JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Protein Secondary Structure Prediction with SPARROW

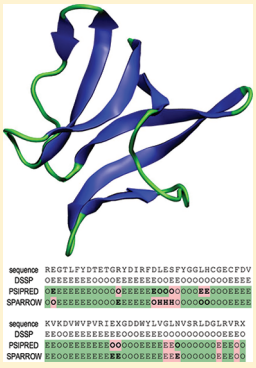
Francesco Bettella,[†,‡,§] Dawid Rasinski,[†,§] and Ernst Walter Knapp*[,†]

[†]Freie Universität Berlin, Institut für Chemie, Fabeckstr. 36a, D-14195 Berlin, Germany

[‡]deCODE genetics, Sturlugata 8, 101 Reykjavik, Iceland

Supporting Information

ABSTRACT: A first step toward predicting the structure of a protein is to determine its secondary structure. The secondary structure information is generally used as starting point to solve protein crystal structures. In the present study, a machine learning approach based on a complete set of two-class scoring functions was used. Such functions discriminate between two specific structural classes or between a single specific class and the rest. The approach uses a hierarchical scheme of scoring functions and a neural network. The parameters are determined by optimizing the recall of learning data. Quality control is performed by predicting separate independent test data. A first set of scoring functions is trained to correlate the secondary structures of residues with profiles of sequence windows of width 15, centered at these residues. The sequence profiles are obtained by multiple sequence alignment with PSI-BLAST. A second set of scoring functions is trained to correlate the secondary structures of the center residues with the secondary structures of all other residues in the sequence windows used in the first step. Finally, a neural network is trained using the results from the second set of scoring functions as input to make a decision on the secondary structure class of the residue in the center of the sequence window. Here, we consider the three-class problem of helix, strand, and other secondary structures. The corresponding prediction scheme "SPARROW" was trained with the ASTRAL40 database, which contains protein domain structures with less than 40% sequence identity. The secondary structures were determined with DSSP. In a loose assignment, the helix class contains all DSSP helix types (α, 3−10, π), the strand class contains β-strand and β-bridge, and the third class contains the other structures. In a tight assignment, the helix and strand classes contain only α-helix and β-strand classes, respectively. A 10-fold cross validation showed less than 0.8% deviation in the fraction of correct structure assignments between true prediction and recall of data used for training. Using sequences of 140,000 residues as a test data set, 80.46% ± 0.35% of secondary structures are predicted correctly in the loose assignment, a prediction performance, which is very close to the best results in the field. Most applications are done with the loose assignment. However, the tight assignment yields 2.25% better prediction performance. With each individual prediction, we also provide a confidence measure providing the probability that the prediction is correct. The SPARROW software can be used and downloaded on the Web page http://agknapp.chemie.fu-berlin.de/sparrow/.



INTRODUCTION

The three-dimensional structure of a protein is a necessary attribute that defines and enables its biological function. Like the sequence, it can be used to explore protein evolution.[1−5] Understanding the relation between sequence and structure in proteins is therefore of great scientific interest.[6] Several experimental techniques enable investigating the three-dimensional structure of proteins, most notably X-ray crystallography, with increasing importance NMR spectroscopy, but also neutron scattering[7] and cryo-electron microscopy.[8] While these experimental techniques have been progressively perfected and modernized, they are still quite time consuming and can hardly keep up with the vast inflow of protein sequence information obtained from large scale DNA sequencing.[9] Prior knowledge of the secondary structure landscape of a protein is generally useful to aid the prediction of its tertiary structure,[10,11] which typically involves methods of *homology modeling*[12] and *protein threading*[13] but also ab initio methods based on energy functions[14,15] that are used for structure optimization[16] or folding[17] simulations. Secondary structure information can be used, for instance, to build safe starting cores to generate a complete protein fold,[18] to set structural constraints for protein threading.[19,20] or to perform homology modeling.[21−23]

The earliest attempts[24−28] of computer-based secondary structure prediction relied on the propensities of specific amino acids to be found within alpha-helices[29] or beta-pleated-sheets,[30] the two secondary structure motifs most frequently observed in proteins. With this approach, only about 60% of the residues could be assigned to the correct class considering three different secondary structure classes [helical, H; extended (strand), E; other (coil, turn ...), O]. Since then, the secondary structure prediction techniques have been refined, and the accuracy has increased considerably, while sticking to the three-class problem. Modern techniques include decision trees,[31] neural networks,[32−38] support vector machines[39−42] or hidden Markov models,[43−45] and other machine learning devices.[46,47]

Virtually, a quantum jump in secondary structure prediction accuracy was made using profiles from protein sequence

Received: July 13, 2011
Published: January 7, 2012

alignments[34,48,49] instead of using the sequences directly. The main advantage of using profiles (generated for whole proteins of the same family, instead of the amino acid sequences or the corresponding values from the BLOSUM substitution matrix[50]) is that these effectively map global sequence information characteristics for the protein into the relatively small sequence window explicitly used for training and prediction of secondary structures. Using sequence profiles, modern prediction tools[35−37,40,45,46,51−62] determine protein secondary structures for the three-class problem on average with more than 80% accuracy.

It is well established that that under physiological conditions three-dimensional native protein structures are determined by their amino acid sequences.[63] However, the protein native structure can easily be influenced by subtle environmental factors. Different "microenvironments" can be due to (1) crystal contacts in the unit cell, (2) residue locations (being surface exposed or buried), (3) different polypeptide chains in a multichain protein, or (4) contacts of residues with cofactors. Protein structures are often only weakly stable such that thermal unfolding of a protein can occur already at temperatures slightly above room temperature. Thus, native protein structures (and therefore also secondary structures) can show some variations that are not governed by the polypeptide sequence alone, setting theoretical upper limits to sequence-based protein structure prediction. Such effects may be enhanced when a rigorous secondary structure assignment to specific classes (for instance DSSP[64]) is used. Assignments at the borderline of different classes can be shifted to another class by very small influences. Thus, depending on the nature of the protein crystal structures and the classification scheme used, there is an accuracy limit of secondary structure prediction. Monitoring the variation of secondary structures of homologous proteins, Rost[65] estimated this limit to be at 88% (for a three-class classification scheme of protein secondary structure using DSSP[64] to define the classes).

In the present study, only those secondary structure prediction programs were considered for comparison that could be downloaded and thus be used in extensive computations. These are PSIPRED_2.6,[35] PROF,[52] Prospect_2,[66] SSpro_4.03,[67] while the popular secondary structure prediction program PredictProtein[68] by Rost and Sander[69] was not considered. Because empirical prediction schemes use only protein structures that were available at the time they were created, an unequivocal ranking among them is not trivial. The EVA project[70−72] for the evaluation of automatic protein secondary structure prediction was a good testing platform for automatic prediction servers. Unfortunately, the EVA service was discontinued some time ago. Competitions like CASP[73] (critical assessment of protein structure prediction) provided among others a test ground for protein secondary structure prediction. In the CASP9[74] contest, 128 protein structures were made available with a total of 28,908 residues. The size of the test set used in the present study is about five times larger and contains only protein domains with less than 40% sequence identity. The CASP9 data set is not representative nor is it large enough to guarantee a statistical error of less than 0.5%.

Despite the many secondary structure prediction tools that are available, we believe that there is still room for improvement, particularly when new machine learning methods are combined with the growing database of protein structures. The new protein secondary structure prediction tool SPARROW (secondary structure prediction using arrays of optimized weights) is the result of this approach.

In the present contribution, we present the most relevant features of SPARROW (http://agknapp.chemie.fu-berlin.de/sparrow/). We also provide the basis for a fair comparison of protein secondary structure prediction software and suggest a standard procedure to benchmark protein secondary structure prediction tools.

A large body of protein structures is solved by X-ray crystallography, while only a few structures are determined in solution by NMR spectroscopy. Hence, the influence of the latter is statistically not relevant. The same conclusion holds for membrane proteins. Therefore, these two types of protein structures are not considered in the present study.

To be concise, the SPARROW software correlates the local secondary structure of a specific residue in a protein with the profile in a sequence window centered on this residue. Thereby, the sequence is represented as a sequence profile generated by multiple sequence alignment with PSI-BLAST.[49] In the first step, scoring functions with optimized parameters are used to relate the profile of a sequence window (of width 15) centered on the considered residue with its secondary structure. In the second step, the resulting secondary structure propensities for all residues in this window are again correlated with the secondary structure of the central residue. In the third step, the results of the scoring functions from the second step are lumped together using a neural network with three output neurons corresponding to the three structure classes. More details of the procedure are given in the Method section.

METHOD

Classification Strategies of SPARROW. Like other sequence-based empirical secondary structure prediction tools, SPARROW employs a statistical learning procedure that fits in the mathematical frame of machine learning outlined by Vapnik.[75] On the basis of existing protein structure data, SPARROW infers a mathematical correlation between sequence windows of 15 consecutive amino acids in a protein and the secondary structure of the *central* residues of these windows. This correlation is established in three stages. The first two stages are based on a multi-linear regression technique.[76] The third stage involves an artificial neural network.

In the first two stages, the correlation procedure is mediated by a complete set of *independent* 2Class schemes. These classification problems can be easily solved using a technique[41] analogue to Fisher's linear discriminant analysis.[77] Appropriate pairs of classes can be obtained with two different strategies: (1) All secondary structures of one specific type are assigned to one class (the reference class), while all other structural data are placed in a corresponding second class (*one-against-rest*: *1-r*). (2) Alternatively, one can select two different classes from the set of all classes, while ignoring the data of the remaining classes (*one-against-one*: *1-1*).[78] In the *1-r* case, for each one of the possible 2Class problems, all data are considered in the training procedure. In the *1-1* case, only data belonging to the two corresponding classes are considered. The different possible classification schemes are depicted in Figure 1 for $m = 3$ classes [helical (H), extended (E), other (O)].

