Use of Variable Selection in Modeling the Secondary Structural Content of Proteins from Their Composition of Amino Acid Residues

Teuta Piližota,† Bono Lučić,* and Nenad Trinajstić

The Rugjer Bošković Institute, P.O. Box 180, HR-10002 Zagreb, Croatia

Received February 27, 2003

The possibility of prediction of protein secondary structure content from composition of their amino acid residues can help in bridging the gap between proteins of known primary sequence having an unknown secondary structure. Almost all recently published models for understanding the relationship between composition (frequency of occurrence) of amino acid residues and secondary structure content of proteins involved composition of all 20 amino acid residues. However, it is well-known that many amino acid residues are mutually similar according to their physicochemical properties (hydrophobicity, hydrophilicity, charge, size, etc.). Because of that, we were motivated to investigate the possibility of reduction of the total number of terms (frequencies of amino acid residues) in the models for describing the relation between the composition of amino acid residues and the percentage of residues belonging to α, *β*, and coil secondary structure. For this purpose, the CROMRsel algorithm (*J. Chem. Inf. Comput. Sci.* **1999**, *39*, 121−132) for selection of a small subset of the most important variables/descriptors into the multiregression (MR) models, i.e., frequency of occurrence of amino acid residues in proteins, was used. Analysis was performed on a data set containing 475 proteins, taken from *Proteins* **1996**, *25*, 157−168. A complete data set was partitioned into a 317-protein training set and 158-protein test set. The best possible linear models containing $I = 1, ..., 20$ frequencies were selected among all 20 frequencies of occurrence of amino acid residues on the 317-protein training set, and were used for performing prediction of the corresponding percentage of secondary structure content on the 158-protein test set. For the 317-protein data set the best selected concise models for the α, *β*, and coil secondary structure contain only 9, 5, and 8 frequencies, respectively. Selected concise models are of the same or better fitted, cross-validated, and predictive statistical parameters than the models containing all 20 frequencies. Additionally, for each I ($I = 1, ..., 20$) 30 the best possible random models were selected. In each case, the best possible real models are much better than each of the best possible random models, showing clearly that there is no risk of a chance correlation (what one could expect due to the application of an exhaustive search for the best model having I frequencies among all $20!/I!(20-I)!$ possible models). Finally, the best selected models on the complete 475-protein data set for the α, *β*, and coil secondary structure contain only 7, 4, and 7 frequencies of amino acid residues, respectively. These models are much simpler and have better fitted and cross-validated errors than the corresponding models from the literature, that were obtained without using a procedure for selection of the most important frequencies of amino acid residues in proteins.

INTRODUCTION

Measurement or prediction of the secondary structure content (percentages of α-helix (α), *β*-strand (*β*), and coil) of a soluble protein can be considered as the first step in getting information on its structure. The protein secondary structure content can be experimentally determined by circular dichroism (CD) spectroscopy in the UV absorption range[1] and IR Raman spectroscopy.[2] In some cases, the accuracy of these experimental methods is not satisfactory. In addition, there are no general methods suitable for every protein.[3] Almost all published theoretical approaches were developed by the multiregression technique and use protein primary sequence information solely, or together with physical and chemical properties of amino acid residues (refs 3−6 and references therein). Additional descriptors, derived from the protein sequence information and properties of amino acid residues, squares, or cross-products of initial frequencies of amino acid residues, that improve models for predicting the protein secondary structure content, were added to the frequency of occurrence of 20 amino acid residues (which have been used in each model).

However, all of these models contain a lot of optimized parameters corresponding to a large number of included descriptors, and, consequently, they are of limited accuracy in predicting the secondary structure content of a new set of proteins. In addition, it is not easy to interpret such models because they are complex and include several strongly intercorrelated descriptors related to very similar amino acid residues (which have similar physical and chemical properties). From the similarity analysis of protein sequence databases, we know that protein structures are redundant (i.e., a large portion of protein primary sequence may be irrelevant for protein function), and that many point mutations do not change protein structure and function. Redundancy of protein

* Corresponding author phone: ++385-1-4680095; fax: ++385-1-4680-245; e-mail: lucic@irb.hr.

† Present address: The Clarendon Laboratory, Department of Physics, University of Oxford, South Parks Road, Oxford OX1 3PU.

10.1021/ci034037p CCC: $27.50
Published on Web 11/12/2003

structures regarding mutations indicates redundancy between 20 amino acid residues. It is known that there is a significant similarity between amino acid structures (and amino acid residue structures, as well) and their physical and chemical properties. Because of that, it is important to find the minimal number of frequencies of amino acid residues needed for describing a global property of proteins such as the protein secondary structure content is. Recently, Zhang et al.[3] found that it is possible to reduce the number of terms in regression models for predicting the protein secondary structure content. However, they performed an approximative selection procedure and the predicted secondary structure content for the three protein classes (alpha, beta and alpha/beta classes) that are defined based on the protein secondary structure content.

To solve this problem, we used a descriptor/variable selection procedure called CROMRsel for selecting a small subset of the most relevant descriptors.[7−9] This modeling procedure has been developed and successfully applied on several data sets in the field of QSAR/QSPR (Quantitative Structure Activity/Property Relationship).[9−14] The CROMRsel descriptor selection method was compared with the robust and concise neural network methods,[7,9,10,14] neural network ensemble methods,[14] selection of descriptors for the multiregression method based on a genetic algorithm,[12,14] and the heuristic method from the CODESSA program.[8,11,12] In all these cases, it has been shown that the models selected by the CROMRsel descriptor selection procedure (1) have equal or better statistical performance measured by the corresponding statistical parameters and (2) are (much) simpler, i.e., contain a smaller number of optimized parameters (descriptors). In this study we wish to show that it is possible to significantly simplify protein secondary structure models.

DATA SETS AND COMPUTATIONAL DETAILS

Data Sets. A data set from ref 4, containing 475 soluble protein structures determined at the resolution of 3.0 Å with a low residue identity (≤35%) among all aligned pairs of sequences, was used in this paper. Protein structures determined by the NMR method and those with an incomplete backbone necessary for derivation of the secondary structural state of the residues were excluded. For this data set the tertiary protein structure selection from the Brookhaven Protein Data Bank (PDB)[15,16] was done automatically with the program OBSTRUCT.[17] These secondary structure assignments were made with the standard method of Kabsch and Sander.[18] Their secondary structural types H, G, and I were classified as helix, and residues marked with E were considered as being part of the sheet. The secondary structure of the remaining residues was assigned as coil as well as all helices shorter than five amino acid residues.

