Prediction of Secondary Structural Content of Proteins From Their Amino Acid Composition Alone. II. The Paradox With Secondary Structural Class

Frank Eisenhaber,[1,2] Cornelius Frömmel,[1] and Patrick Argos[2]
[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*

ABSTRACT The success rates reported for secondary structural class prediction with different methods are contradictory. On one side, the problem of recognizing the secondary structural class of a protein knowing only its amino acid composition appears completely solved by simply applying jury decision with an elliptically scaled distance function. Chou and coworkers repeatedly (see Crit. Rev. Biochem. Mol. Biol. 30:275–349, 1995) published prediction accuracies near 100%. On the other hand, traditional secondary structure prediction techniques achieve success rates of about 70% for the secondary structural state per residue and about 75% for structural class only with extensive input information (full sequence of the query protein, its amino acid composition and length, multiple alignments with homologous sequences).

In this article, we resolve the paradox and consider (1) the question of the secondary structural class definition, (2) the role of the representativity of the test set of protein tertiary structure for the current state of the Protein Data Bank (PDB); and (3) we estimate the real impact of amino acid composition on secondary structural class. We formulate three objective criteria for a reasonable definition of secondary structural classes and show that only the criterion of Nakashima et al. (J. Biochem. 99:153–162, 1986) complies with all of them. Only this definition matches the distribution of secondary structural content in representative PDB subsets, whereas other criteria leave many proteins (up to 65% of all PDB entries) simply unassigned.

We review critically specialized secondary-structural class prediction methods, especially those of Chou and coworkers, which claim almost 100% accuracy using only amino acid composition, and resolve the paradox that these prediction accuracies are better than those from secondary structure predictions from multiple alignments. We show (i) that these techniques rely on a preselection of test sets which removes irregular proteins and other proteins without any class assignment (about 35% of all PDB entries); and (ii) that even for preselected representative test sets, the success rate drops to 60% and lower for a 4-type classification (α, β, α + β, α/β). The prediction accuracies fall to about 50% if the secondary structural class definition of Nakashima et al. is applied and only few irregular proteins are preselected and removed from automatically generated, representative subsets of the PDB.

We have applied two new vector decomposition methods for secondary structural content prediction from amino acid composition alone, with and without consideration of amino acid compositional coupling in the learning set of tertiary structures respectively, to the problem of class prediction and achieve about 60% correct assignment among four classes (α, β, mixed, irregular) as well as single sequence-based secondary structure prediction methods like GORIII and COMBI. Our results demonstrate that 60% correctness is the upper limit for a 4-type class prediction from amino acid composition alone for an unknown query protein and that consideration of compositional coupling does not improve the prediction success. The prediction program SSCP offering secondary structural class assignment for query compositions and sequences has been made available as a World Wide Web and E-mail service.

Key words: **protein structure prediction, prediction of secondary structural class, prediction of folding type, amino acid composition, jackknife analysis**

INTRODUCTION

In the hierarchy of prediction methods, secondary structural class prediction (secondary structure above or below a threshold with classification into folding types all-α, all-β, mixed types (sometimes subclassified in α + β, α/β and irregular forms) cor-

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

responds to a lower level and appears a simpler task compared with secondary structural content prediction (fraction of residues in the three states helix, sheet, and coil) or even traditional secondary structure prediction (state of every residue).

Nishikawa et al.[1,2] discovered that the amino acid composition is tightly related with secondary structural class. Subsequently, both analytical distance criteria in the amino acid composition space[3–15] and neural network methods.[16–19] have been applied for the injury decision between 3–5 folding types.

The output of secondary structure content and secondary structure prediction techniques can also be read in terms of secondary structural class. Especially after three recent publications,[13–15] the paradoxical situation emerged that folding type prediction appears solved (reported prediction accuracies up to 100%) whereas secondary structure prediction, even with multiple alignments, approaches only about 70% accuracy and the success rate of its class prediction is only near 75%.[20–22] This contradiction was also not critically considered in recent reviews.[23,24] In this work, we solve this contradiction by showing (i) that certain structural class definitions leave many proteins with intermediate secondary structural content without assignment to any class (predictions are made only for proteins with extreme contents); (ii) that analytical distance-based jury decision methods (Euclidean distance method,[4] Hamming distance method,[5] vector projection method,[9] Mahalanobis distance method[11–15]) yield only prediction accuracies up to 55% for representative test sets even of extreme proteins; and (iii) we determine the real impact of amino acid composition on secondary structural class.

Why is it so important to compute the accuracy of a prediction technique with a test set comprising as many as possible unrelated structures? With smaller test sets, there is a real chance to miss unfavorable (for the given technique) protein structures and to overestimate, maybe dramatically, the success rate. In this case, the prediction accuracy is not a measure for the correctness of prediction for an unknown protein.

We have applied two new vector decomposition methods for secondary structural content prediction from amino acid composition alone, with and without consideration of amino acid compositional coupling in the learning set of tertiary structures, respectively (see preceding article) to the problem of class prediction and achieve about 60% correct assignment among four classes (α, β, mixed, irregular) as well as single sequence-based secondary structure prediction methods like GORIII and COMBI.[22] Our results demonstrate that about 60% correctness is the upper limit for a 4-type class prediction (α, β, mixed, irregular) from amino acid composition alone for an unknown query protein (given the current limitations of the PDB) and that consideration of compositional coupling does not improve the prediction success.

