PROTEINS: Structure, Function, and Genetics 25:157–168 (1996)

Prediction of Secondary Structural Content of Proteins From Their Amino Acid Composition Alone. I. New Analytic Vector Decomposition Methods

Frank Eisenhaber,[1,2] Federica Imperiale,[2,3] Patrick Argos,[2] and Cornelius Frömmel[1]
[1]Institut für Biochemie der Charité, Medizinische Fakultät, Humboldt-Universität zu Berlin, D-10098 Berlin-Mitte; [2]European Molecular Biology Laboratory, D-69012 Heidelberg, Germany; [3]Department of Mathematics, University of Genova, I-16132 Genova, Italy

ABSTRACT **The predictive limits of the amino acid composition for the secondary structural content (percentage of residues in the secondary structural states helix, sheet, and coil) in proteins are assessed quantitatively.**

For the first time, techniques for prediction of secondary structural content are presented which rely on the amino acid composition as the only information on the query protein. In our first method, the amino acid composition of an unknown protein is represented by the best (in a least square sense) linear combination of the characteristic amino acid compositions of the three secondary structural types computed from a learning set of tertiary structures. The second technique is a generalization of the first one and takes into account also possible compositional couplings between any two sorts of amino acids. Its mathematical formulation results in an eigenvalue/eigenvector problem of the second moment matrix describing the amino acid compositional fluctuations of secondary structural types in various proteins of a learning set. Possible correlations of the principal directions of the eigenspaces with physical properties of the amino acids were also checked. For example, the first two eigenvectors of the helical eigenspace correlate with the size and hydrophobicity of the residue types respectively.

As learning and test sets of tertiary structures, we utilized representative, automatically generated subsets of Protein Data Bank (PDB) consisting of non-homologous protein structures at the resolution thresholds ≤1.8Å, ≤2.0Å, ≤2.5Å, and ≤3.0Å. We show that the consideration of compositional couplings improves prediction accuracy, albeit not dramatically. Whereas in the self-consistency test (learning with the protein to be predicted), a clear decrease of prediction accuracy with worsening resolution is observed, the jackknife test (leave the predicted protein out) yielded best results for the largest dataset (≤3.0 Å, almost no difference to the self-consistency test!), i.e., only this set, with more than 400 proteins, is sufficient for stable computation of the parameters in the prediction function of the second method.

The average absolute error in predicting the fraction of helix, sheet, and coil from amino acid composition of the query protein are 13.7, 12.6, and 11.4%, respectively with r.m.s. deviations in the range of 8.6÷11.8% for the 3.0 Å dataset in a jackknife test. The absolute precision of the average absolute errors is in the range of 1÷3% as measured for other representative subsets of the PDB.

Secondary structural content prediction methods found in the literature have been clustered in accordance with their prediction accuracies. To our surprise, much more complex secondary structure prediction methods utilized for the same purpose of secondary structural content prediction achieve prediction accuracies very similar to those of the present analytic techniques, implying that all the information beyond the amino acid composition is, in fact, mainly utilized for positioning the secondary structural state in the sequence but not for determination of the overall number of residues in a secondary structural type. This result implies that higher prediction accuracies cannot be achieved relying solely on the amino acid composition of an unknown query protein as prediction input. Our prediction program SSCP has been made available as a World Wide Web and E-mail service. © 1996 Wiley-Liss, Inc.

Key words: protein structure prediction, prediction of secondary structural content, amino acid composition, jackknife analysis

INTRODUCTION

Uneven technological progress in resolving protein sequences (automation and genome projects)

Received June 5, 1995; revision accepted December 14, 1995.
Address reprint requests to Frank Eisenhaber, European Molecular Biology Laboratory, Meyerhofstraße 1, Postfach 10.2209, D-69012 Heidelberg, Germany.

and protein structures (necessity of human interference in crystallization and electron density map interpretation in X-ray crystallography or in the NMR assignment problem) has resulted in a widening gap between increasing sequence knowledge (about 44,000 protein sequences in SWISSPROT[1] in May 1995) and slow accumulation of unique protein folds (only a few hundred, mainly for water soluble globular proteins, in the Protein Data Bank[2,3]). As a result and especially in the last decade, enormous scientific efforts have been made to develop computer methods for prediction of structural (and functional) features of proteins from their amino acid sequences; however, decisive steps towards an ultimate breakthrough have yet to be done (for a review, see Eisenhaber et al.[4]).

The prediction of the secondary structural content (i.e., the percentage of residues in different secondary structural states) might be considered as first step in analyzing a new protein sequence and predicting its tertiary architecture. A similar prediction can be attempted utilizing experimentally determined amino acid compositions. Surprisingly, only a few researchers have directly addressed this problem. Early multiple linear regression analysis[5,6] have been repeated and improved by Muskal and Kim[7] relying on a much larger database of protein structures. The amino acid composition, the molecular weight, and the existence of heme in the protein structure were correlated with the contents of helix and strand. Muskal and Kim[7] have also presented neural network systems for secondary structural contents prediction with the same input parameters. Heme as input parameter has been traditional since Krigbaum and Knutton[6] and reflects a bias in the protein set to globins and cytochromes (heme correlates positively with helix content and negatively with sheet content). Other prosthetic groups were not considered. At the same time, for protein sequences obtained in genome projects and also for experimentally determined amino acid compositions of proteins, it is not obvious which type of cofactor is required for biological function.

The characterization of a protein structure by its secondary structural content is an intermediate level between two extremes of the complete description of the secondary structural states of every residue and of the simple assignment of a secondary structural class (folding types all-α, all-β, α + β, α/β, irregular). The output of secondary structure prediction algorithms may be also interpreted in terms of secondary structural content.[7–10]

Nishikawa et al.[11,12] found that folding type, amino acid composition, biological function (enzyme or non-enzyme), intra- or extracellular location, and the number of disulphide bonds are related. The structural class (folding type) of a protein is mainly defined by thresholds of secondary structural content (see section III.D.2 of Eisenhaber et al.[4]). Subsequently, both analytical distance criteria in the amino acid composition space[13–25] and neural network methods[10,26–28] have been applied for the jury decision between 3–5 folding types (α- and β-proteins, one or two types of mixed structures, and sometimes a group of "irregular" forms). It was found that sequence properties such as simple hydrophobic patterns (used in combination with the amino acid composition) add almost nothing to the accuracy of the structural class prediction.[4,29,30]

Especially after three recent publications,[23–25] the paradoxical situation emerged that secondary structural class (folding type) prediction for an unknown query sequence appears solved using only amino acid composition as input and an elliptically scaled distance as decision criterion (reported prediction accuracies up to 100%[23–25]), whereas secondary structure prediction even with multiple alignments methods reaches only about 70% accuracy.[8,9,31] This apparent progress has also been described in recent reviews.[4,32] At the beginning of this work, we were inspired by the optimistic secondary structural class prediction results based exclusively on the amino acid composition of the query protein. We started to investigate the direct impact of the amino acid composition on the secondary structural content (α-helix, β-sheet, coil) of a protein. Our prediction approach is first to rely purely on the amino acid composition of the query protein. We present two methods for prediction of the secondary structural content, not just for the secondary structural class. Both are based on easily comprehended analytic vector decomposition methods. In contrast to the first method, the second does take into account different weighting of amino acids as well as possible compositional couplings between pairs of two residue types. The average amino acid composition of a secondary structure type together with the characteristic amino acid compositional fluctuations were determined from a learning set of protein structures. We show that the consideration of weighting of amino acid types and of coupling between them does improve prediction accuracy; albeit not greatly.