Performing Partition of Data into the Training Set and the Test Set. To test the predictive performance of selected models an external test set was needed. The total data set containing 475 proteins was randomly partitioned into the training set (317 proteins) and the test set (158 proteins). The best models, containing $I = 1, 2, 3, ..., 19$ most informative frequencies of amino acid residues in proteins, were selected *on the training set* according to the best fitted statistical parameters (the standard error of estimate and the correlation coefficient). Additionally, the cross-validated statistical parameters of the models containing $I = 1, 2, 3, ..., 19$ frequencies selected on the training set were calculated. After that, the accuracy of the best models was measured in prediction on the (never seen) test set. Training and test sets are given in Table S1 in Supporting Information as well as on http://www.irb.hr/~lucic/317_158proteins. Models containing all 20 frequencies were not computed for the training set because such models were unstable due to the high intercorrelation between pairs of frequencies of amino acid residues.

Protein Secondary Structure Content. Percentages of amino acid residues belonging to alpha (α), beta (β), or coil secondary structures were computed for each protein by eq 1 from the experimentally determined protein structures (as described in the preceding subsection)

$$\%\alpha = 100 \cdot N_\alpha/N, \quad \%\beta = 100 \cdot N_\beta/N, \quad \%coil = 100 \cdot N_{coil}/N \quad (1)$$

where N_α, N_β, and N_{coil} are the number of amino acid residues belonging to α, β, and coil secondary structures, respectively. N is the total number of amino acid residues in a protein ($N = N_\alpha + N_\beta + N_{coil}$).

Normalization of Predicted Protein Secondary Structure Contents. Each single percentage predicted from the model can be less than 0% and greater than 100%. In addition, the sum of three predicted secondary structure contents can be greater than or less than 100%. Due to this fact, predicted contents were normalized using the following rules: (1) each content predicted to be less than 0% and greater than 100% was assigned to 0% and 100%, respectively, and after that (2) the sum of predicted percentages of secondary structure contents for α, β, and coil were normalized to be 100% (each single content was divided by the sum of contents predicted for α, β, and coil). It is clear that protein secondary structure contents predicted to be less than 0% and greater than 100% are not possible. These normalization rules are independent of the experimental secondary structure content and can be applied to each protein for which experimental structure is not determined. Only the models for the percentage of α, β, and coil that were selected as the best ones were normalized.

Frequencies of Amino Acid Residues. Compositions of the 20 amino acid residues (x_1, x_2, x_3, ..., x_{20}) were computed for each protein according to eq 2:

$$x_i = 20 \cdot N_i/N, \quad i = 1, 2, 3, ..., 20 \quad (2)$$

Here, N_i is the total number of amino acid residues i in the protein sequence having, in total, N amino acid residues. Amino acid residues are ordered alphabetically according to their single letter code. To be self-explanatory, frequencies of amino acid residues (x_i) will be also designated (especially in tables) by their three-letter code, i.e., as x_{ala}, x_{cys}, x_{asp}, ..., x_{trp}, x_{tyr}. In this study, instead of the term "frequencies of 20 amino acid residues" the term "descriptors" will be used.

Statistical Parameters. The quality of the models was expressed by the fitted (S), leave-one-out (LOO) cross-validated (S_{cv}), and predictive (S_{pred}) standard errors of estimates and by the fitted (R), leave-one-out cross-validated (R_{cv}), and predictive (R_{pred}) correlation coefficients. Although the standard error and the mean absolute error give almost the same information, the mean absolute fitted (S_{abs}), leave-

one-out cross-validated ($S_{abs,cv}$), and predictive ($S_{abs,pred}$) errors were also calculated, because of the need for direct comparison with other published models.[4] The total number of proteins in the data set (317 for the training set, 158 for the test set, or 475 for the complete data set) was used in the denominator in equations employed for calculating the standard errors and mean absolute errors. The fitted and cross-validated (CV) statistical parameters were calculated for the training set containing 317 proteins, and the predictive statistical parameters were calculated for the test set containing 158 proteins. Since we wanted to compare the models obtained using methodology described in this paper with the models developed and published on the same set of proteins,[4] the models on the whole data set containing 475 proteins were also developed, and the corresponding fitted and CV statistical parameters were calculated.

Generation of Multiregression Models by CROMRsel. In this paper only the linear multivariate regressions (multiregression, MR) were generated. Descriptors and models were generated by the use of CROMRsel procedure described previously.[7−9] All the MR models in this paper were obtained by selecting "the best possible MR models" according to the highest fitted and cross-validated correlation coefficients (i.e., we did not use stepwise CROMRsel selection procedures from refs 7 and 8).

In short, the CROMRsel procedure used in this study starts with the selection of I descriptors ($I = 1, 2, 3, ..., 19$) into the MR model and orthogonalization of descriptors. After that, the square of the correlation coefficients of a multidescriptor model can be simply calculated as the sum of squares of the correlation coefficients between each orthogonalized descriptor and the modeled property.[7,8] This procedure was repeated for each possible model containing I descriptors that can be selected from the set containing N descriptors in total. For the data set containing $N = 20$ descriptors and for one selected value of I, there were $N!/I!(N-I)!$ possible models. According to the best fitted correlation coefficient, a single best model was selected among all possible models having I descriptors. After that the leave-one-out cross-validated and predictive (prediction on the test set) statistical parameters were calculated for this model.

Performing Randomization Tests. To obtain information about the quality of random models containing I ($I = 1, 2, 3, ..., 19$) descriptors, the experimental percentage of α, β, and coil secondary structure contents were randomized, and the best possible random models containing I descriptors (frequencies of amino acid residues in protein) were selected using CROMRsel. Randomization and selection of the best models was repeated 30 times for each I ($I = 1, 2, 3, ..., 19$) on the 317-protein training set, i.e., 570 the best possible random models were selected for each of the three secondary structure types in total. When the complete 475-protein data set was used, the randomization test was performed only for the models for α, β, and coil secondary structure content that were selected as the best ones, and, in this case, the randomization was repeated 30 times too.

RESULTS AND DISCUSSION

Two classes of models for predicting the percentage of the secondary structure content in soluble proteins for α, β, and coil secondary structures were developed in this paper. The first class was related to the models obtained on the training set consisting of 317 randomly selected proteins that were used for performing prediction on the test set consisting of 158 proteins. The second class of models was obtained on the complete data set containing 475 proteins. Comparison between the models obtained on the 475-protein data set and the models obtained by Analytic Vector Decomposition Methods and published by Eisenhaber et al.[4] for the same data set was performed and discussed.