METHODS

Our program SSCP (Secondary Structural Content Prediction) was utilized for prediction of secondary structural content (helix, sheet, coil) of query proteins from their amino acid composition alone. This implementation contains both algorithms described in the preceding article. The first method utilizes only the amino acid compositions of the characteristic secondary structural types in a learning set of protein structures. The second technique relies also on amino acid compositional couplings in the learning set.

The predicted secondary structural content has been used for secondary structural class assignment. We present the percentage of proteins for which a correct assignment of the structural class of α-, β-, mixed, or irregular type was possible using the definition of Nakashima et al.[4] These authors consider proteins with >15%α and <10%β as α-proteins, with <15%α and >10%β as β-proteins, with >15%α and >10%β as mixed proteins, and the remaining as irregular. We do not distinguish between α + β and α/β proteins within the mixed class since this decision cannot be made from secondary structural content alone.

The protein structure selection from the PDB[25,26] for learning and test sets was performed automatically with program OBSTRUCT.[27] Largest possible subsets of non-homologous structures (residue identity amongst all aligned pairs sequences ≤35%, minimal sequence length 80 residues) did not contain NMR structures or structures with incomplete backbone. The protein datasets are available in electronically readable form on request (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, M is the number of proteins), 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.[28] 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 done without human interference. Bias to certain types of proteins, other than from the limitations of the PDB itself, has been excluded.

Secondary structure assignments were made with the standard method of Kabsch and Sander.[29] For prediction with SSCP, the 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[22] except that we reassigned to coil all helices shorter than five residues and all strands shorter than three residues.[30] Note that Chou[13] considers only the H-type as helix.

The prediction accuracy of both methods has been evaluated both (i) with a self-consistency test (inclusion of the protein to be predicted into the learning set) and (ii) with a "jackknife" procedure (learning step without the protein to be predicted) applied to our four sets of proteins. Used as integral success rate measure, the value Q_{class} is defined as the number of proteins with any correct class assignment divided by the number M of all proteins in the dataset.

RESULTS AND DISCUSSION

Impact of Resolution, Number of Proteins, and Amino Acid Coupling

The secondary structural class prediction results obtained with SSCP based on the self-consistency and the jackknife tests for all four datasets are presented in Table I. As in the case of secondary structural content prediction, the differences between the self-consistency and the jackknife tests for the first method are negligible (compare columns I and II). In contrast to the secondary structural content prediction (see preceding article), the exclusion of the protein to be predicted does not influence the accuracy for the second method much (compare columns III and IV, largest variation for the 2.0 Å dataset, maximal difference in Q_{class} is 3.4%).

Whereas for the first method a small decrease in prediction success with resolution can be observed (from 63.3 to 58.8% in Q_{class}), there is virtually no influence of the dataset size or resolution for the second method. The consideration of amino acid coupling does also not improve prediction accuracy (comparison of columns I and III and columns II and IV in Table I). The results are similar for both methods, even with a favorable tendency in Q_{class} for the first method. The reduction of the secondary structural content predicted with the program SSCP to a simple comparison with thresholds given in the secondary structural class definition of Nakashima et al. levels out the influence of dataset size, resolution, and amino acid coupling (compare with the preceding article). Generally, all-β proteins appear to be predicted best. The overall prediction success rate Q_{class} is about 60% with fluctuations from 56 to 63.3% for different representative subsets of the PDB. The precision of the average prediction accuracy is, therefore, in the order of about 5%.

Prediction of Secondary Structural Class With Specialized Jury Decision Methods

Since the techniques used in specialized jury decision methods for the folding type prediction of a query protein have been considered in detail in section III.D.4 of the review by Eisenhaber et al.,[23] we will concentrate here on (i) the role of the structural class definition, (ii) a classification of analytic decision techniques, and (ii) the role of the representativity of a test set of protein structures for the PDB.

(i) The role of the structural class (folding type) definition

In Table II, we list the structural class definitions published in the literature. Even a first view reveals that there has been considerable arbitrariness by the authors in fixing the thresholds. In part, this can be explained by the smaller number of tertiary structures available just a few years ago. Today, many more structures are known compared with the time when Levitt and Chothia[31] proposed their classification, and the clustering due to particularities in secondary structural content is fading. In our opinion, the notion of secondary structure class is becoming increasingly obsolete, especially that of $\alpha+\beta$ and α/β. Nevertheless, the concept is still widely in use. Here, we formulate objective criteria for reasonable structural class definitions.

First, we asked if all proteins (domains) should be assignable to a structural class. This demand alone disqualifies the criteria 3, 5, and 6[5,13,32] in Table II since they assume that there are empty regions in the α-vs.-β-content distribution between the clusters of proteins. For sufficiently representative protein structure selections, this is not the case (Figs. 1,2). For example, applying the criterion of P.Y. Chou[5] to any of our four datasets (1.8Å, 2.0Å, 2.5Å, and 3.0Å), about two-thirds of all proteins cannot be assigned a folding type. The same happens with about one-third of all proteins, if the criterion of K.-C. Chou[13] is used. Thus, in these secondary structural class prediction attempts, a possibly unintentional preselection of proteins with extreme secondary structural content has taken place. This fact questions seriously the reported prediction accuracies. In principle, another prediction scheme is necessary to recognize an unknown query protein as irregular or as "no-type" protein before the described prediction schemes for secondary structural class jury decision might be applied with similar success on the remaining proteins.