Scoring Functions. The first two stages of the classification procedure for protein secondary structures involve scalar scoring functions. These functions establish a correlation between objects (o_j) belonging to a sample space $o_j \in X$

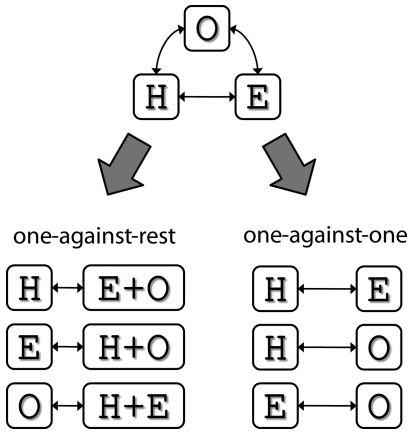


Figure 1. Multi-class classification. A classification problem with three classes (H, E, O) can be solved by splitting it into a set of two-class classification problems (2Class problem). These classification problems can be solved in two ways: either by separating one class from all other classes (*one-against-rest*: *1-r*) or by separating two specific classes from one another ignoring all remaining classes (*one-against-one*: *1-1*). In the case of three classes ($m = 3$) considered here, a total of six (3 + 3) different 2Class problems exist; each one can be solved using a method related to Fisher's linear discriminant analysis.[41,77].

characterized by feature vectors $\mathbf{x}_j \in \mathbb{R}^n$ and the corresponding scalar target values $\hat{y}_j$. The scoring function f involves parameters whose values are determined in an optimization procedure (training) such that the value of the scoring function

$$y_j = f(\mathbf{x}_j) \tag{1}$$

is as close as possible to the proper target value $\hat{y}_j$ of object o_j. For the 2Class problems of the present application the target value, $\hat{y}_j$ is unity if the object o_j belongs to the reference class and zero if not. For the *1-r* multi-class scheme, the reference class is the class that has been singled out from all other classes; for the *1-1* multi-class scheme, it is one of the two explicitly considered classes. Hence, given m classes, we need for classification m different scoring functions for case *1-r* and $m(m - 1)/2$ for case *1-1*. After training, the value of the scoring function $f(\mathbf{x}_j)$ provides in the prediction mode a secondary structure propensity measure for the considered object (o_j). This propensity is an absolute measure for *1-r* and a relative measure for *1-1* classification problems. The scoring functions employed in the present application are quadratic in the feature vector $\mathbf{x}$ but linear in the parameter space, i.e.,

$$f(\mathbf{x};\, \mathbf{W},\, \mathbf{w},\, \mathbf{w}_0) = \mathbf{x}^t \times \mathbf{W} \times \mathbf{x} + \mathbf{x}^t \times \mathbf{w} + w_0 \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^n \times \mathbb{R}^n$, $\mathbf{w} \in \mathbb{R}^n$, and $w_0 \in \mathbb{R}^1$ are the parameters of the scoring function subject to optimization.

Profiles and Structure Propensities. For the first stage scoring function, the independent variable $\mathbf{x}$ of the scoring function is the profile of the respective sequence window (width of 15 residues) centered at the reference residue. The profile is obtained by multiple sequence alignment of the corresponding protein with PSI-BLAST[49] up to the third iteration. In the case in which the 15 residue sequence window is partially outside of the polypeptide sequence of the considered protein, the lacking values of the profile are set to −1. The database against which the proteins under investigation are aligned is the nonredundant aggregate of the GenBank,[79] Swiss-Prot,[80] PIR,[81] PRF,[82] PDB,[7] and NCBI RefSeq[83] databases, containing a total of 4,396,331 protein sequences.

In the second regression stage, the PSI-BLAST profile is replaced by the secondary structure propensities calculated in the first stage using the same sequence window (Figure 2). For sequence windows of 15 residues used in this study, the number of parameters of a single scoring function amounts to 45,451 and 4186 in the first and second stage, respectively, where the symmetry of the quadratic terms is considered.

Optimizing the Scoring Function of Stage 1 and 2. The parameters of the scoring function $f(\mathbf{x}_j)$ are optimized such that its value is as close as possible to the corresponding target value $\hat{y}_j$. Hence, the following objective function L, which is quadratic in the parameters, should be minimized

$$\mathrm{L}(\mathbf{W},\, \mathbf{w},\, w_0) = \frac{1}{2} \sum_j [\hat{y}_j - f(\mathbf{x}_j;\, \mathbf{W},\, \mathbf{w},\, w_0)]^2 \tag{3a}$$

A more sophisticated version of the objective function L, shown below, is actually used in the present study. In L* the sequence samples belonging to different secondary structure motifs are reweighted and regularization terms are included yielding

$$\mathrm{L}^*(\mathbf{W},\, \mathbf{w},\, w_0) = \sum_\sigma \frac{k_\sigma}{2N_\sigma} \sum_{j=1}^{N_\sigma} [\hat{y}_j^{(\sigma)} - f(\mathbf{x}_j;\, \mathbf{W},\, \mathbf{w},\, w_0)]^2 + \lambda_{\mathrm{quad}} \mathbf{W}\!:\, \mathbf{W} + \lambda_{\mathrm{lin}} \mathbf{w}^t \cdot \mathbf{w} \tag{3b}$$

where N_σ is the number of occurrences of secondary structure of class σ ($\sigma \in \{\mathrm{H, E, O}\}$ for the three-class case), the $k_\sigma > 0$ are adjustable class weights, and $\lambda_{\mathrm{quad}} > 0$ and $\lambda_{\mathrm{lin}} > 0$ are the regularization weights of the quadratic and linear terms in the scoring function, eq 2, respectively. Minimization of the quadratic form, eq 3b, results in a set of linear equations where the parameters, $\mathbf{W}$, $\mathbf{w}$, w_0, are the unknowns. The regularization terms have a double purpose: (1) If the regularization weights λ are very small, for example of the order of 10^{-12}, they just avoid a potential singular behavior of the linear equation system, which may occur for spurious linear dependencies. (2) If they are larger, they can control the effective number of features. Unimportant features, which possess for instance only weak or no correlation to target values, can be turned off by setting the corresponding parameters, $\mathbf{W}$, $\mathbf{w}$, w_0, close to zero. This tuning of parameters controls overtraining and prevents learning by heart.

Suitable values of k_σ ($\sigma \in \{\mathrm{H, E, O}\}$), λ_{quad} and λ_{lin} (eq 3b) are obtained through cross validation, employing a smaller database of protein structures as explained below. The sum of class weights, k_σ, is normalized to unity. For the *1-r* classification problem, the optimal class weights were found to be $k_\sigma = 0.4$ for the reference class, and $k_\sigma = 0.3$ for the other two classes. For the *1-1* classification problem, we have $k_\sigma = 0.5$ for both classes. Suitable values for regularization weights λ_{quad} and λ_{lin} are in the first stage 10^{-4} and 10^{-12}, respectively. No parameter suppression appears to be necessary in the second stage in which both regularization weights are set to 10^{-12} to avoid singular behavior of the linear equations.

Neural Network in Stage Three. The propensities obtained with the scoring functions from the first two stages