A. Results Obtained Using the Training/Test Set Partition. The training set containing 317 molecules was used for selecting all the best models containing I ($I = 1, 2, ..., 19$) descriptors for α, β, and coil. Due to the problem related to a high intercorrelation among several pairs of amino acid residues it was not possible to generate models containing all 20 frequencies. For each selected model LOO cross-validated statistical parameters were calculated. Selection of the best models was done by the CROMRsel method described above and in refs 7 and 8. In parallel, predictive abilities of the models (containing I descriptors) selected as the best ones for α, β, and coil secondary structure contents were assessed by performing prediction on the test set containing 158 randomly selected proteins. The standard error and the mean absolute error of prediction was used for measuring the quality of selected models. Standard errors of fit, LOO CV, and prediction are presented for α, β, and coil secondary structures in Figures 1−3, respectively.

Rules for Selecting the Best Models. As an objective criterion for selecting the best models for α, β, and coil, the value of the CV standard error was used as well as the values of the CV standard error for neighboring models, i.e., those containing $I - 1$ and $I + 1$ descriptors. For example, if we consider the model containing I descriptors in Figure 1, and if the CV standard error of the model containing $I + 1$ descriptors insignificantly increases or decreases, then the I-descriptor model is chosen as the better one. Such a criterion is in accordance with the well-known Ockham's Razor (which prefers the model realized with the fewest descriptors, if other things being equal).[19] According to this criterion the best models obtained on the 317-protein training set for α, β, and coil are those containing $I = 9$, $I = 5$, and $I = 8$ frequencies (descriptors), respectively. Details of the best models for the percentage of α, β, and coil secondary structure contents, selected among all the best models containing I ($I = 1, 2, ..., 19$), are included in Table 1.

It is important to note that the selection of the best models containing I descriptors was performed totally unforced, i.e., we did not force inclusion of a specific amino acid residue in the best selected models because of any other reasons than those defined by the statistical criterion used in the CROMRsel method (i.e. the lowest standard error of estimate of fit, to wit, the highest correlation coefficient).

Comparison with Propensities of Amino Acid Residues for α, β, and Coil Secondary Structure (Refs 21 and 22). In the obtained models, relevance of an amino acid residue to form one of the secondary structures is expressed in a somewhat different way than it was described in the propensity values. In discussion we use the propensity values computed in refs 20−22. During selection of the best multivariate regression models, involvement of a specific frequency of amino acid residue (descriptor) in the best

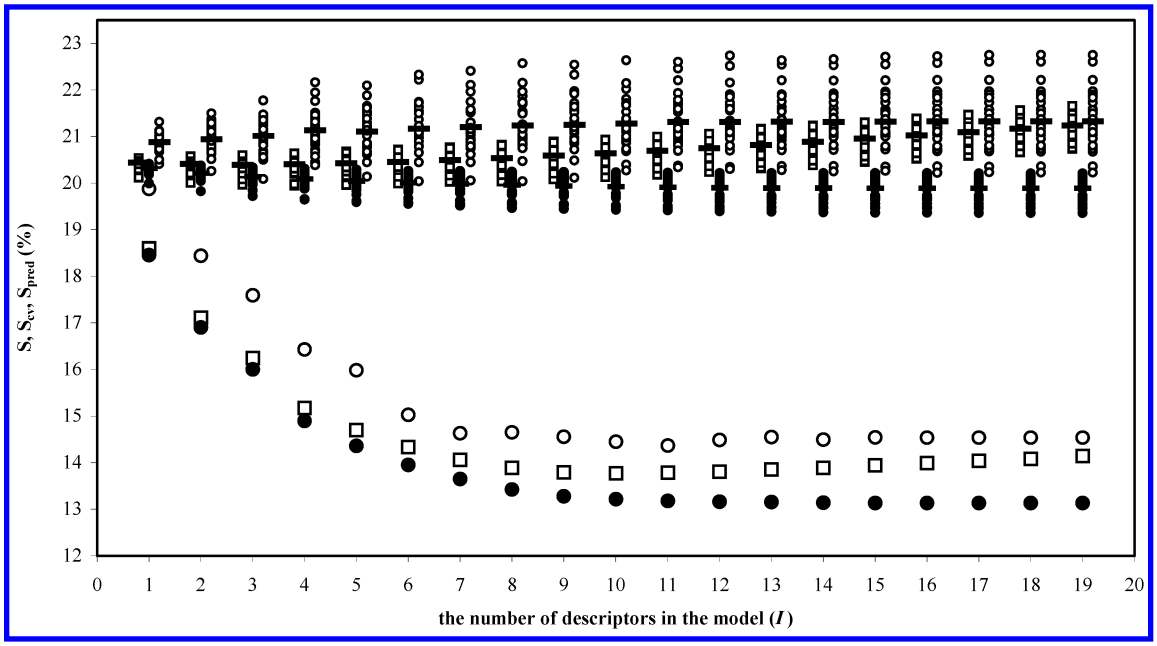

Figure 1. Scatter plot of the fitted (S), cross-validated (S_{cv}) standard errors for the 317-protein training set, and the standard error of prediction (S_{pred}) for the 158-protein test set versus the total number of descriptors involved in the model for the percentage of the α secondary structure content. Larger filled circles, open squares, and open circles are used for denoting S, S_{cv}, and S_{pred} of real models, respectively. Smaller marks given in the upper part of figure are related to the corresponding random models. Averages of standard errors of 30 random models for each I are marked by small bars. To make it more visible, S_{cv} (small open squares) and S_{pred} (small open circles) for random models are placed in front and behind of S (small filled circles), respectively.