The distinction between $\alpha+\beta$ and α/β should also be possible for all mixed protein domains. Few researchers supply quantitative criteria (Table II). K.-C. Chou[13] demands a residue fraction above 60% in parallel or antiparallel strand to make the choice. Two questions arise. How should one treat, for example, a mixed protein with 53% of the residues in one type of strand and the rest in the other as in the case of the entry 1GPB? K.-C. Chou[13] listed it as α/β-protein obviously taking the simple majority of residues as a basis. The other question concerns heavily mixed sheets which seriously question the utility of the $\alpha+\beta$ and α/β classification.[23] Such

TABLE I. Accuracy of Secondary Structural Class Prediction With the Secondary Structure Content Prediction of Our Two Vector Decomposition Methods*

Prediction method		First method (without comp. coupling)		Second method (with comp. coupling)	
Column		I	II	III	IV
Test set	N_{class}	Self-consistency	Jackknife	Self-consistency	Jackknife
Set ≤1.8 Å 166 Proteins					
α	41	58.5	58.5	58.5	56.1
β	54	72.2	72.7	75.9	74.1
Mixed	70	58.6	58.6	40.0	42.9
Irregular	1	100.0	100.0	100.0	0.0
Q_{class}	166	63.3	63.3	56.6	56.0
Set ≤2.0Å 262 Proteins					
α	55	52.7	52.7	54.5	47.3
β	78	74.4	73.1	82.1	76.9
Mixed	127	55.9	55.9	50.4	50.4
Irregular	2	50.0	50.0	50.0	0.0
Q_{class}	262	60.7	60.3	60.7	57.3
Set ≤2.5Å 398 Proteins					
α	84	54.8	54.8	52.4	52.4
β	103	68.0	68.0	72.8	69.9
Mixed	206	56.8	56.8	51.5	51.9
Irregular	5	20.0	20.0	40.0	0.0
Q_{class}	398	58.8	58.8	57.0	56.0
Set 3.0 Å 475 Proteins					
α	99	59.6	58.6	57.6	54.5
β	140	65.7	65.7	77.9	76.4
Mixed	232	56.5	56.0	46.6	47.0
Irregular	4	25.0	25.0	25.0	25.0
Q_{class}	475	59.6	59.2	57.9	57.1

*The success rates of secondary structural class prediction obtained with our two vector decomposition methods for four representative subsets of the PDB with different resolution thresholds (≤1.8Å, ≤2.0Å, ≤2.5Å, and ≤3.0Å) are presented in percent. Secondary structural class (folding type) prediction was performed among the four types α, β, mixed, and irregular. The column N_{class} lists the true number of α-proteins, β-proteins, mixed, and irregular structures in each dataset if the classification criteria of Nakashima et al.[4] are applied. Q_{class} shows the percentage of any proteins for which the class has been assigned correctly.

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).

structures were not known at the time when Levitt and Chothia[31] proposed their secondary structural class classification. A strand may be parallel to the preceding strand in the sheet and antiparallel with respect to the next one. The information about strand directionality is encoded both in the BP1 and BP2 columns of the DSSP-files.[29] It is not consequent to rely only on the BP1 information as was done by K.-C. Chou.[13]

Second, the definition should also comply with the common understanding of mixed domains (high helix and sheet content) and irregular proteins (low secondary structural content). This demand disfavors criterion 1.[3] Definition 4 is also not very convenient since proteins with large sheet-content are sometimes assigned to the irregular class. Nevertheless, it is necessary to express the classification criterion in quantitative terms for objectivity. Reliance on the subjective decision of the researcher after visual inspection of the tertiary protein structures on the graphics screen as reported by Boberg et al.[33] is not a scientific solution of the problem. How should a protein be classified that is missing in their lists and, at the same time, non-homologous to any of them?

Third, it would be desirable that the thresholds for secondary structural content coincide with extrema in the occurrence distribution of α-helices and β-sheets. It can be seen in Figure 3 that the minima are in the contents range 10–25% and at 10% for the two secondary structural types, helix and sheet, respectively. This data also supports the criterion of

TABLE II. Definition of Secondary Structural Classes of Proteins*

	Definition		
Reference	Class	α-Content	β-Content
1) Sheridan et al. (1985)[3]	All-α[‡]	α>β	α<β
	All-β[‡]	α<β	α>β
	Parallel[‡]	With parallel β-sheet	
	Irregular	with >4.5% Cys	
2) Nakashima et al. (1986)[4]	All-α	>15%	<10%
	All-β	<15%	>10%
	Mixed[†]	>15%	>10%
	Irregular	<15%	<10%
3) Klein and DeLisi (1986)[32]	Allα	>40%	<5%
	All-β	<10%	>30%
	Mixed[†]	≥15%	≥15%
	Irregular	α + β<20%	
4) Chou, P.Y. (1989)[5]	All-α	>45%	<5%
	All-β	<5%	>45%
	Mixed[†]	>30%	>20%
	Irregular		?
5) Kneller et al. (1990)[35]	All-α[Σ]	≥30%	≤0.15·(α + β)
	All-β[Σ]	≤10%	
	Mixed[†,Σ]	>15%	>5%
	Irregular	All remaining proteins	
6) Chou, K.-C. (1995)[13]	All-α	>40%	<5%
	All-β	<5%	>40%
	Mixed[†,¶]	>15%	>15%
	Irregular	<10%	<10%

*The definitions are given in terms of secondary structural content in percent.