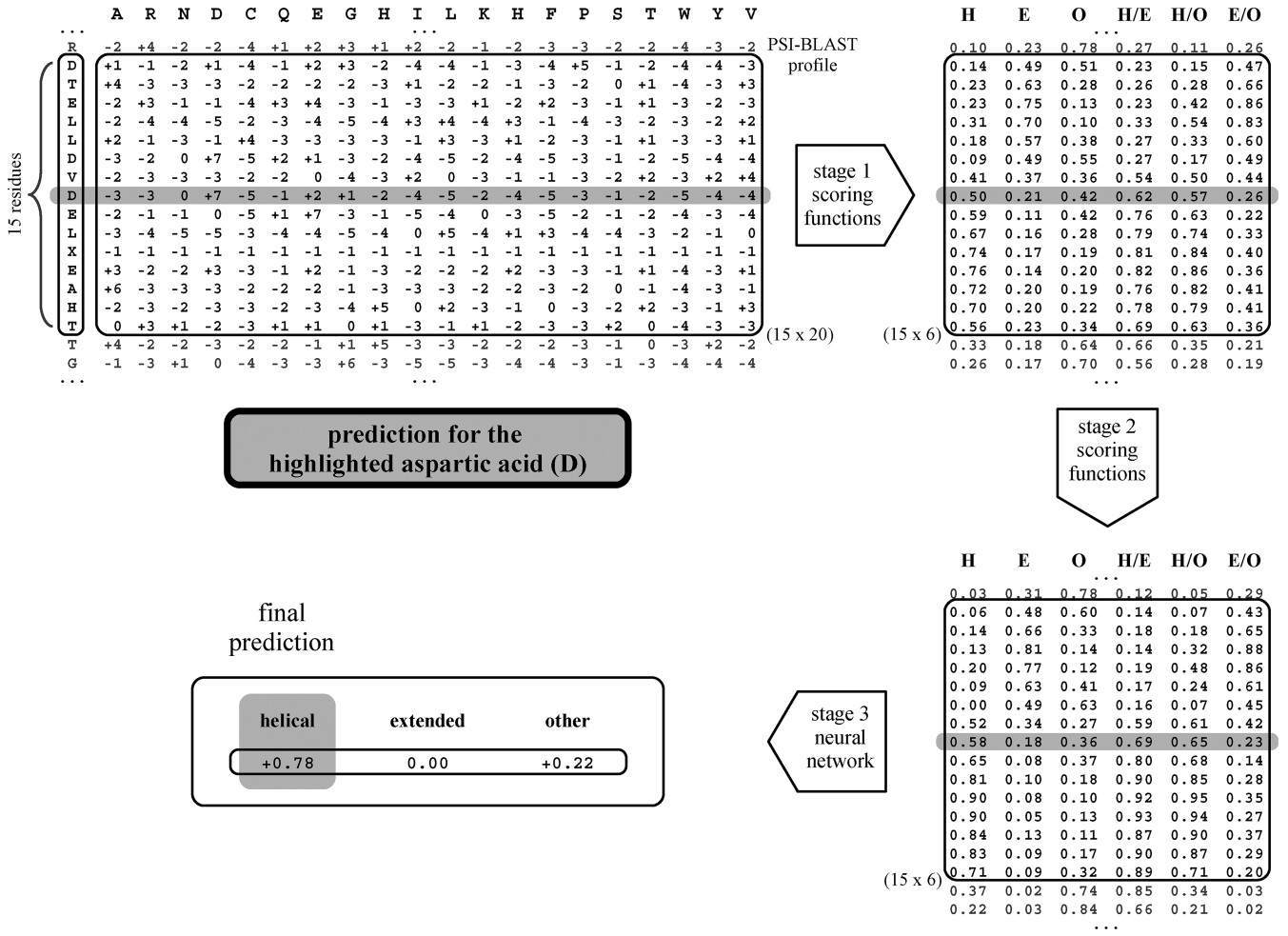


Figure 2. Flow diagram of SPARROW, a protein secondary structure prediction device. To predict the secondary structure for a particular reference residue R [aspartic acid, (D) in the present case], highlighted in gray, a window of the amino acid sequence of 15 residues, width centered at R, is first translated into a PSI-BLAST profile[49] (upper left part). In the first classification stage, these profiles are used as arguments of the first hierarchy of scoring functions (stage 1). They are used to generate an array of secondary structure propensity scores with R in the center of a sequence window of 15 residues (upper right part). These propensities are further processed with structure−structure correlating scoring functions (stage 2) yielding improved secondary structure propensities (lower right part). Finally, in stage 3 a multi-layer perceptron neural network is used to predict the secondary structure at R. It employs the secondary structure propensities obtained in stage 2 for the 15 residues centered at R as values for the input neurons. The three output neurons assign propensities to each of the three secondary structure classes, where the neuron with the highest score determines the class.

can already be used for secondary structure prediction. However, the pairwise classification scheme lacks balancing between the different dual classifiers. To introduce this balance, the propensity scores of all scoring functions from stage 2 are fed to an artificial neural network (Figure 2).

The employed artificial neural network is a multi-layer perceptron trained by back-propagation.[84] The secondary structure propensities obtained with all six stage 2 scoring functions involving the amino acids in the sequence window of 15 residues are fed into the 92 = 15 × 6 + 2 neurons that constitute the neural network input layer. The extra 2 input neurons are used to indicate whether the reference residue is so close to the C- or N-terminus such that the 15 residue window is partially outside of the considered polypeptide sequence. The output layer consists of three neurons (one for each secondary structure class). They yield balanced propensity scores for the three secondary structure classes. Between the input and output layers, there is one hidden layer comprising 100 nodes. Thus, the total number of adjustable parameters of the neural network is 9603 = (92 × 100 + 100) + (100 × 3 + 3), where the extra "100" and "3" account for the threshold values of the neurons in hidden and output layer, respectively.

The values of the scoring functions $f_\sigma(\mathbf{x}_j)$ are approximately in the interval [0, 1] due to the choice of the target values, $\hat{y}_j \in \{0, 1\}$. For technical reasons, the components of the neural network output vectors $[g_H(\mathbf{x}_j), g_E(\mathbf{x}_j), g_O(\mathbf{x}_j)]$ are for the present study approximately in the interval [−1, 1] ([0, 1] in PSIPRED). A two-step transformation

$$\hat{g}_\sigma = [g_\sigma - \min(lb, g_H, g_E, g_O)]/ [\max(1, g_H, g_E, g_O) - \min(lb, g_H, g_E, g_O)] \quad (4a)$$

$$\bar{g}_\sigma = \hat{g}_\sigma / \sum_{\alpha=H,E,O} \hat{g}_\alpha \quad (4b)$$

is applied, where lb is the lower boundary value of the g_σ (−1 for SPARROW and 0 for PSIPRED). The resulting normalized neuron scores $\bar{g}_\sigma$ become pseudoprobabilities obeying $0 < \bar{g}_\sigma < 1$

dx.doi.org/10.1021/ci200321u | J. Chem. Inf. Model. 2012, 52, 545−556

and $\bar{g}_H + \bar{g}_E + \bar{g}_O = 1$. They describe how likely it is for a specific residue to adopt one of the three secondary structure types.

Secondary Structure Databases. The data sets of protein domain structures used for secondary structure prediction are based on ASTRAL40 from the SCOP database.[85−87] Proteins in the ASTRAL40 sets have sequence identities of less than 40%. General information on number of protein domains, number of residues, and release dates of the different versions of ASTRAL40 used in the present study are listed in Table 1.

Table 1. Number of Protein Domains and Residues Taken from ASTRAL40 Databases[85−87] Used for Secondary Structure Prediction

version	1.71[a]	1.73[b]	1.75	1.75−1.73[c]
release date	Jan. 2007	Feb. 2008	June 2009	−
protein domains[d]	7040	9472	10,504	918
residues	1,250,671	1,652,607	1,814,591	146,399

[a]ASTRAL version used to optimize the class and regularization weights k and λ in eq 3b, respectively. [b]ASTRAL version used to optimize the parameters $\mathbf{w}$ and $\mathbf{W}$ of the objective function in eqs 3a and 3b. [c]Test set of protein structures containing all protein domains of the ASTRAL40_1.75 data set that have sequence identities smaller than 40% relative to proteins of the ASTRAL40_1.73 data set. [d]Membrane proteins and proteins for which no consistent DSSP[64] output could be obtained were excluded.

To optimize the regularization weights of the objective function, eq (3b), used for SPARROW, ASTRAL40, version 1.71 (released January 2007) was used, involving 7897 protein domain structures. Partially unresolved structures and transmembrane domains were removed leaving a total of 7040 protein domains as listed in Table 1. These proteins were divided into 10 subsets to carry out cross-validation tests to determine regularization (λ) and class (k) weights used in the objective function, eq 3b, and to measure statistical errors. SPARROW's ability to generalize from the knowledge of training data to unknown test data is also probed on this database.

In order to compare the secondary structure prediction performance of different programs, we trained SPARROW with the newer and larger set of 9472 protein domain structures contained in ASTRAL40_1.73.[85−87] Protein domains belonging to the newest ASTRAL40_1.75 data set (released June 2009), which have sequence similarities of less than 40% with protein domains in ASTRAL40_1.73, were employed as a test set for prediction. This test set will henceforth be called the 1.75−1.73 test set. PDB ids of this test set are listed in Table S3 of the Supporting Information. The choice of not using the newest ASTRAL40 set to train SPARROW is a compromise between using a relatively new and complete set of protein structures and using a well-defined test set of protein structures for prediction. The number of residues in the 1.75−1.73 test set is roughly 10% of the whole ASTRAL40_1.71 set, such that the statistical errors for the prediction obtained from a 10-fold cross validation based on the ASTRAL40_1.71 set can be transferred to the 1.75−1.73 test set.

Secondary Structure Definition with DSSP. The secondary structures were evaluated from the corresponding PDB structure files using DSSP.[64] The eight secondary structure classes of DSSP were merged into three classes in two different ways. The tight assignment considers the following three classes: helical (H, α-helix only), extended (E, β-strand only), and other (O, remaining structure types), while the loose assignment uses helical (H: α, 3−10 and π-helix), extended (E: β-strand and β-bridge), and other (O: remaining structure types). More detailed information regarding the statistics of the 10 subsets is listed in Table S1 of the Supporting Information.

Prediction Quality. The average accuracy of SPARROW at the three stages is reported in Table 2. The quality measures

Table 2. Protein Secondary Structure Prediction with SPARROW for Three Stages Using the Tight Secondary Structure Assignment[a]

quality measure	prediction	recall	stage
Q	0.8011 ± 0.0034	0.8084 ± 0.0005	1
R	0.6969 ± 0.0054	0.7084 ± 0.0008	1
Q	0.8139 ± 0.0037	0.8203 ± 0.0005	2
R	0.7143 ± 0.0057	0.7243 ± 0.0007	2
Q	0.8268 ± 0.0036	0.8342 ± 0.0007	3
R	0.7266 ± 0.0056	0.7384 ± 0.0009	3

[a]The results for prediction (of unknown test data) and recall (of learning data) were computed based on averages of ten-fold cross-validation using the partitioning of ASTRAL40_1.71[85−87] protein domain structures in 10 subsets (Table S1 of the Supporting Information). Nine of the 10 subsets are merged and used for training; the 10th subset is used for prediction. The test sets have an average size of 1.25×10^5 residues, while the training sets contain about 1.15×10^6 residues. Reported are values of the quality measures Q and R, eqs 5a and 5b, averaged over the $M = 10$ subsets according to eq 7 with the corresponding standard deviations evaluated for the training (recall) and test (prediction) subsets. The secondary structure classification is performed with DSSP[64] using the three secondary structure classes of the tight assignment [helical (H): α-helix only; extended (E): β-strand only; other (O): remaining structure types].

considered include accuracy (Q), precision (Prec), and a generalized Matthews correlation coefficient (R).[88,89] All three quality measures can be expressed by the elements of the nonsymmetric confusion matrix C.[90,91] Entry $C_{\sigma\rho}$ of this matrix contains the number of objects that belong to class σ but are predicted to be of class ρ.