Table 1. Best Selected Models for the Percentage of α, *β*, and Coil Secondary Structure Contents Obtained on the 317-Protein Training Set and Tested on the 158-Protein Test Set[a]

I	S	S_{cv}	S_{abs}	$S_{abs,cv}$	details of the model, 317-protein data set[b]
			A. Percentage of the α Secondary Structure Content		
9	13.15	13.58	10.50	10.82	% α = (95.75 ± 6.71) + (9.68 ± 1.18)·x_{ala} + (9.68 ± 1.18)·x_{cys} + (−11.34 ± 1.76)·x_{val} + (−11.18 ± 1.62)·x_{thr} + (−9.75 ± 1.46)·x_{gly} + (−6.64 ± 1.54)·x_{ser} + (−4.87 ± 1.90)·x_{asp} + (−7.46 ± 2.07)·x_{asn} +(−17.16 ± 2.03)·x_{pro}
			prediction of % α on the 158-protein test set using this model: S_{pred} = 14.49, $S_{abs,pred}$ = 11.19		
			B. Percentage of the *β* Secondary Structure Content		
5	11.45	11.73	8.99	9.19	% *β* = (5.39 ± 3.78) + (−7.03 ± 0.95)·x_{ala} + (−5.82 ± 2.83)·x_{met} + (11.20 ± 1.46)·x_{val} + (9.09 ± 1.40)·x_{thr} + (3.97 ± 1.30)·x_{ser}
			prediction of % *β* on the 158-protein test set using this model: S_{pred} = 13.30, $S_{abs,pred}$ = 10.75		
			C. Percentage of the Coil Secondary Structure content		
8	9.20	9.52	6.94	7.14	% coil = (7.61 ± 4.48) + (−2.26 ± 0.81)·x_{ala} + (7.96 ± 1.18)·x_{cys} + (4.41 ± 2.04)·x_{his} + (7.45 ± 1.01)·x_{gly} + (3.17 ± 1.06)·x_{ser} + (4.64 ± 1.31)·x_{asp} + (5.41 ± 1.44)·x_{asn} + (15.40 ± 1.41)·x_{pro}
			prediction of % coil on the 158-protein test set using this model: S_{pred} = 9.85, $S_{abs,pred}$ = 7.57		

[a] The sum of predicted percentages for these three models is normalized to be 100%, as described in the text. Statistical parameters S, S_{cv}, S_{abs}, $S_{abs,cv}$ are obtained on the training set containing 317 proteins. I = the total number of descriptors involved in the multiregression model. [b] Regression coefficient and the corresponding error of regression coefficient are given in parentheses for each frequency of amino acid residue involved in the model. Frequencies of amino acid residues (x_{ala}, x_{met}, ...) are denoted by their three-letter code.

model is dictated both by its correlation with modeled property (the percentage of secondary structure content) and, at the same time, by its intercorrelations with all other frequencies of amino acid residues (descriptors) in the protein data set used for modeling. If the correlation coefficient of a descriptor with modeled property is higher, and, if at the same time, the correlation coefficients of this descriptor with the remaining descriptors are smaller, then its inclusion in the best model is more probable. Finally, selection of models, from 20 initial descriptors, having $I < 20$ descriptors according to the statistical rules described in the preceding subsection (named "Rules for Selecting the Best Models") necessarily lead to the models that do not contain some frequencies of amino acid residues (descriptors) that are classified, according to their propensities, as formers or even as strong formers of α, *β*, and coil secondary structure, because of their relatively high intercorrelation with other frequencies of amino acid residues.

On the other hand, the propensity of amino acid residues for forming/breaking α, *β*, and coil secondary structure, as described by the propensity scales in refs 20−22, are only based on the occurrence of a single amino acid residue in a

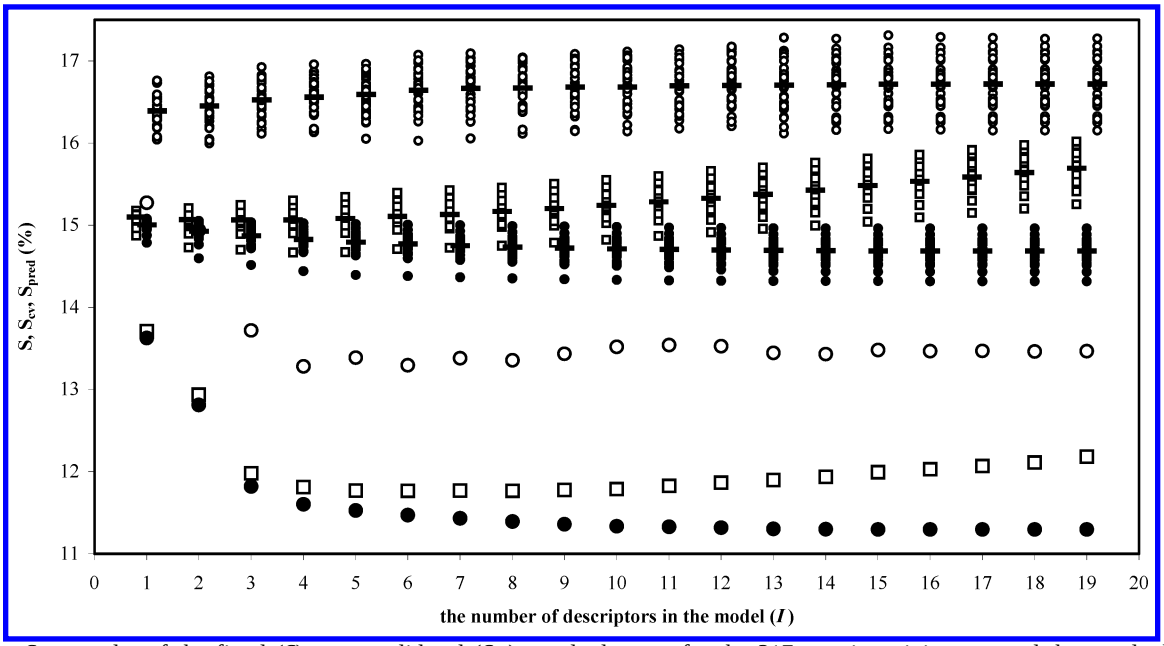

Figure 2. Scatter plot of the fitted (S), cross-validated (S_{cv}) standard errors for the 317-protein training set, and the standard error of prediction (S_{pred}) for the 158-protein test set versus the total number of descriptors involved in the model for the percentage of the β secondary structure content. For additional explanation see the Figure 1 caption.

specific secondary structure. A propensity of an amino acid residue for a secondary structure does not directly take into account the fact that some other similar amino acid residues also may have similar behavior, and consequently, similar propensity value for the same secondary structure as well as similar frequencies of occurrences in a large set of proteins.

Interpretation of the Models Obtained on the 317-Proteins Training Set. The nine-descriptor model for the percentage of α secondary structure content involved alanine (ala) as the most significant descriptor (significance of a descriptor is measured by the ratio between the regression coefficient and its corresponding error). It is well-known that alanine is a strong former of the α-helix secondary structure, and we can see that in this model only alanine has the positive regression coefficient.[20−22] Aspartate (aspartic acid, asn), which is an extremely weak helix former according to Table 11 in ref 21, has the negative regression coefficient although it tends to form the helix secondary structure. Other amino acids are either indifferent for forming helices in soluble proteins or are the helix breakers (see Table 11 in ref 21), and all have negative regression coefficients. This means, if a frequency of amino acid residues belonging to this subgroup (cys, val, thr, gly, ser, asn, pro) in a protein increases, then the percentage of the α-helix secondary structure decreases.