[†] Mixed domains are classified as α/β or α + β depending on the existence of parallel β-sheet or a "sufficient" alternation of secondary structural types (Kneller et al., 1990).

[‡] A smooth transition between α- and β-structure is considered. Mixed domains always contain parallel sheets.

[Σ] The authors define all regular proteins as having a minimal length.

[¶] Mixed domains are classified as α/β if the content of parallel strand is ≥60% and as α + β if the content of antiparallel strand is ≥60%. The content of parallel and antiparallel strand has been calculated from the BP1 entry in DSSP-files or from the SHEET card in PDB-files.

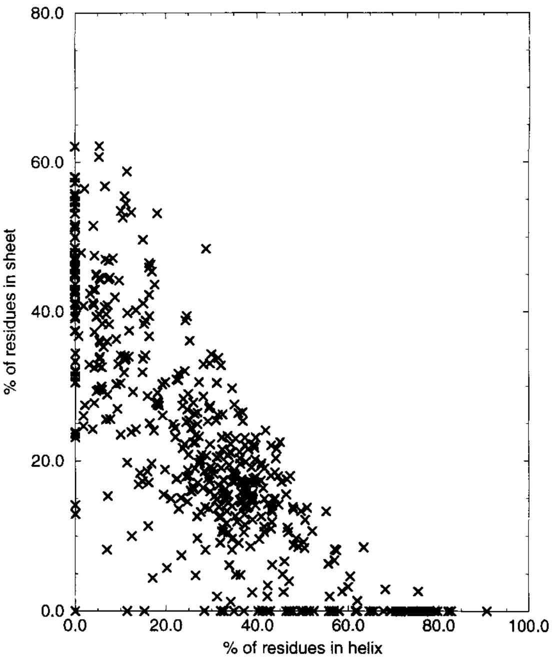

Fig. 1. α-vs.-β-content plot for representative subset of the Protein Data Bank. The plot is shown for the 3.0 Å set of tertiary structures. It can be clearly seen that there are no empty regions in the plot. The region of very low secondary structural content is somewhat less populated than one would expect for selections of polypeptide chains without the condition of minimal chain length (here: 80 residues or more).

Nakashima et al.[4] as the classification rule of choice.

(ii) Classification of analytic decision techniques

The highest prediction accuracies (reaching almost 100%) have been published for analytic decision techniques. In these approaches, a secondary structural class is characterized by a mean amino acid composition. For the class prediction, a distance measure between the amino acid composition of a query protein and the characteristic mean composition of the structural class is used. The smallest mutual distance is taken for class assignment.

The analytical methods for structural class prediction from amino acid composition can be classified into two groups. The first group comprises decision techniques mainly based on the Euclidean distance[4] or mathematically very similar expressions such as the angle between vectors ("vector projection method"[9]), the correlation coefficient,[7] or the maximal component in a vector decomposition.[6] Together with the Hamming distance method[5] and a "fuzzy"-clustering technique (a least Minkowski distance method with variable characteristic mean amino acid compositions of secondary structural classes[34]), all these approaches ignore amino acid type weighting and compositional coupling.

Due to the similar mathematical formulation, uniform prediction accuracies should be expected for all these techniques. Indeed, e.g., for the historic set of 64 selected proteins in a self-consistency test, the prediction accuracies are 83.6% for the correlation coefficient approach,[7,34] 82.8% for the vector decomposition,[6] 79.7% for the Hamming distance technique,[5,34] and 81.3% for the "fuzzy" clustering procedure.[34] Interestingly, several attempts of sec-