Surprisingly, the prediction accuracy of our simple approach is similar to that of single sequence secondary structure prediction methods applied for the task of predicting the secondary structural content. This result implies that all information additional to the knowledge of amino acid composition is utilized by these methods mainly for positioning the secondary structural state in the sequence but not for determining the absolute number of residues in a specific secondary structural state. This result implies also that higher prediction accuracies cannot be achieved relying solely on the amino acid composition of an unknown query protein as prediction input. In a subsequent article, we discuss the paradoxical situation with the secondary structure and

secondary structural class prediction success rates and resolve the contradiction.

METHODS

First Method (Without Compositional Couplings)

This simple technique is designed to calculate the secondary structural content of a query protein from its amino acid composition utilizing the amino acid compositions of secondary structural types in known tertiary structures of proteins. Various amino acids have differing propensities for α-helix and β-sheet. Assuming relatively constant amino acid compositions for all secondary structural types (independent of the individual proteins), the overall amino acid composition of a protein can be expressed as their linear combination with specific coefficients corresponding to the fraction of a secondary structural type in the tertiary structure.

Thus, the residue composition of the query protein P (represented by a 20-dimensional vector of occurrence frequencies of amino acid types in the amino acid type fraction space)

$$P = (c_1, c_2, \ldots, c_{20}) \text{ with } \sum_{k=1}^{20} c_k = 1. \quad (1)$$

is approximated by a linear combination

$$P \approx \sum_{i=1}^{n} \alpha_i P_i \quad (2)$$

of the typical amino acid compositions P_i of the $n=3$ secondary structural types ($i=1$ α-helix, $i=2$ β-sheet, $i=3$ coil)

$$P_i = (c_{1,i}, c_{2,i}, \ldots, c_{20,i}) \text{ with } \sum_{k=1}^{20} c_{k,i} = 1 \quad (3)$$

as observed in some learning set of structures. The non-negative coefficients α_i (barycentric coordinates) in this linear combination would describe the weight (the portion) of the corresponding secondary structural type in the overall structure of the query protein.

Barycentric coordinates are a traditional tool for describing the relative influence of a set of n basis vectors P_i ($i=1, \ldots, n$; here $n=3$) spanning a subspace in a space of dimension N ($n \leq N$; here $N=20$) on a vector P in this subspace. If the basis set is complete and non-redundant, scalar coefficients α_i of the linear combination of basis vectors are defined unambiguously in

$$P = \sum_{i=1}^{n} \alpha_i P_i \text{ with } \sum_{i=1}^{n} \alpha_i = 1. \quad (4)$$

Pairs or triples of independent vectors the end point of which span subspaces of a line or a plane respectively are well-known examples in the Euclidean space. If the vector P is not contained in the subspace spanned by the basis vectors, a component decomposition similar to Eq. (3) is still possible in the sense of an optimization criterion, for example, the least square condition

$$Q = \min\left(P - \sum_{i=1}^{n} \alpha_i P_i \right)^2. \quad (5)$$

The minimum has to be determined under the condition of the following three constraints

$$\begin{cases} \sum \alpha_i = 1, \\ \alpha_i = a_i^2, \text{ and} \\ \alpha_i = 1 - b_i^2. \end{cases} \quad (6)$$

The equations [Eq. (6)] express in a formal manner that all coefficients α_i are non-negative and not larger than 1 and that all coefficients sum up to unity. We use the standard method of Lagrange multipliers for the minimization problem in Eq. (5). The symbols λ, λ_i^-, λ_i^+ in the function

$$Q' = \left(P - \sum_{i=1}^{n} \alpha_i P_i \right)^2 + \lambda \left(1 - \sum_{i=1}^{n} \alpha_i \right) + \sum_{i=1}^{n} \left[\lambda_i^- (a_i^2 - \alpha_i) + \lambda_i^+ (1 - b_i^2 - \alpha_i) \right] \quad (7)$$

are Lagrange multipliers for the three types of constraints. Eq. (5) is the basis for our first prediction method. In this approach, each amino acid has the same weighting, and compositional couplings between amino acid types are neglected. The partial derivatives of Q′ with respect to α_i and to the Lagrange multipliers in Eq. (7) are equal to zero. In a mathematically similar way as in the appendix of Zhang and Chou,[16] a linear equation system for the coefficients α_i can be derived. These barycentric coordinates serve as prediction for the secondary structural content. For solving the system of linear equations derived for calculating the minimum of Q′ in Eq. (7), numerical recipes from Press et al.[33] were applied.

Second Method (With Compositional Couplings)

In the following, we motivate a generalization of Eq. (5) to allow different weightings of pairs for amino acid types. Eq. (5) is equivalent with

$$Q = \min\left(\sum_{i=1}^{n} \alpha_i f(P, P_i) \right)^2 \quad (8)$$

using the condition of the sum of coefficients α_i being unity. The symbol f denotes the function $f(A,B) = A - B$. Now, we think about possibly more complicated functional forms of the function f. The amino acid compositions of a secondary structural type in various proteins will be represented by a cloud of points in the 20-dimensional amino acid type frac-

tion space in the vicinity of an average amino acid composition of the secondary structural type considered. This cloud will deviate from a spherical point density if there are characteristic couplings between amino acid types and/or if some amino acid types have different weight for this secondary structural type. As a result, the cloud is better described by a multidimensional ellipsoid. Now, consider the difference vector between a composition point representing a query sequence and the center of the ellipsoid. Its projection on the eigenvectors of the ellipsoid becomes more significant, the smaller the corresponding semi-axis of the ellipsoid (length of the principal axis). Therefore, the lengths of the semi-axes may be used as normalizing measures. This is the essence of the second method.

