fold class predictions. Nucleic Acids Res. 22:3616–3619, 1994.
8. Nishikawa, K., Ooi, T. Correlation of the amino acid composition of a protein to its structural and biological characters. J. Biochem. 91:1821–1824, 1982.
9. Nishikawa, K., Kubota, Y., Ooi, T. Classification of proteins into groups based on amino acid composition and other characters. I. Angular distribution. J. Biochem. 94:981–995, 1983.
10. Nishikawa, K., Kubota, Y., Ooi, T. Classification of proteins into groups based on amino acid composition and other characters. II. Grouping into four types. J. Biochem. 94:997–1007, 1983.
11. Nakashima, H., Nishikawa, K., Ooi, T. The folding type of a protein is relevant to the amino acid composition. J. Biochem. 99:153–162, 1986.
12. Chou, P.Y. Prediction of protein structural classes from amino acid composition. In: 'Prediction of Protein Structure and Principles of Protein Conformation.' Fasman, G.D. (ed.). New York: Plenum Press, 1989:549–586.
13. Eisenhaber, F., Frommel, C., Argos, P. Prediction of secondary structural content of proteins from their amino acid composition alone. II. The paradox with secondary structural class. Proteins 25:169–179, 1996.
14. Berry, M.W., Dumais, S.T., O'Brien, G.W. Using linear algebra for intelligent information retrieval. SIAM Rev. 37:573–595, 1995.
15. Harary, F. 'Graph Theory.' Reading MA: Addison-Wesley, 1971.
16. Jernigan, R.L. Generating general shapes and conformations with regular lattices, for compact proteins. In: 'Structure & Function' R.H.Sarma, Sarma, M.H. (ed.). Adenine Press, Schenectady, NY, Vol. 2 1992:169–182.
17. Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D., Chan, H.S. Principles of protein folding. A perspective from simple exact models. Protein Sci. 4:561–602, 1995.
18. Li, H., Helling, R., Tang, C., Wingreen, N. Emergence of preferred structures in a simple model of protein folding. Science 273:666–669, 1996.
19. Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Brice, M., Rodgers, J., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Databank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
20. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., Weng, J. 'Protein Data Bank, Crystallographic Databases-Information Content Software Systems, Scientific Applications' Allen, F.H., Bergerhoff, G., Sievers, R. (eds.). Bonn, Cambridge, and Chester: Data Commission of the International Union of Crystallography, 1987:107.
21. Chou, K., Zhang, C. Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30:275–439, 1995.
22. Jernigan, R.L., Bahar, I. Structure-derived potentials and protein simulations. Curr. Opin. Struct. Biol. 6:195–209, 1996.
23. Bahar, I., Jernigan, R.L. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. J. Mol. Biol. 266:195–214, 1997.
24. Bahar, I., Jernigan, R.L. Cooperative structural transitions induced by non-homogeneous intramolecular interactions in compact globular proteins. Biophys. J. 66:467–477, 1994.
25. Schuler, K.E. (ed.) Stochastic processes in chemical physics. 'Advances in Chemical Physics.' Prigogine, I., Rice, S. (series eds.). Vol. 15. New York: Interscience, 1969.
26. Miyazawa, S., Jernigan, R.L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J. Mol. Biol. 256:623–644, 1996.
27. Garnier, J., Robson, B. The GOR method for predicting secondary structures in proteins. In: 'Prediction of Protein Structure and Principles of Protein Conformation.' Fasman, G.D. (ed.). New York: Plenum Press, 1989:417–465.
28. Deléage, G., Roux, B. Use of class prediction to improve protein secondary structure prediction. In: 'Prediction of Protein Structure and Principles of Protein Conformation.' Fasman, G.D. (ed.). New York: Plenum Press, 1989:587–597.
29. Eisenhaber, F., Imperiale, F., Argos, P., Frommel, C. Prediction of secondary structural content of proteins from their amino acid composition alone. I. New analytic vector decomposition methods. Proteins 25:157–168, 1996.
30. Naor, D., Fischer, D., Jernigan, R.L., Wolfson, H.J., Nussinov, R. Amino acid pair interchanges at spatially conserved locations. J. Mol. Biol. 256:924–938, 1996.
31. Chandonia, J.-M., Karplus, M. Neural networks for secondary structure and structural class predictions. Protein Sci. 4:275–285, 1995.
32. Rahman, R.S., Rackovsky, S. Protein sequence randomness and sequence/structure correlations. Biophys. J. 68:1531–1539, 1995.
33. Pillai, K.C.S. Mahalanobis, Prasanta Chandra. In: Encyclopedia of Statistical Sciences.' Kotz, S., Johnson, N.L. (ed.). New York: John Wiley & Sons, 1985:176–181.