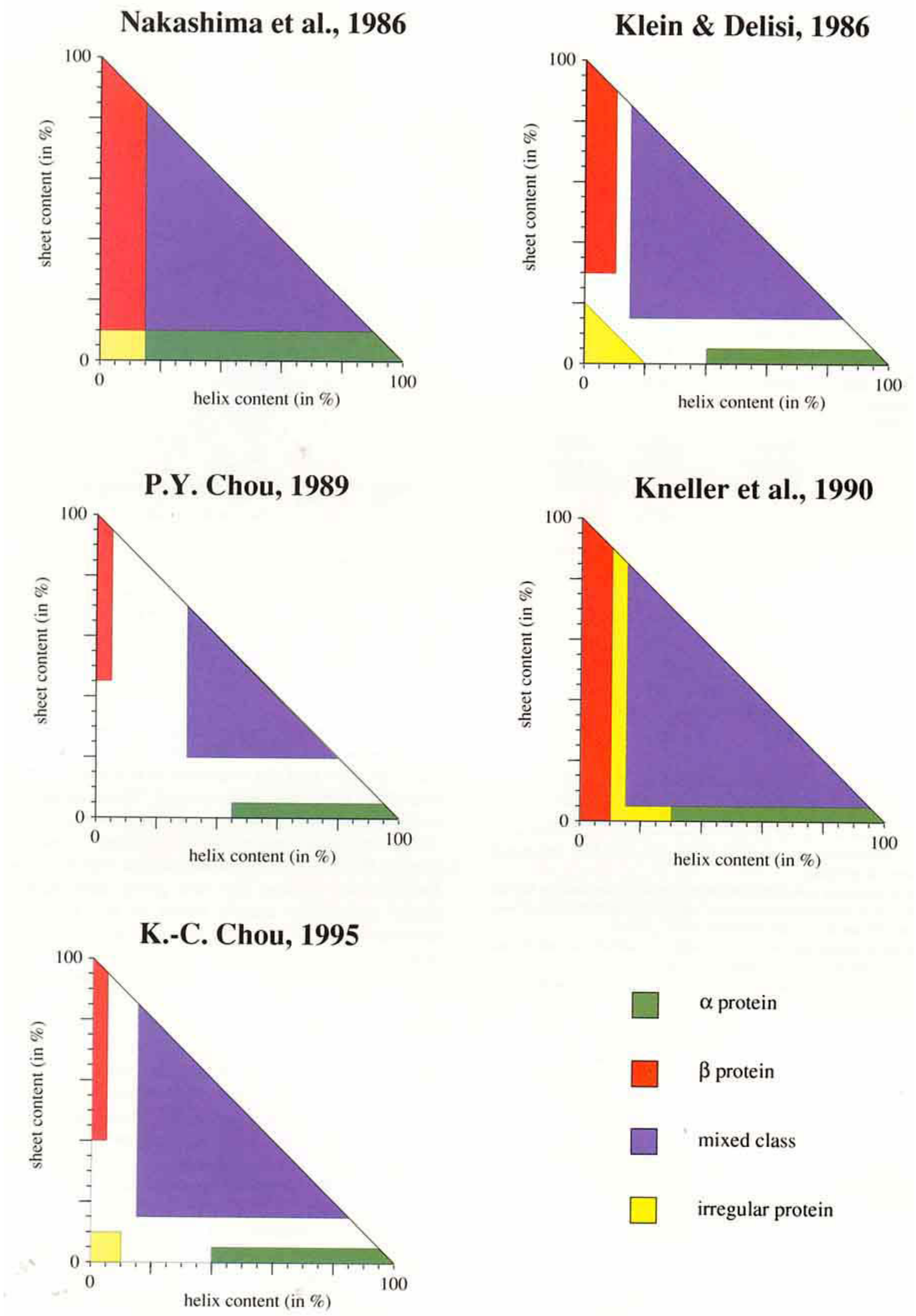

Fig. 2. Secondary structural class definitions. The definitions are visualized in form of α-vs.-β-content trigonal plots to demonstrate the white regions which result in proteins structures which are not assignable to any secondary structural class (folding type). Some definitions are only applicable to the small fraction of proteins with extreme secondary structural content.

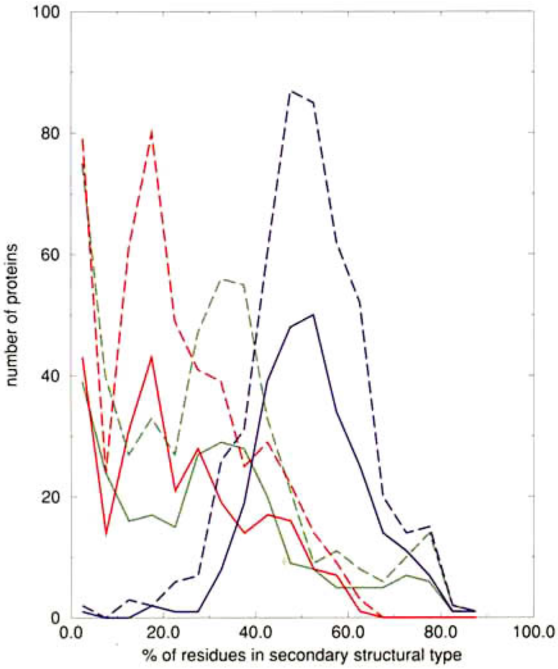

Fig. 3. Distribution of secondary structural content in a representative subset of the Protein Data Bank. The distribution of protein chains as function of the secondary structure content (green, α-content; red, β-content; blue, the sum of both values) is shown for the 2.0 Å set (continuous line) and for the 3.0 Å set (dashed line) of tertiary structures. In most proteins, between 30 and 80% of all residues are involved in helix or sheet formation. The β-content distribution has a clear minimum at about 10%. For helix content, a valley in the region 10–25% is observed. The location of the extrema is identical for all of our four protein sets.

ondary structural class prediction with neural networks from amino acid composition and simple hydrophobic sequence patterns resulted also in up to 80.2% prediction accuracy for the same 64-protein set.[16] These differences are really small compared with fluctuations due to changes in datasets. For example, the vector projection method has a success rate of 83.6%[9] for the set of 64 proteins but only 66.7% for another set of 120 proteins,[13] a decrease by 16.9%! The results are even lower for representative datasets, as is illustrated in Table III (vide infra).

In the second group, all methods rely on elliptical scaling of the Euclidean distance components which results finally in an eigenvector problem of the second moment matrices of compositional fluctuations within each secondary structural class of proteins. In its simplest form, only the weighting of amino acids has been considered[8] and prediction accuracy reached 100% in the self-consistency test. The more consequent formulation of the eigenvector problem which has been published repeatedly with differing phraseology ("elliptically scaled distance,"[10] "Mahalanobis distance"[11–15]) also takes into account compositional couplings between amino acid types. The prediction accuracies reported are also near 100%. Thus, one should assume that the compositional coupling of different amino acid types (in contrast to weighting = coupling of an amino acid type with itself) is not very important. We will demonstrate with the data in Table III (vide infra) that, at the level of secondary structural class, any compositional coupling is of little influence.