From a mathematical point of view, the above interpretation is equivalent to first transforming $P-P_i$ from the amino acid type fraction space into the eigenvector space of the second moment matrix $M_i^{(2)}$ of the point cloud of the secondary structural type i, and then scaling all components by the square roots of the corresponding eigenvalues $b_{k,i}^2$, followed transforming the resulting vector back into the amino acid type fraction space. The second moment matrix $M_i^{(2)}$ is obtained as

$$M_i^{(2)} = \sum_{m=1}^{M} (P_i - P_i^{(m)})^T \cdot (P_i - P_i^{(m)}) \qquad (9)$$

where $P_i^{(m)}$ is the composition vector for the secondary structural type i of protein m (m = 1,2, . . . ,M). The eigenvalues of this symmetric matrix are real and non-negative numbers. The function f can be written as

$$f(P, P_i) = \left(\sum_{k=1;k\neq k'}^{20} \Delta c_i \, E_{k,i}^T \, e_k / b_{k,i} \right) U_i \qquad (10)$$

where U_i is the transformation matrix from the eigenvector space to the amino acid fraction space with the eigenvectors $E_{k,i}$ of the second moment matrix $M_i^{(2)}$ as rows, the superscript T denotes matrix transposition, e_k are the unit vectors of the k-th dimension, $b_{k,i}$ are the square roots of the k-th eigenvalues of secondary structural type i, and

$$\Delta c_i = (c_1 - c_{1,i}, c_2 - c_{2,i}, \ldots, c_{20} - c_{20,i}). \qquad (11)$$

The scalar product is executed between Δc_i and $E_{k,i}^T$. The number k′ is the index of the zero eigenvalue corresponding to the eigenvector perpendicular to the "composition" hyperplane. Obviously, the point cloud and the corresponding hyperellipsoid are only 19-dimensional since all points represent compositions (the sum of the coordinates is unity), and the set of points is, therefore, contained in a 19-dimensional "composition" hyperplane. For the derivation of Eq. (10), it should be noted that the second moment matrix of an ellipsoid is inverse to the matrix describing the quadratic form (the analytic equation) of an ellipsoid. For solving the system of linear equations derived for calculating the minimum of Q in Eq. (8), and for computing eigenvectors and eigenvalues of the second moment matrix $M_i^{(2)}$, numerical recipes from Press et al.[33] were applied.

Protein Data Sets and Algorithm Implementation

The algorithm has been implemented in a computer program SSCP (Secondary Structural Content Prediction) written in the programming language C. The values of the average composition vectors P_i, the eigenvectors $E_{k,i}$, and the eigenvalues $b_{k,i}^2$ of the second moment matrix $M_i^{(2)}$ for the secondary structural type i (i=1 α-helix, i=2 β-sheet, i=3 coil) have been computed from representative learning sets of M protein structures with low residue identity amongst all aligned pairs sequences (≤35%). We excluded NMR structures since there is no generally recognized measure of structure quality comparable with the resolution of X-ray crystal structures. We did not consider also structures with incomplete backbone necessary for derivation of the secondary structural state of the residues. The tertiary structure selection from the Brookhaven Protein Data Bank (PDB)[2,3] was done automatically with the program OBSTRUCT.[34] All four protein datasets are available in electronically readable form on request (copy file pdbselection.tar.Z with FTP phenix.EMBL-Heidelberg.DE:pub/ASC.21 or contact F.E. via E-mail Eisenhaber@EMBL-Heidelberg.DE on Internet). Four protein sets with resolution better than or equal to 1.8 Å (M=166), 2.0 Å (M=262), 2.5 Å (M=398), and 3.0 Å (M=475) were studied. At the best resolution level of 1.8 Å, side-chain conformations can be reliably determined.[35] At the worst resolution of 3 Å, secondary structure elements are located only as rigid bodies without clear recognition of their terminal residues in the protein sequence. We emphasize that the protein selection was performed without human interference. Therefore, bias to certain types of proteins other than from the limitations of the PDB itself has been excluded and the protein datasets can be considered as representative for the current state of the PDB.

Definition of Secondary Structure

Secondary structure assignments were made with the standard method of Kabsch and Sander.[36] Their secondary structural types H, G, and I were classified as helix, residues marked with E were considered as being part of sheet. The remaining residues were regarded as coil. This definition is essentially identical with that of Rost and Sander.[9] In contrast, Chou[23] considers only the H-type as helix. In our

opinion, it is justified to treat the 3_{10} structure as helix since it is often appended to existing α-helices.

The learning results of the second method were found very sensitive to noise. Two sources were detected. (i) The Kabsch and Sander method produces an artificially large number of small helices and strands.[37] We reassigned to coil all helices shorter than five residues and all strands shorter than three residues. (ii) Proteins having only very few residues in a certain type of secondary structure produce extremely large fluctuations from the average amino acid composition characteristic for this type of secondary structure and, consequently, introduce a stochastic variation into the respective second moment matrix. Therefore, only proteins which are representative for a certain type of secondary structure were considered for the computation of the second moment matrix: sufficiently long polypeptide chains (with 80 residues or more, this was one of the input requirements for the selection program OBSTRUCT) and with secondary structure fractions above some threshold ($t_\alpha = t_\beta = 15\%$ for α-helix and β-sheet, $t_c = 30\%$ for coil). We tested also two other sets ($t_\alpha = 15\%$ and $t_\beta = 10\%$, $t_\alpha = 20\%$ and $t_\beta = 20\%$). In the first case, the prediction accuracy for the second method was almost as good as for the original thresholds ($t_\alpha = 15\%$ and $t_\beta = 15\%$). In the second variant, the predictions were clearly worse. Nevertheless, the secondary structural content thresholds are not simply parameters for optimization of the prediction method but reflect the distribution of secondary structural contents in real tertiary structures. As can be clearly seen in Figure 3 of the subsequent paper, the distribution of helical structure in proteins has a valley in the region $10 \div 25\%$. A similar minimum for the contents of β-strands is at 10%.

Accuracy Measures

The prediction accuracy of the two methods has been evaluated both (i) with a self-consistency test and (ii) with a "jackknife" procedure applied to our four sets of proteins, where in contrast to the first check, the learning step was performed with all structures of a given dataset except the one for which the secondary structural content is to be predicted. This procedure was repeated for all proteins. The results are presented in Tables I–IV. The integral prediction accuracy over a dataset was estimated with the following measures:

(i) The mean absolute prediction errors Δα, Δβ, and Δcoil and their standard deviations (given in parentheses in Tables I–IV) between real and predicted contents have been calculated for each secondary structural type. For example, in the case of helical structure, the mean value is calculated as

$$\Delta\alpha = M^{-1} \sum_{m=1}^{M} \left| \alpha_{real}^{(m)} - \alpha_{pred}^{(m)} \right|. \quad (12)$$

M denotes the number of proteins, $\alpha_{real}^{(m)}$ and $\alpha_{pred}^{(m)}$ are the real and predicted helix contents of the m-th protein given in percent of the overall number of protein residues. The corresponding standard deviation sd(Δα) can be obtained with

$$sd(\Delta\alpha) = \sqrt{M^{-1} \sum (\alpha_{real}^{(m)} - \alpha_{pred}^{(m)})\,(\alpha_{real}^{(m)} - \alpha_{pred}^{(m)}) - (\Delta\alpha)^2}. \quad (13)$$

(ii) The integral absolute error IE over all three secondary structural types for a given protein was computed as

$$IE = \sqrt{[(\Delta\alpha)^2 + (\Delta\beta)^2 + (\Delta coil)^2]/3}. \quad (14)$$

This measure has been used for the first time in this work.