In the five-descriptor model for predicting the percentage of the β secondary structure content valine (val), alanine, and threonine (thr) are the most significant descriptors (Table 1). Valine is a strong former, while threonine is a weak former of the beta structure (Table 11 in ref 21), and both have the positive regression coefficient. Alanine participates in the model for predicting the β secondary structure content as a β breaker and has the negative regression coefficient. Frequencies for methionine and serine, which are classified as indifferent for forming and breaking the β structure, correct estimated values obtained by the three most significant amino acid residues.

In the eight-descriptor model for predicting the percentage of the coil secondary structure (Table 1), proline, glycine, and cysteine are the most significant descriptors that tend to form an undefined structure (i.e., these amino acid residues tend to break the regular α and β secondary structure), and all have the positive regression coefficient. Such properties of proline and glycine are known and are in accordance with their conformational preferences given in refs 20−22 (Tables I, XI, and I, respectively). However, in refs 20−21 cysteine was classified as an amino acid residue indifferent for α and even as a weak β-former. This difference can also be caused by the fact that this 317-protein data set is much larger than those used for calculating conformational parameters in refs 20−22. Contributions of other amino acid residues involved in the model for coil can be interpreted as correcting ones.

Random Models. In addition, for each type of secondary structure, and for each I, 30 of the best possible random models were selected by the CROMRsel. The random models were obtained starting, each time, from a randomly ordered column/vector containing experimental percentages of the secondary structure content of proteins. Standard errors of fit and LOO CV of random models for the training set containing 317 proteins and the standard error of prediction of random models for the test set containing 158 proteins are also given in Figures 1−3 for α, β, and coil, respectively. Standard errors related to the random models in Figures 1−3 are designated by small symbols (filled circles are for fit (S), open squares for LOO CV (S_{cv}), and open circles for prediction (S_{pred}) on the test set). In addition, for each I standard errors of 30 random models are given as well as their averages (designated as bars). Each type of the standard error of random models should be compared with the corresponding fit, CV, or predictive standard error of the best real models.

We can see from Figures 1−3 that there is no risk of a chance correlation, i.e., that obtained models are far from being random. This is also confirmed by a significant agreement between the conformational parameters given in

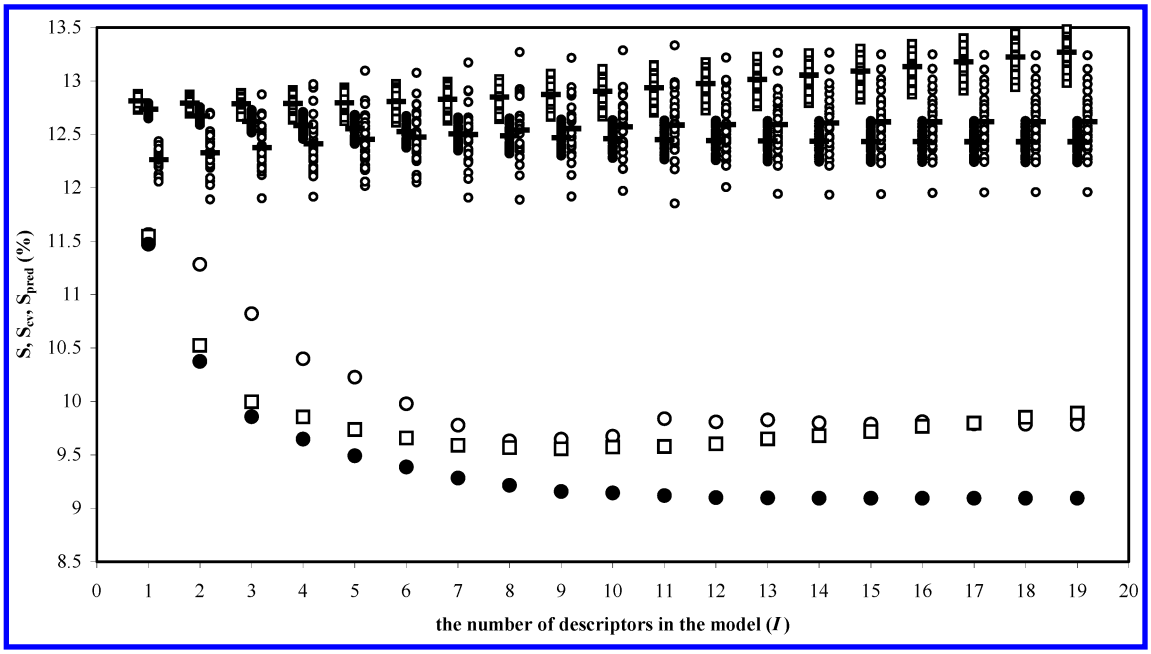

Figure 3. Scatter plot of the fitted (S), cross-validated (S_{cv}) standard errors for the 317-protein training set, and the standard error of prediction (S_{pred}) for the 158-protein test set versus the total number of descriptors involved in the model for the percentage of the coil secondary structure content. For additional explanation see the Figure 1 caption.

refs 20−22, which were derived from three different data sets of proteins, and significance of frequencies of amino acid residues (descriptors) involved in models given in Table 1, and discussed above.

The averages of the fit standard errors (S) of random models for α, β, and coil (Figures 1−3) are constant or decrease slowly by increasing the number of frequencies of amino acid residues (I). The averages of the LOO CV standard errors (S_{cv}) of random models for α, β, and coil increase slowly by increasing I, and the averages of the predictive standard errors (S_{pred}) are almost constant. These facts only indicate an extremely low sensitivity of random models on increasing the number of optimized parameters in the models. However, the standard error of random models selected in this study does not depend on the total number of possible models (which is equal to $20/I!(20-I)!$) that should be checked in order to select a single best model having I descriptors chosen among 20 descriptors. At the same time, we can see that all the corresponding real standard errors reach their minimum values when 5−9 frequencies are involved in the models. If the accuracy of real models is dependent on the number of models that should be checked in order to select the best possible model (according to the highest correlation coefficient), than we expect the same behavior for random models too. After performing this comparative analysis of real and random models, we can safely state that there is no evidence that real models selected by the CROMRsel procedure are obtained by chance.