(iii) The role of the representativity of the test set of protein structures

At this place, we want to emphasize that a success rate characterizing a prediction technique should assess the probability of a correct prediction for an unknown protein. It is of no use to publish lists of proteins for which a new method works well. The most stringent test set of tertiary structures available would consist of all unrelated proteins, the structure of which is known. Smaller test sets always open the theoretical possibility that just the proteins which would be badly predicted have been missed and that the prediction rates are overestimated. No crossvalidation procedure can compensate for such limitations of a test set. Therefore, it is extremely important that success rates are calculated from representative subsets of the PDB. Even in this case, the prediction accuracies should be considered as upper limits which will probably decrease if more new structures are available.

There are deep contradictions amongst the success rates published for protein structure prediction methods as seen from a literature study. This became especially evident after three recent publications.[13–15] On one side, secondary structure prediction methods achieve, with difficulty, prediction accuracies near 70% for the state per residue and up to 75% for secondary structural class[20–22] and require extensive input information (local sequence, multiple alignments, amino acid composition, sequence length, etc.). At the same time, structural class prediction accuracies near 100% have been repeatedly reported,[10–15] although doubts have been thrown upon these results by other researchers.[33] An outside observer gets the ultimate impression that, in contrast to predicting the location of secondary structures in protein sequences, the prediction problem of the secondary structural class has already been solved elegantly and solely with the amino acid composition and an elliptically scaled distance as decision criteria.

We attempted to reproduce the high prediction level with the Mahalanobis distance method.[13] Some deviations in numerical values from the published results are usually a sign or of programming errors or of silent assumptions not described in the research report (in this case, rounding of composition values to 1% before treatment in classification rules and in the prediction algorithm). We cordially thank K.-C. Chou for offering his program

TABLE III. Prediction Accuracies for Structural Class (α, β, α + β, α/β) With Four Specialized Jury Decision Methods[†]

Test set of protein structures		r (Å)	1.8	2.0	2.5	3.0
		N	166	262	398	475
Definition of structural class K.-C. Chou (1995)[13]		Irregular	1	1	2	2
		"Other"	55	81	127	163
		N*	110	180	269	310
Learning set	Prediction method					
4 × 30 = 120 Proteins of K.-C. Chou (1995)[13]	Nakashima et al. (1986)[4]	N_c	56	88	122	145
		% of N*	50.9	48.9	45.4	46.8
	P.Y. Chou (1989)[5]	N_c	57	86	121	144
		% of N*	51.8	47.8	45.0	46.5
	Chou and Zhang (1993)[9]	N_c	56	88	124	145
		% of N*	50.9	48.9	46.1	46.8
	K.-C. Chou (1995)[13]	N_c	60	89	115	137
		% of N*	54.5	49.4	42.8	44.2
Definition of structural class Nakashima et al. (1986)[4]		Irregular	1	1	3	2
		"Other"	0	0	0	0
		N*	165	261	395	473
Learning set	Prediction method					
4 × 30 = 120 Proteins of K.-C. Chou (1995)[13]	Nakashima et al. (1986)[4]	N_c	75	122	169	221
		% of N*	45.5	46.7	42.8	46.7
	P.Y. Chou (1989)[5]	N_c	83	124	169	214
		% of N*	50.3	47.5	42.8	45.2
	Chou and Zhang (1993)[9]	N_c	77	123	173	219
		% of N*	46.7	47.1	43.8	46.3
	K.-C. Chou (1995)[13]	N_c	84	123	167	205
		% of N*	50.9	47.1	42.3	43.3
Set 2.0 Å	K.-C. Chou (1995)[13]	N_c	105	173	210	251
M = 261		% of N*	63.6	66.3	53.2	53.1

[†] The absolute number and the percentage of proteins with correctly predicted structural class among the four types (α, β, α + β, α/β) for the test sets 1.8Å, 2.0Å, 2.5Å, and 3.0Å are presented. The predictions have been produced with the following four methods:
1. the Euclidean distance method of Nakashima et al.,[4]
2. the Hamming distance method of P.Y. Chou,[5]
3. the vector projection method of Chou and Zhang,[9] and
4. the Mahalanobis distance method of K.-C. Chou[11–15]

in the form as implemented in the routine "4WAY2" or K.-C. Chou. The real secondary structural content of the proteins was computed on the basis of DSSP-files in the same way as described by Chou[13] (only states H and E were considered as helix and strand respectively). In the upper part of the table, the structural class definition of K.-C. Chou[13] (see definition 6 in Table II) was applied. For the lower part, the definition of Nakashima et al.[4] (definition 2 in Table II) has been used. The learning set tertiary structures was in all cases identical with the database of 120 proteins[13] except for the results in the last row when our dataset 2.0 Å (without the single irregular structure in this set) was used for the learning phase.