(iii) The correlation coefficients C_α, C_β, and C_{coil} have been calculated between the two vectors of real and predicted secondary structural contents over a dataset.

(iv) To assess how many proteins are well predicted, we computed the percentage of proteins of the dataset for which the absolute errors Δα, Δβ, Δcoil, or IE are less than 20%. This measure has been used for the first time in this work.

RESULTS AND DISCUSSION

Impact of Resolution and the Number of Proteins

The prediction results based on the self-consistency and the jackknife tests for all four datasets are presented in Tables I–IV. There is some, but generally undramatic, influence of resolution on prediction accuracy.

As judged from the self-consistency test, the prediction results are better with improved resolution (compare columns I in Tables I–IV for the first method and column III in Tables I–IV for the second method). This is especially obvious if the fraction of proteins with an average error IE below 20% is considered. Whereas 88.0 and 89.8% of all proteins of the 1.8 Å dataset are predicted with less than 20% absolute error with the first and the second method respectively, these numbers decrease to 84.0% and 85.7% for the 3.0 Å protein set. We think that at the 3.0 Å resolution level, the terminal regions of secondary structures in the sequence might not be always well assigned, resulting some noise. Additionally, the self-consistency test clearly favors the second method utilizing compositional couplings (compare columns I and III in Tables I–IV). The absolute error decreases by up to 2.4%, the percentage of proteins with less than 20% error increases up to 11.8%, and the correlation coefficients improve up to 0.10.

With full right, the self-consistency test is not considered as appropriate check for the success rate of a prediction method since the information about the protein to be predicted is contained in the learning

TABLE I. Accuracy of Secondary Structural Content Prediction (Dataset 1.8 Å Resolution With 166 Protein Tertiary Structures)*

Prediction method	First (without comp. coupling)		Second (with comp. coupling)	
Column	I	II	III	IV
Test	Self-consistency	Jackknife	Self-consistency	Jackknife
Absolute error in %, mean and standard deviation				
$\Delta\alpha$	14.6 (11.1)	14.7 (11.2)	12.6 (10.3)	14.3 (11.7)
$\Delta\beta$	11.5 (9.3)	11.7 (9.5)	10.8 (9.1)	12.3 (10.4)
Δcoil	13.1 (9.7)	13.2 (9.8)	10.7 (8.0)	12.0 (8.9)
IE	11.7 (6.8)	11.9 (6.9)	10.5 (6.8)	11.9 (7.6)
Percentage of all proteins with maximally 20% absolute error				
α	71.7	70.5	77.7	71.7
β	80.7	80.1	82.5	77.7
coil	75.9	75.9	86.7	83.7
IE	88.0	86.7	89.8	85.5
Correlation between real and predicted content				
C_α	0.77	0.76	0.77	0.71
C_β	0.60	0.58	0.65	0.54
C_{coil}	0.58	0.57	0.67	0.57

*The 1.8 Å resolution protein structure selection comprises 166 proteins. Applying the classification criteria of Nakashima et al.,[14] 41 and 54 structures are α-proteins and β-proteins respectively. Seventy proteins belong to the mixed class. One protein is irregular.

We present the prediction accuracies utilizing method 1 (without amino acid compositional couplings) and method 2 (with amino acid compositional couplings) for both the self-consistency test (with the predicted protein included into the learning set of structures) and the jackknife test (the predicted protein has not been included into the learning set).

The prediction accuracy measures are described in detail in the section "Accuracy Measures" of the "Methods." The equations (11) and (12) show how the absolute prediction error and its standard deviation are calculated. IE is the integrated error over all three secondary structural type [Eq. (14)] averaged over all proteins of the dataset.

set. A good compromise between the demand of non-overlapping learn and test set of structures and the limited number of non-related tertiary structures in the PDB is the so-called jackknife test. All proteins of a given structure selection are used for learning except the single one that is to be predicted. It is not surprising that there is little difference in accuracy for the first method with or without applicating the jackknife procedure (compare columns I and II in Tables I–IV). The average amino acid compositions of the three secondary structural types (altogether three types of structure × 20 vector components − 3 constraints = 57 independent parameters) are quite stable beginning with learning sets of about 100 randomly selected non-homologous proteins. For the second method, the difference between the self-consistency and jackknife tests is much more pronounced (compare columns III and IV in Tables I–IV) since many more independent parameters need be determined (19 unitary eigenvectors and 19 eigenvalues in addition to the average composition for each secondary structural type = 247 independent parameters). The "jackknife" procedure results in a 0–2% increase both in the average absolute error and its root mean square deviation; it also reduces the percentage of proteins within the 20% error margin up to 6%. The influence of the "jackknife" procedure diminishes with the number of proteins in the learning set and almost vanishes also in the case of the 2nd method for the 3.0 Å dataset. At very high resolution (1.8 Å), the number of proteins is not sufficiently large to explore the composition space. Therefore, the orientations of the eigenspaces computed from the second moment matrices are not stable as revealed by the jackknife test; in other words, there is memorization towards the small learning set. For the jackknife test, the 3.0 Å dataset has best results for average absolute prediction errors. Surprisingly, the larger error in structure determination due to lower resolution is more than compensated by the better independence of the parameters in the prediction function from the different learning sets (consisting of all proteins of the 3.0 Å selection except the varying protein that is to be predicted) due to the larger number of non-related proteins.

We do not see clear preferences such that one type of secondary structure might be better predicted than any other. Although the absolute error for helix content is somewhat larger than that of sheet or coil contents, the correlation coefficients are best for helix content.

We also checked which proteins were grossly mispredicted. The prediction algorithm has problems with unusual amino acid compositions for given secondary structural types. For example, the protein 1LIS (fertilization protein from *Haliotis rufescens*) contains seven glycines of which six are in helix. Both vector decomposition methods largely underpredict the α-content. Considerable prediction errors also occur if the structure is strongly influenced by non-polypeptide components; for example,

TABLE II. Accuracy of Secondary Structural Content Prediction (Dataset 2.0 Å Resolution With 262 Protein Tertiary Structures)*

Prediction method	First (without comp. coupling)		Second (with comp. coupling)	
Column	I	II	III	IV
Test	Self-consistency	Jackknife	Self-consistency	Jackknife
Absolute error in %, mean and standard deviation				
$\Delta\alpha$	14.4 (11.7)	14.5 (11.7)	13.2 (10.9)	14.2 (11.5)
$\Delta\beta$	11.8 (9.6)	12.0 (9.7)	11.4 (9.6)	12.4 (10.4)
Δcoil	12.6 (9.7)	12.7 (9.8)	10.7 (7.6)	11.3 (8.3)
IE	11.8 (7.2)	11.9 (7.3)	11.0 (7.2)	11.8 (7.7)
Percentage of all proteins with maximally 20% absolute error				
α	71.8	71.8	76.0	74.0
β	83.2	82.4	85.9	82.4
coil	79.8	79.4	88.5	86.6
IE	86.3	85.9	88.5	85.1
Correlation between real and predicted content				
C_α	0.70	0.70	0.72	0.68
C_β	0.53	0.52	0.57	0.49
C_{coil}	0.54	0.53	0.62	0.56

*The 2.0 Å resolution protein structure selection comprises 262 proteins. Applying the classification criteria of Nakashima et al.,[14] 55 and 78 structures are α-proteins and β-proteins, respectively. One hundred and twenty-seven proteins belong to the mixed class. Two proteins are irregular.