B. Results Obtained Using the 475-Protein Data Set from Ref 4. Using the CROMRsel program all the best possible models containing I descriptors (I = 1, 2, ..., 20) were selected from the complete data set for the α, β, and coil secondary structure content. Since we wanted to perform a strict comparison with the published models for the 475-protein data set from ref 4, fitted (S_{abs}) and LOO cross-validated ($S_{abs,cv}$) mean absolute errors were calculated for each model and are given in Figure 4.

S_{abs} (filled symbols) and $S_{abs,cv}$ (open symbols) in Figure 4 are designated as circles, squares, and diamonds for α, β, and coil, respectively. Additionally, for each model in Figure 4 corresponding values of standard errors are also given (Figure 5).

Based on an analysis of the fitted statistical parameters for the models containing I descriptors (I = 1, 2, ..., 20) in Figures 4 and 5, we selected models containing seven (I = 7), four (I = 4), and seven (I = 7) descriptors as the best final models for the α, β, and coil secondary structure content, respectively. Details of these models are given in Table 2, together with the statistical parameters of the corresponding models published in ref 4. Frequencies involved in the best models are closely related, for each secondary structure type, to those obtained and discussed for the 317-protein training set.

From the comparison of statistical parameters, we can see that the models obtained in this work are superior to those from in ref 4, for the same (475-protein) data set. Additionally, the models selected by the CROMRsel program in this study are much simpler. For example, the linear models for α, β, and coil in ref 4 involved 57 optimized parameters, altogether. Corresponding nonlinear models in ref 4 for predicting the percentage of three secondary structures contain 247 optimized parameters. Linear models for percentage of α, β, and coil developed on the 475-protein data set in this study involve 18 descriptors and 21 optimized parameters, altogether, and are better than corresponding linear and nonlinear models from ref 4 (see Table 2). Simplicity of the developed models facilitate analysis and understanding of relationships between information contained in the protein sequence and global structural information, like the percentage of α, β, and coil secondary structure contents in soluble proteins.

In addition, it is self-understandable that it is better to have more concise than more complex models. Secondary structure contents (as well as each other's protein property) cannot

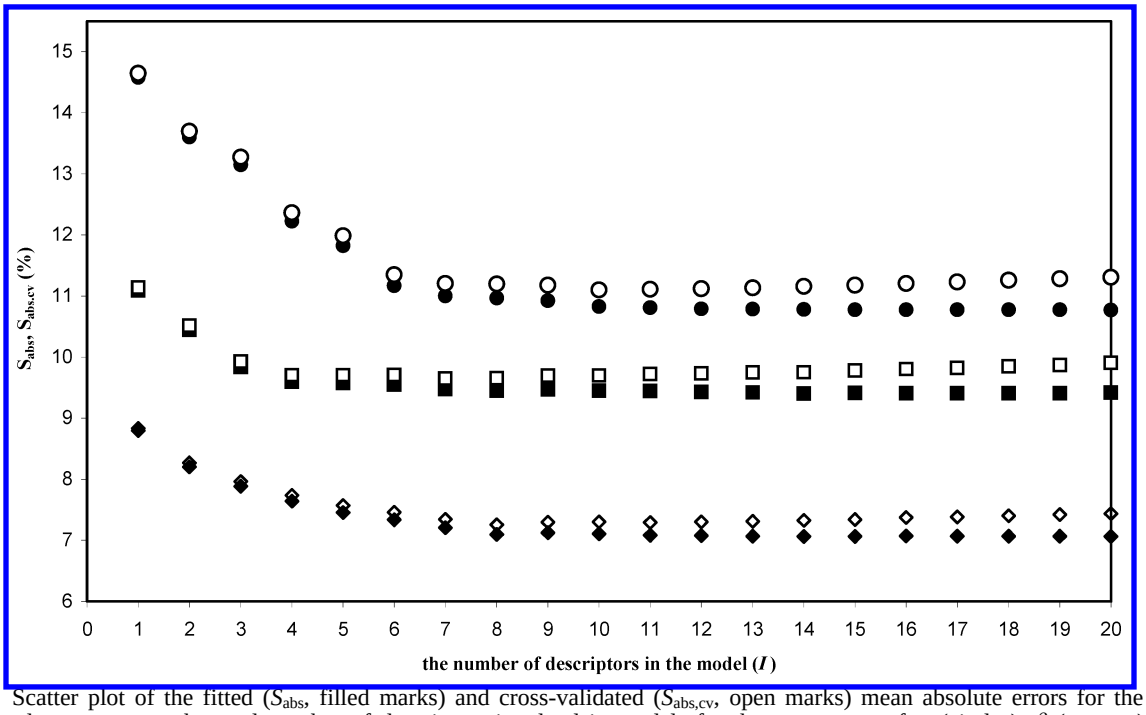

Figure 4. Scatter plot of the fitted (S_{abs}, filled marks) and cross-validated ($S_{abs,cv}$, open marks) mean absolute errors for the complete 475-protein data set versus the total number of descriptors involved in models for the percentage of α (circles), β (squares), and coil (diamonds) secondary structure contents.