Since irregular proteins cannot be predicted with the four algorithms, we removed them from the test set (table entry "irregular"). In the case of the structural class definition of K.-C. Chou,[13] about 35% of all proteins cannot be assigned any type. Their number is given in the table entry "other." N* is the number of proteins (after substraction of irregular and not assignable proteins from originally N proteins) for which finally a prediction is attempted. N_c is the absolute number of proteins with correctly predicted structural class. To assess the quality of the prediction, the overlap of learning and test sets are of interest. Chou's database of 120 polypeptide chains has 15, 19, 22, and 28 identical entries with our sets 1.8Å, 2.0Å, 2.5Å, and 3.0Å, respectively. We did not calculate the number of homologous polypeptide sequences.

Since all proteins being assigned to a structural class with the definition of Chou belong to the same class with the definition of Nakashima et al., we can use the set of four groups of 30 proteins (120 proteins) as learning data in both cases. Since this dataset is not representative for the current status of the Protein Data Bank, we executed the learning step also with the 2.0 Å set of tertiary structures for the Mahalanobis distance method.

"4WAY2," accompanying routines and protein lists[13] in electronic form which enabled us to reproduce his results for objective comparison.

We tested the predictive power of analytical distance based jury decision methods with our four representative selections of protein structures. In Table III, we present the prediction accuracies obtained with the following methods as implemented in the program "4WAY2":

- the least Euclidean distance method,[4]
- the least Hamming distance method,[5]
- the so-called vector projection method (angle between the amino acid composition vector of the query protein and one of the average composition vectors representing a structural class),[9] and
- the Mahalanobis distance method.[11–15]

Taking into account the comments on method classification, the verification of these four techniques is sufficient to test the power of the general approach.

First, we summarize the published prediction accuracies of the four methods. The self-consistency test for a set of 120 proteins for the Mahalanobis distance method yields an accuracy of 96.2%; for a test set of 64 other proteins are, after learning from the previous set of 120 proteins, 95.3% correctly predicted[13] (the four polypeptide chains 1DNKA, 1POC, 1GPB, and 1SBP are identical in both the 64 set and the 120 set). For the same tests, the results for the other three methods are uniformly lower and are in the range of 66–69%. For the Euclidean and the Hamming distance methods, these accuracies were published by Chou.[13] For the vector projection method, we have calculated ourselves the prediction success rate and obtained the same result as Chou for the Euclidean distance method. Surprisingly, if the set of 64 proteins is used both for learning and prediction, the success rate of the Hamming distance method is 79.7% and that of the vector projection method is even 83.6%![9] Thus, there are obviously dramatic fluctuations in the prediction accuracies depending on the test protocol. The memorization effects should be even more dramatic for the method utilizing elliptically scaled distances since the number of parameters in the prediction function is much larger.

As we have seen in the discussion of the structural class definitions, the classification by definition 6 (Table II) requires the removal of irregular and "other" (no-type) proteins from our test sets ≤1.8Å, ≤2.0Å, ≤2.5Å, and ≤3.0Å. Due to this preselection, the size of the test sets decreases by about 35%. The reduction of the number of proteins in the test sets due to the application of the secondary structural class criterion of K.-C. Chou[13] is described in detail in the upper part of Table III. As a learning set, the 120 proteins of the same author have been used. Nevertheless, even for "good" proteins (i.e., proteins with extreme secondary structural content), the prediction accuracies are quite low (45–51% for the three simple distance techniques and 42–54% for the Mahalanobis method). The low success rate is the result of many false predictions throughout all four classes, not just in a single one (data not shown).

In the lower part of Table III, the secondary structural class definition of Nakashima et al.[4] (definition 2 in Table II) has been applied. For this classification rule, only the irregular proteins have to be preselected and removed from the test sets. As learning set, the 120 proteins of Chou have been used. The prediction accuracies clearly decrease compared with the upper part of the table and are in the range 42.3–50.9%. Since it may be objected that the learning set of 120 proteins is too small, we performed predictions also with learning from the 2.0 Å protein selection for the Mahalanobis distance method ($M = 261$ proteins since one irregular structure has to be removed). The prediction accuracies are improved for the 1.8 Å and 2.0 Å test sets (about 65%). In fact, these are self-consistency tests for datasets representing the current status of the PDB. The success rate drops immediately to the previous value near 50% for the larger test sets 2.5 Å and 3.0 Å which contain many entries not present in the learning set.

We conclude from these computations that a jury decision among four structural classes (all-α, all-β, $\alpha + \beta$, α/β) based only on the amino acid composition of the query protein is at best 50–60% correct. The consideration of compositional couplings improves the prediction accuracy not much and not for all test protocols and is, therefore, of minor importance.

We have also carried out 3-type (α, β, mixed) and 5-type (all-α, all-β, $\alpha + \beta$, α/β, and the class of irregular proteins) predictions. As learning and test sets, the 120 and 64 protein selections of Chou, respectively, were used. For 3-type predictions, the success rate of the Euclidean and the Hamming distance methods was 60.9%, that of the vector projection method was equal to 62.9%. The Mahalanobis distance method predicted all proteins as mixed. For the 5-type prediction, only 42.2% of all 64 proteins were correctly classified by the Mahalanobis distance method. For our four large, representative subsets of the PDB, even worse predictions were obtained.