We present the prediction accuracies utilizing method 1 (without amino acid compositional couplings) and method 2 (with amino acid compositional couplings) for both the self-consistency test (with the predicted protein included into the learning set of structures) and the jackknife test (the predicted protein has not been included into the learning set). For the description of the accuracy measures, see legend of Table I and the section "Accuracy Measures" of the "Methods."

in the case of proteins with iron-sulfur complexes (e.g., 1FDD ferredoxin mutant from *Azotobacter vinelandii,* 1HIP high potential iron protein from *Chromatium vinosum*).

Prediction Improvement by Residue Type Weighting and Compositional Couplings

Method 2 which considers weighting of amino acid types and compositional couplings in secondary structures achieves better predictions if absolute errors $\Delta\alpha$, Δcoil, and IE are considered, albeit the improvement is not dramatic. For other error measures (for example, absolute error for sheet $\Delta\beta$ and the correlation coefficients C_α and C_β), the first method is sometimes advantageous. We conclude that weighting and compositional couplings have a modulating role, the average amino acid composition of the secondary structural types being the primary factor. It is quite improbable that the learning data are insufficient to reveal the compositional effects since changes in prediction accuracy are similar in size across all four datasets.

The possibility to find physical interpretations of the parameters used in the prediction function is a considerable advantage of analytic methods compared with neural network techniques.[7,9,10] We wanted to see which physical properties of the amino acid types are responsible for the principal coordinate axes in the eigenspaces corresponding to the three structural types α-helix, β-sheet, and coil. Therefore, we computed correlation coefficients of the eigenvectors with the scales of more than 220 amino acid properties enlisted in the database of Nakai et al.[38] For each eigenvector, we listed the 10 highest scoring amino acid properties. In general, the correlation coefficients are not very high in absolute value (mostly 0.4–0.7, only a few up to 0.9) and the results are often not easy to interpret if many different properties rank among the top ten. We focused our attention on cases when many scales (five or more) pointing to a similar property appear for all datasets.

Large eigenvalues correspond to properties tolerated in a secondary structural type. For example, the main principal axes in the α-helical eigenspace correlates with the size of the amino acids (e.g., −0.6 with the radius of gyration of side chain[39]). Therefore, small amino acids are coupled with other small ones while large amino acids have a tendency to occur together with other large residues in a given helix. Obviously, helices have a tendency to be cylindrical. The radius of individual helices may be different, but the amino acids prefer to have similar size in one helix. This property explains the success of the planar "helical lattice superposition model" for helix docking.[40] The second principal axis in the α-helical eigenspace corresponds to solvation properties of the amino acids (e.g., 0.6 with the hydration number in[41]).

Small eigenvalues denote properties with small tolerance ranges. For example, the lowest non-zero eigenvalue of the β-eigenspace corresponds to the

TABLE III. Accuracy of Secondary Structural Content Prediction (Dataset 2.5 Å Resolution With 398 Protein Tertiary Structures)*

Prediction method	First (without comp. coupling)		Second (with comp. coupling)	
Column	I	II	III	IV
Test	Self-consistency	Jackknife	Self-consistency	Jackknife
Absolute error in %, mean and standard deviation				
$\Delta\alpha$	15.2 (12.1)	15.2 (12.1)	14.2 (11.6)	14.8 (12.1)
$\Delta\beta$	12.0 (10.1)	12.1 (10.2)	12.4 (10.5)	13.1 (11.0)
Δcoil	12.7 (9.6)	12.8 (9.7)	11.0 (8.2)	11.3 (8.5)
IE	12.2 (7.7)	12.2 (7.7)	11.8 (7.9)	12.3 (8.2)
Percentage of all proteins with maximally 20% absolute error				
α	69.6	69.3	74.9	73.1
β	81.7	80.9	80.9	79.4
coil	82.2	82.4	86.7	85.9
IE	83.9	83.7	86.2	84.7
Correlation between real and predicted content				
C_α	0.67	0.66	0.67	0.64
C_β	0.45	0.44	0.45	0.40
C_{coil}	0.57	0.56	0.62	0.59

*The 2.5 Å resolution protein structure selection comprises 398 proteins. Applying the classification criteria of Nakashima et al.,[14] 84 and 103 structures are α-proteins and β-proteins, respectively. Two hundred and six proteins belong to the mixed class. Five proteins are irregular.

We present the prediction accuracies utilizing method 1 (without amino acid compositional couplings) and method 2 (with amino acid compositional couplings) for both the self-consistency test (with the predicted protein included into the learning set of structures) and the jackknife test (the predicted protein has not been included into the learning set). For the description of the accuracy measures, see legend of Table I and the section "Accuracy Measures" of the "Methods."

direction correlating with turn- and coil-propensities. In the case of the α-helical and coil-eigenspace, this axis selects the tryptophan content. It is known from a SWISSPROT analysis that tryptophan is the rarest amino acid in proteins.[42]

Comparison With Prediction Accuracies for Secondary Structural Content From Other Methods

Our techniques for secondary structural content prediction are the first which rely only on the knowledge of amino acid composition of a query protein. Other methods described in the literature require more information, and a higher prediction accuracy should be expected. In Table V, we present results of a literature survey. Accuracies in secondary structural content predictions were obtained with many approaches: theoretical (random prediction RAN); multiple linear regression with composition, heme, and molecular weight (MLR[7]); single chain secondary structure prediction methods (GORIII,[43] COMBI,[44] QS[45]); network predictions (NHN and TN,[7] PHD3[9]) with substantially more input information in addition to amino acid composition); and experimental (UV-CD[46]) methods. The accuracy of the various prediction methods is difficult to estimate since it depends (i) on the type of input information for the prediction function and (ii) on the number and selection of protein structures in the learning and test sets. We grouped the methods in Table V by the type of input information on the query protein required. We describe, also, the protein test sets used and the validation procedure, if any is reported.