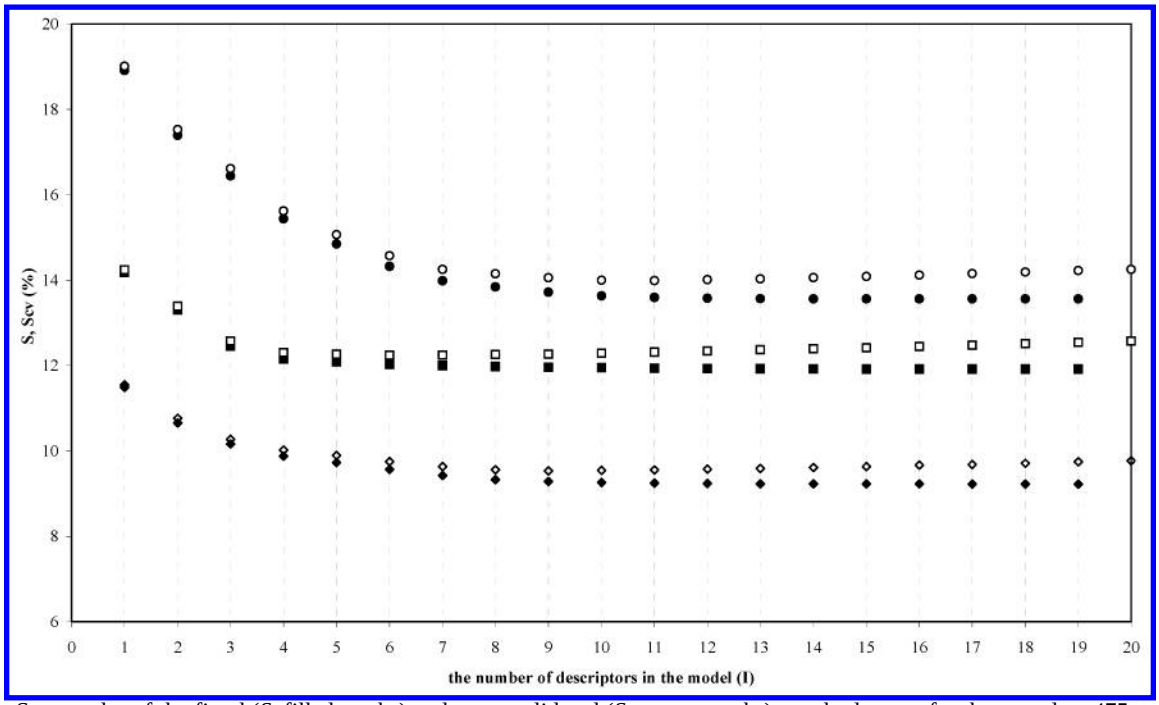

Figure 5. Scatter plot of the fitted (*S*, filled marks) and cross-validated (S_{cv}, open marks) standard errors for the complete 475-protein data set versus the total number of descriptors involved in models for the percentage of α (circles), β (squares), and coil (diamonds) secondary structure contents.

be modeled without errors. The protein secondary structure content is only one of the properties of protein structure, and we show in this study that, to the some extent, they can be modeled without taking into account all 20 amino acid residues. Simplification of the protein secondary structure models because of the statistical reasons and exclusion of some amino acid residues from the models (although they are classified in refs 20−22 as formers of a specific secondary structure type) does not mean that these amino acid residues are not important for the complete protein structure or for other aspect/properties of protein structure. This simply means that there is redundancy between the frequencies of occurrences of amino acid residues. Because of that, by inclusion of additional amino acid residues although there are no valid statistical reasons for enlargement of the models, obtained models become more complex and more difficult for analysis, and the quality (accuracy) of the final models is not improved.

CONCLUSION

It is clear that easily interpretable models should be preferred over the complex ones. This is more easily achieved

Table 2. Best Selected Models for the Percentage of α, *β*, and Coil Secondary Structure Contents Obtained on the Complete 475-Protein Set in This Study and in Ref 4[a]

I	*S*	S_{cv}	S_{abs}	$S_{abs,cv}$	details of the model, this study[b]
			A. Percentage of the α Secondary Structure Content		
7	13.76	13.99	10.76	10.93	$\% \alpha = (79.22 \pm 3.97) + (10.00 \pm 0.98)\cdot x_{ala} + (-7.16 \pm 1.50)\cdot x_{cys} + (-10.11 \pm 1.49)\cdot x_{val} + (-10.31 \pm 1.38)\cdot x_{thr} + (-9.12 \pm 1.23)\cdot x_{gly} + (-7.96 \pm 1.30)\cdot x_{ser} + (-15.07 \pm 1.75)\cdot x_{pro}$
19[c]			14.7	14.7	linear model in ref 4 (Table 4, page 165)
~82[c]			13.3	13.7	nonlinear model in ref 4 (Table 4, page 165)
			B. Percentage of the *β* Secondary Structure Content		
4	12.12	12.28	9.52	9.64	$\% \beta = (-1.87 \pm 2.85) + (-6.39 \pm 0.80)\cdot x_{ala} + (11.86 \pm 1.26)\cdot x_{val} + (9.39 \pm 1.19)\cdot x_{thr} + (5.46 \pm 1.12)\cdot x_{ser}$
19[c]			12.0	12.1	linear model in ref 4 (Table 4, page 165)
~82[c]			12.0	12.6	nonlinear model in ref 4 (Table 4, page 165)
			C. Percentage of the Coil Secondary Structure Content		
7	9.46	9.64	7.28	7.40	$\%\ \text{coil} = (12.35 \pm 3.34) + (-2.69 \pm 0.66)\cdot x_{ala} + (8.04 \pm 0.99)\cdot x_{cys} + (6.77 \pm 0.82)\cdot x_{gly} + (3.29 \pm 0.88)\cdot x_{ser} + (5.08 \pm 1.08)\cdot x_{asp} + (4.70 \pm 1.18)\cdot x_{asn} + (14.50 \pm 1.20)\cdot x_{pro}$
19[c]			12.7	12.8	linear model in ref 4 (Table 4, page 165)
~82[c]			11.2	11.4	nonlinear model in ref 4 (Table 4, page 165)

I	S_{av}, (S_{min},S_{max})	$S_{cv,av}$, ($S_{cv,min}$, $S_{cv,max}$)	this study
	D. Standard Errors of 30 Random Models for the %α, %*β*, and %coil Secondary Structure Content[d]		
7	20.21, (19.98, 20.40)	20.55, (20.32, 20.78)	random models for α
4	15.29, (15.16, 15.41)	15.46, (15.33, 15.58)	random models for *β*
7	15.26, (15.10, 15.42)	15.52, (15.37, 15.68)	random models for coil

[a] The sum of predicted percentages for these three models are normalized to be 100%, as described in the text. Statistical parameters *S*, S_{cv}, S_{abs}, and $S_{abs,cv}$ are obtained on the 475-protein set. I = the total number of descriptors involved in the multiregression model. The total number of optimized parameters in multiregression models selected in this study is $I + 1$. [b] Regression coefficient and the corresponding error of regression coefficient are given in parentheses for each frequency of amino acid residue involved in the model. [c] In ref 4 (p 162), it was calculated that there are 57 independent parameters (optimized parameters) for all three secondary structure types in linear and 247 optimized parameters in nonlinear models, respectively. From these numbers, we calculated that there were, in average, $57/3 = 19$ optimized parameters in linear and $247/3 \approx 82$ optimized parameters in nonlinear models for one secondary structure type. [d] For each secondary structure type 30 "the best possible" random models containing 7, 4, and 7 descriptors were selected from 20 initial descriptors for the percentage of α, *β*, and coil secondary structure content, respectively. Fitted and cross-validated average (S_{av}, $S_{cv,av}$), minimal (S_{min}, $S_{cv,min}$), and maximal (S_{max}, $S_{cv,max}$) values of standard errors were calculated (cross-validated values are denoted by cv).

if one uses the simple functional form of models relating structural parameters (descriptors) with the property/activity of molecules. As it is shown in this study, the linear multiregression, as the simplest modeling procedure giving the simplest functional form of the model, together with an efficient procedure for selection of the most important descriptors, can produce very concise and very good models.