Our two vector decomposition methods arrive at 56–63% prediction accuracy for a 4-type decision (α, β, mixed, irregular, see Q_{class} in Table I) over several large, automatically generated, and representative subsets of the PDB. Our result is comparable with the prediction power of single sequence techniques (GORIII and COMBI yield 66% but for a test set of only 124 proteins[22]). A neural network technique of Chandonia and Karplus[19] achieves 62.3% without and 73.9% in combination with traditional secondary structure prediction for every residue. As input information, amino acid composition, sequence length, and local sequence were utilized. Given the small set of only 64 proteins among which are sequences with pairwise residue identity even up to 47%, the success rate is certainly an upper boundary estimate. With information from multiple sequence alignments, PHD3[22] predicts 75% of 124 proteins correctly. This increase of prediction accuracy can be attributed to the improved prediction input. Therefore, we conclude that amino acid composition alone determines secondary structural class to about 60% given the current limitations of the PDB. Higher prediction accuracies published elsewhere are a result of a limited structure selection and/or test protocol.

CONCLUSION

The expected secondary structure prediction accuracies for an unknown query protein can be summarized in the following way. If only the amino acid composition of its sequence is given, the secondary structural content can be predicted with an absolute error of about 13% (see preceding article). The secondary structural class prediction (all-α, all-β, mixed, irregular) is correct in about 60% of all cases (results in this article). These prediction accuracies improve very little if also the local sequence information is available, as judged by comparison with GORII, COMBI, and QS (preceding and this article). In this case, the secondary structural state per residue is correctly predicted in about 60%.[23] The use of information from multiple alignments with homologous sequences apparently improves the prediction accuracy. The secondary structural class can be estimated with about 75% correctness. The absolute error of the secondary structural contents is in the range 7–8% and the state of each residue is predicted with about 70% confidence.[22]

AVAILABILITY OF PROGRAMS

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/

Upon input of amino acid sequence or composition, the prediction of secondary structural content and secondary structural class (folding type) is returned. 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 are gratefully acknowledged. F.E. is grateful for financial support from the Fund "Wissenschaftler-Integrationsprogramm" (grant 020386/B) jointly administered by the East-German Länder and the government of the Federal Republic of Germany. We acknowledge that the success rate comparisons would not have been described in this detail without the explicit demand of two referees.

REFERENCES

1. 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.
2. 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.
3. 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.
4. 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.
5. Chou PY; Prediction of protein structural classes from amino acid composition. In: Prediction of Protein Structure." Fasman, G.D. (ed.). New York: Plenum Press, 1989: 549–586.
6. Zhang, C.-T., Chou, K.-C. An optimization approach to predicting protein structural class from amino acid composition. Protein Sci. 1:401–408, 1992.
7. 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.
8. 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.
9. Chou, K.-C., Zhang, C.-T. A new approach to prediction protein folding types. J. Prot. Chem. 12:169–178, 1993.
10. 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.
11. 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.
12. Chou, K.-C. Does the folding type depend on its amino acid composition? FEBS Lett. 363:127–131, 1995.
13. 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.
14. Zhang, C.-T., Chou, K.-C. An eigenvalue-eigenvector approach to predicting protein folding types. J. Prot. Chem. 14:309–326, 1995.
15. Chou, K.-C., Zhang, C.-T. Prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30:275–349, 1995.
16. 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.
17. Reczko, M., Cohr, H., Subramaniam, S., Pamigighantam, S., Hatzigeorgiou, A. Fold-class prediction by neural networks. In: "Protein Structure by Distance Analysis." Bohr, H., Brunak, S. (eds.). Amsterdam, Tokyo: IOS Press, Ohmsha, 1994:277–286.
18. Reczko, M., Bohr, H. The DEF data base of sequence based protein fold class predictions. Nucleic Acids Res. 22:3616–3619, 1994.
19. Chandonia, J.-M., Karplus, M. Neural networks for secondary structure and structural class predictions. Protein Sci. 4:275–285, 1995.
20. Rost, B., Sander, C. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232:584–599, 1993.
21. Levin, J.M., Pascarella, S., Argos, P., Garnier, J. Quanti-

fication of secondary structure prediction improvement using multiple alignments. Protein Eng. 6:849–854, 1993.

22. Rost, B., Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19:55–72, 1994.
23. 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.
24. Barton, G.J. Protein secondary structure prediction. Curr. Opin. Struct. Biol. 5:372–376, 1995.
25. 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.
26. 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.
27. Heringa, J., Sommerfeldt, H., Higgins, D., Argos, P. OBSTRUCT: 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.
28. 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.
29. Kabsch, W., Sander, C. Dictionary of protein secondary structures: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
30. 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.
31. Levitt, M., Chothia, C. Structural patterns in globular proteins. Nature 261:552–558, 1976.
32. Klein, P., DeLisi, C. Prediction of protein structural class from the amino acid sequence. Biopolymers 25:1659–1672, 1986.
33. Boberg, J., Salakoski, T., Vihinen, M. Accurate prediction of protein secondary structural class with fuzzy structural vectors. Protein. Eng. 8:505–512, 1995.
34. Zhang, C.-T., Chou, K.-C., Maggiora, G.M. Predicting protein structural class from amino acid composition: Application of fuzzy clustering. Protein Eng. 8:425–435, 1995.
35. Kneller, D.G., Cohen, F.E., Langridge, R. Improvements in secondary structure prediction by enhanced neural networks. J. Mol. Biol. 214:171–182.