The results in Tables I–IV (columns III and IV) can be used to estimate fluctuations in the prediction accuracies due to different representative protein datasets and test protocols. The range of the absolute error is about 2% and that of the correlation coefficients up to 0.25. These values may be larger if small, non-representative test sets have been used. For example, we removed only 20 badly predicted proteins from the 2.0 Å dataset (only about 8% of the whole dataset); the remaining 242 proteins constitute still an impressively large set of non-homologous structures. In this case, the prediction results improve drastically both in the self-consistency and in the jackknife test. The absolute errors for α- and β-contents are 12 and 11%, respectively, with standard deviations of about 9%. The 20% error margin is not exceeded for more than 90% of all proteins.

As a tendency, smaller test sets will generally result in better prediction accuracies. There is always the theoretical possibility that just the unfavorable proteins have been missed and, therefore, the prediction accuracies are overestimated. This difficulty can also not be avoided with any crossvalidation such as jackknifing. Thus, it is extremely important that success rates are calculated from representative test sets, as large as possible, which contain all or most of the structural variants known today.

TABLE IV. Accuracy of Secondary Structural Content Prediction (Dataset 3.0 Å Resolution With 475 Protein Tertiary Structures)*

Prediction method	First (without comp. coupling)		Second (with comp. coupling)	
Column	I	II	III	IV
Test	Self-consistency	Jackknife	Self-consistency	Jackknife
Absolute error in %, mean and standard deviation				
$\Delta\alpha$	14.7 (11.9)	14.7 (12.0)	13.3 (11.4)	13.7 (11.8)
$\Delta\beta$	12.0 (9.9)	12.1 (10.0)	12.0 (10.4)	12.6 (10.8)
Δcoil	12.7 (9.3)	12.8 (9.4)	11.2 (8.4)	11.4 (8.6)
IE	12.0 (7.5)	12.0 (7.5)	11.4 (7.7)	11.8 (8.0)
Percentage of all proteins with maximally 20% absolute error				
α	70.7	70.9	76.2	74.7
β	80.6	80.4	81.1	79.4
coil	81.5	81.5	84.8	83.8
IE	84.0	84.0	85.7	84.6
Correlation between real and predicted content				
C_α	0.69	0.69	0.70	0.68
C_β	0.49	0.48	0.48	0.44
C_{coil}	0.57	0.57	0.60	0.60

*The 3.0 Å resolution protein structure selection comprises 475 proteins. Applying the classification criteria of Nakashima et al.,[14] 99 and 140 structures are α-proteins and β-proteins, respectively. Two hundred and thirty-two proteins belong to the mixed class. Four proteins are irregular.

We present the prediction accuracies utilizing method 1 (without amino acid compositional couplings) and method 2 (with amino acid compositional couplings) for both the self-consistency test (with the predicted protein included into the learning set of structures) and the jackknife test (the predicted protein has not been included into the learning set). For the description of the accuracy measures, see legend of Table I and the section "Accuracy Measures" of the "Methods."

Even in this case, the prediction accuracies should be considered only as upper boundaries given the current limitations of the PDB.

Accepting the error intervals obtained above, the methods listed in Table V cluster clearly in accordance with the prediction accuracy in three groups:

1. TN and PDH3;
2. MLR, GORRIII, COMBI, QS, NHN, our two methods, and also the spectroscopic techniques; and
3. the random guess, which is obviously worse than any other prediction approach.

The prediction errors for the contents of helix and strand are in the range of $10.2 \div 13.2\%$ with standard deviations between 9.2 and 12.0% for the single sequence secondary structure prediction methods GORIII, COMBI, and QS (see Table II). Surprisingly, the prediction accuracy of our simple approach, relying only on composition of the query protein, is comparable. We can conclude that all sequence information, in addition to the knowledge of amino acid composition, is utilized by these methods mainly for positioning the secondary structural state in the sequence, but not for determining the absolute number of residues in a specific secondary structural state. *This result implies also that a further improvement of the accuracy of secondary structural content prediction, based purely on the knowledge of amino acid content of the query protein, is very unlikely.* Better predictions would require more input information.

Our method is also comparable with MLR, NHN, and simple spectroscopic methods (UV-CD). The prediction errors are about 5% better for PHD3. In this case, information from multiple alignments with the query sequence is used as input information. This has proven to give a significant increase in the accuracy of secondary structure prediction.[8,9,31] The tandem network TN[7] uses molecular weight and heme as input in addition to amino acid composition. The very high accuracy, as reported by Muskal and Kim,[7] seems suspicious since it is much better than that of PHD3, which uses more input information and it is also better than that of our two methods which are similar in input information to TN (except of molecular weight and heme). The reported prediction accuracy of the TN technique appears to be not representative since only 15 proteins have been included in the test set. It is well known that the prediction errors vary widely among small subsets of protein structures.

An error margin of 10% in secondary structural content prediction is considered desirable for reliable folding type prediction.[7–9] As seen from Table II, all secondary structure content prediction methods, except PHD3, do not reliably approach this value. This is partly also the case for experimental methods.[46–51] A higher accuracy of content predictions appears to need more sequence information in addition to the amino acid composition.

It is worth noting that some secondary structure prediction techniques do not exploit the information

TABLE V. Comparison of the Accuracies in Secondary Structural Content Prediction Obtained by Various Theoretical and Experimental Methods

Input information	Prediction method	Proteins in test set	Statistical significance	Δα	Δβ	Δcoil	C_α	C_β	C_{coil}
Random	RAN*	—	—	32.1 (20.8)	21.3 (14.5)	—	−0.36	−0.22	—
Single sequence	GORIII*	124	—	11.3 (9.4)	10.6 (9.8)	—	0.78	0.46	—
	COMBI*	124	—	11.2 (9.0)	13.2 (10.6)	—	0.83	0.51	—
	QS‡	104	—	12.5 (12.0)	10.2 (9.2)	—	—	—	—
AAC, MW, heme**	MLR‡	104	Jackknife	12.9 (11.1)	9.0 (8.5)	—	—	—	—
	NHN‡	104	Jackknife	12.5 (10.9)	9.6 (8.5)	—	—	—	—
	TN‡	15	—	5.0 (3.4)	5.6 (4.9)	—	—	—	—
Multiple alignment, AAC, NAA**	PHD3*	124	Partly cross-validated	7.8 (6.8)	7.3 (7.9)	—	0.91	0.73	0.68†
Spectroscopy	UV-CD§	22	—	—	—	—	0.84	0.39	0.56
AAC**	This work method 1¶	475	Self-consistency	14.7 (11.9)	12.0 (9.9)	12.7 (9.3)	0.69	0.49	0.57
			Jackknife	14.7 (11.9)	12.1 (10.0)	12.8 (9.4)	0.70	0.48	0.57
	This work method 2¶	475	Self-consistency	13.3 (11.4)	12.0 (10.4)	11.2 (8.4)	0.70	0.48	0.62
			Jackknife	13.7 (11.8)	12.6 (10.8)	11.4 (8.6)	0.68	0.44	0.60

*Data from Rost and Sander.[9] The performance of the prediction techniques with respect to the 124 proteins in their set 2 is shown. RAN is a random prediction. GORIII and COMBI are the secondary structure prediction methods of Gibrat et al.[43] and Biou et al.,[44] respectively. PHD3 is the three-level neural network predictor of Burkhard Rost. All methods use local sequence information from the query sequence as input. In the case of PHD3, the method relies also on multiple alignments of the query sequence with other sequences.