The accuracy Q of a prediction device measures the fraction of all objects (N_{total}) that are correctly assigned to their own classes, i.e., the fraction of correct predictions[92]

$$Q = \frac{\sum_\sigma N_\sigma^{(\text{true positives})}}{N_{\text{total}}} = \frac{\sum_\sigma C_{\sigma\sigma}}{\sum_{\rho,\sigma} C_{\rho\sigma}} \tag{5a}$$

Even with random predictions, Q is larger than zero. Q can be close to unity if there is one large class and the prediction device assigns all objects to this class ignoring all other classes. A quality measure for prediction that does not suffer from such drawbacks is the Matthews correlation coefficient R originally defined for 2Class problems.[88] R varies between minus and plus unity where unity (minus unity) means perfect correlation (anti-correlation) between predicted and actual class and zero for random class assignment. A generalized form of this correlation coefficient for more than two classes is defined as[89]

$$R = \frac{\sum_{\sigma,\rho,\mu} (C_{\sigma\sigma}C_{\rho\mu} - C_{\sigma\rho}C_{\mu\sigma})}{\sqrt{\sum_{\sigma,\rho} (C_{\sigma\rho} \sum_{\mu,\nu\neq\sigma} C_{\nu\mu})} \sqrt{\sum_{\sigma,\rho} (C_{\rho\sigma} \sum_{\mu,\nu\neq\sigma} C_{\mu\nu})}} \tag{5b}$$

Neither criterium of prediction quality (Q and R) is class specific. A prediction quality measure that is class specific is for

dx.doi.org/10.1021/ci200321u | J. Chem. Inf. Model. 2012, 52, 545−556

instance the sensitivity (or class specific accuracy)

$$\mathrm{Sens}(\sigma) = \frac{N_\sigma^{(\text{true positives})}}{N_\sigma^{(\text{true positives})} + N_\sigma^{(\text{false negatives})}} = \frac{C_{\sigma\sigma}}{\sum_\rho C_{\sigma\rho}} = \frac{C_{\sigma\sigma}}{N_\sigma} \quad (5c)$$

Sens(σ) does not consider false positive predictions. As a consequence, a prediction device that is very optimistic with respect to class σ would yield Sens(σ) close to unity, but small values for the other classes.

Another prediction quality measure is precision. The precision to predict class σ [Prec(σ)] measures the quality of a prediction device by monitoring the ratio of the number of objects properly assigned to class σ ($N_\sigma^{(\text{true positives})}$) relative to the number of all objects assigned to class σ ($N_\sigma^{(\text{predicted positives})}$) including also the objects falsely assigned to class σ ($N_\sigma^{(\text{false positives})}$)

$$\mathrm{Prec}(\sigma) = \frac{N_\sigma^{(\text{true positives})}}{N_\sigma^{(\text{true positives})} + N_\sigma^{(\text{false positives})}} = \frac{N_\sigma^{(\text{true positives})}}{N_\sigma^{(\text{predicted positives})}} = \frac{C_{\sigma\sigma}}{\sum_\rho C_{\rho\sigma}} \quad (5d)$$

A pessimistic prediction device with respect to class σ would yield Prec(σ) close to unity, but only very few true positives would be found. As a consequence, precisions (Prec) of the other classes would be low. Sens and Prec, eqs 5c and 5d, range between zero and unity. Both may yield a biased view of the prediction quality under certain conditions. While the σ sensitivity considers only prediction events of objects belonging to class σ, the σ precision considers with false positives also objects from the other classes. Therefore, for a multi-class prediction problem it may be more useful to monitor precision rather than sensitivity.

Normalized quality measures for $X \in (Q, \mathrm{Prec})$ are defined in analogy to the normalized neural network output scores, eqs 4a and 4b. These are unity for a perfect prediction device and zero for an intelligent but random prediction, which uses the information about the relative sizes of the different classes $p_0(\sigma) = N_\sigma/N_{\text{total}}$ as follows

$$\bar{X} = (X - X_0)/(1 - X_0) \quad (6)$$

where for $X = Q$, $Q_0 = \sum_{\sigma=\mathrm{H,E,O}} p_0^2(\sigma)$ and for $X = \mathrm{Prec}(\sigma)$, $\mathrm{Prec}_0(\sigma) = p_0(\sigma)$. For the three-class secondary structure prediction, the relative sizes of the classes are $p_0 = 0.33, 0.21, 0.46$ for tight assignment and $p_0 = 0.36, 0.22, 0.41$ for loose assignment, considering helix, strand, and other secondary structures, respectively. Hence, $Q_0 = 0.36$ for tight assignment (0.35 for loose assignment).

Averages $\langle X \rangle$ and standard deviations $\langle \Delta X^2 \rangle$ of the above quantities [$X \in (Q, R, \mathrm{Prec})$] are computed using the partitioning of protein structures in $M = 10$ subsets (Table S2 of the Supporting Information) according to

$$\langle X \rangle = \frac{1}{M}\sum_{k=1}^{M} X^{(k)}, \quad \langle \Delta X^2 \rangle = \langle (X - \langle X \rangle)^2 \rangle \quad (7)$$

Reliability of Individual Prediction. Besides the overall quality of a prediction device, which is based on the analysis of a large data set or even the total database, it is desirable to have an independent measure for the reliability of an individual prediction. For this purpose, we introduce a confidence measure conf($\mathbf{x}_i$) that a specific residue (i) centered in the sequence window $\mathbf{x}_i$ belongs to a specific secondary structure class. We first sort the probabilities $\bar{g}_\sigma(\mathbf{x}_i)$ obtained from eq 4b according to magnitude: $\bar{g}_{\sigma_1}(\mathbf{x}_i) > \bar{g}_{\sigma_2}(\mathbf{x}_i) > \ldots > \bar{g}_{\sigma_m}(\mathbf{x}_i)$, then we define the confidence

$$\mathrm{conf}(\mathbf{x}_i) = \bar{g}_{\sigma_1}(\mathbf{x}_i) - \bar{g}_{\sigma_2}(\mathbf{x}_i) \quad (8)$$

This quantity is the difference between the two largest values of $\bar{g}_\sigma(\mathbf{x}_i)$ ($\sigma \in \{\mathrm{H, E, O}\}$), and it measures the shortest distance to the border with another class.[93] The confidence ranges between zero and unity. It vanishes if the two larger components are nearly equal and is close to unity if one of the components is much larger than the other two. Ideally, for a good prediction device and an evaluation with a large database, appropriately defined confidences, accuracies, and precisions should yield nearly identical results.

Computational Methods and CPU Time Considerations. In a first step, the sequence profiles of all considered proteins must be computed with the program PSI-BLAST.[49] This requires several tens of hours of CPU time when a large number of proteins is considered. To compute the optimized parameters, the objective function, eq 3b, must be minimized. Because the objective function is a quadratic form in the parameters, minimization requires solving a system of linear equations whose number is as large as the number of parameters.[41] The symmetric coefficient matrix of the linear equations consists of all possible covariances of the feature vectors $\langle \mathbf{x}:\mathbf{x} \rangle$,[41] which requires extensive averaging over millions of sequence profiles. In particular, in the first stage, the coefficient matrices are very large (45451×45451) such that the generation of these matrix elements requires substantial disk space and takes many weeks on a single CPU but can be efficiently parallelized. Because the coefficient matrices are symmetric, the efficient Cholesky algorithm can be used to solve for the roots of the linear equations, needing about one day of CPU time per set of linear equations. The CPU requirements for the linear equations in stage 2 and the neural network in stage 3 are in comparison very moderate. The CPU requirements for usage of the secondary structure prediction tool SPARROW is very moderat, too. By far the largest amount of CPU time is needed to generate the sequence profile for a protein (typically about one minute CPU time per protein) whose secondary structure should be predicted. However, this limitation is typical for all other secondary structure prediction devices that use sequence profiles.

RESULTS AND DISCUSSIONS

Results for Three Stages in SPARROW. To show how SPARROW performs and improves over the three stages, we can monitor the accuracy Q, eq 5a, and the generalized Matthews correlation coefficient, R, eq 5b. In this demonstration, SPARROW used the small ASTRAL40_1.71[85−87] database of protein domains for training and testing using the tight assignment of secondary structures. For a more detailed description of the ASTRAL40 databases see the Methods section, Secondary Structure Databases. Table 2 shows results for prediction (SPARROW applied to test data not used for learning) and recall (SPARROW applied to the training data). More details of prediction performances are provided in Table

dx.doi.org/10.1021/ci200321u | J. Chem. Inf. Model. 2012, 52, 545−556

S2 of the Supporting Information. By comparing the values of the quality measures for the different stages, one can observe small but significant increases in quality. In the recall phase, i.e., predicting the trained data, the quality index Q is about 1.2% (1.4%) larger in stage 2 (stage 3) relative to the preceding stage. In the true prediction phase, the increases in prediction quality with the stages are quite similar but subject to larger statistical errors. The latter is due to the size of the data sets for prediction, which are about nine times smaller than those used for recall (Table S1 of the Supporting Information). In spite of the larger statistical errors in the prediction phase, one can recognize that the Q values in the recall phase are systematically higher by about 0.7% than in prediction phase for all three stages. These small systematic deviations are a trace of overtraining, which is very low thanks to the regularization terms in the objective function, eq 3b. Hence, these terms have been successfully optimized to avoid overtraining. Similar conclusions can be derived considering the quality measure R, eq 5b.