The CROMRsel procedure developed and tested in the field of QSAR/QSPR is successfully applied to the problems belonging to the field of biophysics/bioinformatics. We show that less than half of the amino acid residues are enough to obtain good predictive models. Including additional frequencies of amino acid residues does not improve neither fit nor cross-validated or predictive performance of the models for modeling the percentage of α, *β*, and coil secondary structure contents in soluble proteins.

It is also shown by performing partition of the complete data set into the training and test sets and by repeated randomization tests that there is no risk of random correlation, i.e., the probability that the models obtained in this paper were selected by chance among a lot of checked models is very small.

The models obtained for the 475-protein set in this study are strictly compared with the published models obtained in ref 4 for the same data set. Models containing only 7, 4, and 7 frequencies for α, *β*, and coil, respectively, are much simpler than the corresponding models in ref 4 and much better according to the same statistical parameters. These models based only on linear frequencies of amino acid residues as descriptors are even better than nonlinear models in ref 4 (that take also into account compositional couplings, i.e., cross-products of frequencies).

ACKNOWLEDGMENT

We are grateful to Frank Eisenhaber for providing the data set containing 475 proteins used in ref 4 and in this paper. The authors thank anonymous reviewers for their valuable and useful comments that resulted in improvement of this paper. This work was supported by the Ministry of Science and Technology of the Republic of Croatia through Grants 0098034 (N.T. and B.L.).

Supporting Information Available: Experimental percentage of α, *β*, and coil secondary structure contents and frequencies of 20 amino acid residues of 475 proteins including details related to the partition of data set into the 317-protein training and 158-protein test sets (Table S1). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Sreerama, N.; Woody, R. W. Protein secondary structure from circular dichroism spectroscopy. *J. Mol. Biol.* **1994**, *242*, 497−507.

(2) Bussian, B. M.; Sander, C. How to Determine Protein Secondary Structure in Solution by Raman Spectroscopy: Practical Guide and Test Case DNsae. *Biochemistry* **1989**, *28*, 4271−4277.
(3) Zhang, Z.; Sun, Z.-R.; Zhang, C.-T. A New Approach to Predict the Helix/Strand Content of Globular Proteins. *J. Theor. Biol*. **2001**, *208*, 65−78.
(4) Eisenhaber, F.; Imperiale, F.; Argos, P.; Frömmel, C. Prediction of Secondary Structural Content of Proteins from Their Amino Acid Composition Alone. I. New Analytic Vector Decomposition Methods. *Proteins* **1996**, *25*, 157−168.
(5) Zhang, C.-T.; Lin, Z.-S.; Zhang, Z.; Yan, M. Prediction of the Helix/Strand Content of Globular Proteins Based on Their Primary Strucutre. *Protein Eng*. **1998**, *11*, 971−979.
(6) Chou K. C.; Liu, W.-M. Prediction of Protein Secondary Strucutre Content. *Protein Eng*. **1999**, *12*, 1041−1050.
(7) Lučić, B.; Trinajstić, N. Multivariate Regression Outperforms Several Robust Architectures of Neural Networks in QSAR Modeling. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 121−132.
(8) Lučić, B.; Trinajstić, N.; Sild, S.; Karelson, M.; Katritzky, A. R. A New Efficient Approach for Variable Selection Based on Multiregression: Prediction of Gas Chromatographic Retention Times and Response Factors. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 610−621.
(9) Lučić, B.; Amić, D.; Trinajstić, N. Nonlinear Multivariate Regression Outperforms Several Concisely Designed Neural Networks in QSAR Modeling. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 403−413.
(10) Basak, S. C.; Gute, B. D.; Lučić, B.; Nikolić, S.; Trinajstić, N. A Comparative QSAR Study of Benzamidines Complement-Inhibitory Activity and Benzene Derivatives Acute Toxicities. *Comput. Chem*. **2000**, *24*, 181−191.
(11) Katritzky, A. R.; Chen, K.; Wang, Y.; Karelson, M.; Lučić, B.; Trinajstić, N.; Suzuki, T.; Schüürmann, G. Prediction of Liquid Viscosity for Organic Compounds by a Quantitative Structure−Property Relationship. *J. Phys. Org. Chem*. **2000**, *13*, 80−86.
(12) Lučić, B.; Bašic, I.; Nadramija, D.; Miličević, A.; Trinajstić, N.; Suzuki, T.; Petrukhin, R.; Karelson, M.; Katritzky, A. R. Correlation of Liquid Viscosity with Molecular Structure for Organic Compounds Using Different Variable Selection Methods. *Arkivoc* **2002**, (IV), 45−59 (http://www.arkat-usa.org/ark/journal/2002/Sunko/DS-381D/DS-381D.pdf).
(13) Lučić, B.; Miličević, A.; Nikolić, S.; Trinajstić, N. Harary Index − Twelve Years Later. *Croat. Chem. Acta* **2002**, *4*, 847−868.
(14) Lučić, B.; Nadramija, D.; Bašic, I.; Trinajstić, N. Towards Generating Simpler QSAR Models: Nonlinear Multivariate Regression versus Several Neural Network Ensembles and Some Related Methods. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1094−1102.
(15) 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*. **1977**, *112*, 535−542.
(16) 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; pp 107−132.
(17) 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.* **1992**, *8*, 599−600.
(18) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structures: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577−2637.
(19) Hoffmann, R.; Minkin, V. I.; Carpenter, B. K. Ockham's Razor and Chemistry. *Bull. Soc. Chim. Fr*. **1996**, *133*, 117−130.
(20) Prevelige, P., Jr.; Fasman, G. D. Chou-Fasman Prediction of the Secondary Structure of Proteins − The Chou-Fasman-Prevelige Algorithm, in: *Prediction of Protein Structure and the Principles of Protein Conformation*; Fasman, G. D., Ed.; Plenum Press: New York and London, 1989; pp 391−916.
(21) Chou, P. Y. Prediction of Protein Structural Classes from Amino Acid Compositions, in *Prediction of Protein Structure and the Principles of Protein Conformation*; Fasman, G. D., Ed.; Plenum Press: New York and London, 1989; pp 549−586.
(22) Deléage, G.; Roux, B. Use of Class Prediction to Improve Protein Secondary Structure Prediction. In *Prediction of Protein Structure and the Principles of Protein Conformation*; Fasman, G. D., Ed.; Plenum Press: New York and London, 1989; pp 587−597.

CI034037P