†The correlation coefficient between the real and predicted values of coil content was taken from Rost and Sander.[8]

‡Data from Muskal and Kim (1992).[7] The calculated prediction accuracies are shown with respect to their set of 104 protein structures in OPTBASE. QS is the secondary structure prediction network technique of Qian and Sejnowski[45] which relies on local sequence information from the query sequence. All the other methods, MLR, NHN, and TN, use amino acid composition, the molecular weight, and the existence of heme as input. MLR is a multiple linear regression analysis. NHN represents a network technique without hidden nodes. The prediction accuracies for MLR and NHN were calculated with a "jackknife" procedure. TN is a tandem network system to suppress memorization effects. Since the latter was trained on OPTBASE, we show here the prediction accuracies of TN for EVALBASE, a set of 15 proteins with low homology to OPTBASE.

§Data from Perczel et al.[46] with CD measurements. The correlation coefficients for parallel and antiparallel sheet were 0.37 and 0.41, respectively. The value for β-structure in the table is the arithmetical average of both.

¶Results from this work. Prediction accuracies shown for the protein set with 3.0 Å resolution have been calculated with a self-consistency test and a jackknife procedure.

**Abbreviations: AAC, amino acid composition; MW, molecular weight.

additional to amino acid composition very efficiently. It was shown by Zhu[52] that many single sequence secondary structure prediction methods, which achieve about 60% accuracy in the per residue prediction of secondary structural state, show poor performance in locating and identifying segments of secondary structure in the sequence.

On the Impact of Amino Acid Composition for Protein Structure

We can conclude that the secondary structural content of a protein is determined mainly by the amino acid composition (for about 80% of all proteins within the absolute error margin of 20%). At the same time, the influence of amino acid composition is far from complete. Other sequence properties play a minor but important role.

Several other experimental and theoretical findings suggest that only certain sequence properties are important for the protein structure, not the whole sequence *per se*. Small-scale mutations in protein sequences result mostly only in local structural changes.[53] It is known that the order of secondary structural elements does not determine the protein fold for some proteins but does influence folding kinetics or, maybe, stability (see Viguera et al.,[54] Predki and Regan[55] and references therein). The plot of the occurrence frequency of pairs of amino acid types vs. distance between their C_β positions in protein crystal structures shows a dependency on sequential separations only up to about 20 residues.[56] Evolutionary trees and sequence/structure correlations derived from similarities in the distribution of sequence words (fragments) agree well with results from full sequence alignment methods.[57–59] Molten globular folds containing secondary structural elements and displaying globularity but lacking specific side chain interactions and packing have been frequently observed.[60] From the viewpoint of tertiary interactions between residues and/or oligopeptide segments, the linear covalent link of the backbone seems not to be essential for their specific association. Such insight may help to understand the degeneracy observed when moving

from the sequence space to the structure space of proteins.

AVAILABILITY OF PROGRAMS

Interested readers can try the algorithms described in this work on their own query sequences. The computer program SSCP has been made available as a World Wide Web service. Please find the hyperlink to SSCP on

http://www.embl-heidelberg.de/~eisenhab/

On input of amino acid sequence or composition, the prediction of secondary structural content and secondary structural class (folding type) is returned. The prediction program can also obtained together with a file containing the results of learning from a protein dataset. The program SSCP is also available as E-mail server. Send an E-mail to SSCP@ EMBL-Heidelberg.DE with the subject HELP. Please contact F.E. by E-mail (Eisenhaber@EMBL-Heidelberg.DE) or by normal post for remarks and suggestions.

ACKNOWLEDGMENTS

The authors thank Jaap Heringa for the determination of sets of non-homologous protein structures from the PDB with his program OBSTRUCT. The discussions with Burkhard Rost and Dirk Walther are gratefully acknowledged. The authors are grateful for financial support from the Fund "Wissenschaftler-Integrationsprogramm" (grant 020386/B to F.E.), jointly administered by the East-German Länder and the government of the Federal Republic of Germany, and from the European Union (grant 18/94 to F.I. by Comett.Li.Sa, Genova).

REFERENCES

1. Bairoch, A., Boeckmann, B. The SWISS-PROT protein sequence data bank: Current status. Nucleic Acids Res. 22: 3578–3580, 1994.
2. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. Protein data bank: A computer based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
3. 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/Chester: Data Commission of the International Union of Crystallography, 1987:107–132.
4. Eisenhaber, F., Persson, B., Argos, P. Protein structure prediction: Recognition of primary, secondary, and tertiary structural features from amino acid sequence. Crit. Rev. Biochem. Mol. Biol. 30:1–94, 1995.
5. Davies, D. A correlation between amino acid composition and protein structure. J. Mol. Biol. 9:605–609, 1964.
6. Krigbaum, W.R., Knutton, S.P. Prediction of the amount of secondary structure in a globular protein from its amino acid composition. Proc. Natl. Acad. Sci. U.S.A. 70:2809–2813, 1973.
7. Muskal, S.M., Kim, S.-H. Predicting protein secondary structure content: A tandem neural network approach. J. Mol. Biol. 225:713–727, 1992.
8. Rost, B., Sander, C. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232:584–599, 1993.
9. Rost, B., Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19:55–72, 1994.
10. Chandonia, J.-M., Karplus, M. Neural networks for secondary structure and structural class predictions. Protein Sci. 4:275–285, 1995.
11. 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.
12. 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.
13. Sheridan, R.P., Dixon, J.S., Venkataraghavan, R., Kuntz, I.D., Scott, K.P. Amino acid composition and hydrophobicity patterns of protein domains correlate with their structures. Biopolymers 24:1995–2023, 1985.
14. 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.
15. Chou, P.Y. "Prediction of Protein Structural Classes From Amino Acid Composition, Prediction of Protein Structure." Fasman, G.D. (ed.). New York: Plenum Press, 1989:549–586.
16. Zhang, C.-T., Chou, K.-C. An optimization approach to predicting protein structural class from amino acid composition. Protein Sci. 1:401–408, 1992.
17. Chou, K.-C., Zhang, C.-T. A correlation-coefficient method to predicting protein-structural classes from amino acid composition. Eur. J. Biochem. 207:429–433, 1992.
18. Zhou, G., Xu, X., Zhang, C.-T. A weighting method for prediction of protein structural class from amino acid composition. Eur. J. Biochem. 210:747–749, 1992.
19. Chou, K.-C., Zhang, C.-T. A new approach to prediction protein folding types. J. Protein Chem. 12:169–178, 1993.
20. Mao, B., Chou, K.-C., Zhang, C.-T. Protein folding classes: A geometric interpretation of the amino acid composition of globular proteins. Protein Eng. 7:319–330, 1994.
21. Chou, K.-C., Zhang, C.-T. Predicting folding types by distance functions that make allowance for amino acid interactions. J. Biol. Chem. 269:22014–22020, 1994.
22. Chou, K.-C. Does the folding type depend on its amino acid composition? FEBS Lett. 363:127–131, 1995.
23. 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.
24. Zhang, C.-T., Chou, K.-C. An eigenvalue-eigenvector approach to predicting protein folding types. J. Protein Chem. 14:309–326, 1995.
25. Chou, K.-C., Zhang, C.-T. Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30:275–349, 1995.
26. Metfessel, B.A., Saurugger, P.N., Connelly, D.P., Rich, S.S. Cross-validation of protein structural class prediction using statistical clustering and neural networks. Protein Sci. 2:1171–1182, 1993.
27. Reczko, M., Bohr, H., Subramaniam, S., Pamigighantam, S., Hatzigeorgiou, A. "Fold-Class Prediction by Neural Networks, Protein Structure by Distance Analysis." Bohr, H., Brunak, S. (eds.). Amsterdam, Tokyo: IOS Press, Ohmsha, 1994:277–286.
28. Reczko, M., Bohr, H. The DEF data base of sequence based protein fold class predictions. Nucleic Acids Res. 22:3616–3619, 1994.
29. Klein, P. Prediction of protein structural class by discriminant analysis. Biochim. Biophys. Acta 874:205–215, 1986.
30. Kneller, D.G., Cohen, F.E., Langridge, R. Improvements in protein secondary structure prediction by enhanced neural networks. J. Mol. Biol. 214:171–182, 1990.
31. Levin, J.M., Pascarella, S., Argos, P., Garnier, J. Quantification of secondary structure prediction improvement using multiple alignments. Protein Eng. 6:849–854, 1993.
32. Barton, G.J. Protein secondary structure prediction. Curr. Opin. Struct. Biol. 5:372–376, 1995.
33. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P. "Numerical Recipes in C. The Art of Scientific Computing." Cambridge: Cambridge University Press, 1992.
34. Heringa, J., Sommerfeldt, H., Higgins, D., Argos, P. OB-