In the recall phase, the error margins for Q and R in stage 3 are well below 0.001 (less than 0.1%). These error margins can be considered to be valid also for the other quantities monitoring performances if the data size is similar, i.e., 9/10 of ASTRAL40_171 containing about 1.15×10^6 residues. In the prediction phase, the error for Q in stage 3 is about 4 times larger due to the smaller sizes of the used test data sets involving about 1.25×10^5 residues (see Table S1 of the Supporting Information). Extrapolating these error margins to the case of the 1.75−1.73 test set involving 1.4×10^5 residues, we expect approximately the same order of magnitude of statistical errors, namely 0.3%.

Comparison of Secondary Structure Prediction Devices. To compare the performance of SPARROW with other prediction devices, SPARROW was trained with the ASTRAL40_1.73 database. In this comparison, the loose assignment was used for the three secondary structure classes, the most common classification scheme, with which the majority of all other prediction devices have been trained. A test data set was created by including all protein domains belonging to the ASTRAL40_1.75 database that have less sequence than 40% similarity with any protein domain belonging to the training set ASTRAL40_1.73. The results are listed in Table 3 monitoring the accuracy Q of the prediction devices PROF,[52] Prospect_2,[66] SSpro_4.03,[67] SPARROW, and PSIPRED_2.6.[35] These prediction devices were chosen because they are available for download. The results of the comparison are listed in Table 3. By comparing the results obtained with SPARROW in Tables 2 and 3, we observe a lower performance by about 1.5% in the latter case. This diminished performance is due to the difference between tight and loose assignment of the eight secondary structure classes of DSSP[64] to the three classes H, E, O and not due to the different ASTRAL40 data sets used for training.

Table 3. Performance of Different Secondary Prediction Programs Compared with SPARROW Monitored by Accuracy Q, eq 5a[a]

program[b]	train date[c]	accuracy Q
PROF[52]	April 2004	0.7311
Prospect_2[66]	June 2002	0.7605
SSpro_4.03[67]	January 2006	0.7768
SPARROW	February 2008	0.8046
PSIPRED_2.6[35]	March 2008	0.8103

[a]Here, SPARROW was trained with the ASTRAL40_1.73 database[85−87] using the loose assignment. The test set 1.75−1.73 used for prediction consists of 918 protein domains containing 146,399 residues. [b]Programs ordered according to improved performance. [c]Approximate date when training was performed. For SPARROW, we used the date when the employed training set ASTRAL40_1.73 was first released.

Although the results of SPARROW are superior relative to PROF,[52] Prospect_2,[66] and SSpro_4.03,[67] the Q value is about 0.6% smaller than obtained from PSIPRED_2.6.[35] Because the error margin estimated for the size of the 1.75−1.73 test set is only 0.3%, the discrepancy is significant. One reason for the lower performance of PROF, Prospect 2, and SSpro 4.03 may be that SPARROW is a newer development and had therefore access to larger databases of secondary structures for training. This makes a fair performance comparison difficult. SPARROW uses a test set in which the sequence identities are lower than 40% relative to the sequences of the ASTRAL40_1.73[85−87] training set. Because ASTRAL40_1.73 was released before February 2008, PSIPRED_2.6 has probably used the same data, but it is not clearly documented. It should be noted that protein structures used in the test set of SPARROW may have been available long before the release date (June 2009) of the corresponding ASTRAL40_1.75 database. Therefore, it is possible that some of these proteins may have been considered even in the programs developed prior to SPARROW.

Secondary Structure Prediction with SPARROW and PSIPRED for the CASP9 Data Set. To further investigate the relative performance of SPARROW (trained on ASTRAL40_1.73) and PSIPRED_2.6 in a realistic prediction scenario, we used the protein structures of the CASP9 experiment in 2010.[74] The protein structures of the CSAP9 data set are not contained in the training data set of SPARROW and probably also not in that of PSIPRED, allowing unbiased prediction. From the polypeptide chains in multi-chain proteins, only those chains were kept whose sequences had less than 90% sequence identity. This procedure leads to a data set of 129 polypeptide chains with a total of 23,640 residues. Of these, 8922 are in helices, 5575 in strands, and 9143 in coils, considering the loose secondary structure assignment with DSSP.[64] Secondary structure prediction was performed for this data set with SPARROW and PSIPRED_2.6 leading to accuracies [Q, eq 5a] of 0.8067 and 0.8116, respectively. For more details see Table S4 in the Supporting Information. Accordingly, PSIPRED_2.6 would perform slightly better than SPARROW, but the difference in Q is not significant because it is less than an estimated error margin of 0.6%.

Figure 3 shows secondary structure predictions for one polypeptide chain from the CASP9 data set (PDB id: 3NRL, chain A) using the online Web servers of SPARROW[94] and PSIPRED.[95,96] Correct predictions are highlighted in green, wrong predictions in pink. Interestingly, both programs show identical performance regarding the number of correctly predicted residues, namely, 58 out of 69 residues. In detail, however, both predictions differ in 11 residues (15.94% of the whole chain). For 5 of the 11 residues, only one of the two secondary structure predictors fails. Hence, there is room for improvement combining the two prediction devices.

Correlating Accuracy and Precision with Prediction confidence. A fully functional prediction tool should not only provide prediction results but also provide a measure of the reliability of the individual predictions, i.e., a measure of the

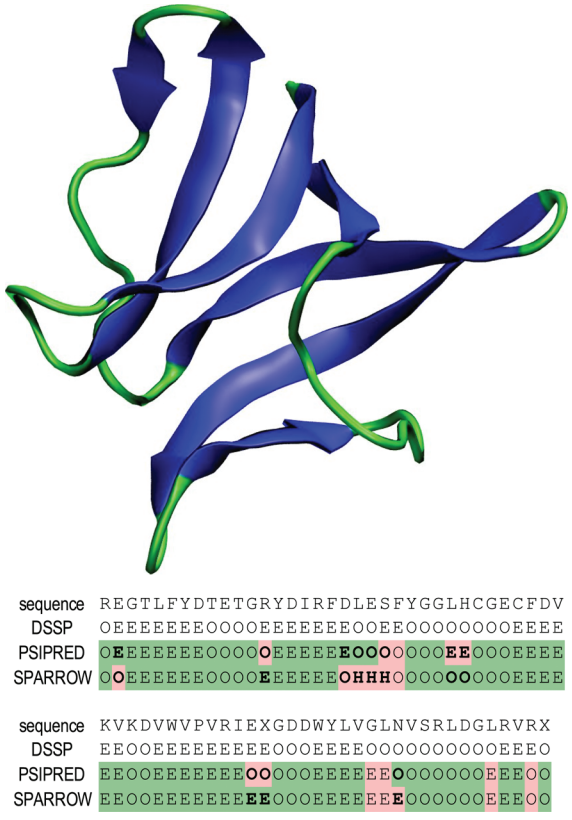


Figure 3. Example of a prediction using the actual Web server of SPARROW[94] and PSIPRED[95,96] for a polypeptide chain from the CASP9 data set (PDB id: 3NRL, chain A). Upper part: 3-D structure with β-strands highlighted in blue and loops in green according to the loose secondary structure assignment. Lower part: polypeptides sequence and secondary structure are shown. Correct predictions are highlighted in green, wrong predictions in pink. Both programs predict 58 out of 69 residues correctly, while differing in secondary structure prediction for 11 residues. For six residues, both predictions are wrong.

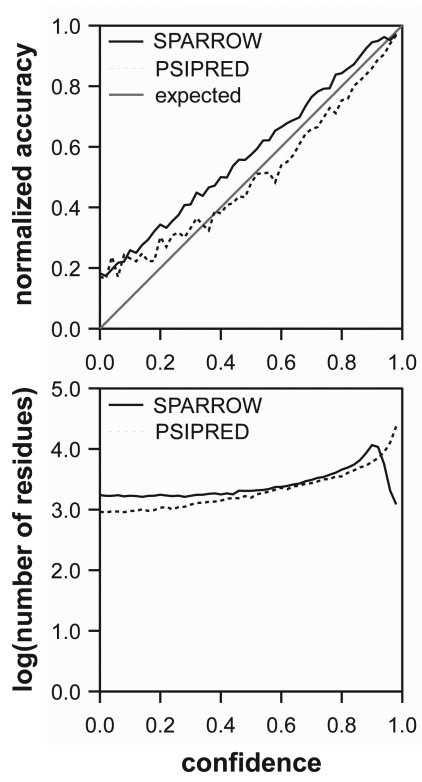


Figure 4. Risk intelligence plot[97] for SPARROW (solid black line) and PSIPRED_2.6 (dashed line) using the 1.75−1.73 test set. In the upper histogram, the normalized accuracy $\overline{Q}$, eq 5a combined with eq 6, is plotted as a function of prediction confidence, eq 8. To generate the plots, the $\overline{Q}$ values were averaged in confidence bins of width 0.01. The solid gray line shows ideal risk intelligence dependence. In the lower histogram, the number of secondary structure prediction events belonging to specific confidence bins is plotted on a logarithmic scale as a function of the confidence value.

likelihood that the individual prediction is actually correct. For that purpose we use the prediction confidence, eq 8, introduced in the Method section. It consists of the difference of the two largest normalized output scores from the neural network $\overline{g}_\sigma$, eq 4b. A plot of the normalized accuracy $\overline{Q}$, eq 6, as a function of the confidence (a so-called risk−intelligence plot[97]) is shown in Figure 4. $\overline{Q}$ measures the ability of a prediction device to perform deterministic predictions. It varies between zero for intelligent random guesses that consider just the abundances of the objects in the different classes and unity for a perfect prediction device. The confidence, eq 8, estimates the probability that an individual prediction is correct. A proper definition of the confidence (conf) and prediction accuracy (for instance $\overline{Q}$) should yield identical values (straight line in top part of Figure 4).