## APPENDIX

### SVD Method for Characterizing Structural Classes on the Basis of Amino Acid Composition

The protein vectors, $\Delta\mathbf{r}_i$, given by Equation (1), are arranged in a matrix $\mathbf{A}_{[19\times n]}$. The rows of this matrix identify the 19 independent residue fractions, and the n columns refer to the proteins under investigation. The matrix $\mathbf{A}$ is decomposed into a product of three matrices by the SVD technique as

$$\mathbf{A}_{[19\times n]} = \mathbf{U}_{[19\times 19]}\, \mathbf{\Lambda}_{[19\times 19]}\, \mathbf{V}^{T}_{[19\times n]} \qquad \text{(A1)}$$

where $\mathbf{U}$ and $\mathbf{V}$ are the matrices of the left singular vectors (or principal axes) ($\mathbf{u}_i$), and right singular vectors ($\mathbf{v}_i$), respectively, of $\mathbf{A}$, $\Lambda$ is the diagonal matrix of the singular values. The details of the SVD technique may be found in Reference 14. The columns of $\mathbf{V}^T$ are the protein vectors $\Delta\hat{\mathbf{r}}_i$ expressed in the singular frame spanned by the axes $\mathbf{u}_i$, $1 \le i \le 19$. The superposed hat in $\Delta\hat{\mathbf{r}}_i$ designates the representation in the singular space. Thus, the matrix $\Lambda^{-1}\mathbf{U}^T$ when operated on $\mathbf{A}$ rotates each of the n original protein vectors into $\Delta\hat{\mathbf{r}}_i$ following the expression

$$\Delta\hat{\mathbf{r}}_i = \mathbf{\Lambda}^{-1}\, \mathbf{U}^T\, \Delta\mathbf{r}_i \qquad \text{(A2)}$$

The singular space gives the best representation of proteins insofar as their amino acid compositions are concerned, the orthonormal axes $\mathbf{u}_i$ being automatically chosen to magnify the differences in composi-

tion fluctuations. Thus, the distance $d_i$ of protein i from the center or the norm $\bar{\mathbf{r}}$ of a structural class is best accounted for by

$$d_i = [\Delta\hat{\mathbf{r}}_i \cdot \Delta\hat{\mathbf{r}}_i]^{1/2} \quad \text{(A3)}$$

where the dot denotes the scalar product. If $d_i$ is small, then the protein i is similar to those participating in the considered class. The extent of similarity is of course based on which criterion the comparison is made. In the present analysis, proteins are grouped according to their residue fractions, whereas a similar analysis was performed by Rahman and Rackovsky[32] on the basis of 10 characteristic properties.

By classifying the residues and the proteins according to the distance between their respective vectors, it becomes possible to make a cluster analysis. We note that the dot product in brackets in Equation (A3) may be written as

$$\Delta\hat{\mathbf{r}}_i \cdot \Delta\hat{\mathbf{r}}_i = \Delta\mathbf{r}_i{}^{\mathrm{T}}\mathbf{U}\,\mathbf{\Lambda}^{-2}\mathbf{U}^{\mathrm{T}}\,\Delta\mathbf{r}_i = \Delta\mathbf{r}_i^{\mathrm{T}}\,\mathbf{S}^{-1}\,\Delta\mathbf{r}_i \quad \text{(A4)}$$

by using Equation (A2), and the covariance matrix **S** defined by

$$\mathbf{S} = \mathbf{A}\,\mathbf{A}^{\mathrm{T}} = \mathbf{U}\,\mathbf{\Lambda}^{-2}\,\mathbf{U}^{\mathrm{T}}. \quad \text{(A5)}$$

The above two equations establish the connection between our SVD analysis and the algorithm adopted by Chou.[3,5] The Mahalanobis distance,[33] referred to in the latter study, is nothing else than the distance in the singular space defined by Equation (A3).

The singular values $\lambda_i$ of **A** are conventionally written in descending order along the diagonal of $\Lambda$. Equation (1) may be rewritten in terms of the $p < 19$ dominant singular values of **A** as

$$\mathbf{A}_{[19\times n]} = \mathbf{U}_{[19\times p]}\,\mathbf{\Lambda}_{[p\times p]}\,\mathbf{V}^{\mathrm{T}}_{[p\times n]} \quad \text{(A6)}$$

provided that the singular values $\lambda_j$ in the range $j >$ p are negligibly small. In this approximation, the analysis may be conducted in the p-dimensional subspace of residues.