STRUCT: A program to obtain largest cliques from a protein sequence set according to structural resolution and sequence similarity. Comput. Appl. Biosci. 8:599–600, 1992.
35. Schrauber, H., Eisenhaber, F., Argos, P. Rotamers: To be or not to be? An analysis of amino acid side-chain conformations in globular proteins. J. Mol. Biol. 230:592–612, 1993.
36. Kabsch, W., Sander, C. Dictionary of protein secondary structures: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
37. Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B., Mornon, J.-P. Comparison of three algorithms for the assignment of secondary structure in proteins: The advantage of a consensus assignment. Protein Eng. 6:377–382, 1993.
38. Nakai, K., Kidera, A., Kanehisa, M. Cluster analysis of amino acid indices for prediction of protein structure and function. Protein Eng. 2:93–100, 1988.
39. Levitt, M. A simplified representation of protein conformations for rapid simulation of protein folding. J. Mol. Biol. 104:59–107, 1976.
40. Walther, D., Eisenhaber, F., Argos, P. Principles of helix-helix packing in proteins: The helical lattice superposition model. J. Mol. Biol. 255:536–553, 1996.
41. Hopfinger, A.J. Polymer-solvent interactions for homopolymers in aqueous solutions. Macromolecules 4:731–737, 1971.
42. Barrai, I., Volinia, S., Scapoli, C. The usage of oligopeptides in proteins correlates negatively with molecular weight. Int. J. Pept. Protein Res. 45:326–331, 1995.
43. Gibrat, J.-F., Garnier, J., Robson, B. Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. J. Mol. Biol. 198:425–443, 1987.
44. Biou, V., Gibrat, J.F., Levin, J.M., Robson, B., Garnier, J. Secondary structure prediction: Combination of three different methods. Protein Eng. 2:185–191, 1988.
45. Qian, N., Sejnowski, T.J. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. 202:865–884, 1988.
46. Perczel, A., Park, K., Fasman, G.D. Deconvolution of the circular dichroism spectra of proteins: The circular dichroism spectra of antiparallel β-sheet in proteins. Proteins 13:57–69, 1992.
47. Garnier, J., Salesse, R., Rerat, B., Rerat, C., Blake, C. Comparison of X-ray data to estimated secondary structures from amino acid sequence and circular dichroism of human prealbumin. J. Chim. Phys. 73:1018–1023, 1976.
48. Bussian, B.M., Sander, C. How to determine protein secondary structure in solution by Raman spectroscopy: Practical guide and test case DNase I. Biochemistry 28:4271–4277, 1989.
49. Johnson, W.C., Jr. Protein secondary structure and circular dichroism: A practical guide. Proteins 7:205–214, 1990.
50. Jackson, M., Mantsch, H.H. The use and misuse of FTIR spectroscopy in the determination of protein structure. Crit. Rev. Biochem. Mol. Biol. 30:95–120, 1995.
51. Pancoska, P., Bitto, E., Janota, V., Urbanova, M., Gupta, V.P., Keiderling, T.A. Comparison of and limits of accuracy for statistical analysis of vibrational and electronic circular dichroism spectra in terms of correlations to and predictions of protein secondary structure. Protein Sci. 4:1384–1401, 1995.
52. Zhu, Z.-Y. A new approach to the evaluation of protein secondary structure predictions at the level of elements of secondary structure. Protein Eng. 8:103–108, 1995.
53. Heinz, D.W., Baase, W.A., Zhang, X.-J., Blaber, M., Dahlquist, F.W., Matthews, B.W. Accomodation of amino acid insertions in an α-helix of T4-lysozyme. J. Mol. Biol. 236: 869–886, 1994.
54. Viguera, A.R., Blanco, F.J., Serrano, L. The order of secondary structural elements does not determine the structure of a protein but does affect its folding kinetics. J. Mol. Biol. 247:670–681, 1995.
55. Predki, P.F., Regan, L. Redesigning the topology of a four-helix bundle protein: Monomeric rop. Biochemistry 34: 9834–9839, 1995.
56. Casari, G., Sippl, M.J. Structure-derived hydrophobic potential. Hydrophobic potential derived from X-ray structures of globular proteins are able to identify native folds. J. Mol. Biol. 224:725–732, 1992.
57. Blaisdell, B.E. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc. Natl. Acad. Sci. U.S.A. 83:5155–5159, 1986.
58. van Heel, M. A new family of powerful multivariate statistical sequence analysis (MSSA) techniques. J. Mol. Biol. 216:877–887, 1992.
59. Rahman, R.S., Rackovsky, S. Protein sequence randomness and sequence/structure correlations. Biophys. J. 68: 1531–1539, 1995.
60. Ptitsyn, O.B. Structure of folding intermediates. Curr. Opin. Struct. Biol. 5:74–78, 1995.