Both prediction devices, SPARROW and PSIPRED_2.6, start at low confidence values with normalized accuracies $\overline{Q}$ of about 0.2, well above the ideal straight line (top part of Figure 4). Thus, even for vanishing confidence values, the actual prediction accuracy is 0.2, well above results from random guesses. This deviation is likely due to the definition of the confidence, eq 8, which is too cautious in estimating prediction accuracies in this confidence regime. The confidence vanishes, for instance, if the two largest normalized scores from $\overline{g}_H$, $\overline{g}_E$, $\overline{g}_O$, eqs 4a and 4b, are nearly of identical values, while the score for the third class is significantly smaller. In this case, a properly working prediction device will chose only between the two classes with large scores and ignore the class with small score. Therefore, a prediction made by a trained predictor will be more reliable than an intelligent random guess that uses only the information on the sizes of the different classes, thus yielding $\overline{Q} = 0$ (see discussion after eq 6). However, in the high confidence regime, the problem of the defined confidence is no longer relevant.

With increasing confidence, the normalized prediction accuracy for SPARROW remains above the ideal straight line but approaches it, while for PSIPRED_2.6 it crosses the ideal line at a confidence of 0.4 and remains below the ideal line for larger confidence values (Figure 4 top part). Thus, using the present confidence measure, secondary structure predictions with SPARROW are done too cautiously, while PSIPRED_2.6 has a slight tendency of being too confident in the high confidence regime. Parallel to this behavior one can observe that SPARROW has more predictions in the regime of lower confidence as compared to PSIPRED_2.6 (Figure 4 bottom part). Risk−intelligence plots analogous to Figure 4 can also be made for class specific normalized precisions, as defined in eq 5d in conjunction with eq 6 (Figure S1 of the Supporting Information).

Using the prediction confidence, eq 8, computed for a set residues ($\mathbf{x}_i$) one can introduce a value of trust $\langle$conf$\rangle$, which is a comprehensive measure of the overall quality of a prediction

dx.doi.org/10.1021/ci200321u | J. Chem. Inf. Model. 2012, 52, 545−556

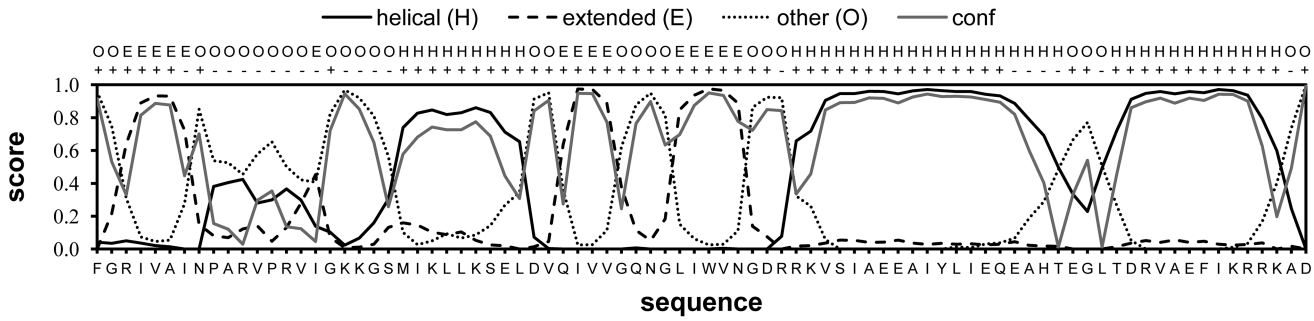


Figure 5. SPARROW output sample from the Web page (http://agknapp.chemie.fu-berlin.de/sparrow/). The application of SPARROW to a chain of archaeal exosome core (PDB id: 2GA0) yields the normalized scores $\bar{g}_H$, $\bar{g}_E$, $\bar{g}_O$, eqs 4a and 4b. These are pseudoprobabilities that the secondary structure of a residue is of the type helix (H), strand (E), or other (O). Besides these scores, the prediction confidence, eq 8, is also given (gray solid line). On the abscissa, the amino acid types are denoted in one-letter code. The information above the diagram indicates the predicted type of secondary structure (H, E, or O). In contrast to the output on the Web page, the diagram also indicates whether the prediction was successful (+) or not (−), which can only be provided if, as in this case, the correct secondary structure is available.

device

$$\langle \mathrm{conf} \rangle = \sum_{i}^{N_{\mathrm{total}}} y(\mathbf{x}_i)\mathrm{conf}(\mathbf{x}_i) / \sum_{i}^{N_{\mathrm{total}}} \mathrm{conf}(\mathbf{x}_i) \tag{9}$$

In eq 9, the target value $y(\mathbf{x}_i)$ is either +1 or −1 depending on whether the prediction is correct or not. The value of trust ⟨conf⟩ counts confidences for correct secondary structure predictions positive, while confidences of wrong predictions are counted negative. It provides a measure of how strong large (small) confidence values correlate with positive (negative) results of prediction. The accuracy Q and the generalized Matthews correlation coefficient R measure the prediction quality. On the other hand, ⟨conf⟩ measures how trustworthy predictions are that were obtained with a specific prediction tool. Using the 1.75−1.73 test set, the value of trust ⟨conf⟩ for secondary structure prediction is 0.7333 and 0.7498 for PSIPRED_2.6 and SPARROW, respectively. Surprisingly, ⟨conf ⟩ is slightly larger for SPARROW than for PSIPRED_2.6, even if the prediction quality is slightly better for PSIPRED_2.6 than for SPARROW. Hence, the employed expression of confidence correlates better with SPARROW than with PSIPRED, despite not having been optimized for either one or the other.

Monitoring Normalized Scores and Confidences for Secondary Structure Prediction. The SPARROW services have recently been made available on the Web (http://agknapp.chemie.fu-berlin.de/sparrow/). On this Web page, the amino acid sequence of a protein is introduced as one-letter code. As output, one obtains the result of secondary structure prediction as a corresponding letter string containing H, E, and O together with a confidence measure, which is a discrete integer varying from 0 to 9, proportional to the confidence value introduced in eq 8. These results are also plotted and can be downloaded. An example of a prediction result for a domain in chain A of archaeal exosome core (PDB id: 2GA0) is shown in Figure 5.

CONCLUSIONS

A classification technique based on machine learning methods that use scoring functions was applied to the problem of protein secondary structure prediction. 2Class discriminant techniques provide secondary structure propensity scores at two subsequent levels of correlation. These scores are then used by an artificial neural network acting as an "intelligent" decision maker. The resulting secondary structure predictor, SPARROW, is an efficient prediction tool, superior in performance to several other methods available as online applications and nearly as accurate as the well established program PSIPRED.[35] The risk−intelligence plots correlate prediction accuracy or precision with a confidence measure. On the basis of these plots, SPARROW has a tendency to be slightly more cautious in predictions than PSIPRED. Considering the newly introduced value of trust parameter, eq 9, which weights the confidences of predictions positive if successful and negative if wrong. SPARROW may be slightly more reliable than PSIPRED.

Interesting aspects are to which extent the prediction tools PSIPRED and SPARROW yield the same secondary structure classification result, how often they differ, and how often they both fail. According to Table 4, the secondary structure prediction carried out with PSIPRED and SPARROW agrees on average for 77.00% of all residues; both fail in 15.47% of all predictions. They differ in 7.53% of the predictions in which one method makes correct predictions, while the other method fails. The slightly better performance of PSIPRED can be deduced from the small difference of 0.6% in the percentages of predictions in which one method fails while the other is successful. Thus, the complementary use of the two methods should improve the final prediction performance. The effect that one can expect from such a combination should be related to the 7.53% of cases in which one of the methods makes a successful prediction. An example of a successful combination of two different protein secondary structure prediction schemes can be found in ref 98.

Table 4. Correlation of the Fraction of Secondary Structures Predicted Correctly or Incorrectly with PSIPRED and SPARROW[a]

	PSIPRED correct (%)	PSIPRED wrong (%)
SPARROW correct	77.00	3.46
SPARROW wrong	4.07	15.47

[a]For instance, 3.46% of all structures are predicted correctly with SPARROW but incorrectly with PSIPRED on the basis of the comparison test set 1.75−1.73 of protein domain structures. Summing up the percentages in rows and columns yields 100%.

dx.doi.org/10.1021/ci200321u | J. Chem. Inf. Model. 2012, 52, 545−556

SPARROW and PSIPRED are based on different methods and nearly of equal prediction quality. Therefore, they provide useful and largely independent expert opinions on protein secondary structure prediction. In addition, SPARROW yields a reliable confidence value in judging the quality of each prediction event.

The ASTRAL40-based benchmark tests were designed with the aim of providing a neutral ground to comparing different programs on a fair basis. While some of the programs do not appear to be regularly updated, others, like PSIPRED, are constantly retrained and updated such that it is hard to determine exactly how large their knowledge base is. Although one can trace back the sources of structural information to the Protein Data Bank, most learning machines make use of selected data sets which need not be compiled or released in a synchronous fashion. Training and testing of SPARROW is based on protein structures from specific versions of the ASTRAL40 database. These steps are taken in an attempt to standardize learning and testing schemes and thus facilitate comparison to protein secondary structure predictors developed in the future.

■ ASSOCIATED CONTENT

Supporting Information

Tables S1−S4 and Figure S1. This information is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +49 (30) 83854387. Fax: +49 (30) 83856921. E-mail: knapp@chemie.fu-berlin.de.

Author Contributions

§Both authors contributed equally.

■ ACKNOWLEDGMENTS

The authors thank Oezguer Demir for useful discussions. This work is financially supported by the Deutsche Forschungsgemeinschaft with the following programs: International Research Training Group (IRTG) on "Genomics and Systems Biology of Molecular Networks" (GRK1360) and Computational Systems Biology(GRK1772).

■ ABBREVIATIONS

2Class problem = two-class classification problem
H = helix class
E = strand class
O = class consisting of other secondary structures other than helix and strand
CASP = critical assessment of protein structure prediction
DSSP = define secondary structure of proteins
SCOP = structural classification of proteins
SPARROW = secondary structure prediction using arrays of optimized weights
1-r = *one-against-rest*: complete set of two-class classification problems to solve a multiclass problem separating one particular class from all other classes in all combinations
1-1 = *one-against-one*: complete set of two-class classification problems to solve a multiclass problem separating two specific classes from all others in all combinations
Q = prediction accuracy
Prec = class-specific precision
R = generalized Matthews correlation coefficient

■ REFERENCES

(1) Graur, D.; Li, W. H. *Fundamentals of Molecular Evolution*; Sinauer Associates: Sunderland, MA, 2000.
(2) Thornton, J. M.; Orengo, C. A.; Todd, A. E.; Pearl, F. M. Protein folds, function and evolution. *J. Mol. Biol.* **1999**, *293*, 333−342.
(3) Nagao, C.; Terada, T. P.; Yomo, T.; Sasai, M. Correlation between evolutionary structural development and protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 18950−18955.
(4) Pal, C.; Papp, B.; Lercher, M. J. An integrated view of protein evolution. *Nature Rev. Gen.* **2006**, *7*, 337−348.
(5) Söding, J.; Lupas, A. N. More than the sum of their parts: On the evolution of proteins from peptides. *Bioessays* **2003**, *25*, 837−846.
(6) Baker, D.; Sali, A. Protein structure prediction and structural genomics. *Science* **2001**, *294*, 93−96.
(7) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(8) van Heel, M. Single-particle electron cryo-microscopy: Towards atomic resolution. *Q. Rev. Biophys.* **2000**, *33*, 307−369.
(9) Grabowski, M.; Joachimiak, A.; Otwinowski, Z.; Minor, W. Structural genomics: Keeping up with expanding knowledge of the protein universe. *Curr. Opin. Struct. Biol.* **2007**, *17*, 347−353.
(10) Jones, D. T.; Taylor, W. R.; Thornton, J. M. A new approach to protein fold recognition. *Nature* **1992**, *358*, 86−89.
(11) Solis, A. D.; Rackovsky, S. On the use of secondary structure in protein structure prediction: A bioinformatic analysis. *Polymer* **2004**, *45*, 525−546.
(12) Zhang, Y.; Skolnick, J. The protein structure prediction problem could be solved using the current PDB library. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 1029−1034.
(13) Bowie, J. U.; Luthy, R.; Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **1991**, *253*, 164−170.
(14) Hendlich, M.; Lackner, P.; Weitckus, S.; Floeckner, H.; Froschauer, R.; Gottsbacher, K.; Casari, G.; Sippl, M. J. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* **1990**, *216*, 167−180.
(15) Miyazawa, S.; Jernigan, R. L. Residue−residue potentials with a favorable contact pair term and an unfavorable high packing density term for simulation and threading. *J. Mol. Biol.* **1996**, *256*, 623−644.
(16) Yue, K.; Dill, K. A. Folding proteins with a simple energy function and extensive conformational searching. *Protein Sci.* **1996**, *5*, 254−261.
(17) Rohl, C. A.; Strauss, C. E.; Misura, K. M. S.; Baker, D. Protein structure prediction using Rosetta. *Method Enzymol.* **2004**, *383*, 66−93.
(18) Colubri, A. D.; Rackovsky, S. Prediction of protein structure by simulating coarse-grained folding pathways: A preliminary report. *J. Biomol. Struct. Dyn.* **2004**, *21*, 625−638.
(19) Fischer, D.; Eisenberg, D. Protein fold recognition using sequence-derived predictions. *Protein Sci.* **1996**, *5*, 947−955.
(20) Rost, B.; Schneider, R.; Sander, C. Protein fold recognition by prediction-based threading. *J. Mol. Biol.* **1997**, *270*, 471−480.
(21) Russell, R. B.; Copley, R. R.; Barton, G. J. Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol.* **1996**, *259*, 349−365.
(22) Karchin, R.; Cline, M.; Gutfreund, M.-Y.; Karplus, K. Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins: Struct., Funct., Genet.* **2003**, *51*, 504−514.
(23) Sim, J.; Kim, S. Y.; Lee, J.; Yoo, A. Predicting the three-dimensional structures of proteins: Combined alignment approach. *J. Korean Phys. Soc.* **2004**, *44*, 611−616.
(24) Burgess, A. W.; Ponnuswamy, P. K.; Sheraga, H. A. Analysis of conformations of amino acid residues and prediction of backbone topography in proteins. *J. Israel Chem.* **1974**, *12*, 239−286.
(25) Chou, P. Y.; Fasman, G. D. Prediction of protein conformation. *Biochemistry* **1974**, *13*, 222−245.

(26) Chou, P. Y.; Fasman, G. D. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.* **1978**, *47*, 45−148.
(27) Garnier, J.; Osguthorpe, D. J.; Robson, B. Analysis and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **1978**, *120*, 97−120.
(28) Garnier, J.; Gibrat, J. F.; Robson, B. GOR method for predicting protein secondary structure from amino acid sequence. *Method Enzymol.* **1996**, *266*, 540−553.
(29) Pauling, L.; Corey, R. B.; Branson, H. R. Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U.S.A.* **1951**, *37*, 205−211.
(30) Pauling, L.; Corey, R. B. Configurations of polypeptide chains with favored orientations of the polypeptide around single bonds: Two pleated sheets. *Proc. Natl. Acad. Sci. U.S.A.* **1951**, *37*, 729−740.
(31) Selbig, J.; Mevissen, T.; Lengauer, T. Decision tree-based formation of consensus protein secondary structure prediction. *Bioinformatics* **1999**, *15*, 1039−1046.
(32) Holley, L. H.; Karplus, M. Protein secondary structure prediction with a neural network. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 152−156.
(33) Frishman, D.; Argos, P. Recognition of distantly related protein sequences using conserved motifs and neural networks. *J. Mol. Biol.* **1992**, *228*, 951−962.
(34) Rost, B.; Sander, C. Improved prediction of protein secondary structure by use sequence profiles and neural networks. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 7558−7562.
(35) Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **1999**, *292*, 195−202.
(36) Pollastri, G.; Przybylski, D.; Rost, B.; Baldi, P. Improving the prediction of secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 228−235.
(37) Lin, K.; Simossis, V. A.; Taylor, W. R.; Heringa, J. A simple and fast secondary structure prediction method using hidden neural networks. *Bioinformatics* **2005**, *21*, 152−159.
(38) Chen, J.; Chaudhari, N. S. Bidirectional segmented-memory recurrent neural network for protein secondary structure prediction. *Soft Comput.* **2006**, *10*, 315−324.
(39) Hua, S.; Sun, Z. A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach. *J. Mol. Biol.* **2001**, *308*, 397−407.
(40) Ward, J. J.; McGuffin, L. J.; Buxton, B. F.; Jones, D. T. Secondary structure prediction with support vector machines. *Bioinformatics* **2003**, *19*, 1650−1655.
(41) Riedesel, H.; Kolbeck, B.; Schmetzer, O.; Knapp, E. W. Peptide binding at class I major histocompatibility complex scored with linear functions and support vector machines. *Genome Inform.* **2004**, *15*, 198−212.
(42) Pham, T. H.; Satou, K.; Ho, T. B. Support vector machines for prediction and analysis of beta and gamma-turns in proteins. *J. Bioinform. Comput. Biol.* **2005**, *3*, 343−358.
(43) Asai, K.; Hayamizu, S.; Handa, K. Prediction of protein secondary structure by the hidden Markov model. *Comput. Appl. Biosci.* **1993**, *9*, 141−149.
(44) Aydin, Z.; Altunbasak, Y.; Borodovsky, M. Protein secondary structure prediction for a single-sequence using hidden semi-Markov models. *BMC Bioinf.* **2006**, *7*, 178.
(45) Martin, J.; Gibrat, J. F.; Rodolphe, F. Analysis of an optimal hidden Markov model for secondary structure prediction. *BMC Struct. Biol.* **2006**, *6*, 25.
(46) Abe, N.; Mamitsuka, H. Predicting protein secondary structure using stochastic tree grammars. *Mach. Learn.* **1997**, *29*, 275−301.
(47) Cheng, J.; Tegge, A. N.; Baldi, P. Machine learning methods for protein structure prediction. *IEEE Rev. Biomed. Eng.* **2008**, *1*, 41−49.
(48) Sander, B. R. C. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **1993**, *232*, 584−599.
(49) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(50) Henikoff, S. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 10915−10919.
(51) Cuff, J. A.; Barton, G. J. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 508−519.
(52) Ouali, M.; King, R. D. Cascaded multiple classifiers for secondary structure prediction. *Protein Sci.* **2000**, *9*, 1162−1176.
(53) Petersen, T. N.; Lundegaard, C.; Nielsen, M.; Bohr, H.; Bohr, J.; Brunak, S.; Gippert, G. P.; Lund, O. Prediction of Protein Secondary Structure at 80% Accuracy. *Proteins: Struct., Funct., Genet.* **2000**, *41*, 17−20.
(54) Guermeur, Y.; Pollastri, G.; Elisseeff, A.; Zelus, D.; Moisy, P.-H.; Baldi, P. Combining protein secondary structure prediction models with ensemble methods of optimal complexity. *Neurocomputing* **2004**, *56*, 305−327.
(55) Armano, G.; Mancosu, G.; Milanesi, L.; Orro, A.; Saba, M.; Vargiu, E. A hybrid genetic-neural system for predicting protein secondary structure. *BMC Bioinf.* **2005**, *6*, S3.
(56) Pollastri, G.; McLysaght, A. Porter: A new, accurate server for protein secondary structure prediction. *Bioinformatics* **2005**, *21*, 1719−1720.
(57) Wood, M. J.; Hirst, J. D. Protein secondary structure prediction with dihedral angles. *Proteins: Struct., Func., Bioinf.* **2005**, *59*, 476−481.
(58) Zhang, G. Z.; Huang, D. S.; Zhu, Y. P.; Li, Y. X. Improving protein secondary structure prediction by using the residue conformational classes. *Pattern Recognit. Lett.* **2005**, *26*, 2346−2352.
(59) Bondugula, R.; Xu, D. MUPRED: A tool for bridging the gap between template based methods and sequence profile based methods for protein secondary structure prediction. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 664−670.
(60) Dor, O.; Zhou, Y. Q. Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. *Proteins: Struct., Funct., Bioinf.* **2007**, *66*, 838−845.
(61) Montgomerie, S.; Cruz, J. A.; Shrivastava, S.; Arndt, D.; Berjanskii, M.; Wishart, D. S. PROTEUS2: A web server for comprehensive protein structure prediction and structure-based annotation. *Nucleic Acids Res.* **2008**, *36*, W202−W209.
(62) Yao, X. Q.; Zhu, H. Q.; She, Z. S. A dynamic Bayesian network approach to protein secondary structure prediction. *BMC Bioinf.* **2008**, *9*, 49.
(63) Anfinsen, C. B. Principles that govern the folding of polypeptide chains. *Science* **1973**, *181*, 223−230.
(64) Kabsch, W.; Sander, C. A dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.
(65) Rost, B. Rising Accuracy of Protein Secondary Structure Prediction. In *Protein Structure Determination, Analysis, and Modeling for Drug Discovery*; Chasman, D., Ed.; Dekker: New York, 2003; pp 207−249.
(66) Xu, Y.; Xu, D. Protein threading using PROSPECT: Design and evaluation. *Proteins: Struct., Funct., Genet.* **2000**, *40*, 343−354.
(67) Cheng, J.; Randall, A.; Sweredoski, M.; Baldi, P. SCRATCH: A protein structure and structural feature prediction server. *Nucleic Acids Res.* **2005**, *33*, W72−W76.
(68) Rost, B.; Yachdav, G.; Liu, J. The PredictProtein server. *Nucleic Acids Res.* **2004**, *32*, W321−W326.
(69) Rost, B.; Sander, C.; Schneider, R. Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.* **1994**, *235*, 13−26.
(70) Rost, B.; Eyrich, V. A. EVA: Large-scale analysis of secondary structure prediction. *Proteins: Struct., Funct., Genet.* **2001**, *45*, S192−S199.
(71) Eyrich, V. A.; Marti-Renom, M. A.; Przybylski, D.; Madhusudhan, M. S.; Fiser, A.; Pazos, F.; Valencia, A.; Sali, A.; Rost, B. EVA: Continuous automatic evaluation of protein structure prediction servers. *Bioinformatics* **2001**, *17*, 1242−1243.
(72) Koh, I. Y.; Eyrich, V. A.; Marti-Renom, M. A.; Przybylski, D.; Madhusudhan, M. S.; Eswar, N.; Graña, O.; Pazos, F.; Valencia, A.;

Sali, A.; Rost, B. EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Res.* **2003**, *31*, 3311−3315.
(73) Moult, J.; Pedersen, J.; Judson, R.; Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins* **1995**, *23*, ii−v.
(74) 9th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction. Protein Structure Prediction Center, 2010. http://predictioncenter.org/casp9/ (accessed September 20, 2011).
(75) Vapnik, V. N. *Statistical Learning Theory*; Wiley: New York, 1998.
(76) Pan, X. Multiple linear regression for protein secondary structure prediction. *Proteins: Struct., Funct., Genet.* **2001**, *43*, 256−259.
(77) Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annals Eugenics* **1936**, *7*, 179−188.
(78) Tax, D. M. J.; Duin, R Using two-class classifiers for multiclass classification. *IEEE Proc.* **2002**, *2*, 124−127.
(79) Benson, D. A.; Karsch-Mizrachi, I.; Lipman, D. J.; Ostell, J.; Wheeler, D. L. GenBank. *Nucleic Acids Res.* **2008**, *36*, 25−30.
(80) Bairoch, A.; Boeckmann, B.; Ferro, S.; Gasteiger, E. Swiss-Prot: Juggling between evolution and stability. *Briefings Bioinf.* **2004**, *5*, 39−55.
(81) Wu, C. H.; Yeh, L. S.; Huang, H.; Arminski, L.; Castro-Alvear, J.; Chen, Y.; Hu, Z. Z.; Ledley, R. S.; Kourtesis, P.; Suzek, B. E.; Vinayaka, C. R.; Zhang, J.; Barker, W. C. The Protein Information Resource. *Nucleic Acids Res.* **2003**, *31*, 345−347.
(82) Aimoto, S.; Ono, S. Peptide Science 2010. Protein Research Foundation. http://www.prf.or.jp/index-e.html (accessed October 7, 2011).
(83) Pruitt, K. D.; Tatusova, T.; Maglott, D. R. NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **2007**, *35*, D61−D65.
(84) Haykin, S. *Neural Networks: A Comprehensive Foundry*, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, 1999.
(85) Brenner, S. E.; Koehl, P.; Levitt, M. The ASTRAL compendium for sequence and structure analysis. *Nucleic Acids Res.* **2000**, *28*, 254−256.
(86) Chandonia, J. M.; Hon, G.; Walker, N. S.; Conte, L. L.; Koehl, P.; Levitt, M.; Brenner, S. E. The ASTRAL compendium in 2004. *Nucleic Acids Res.* **2004**, *32*, 189−192.
(87) Chandonia, J. M.; Walker, N. S.; Conte, L. L.; Koehl, P.; Levitt, M.; Brenner, S. E. ASTRAL compendium enhancements. *Nucleic Acids Res.* **2002**, *30*, 260−263.
(88) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442−451.
(89) Gorodkin, J. Comparing two K-category assignments by a K-category correlation coefficient. *Comp. Biol. Chem.* **2004**, *28*, 367−374.
(90) Sudheep, E. M.; Sumam, M. I.; Joseph, A. Design and performance analysis of data mining techniques based on decision trees and naive bayes classifier for employment chance prediction. *J. Convergence Inf. Technol.* **2011**, *6*, 89−98.
(91) Kohavi, R.; Provost, F. Glossary of Terms. *Mach. Learn.* **1998**, *30*, 271−274.
(92) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, *16*, 412−424.
(93) Aßfalg, J.; Kriegel, H. P.; Pryakhin, A.; Schubert, M. Multi-Represented Classification Based on Confidence Estimation. In *Advances in Knowledge Discovery and Data Mining, Lecture Notes in Computer Science 4426*, Proceedings of 11th Pacific-Asia Conference, PAKDD, Nanjing, China, 2007; Zhou, Z.-H.; Li, H.; Yang, Q., Eds.; Springer: New York/Heidelberg, 2007, pp 23−24.
(94) Bettella, F.; Rasinski, D.; Knapp, E. W. Protein Secondary Structure Prediction with SPARROW. Institute of Chemistry and Biochemistry, Freie Universität Berlin. http://agknapp.chemie.fu-berlin.de/sparrow/ (accessed October 7, 2011).
(95) Bryson, K.; McGuffin, L. J.; Marsden, R. L.; Ward, J. J.; Sodhi, J. S.; Jones, D. T. Protein structure prediction servers at University College London. *Nucleic Acids Res.* **2005**, *33*, W36−38.
(96) Bryson, K.; McGuffin, L. J.; Marsden, R. L.; Ward, J. J.; Sodhi, J. S.; Jones, D. T. The PSIPRED Protein Structure Prediction Server. Department of Computer Science, Bioinformatics Group, University College London. http://bioinf.cs.ucl.ac.uk/psipred/ (accessed October 7, 2011).
(97) Apgar, D. *Risk Intelligence.*: Harvard Business School Press: Boston, MA, 2006.
(98) Lin, H.; Sung, T.; Ho, S.; Hsu, W. Improving protein secondary structure prediction based on short subsequences with local structure similarity. *BMC Genomics* **2010**, *11*, S4.