PROTEINS: Structure, Function, and Bioinformatics 59:467–475 (2005)

Combining Prediction of Secondary Structure and Solvent Accessibility in Proteins

Rafał Adamczak,[1] Aleksey Porollo,[1] and Jarosław Meller[1,2]*
[1]*Biomedical Informatics, Children's Hospital Research Foundation, Cincinnati, Ohio*
[2]*Department of Informatics, Nicholas Copernicus University, Toruń, Poland*

ABSTRACT **Owing to the use of evolutionary information and advanced machine learning protocols, secondary structures of amino acid residues in proteins can be predicted from the primary sequence with more than 75% per-residue accuracy for the 3-state (i.e., helix, β-strand, and coil) classification problem. In this work we investigate whether further progress may be achieved by incorporating the relative solvent accessibility (RSA) of an amino acid residue as a fingerprint of the overall topology of the protein. Toward that goal, we developed a novel method for secondary structure prediction that uses predicted RSA in addition to attributes derived from evolutionary profiles. Our general approach follows the 2-stage protocol of Rost and Sander, with a number of Elman-type recurrent neural networks (NNs) combined into a consensus predictor. The RSA is predicted using our recently developed regression-based method that provides real-valued RSA, with the overall correlation coefficients between the actual and predicted RSA of about 0.66 in rigorous tests on independent control sets. Using the predicted RSA, we were able to improve the performance of our secondary structure prediction by up to 1.4% and achieved the overall per-residue accuracy between 77.0% and 78.4% for the 3-state classification problem on different control sets comprising, together, 603 proteins without homology to proteins included in the training. The effects of including solvent accessibility depend on the quality of RSA prediction. In the limit of perfect prediction (i.e., when using the actual RSA values derived from known protein structures), the accuracy of secondary structure prediction increases by up to 4%. We also observed that projecting real-valued RSA into 2 discrete classes with the commonly used threshold of 25% RSA decreases the classification accuracy for secondary structure prediction. While the level of improvement of secondary structure prediction may be different for prediction protocols that implicitly account for RSA in other ways, we conclude that an increase in the 3-state classification accuracy may be achieved when combining RSA with a state-of-the-art protocol utilizing evolutionary profiles. The new method is available through a Web server at http://sable.cchmc.org. Proteins 2005;59:467–475.**

Key words: secondary structure; neural networks; classification; protein structure prediction; relative solvent accessibility; SABLE

INTRODUCTION

Increasingly accurate methods for secondary structure prediction have significantly contributed to the improved performance of fold recognition algorithms and *de novo* folding simulations.[1–3] According to continuous EVAluation of secondary structure prediction methods by the EVA metaserver,[4] the state-of-the-art methods that utilize evolutionary information, such as neural networks (NN)-based Psi-PRED[5] or hidden Markov model (HMM)-based SAM-T99sec[6] methods, can classify correctly into 1 of 3 states, corresponding to helical, extended (or β-strand) and other conformations, more than 75% residues in rigorous tests on independent sets of structures derived from new submission to the Protein Data Bank (PDB).[7]

Several attempts have been recently made to further improve the accuracy of secondary structure prediction by using more accurate multiple sequence alignments,[8] advanced machine learning techniques,[9] or information obtained when building the overall 3-dimensional (3D) model of a protein.[10] For a general overview of progress in this field, the reader is referred to recent reviews.[11,12] In this work, we focus our efforts on "local" fingerprints of the 3D packing that may indicate sites where sequence propensities for secondary structure elements are likely to be influenced by long-range contacts. Specifically, we investigate whether including solvent-exposed surface area of an amino acid resides in a protein structure [measured in relative terms by relative solvent accessibility (RSA)] as a local fingerprint of long-range interactions and the overall packing of a protein structure could be beneficial for secondary structure prediction.

The solvent exposure of amino acid residues in proteins has been considered in the context of secondary structure prediction before.[13,14] These previous studies provide support for the idea of differentiating between buried and exposed amino acid residues in order to obtain more accurate secondary structure prediction. However, while

*Correspondence to: Jaroslaw Meller, Biomedical Informatics, Children's Hospital Research Foundation, 3333 Burnet Avenue, Cincinnati, OH 45229. E-mail: jmeller@chmcc.org

Received 4 August 2004; Accepted 29 October 2004

Published online 14 March 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20441

improvements have been shown in the context of single-sequence methods when using solvent accessibilities derived from known protein structures,[13,14] it is not clear if the same holds for the state-of-the-art prediction protocols based on evolutionary profiles. Moreover, it remains to be seen if the solvent exposure predicted from the amino acid sequence may be used to improve the secondary structure prediction as well.

In analogy with secondary structure prediction, the RSA prediction problem is typically cast as a classification problem (e.g., with 2 classes of buried and exposed residues) defined by an arbitrarily chosen threshold in terms of RSA.[15–17] However, RSA of structurally equivalent residues in families of homologous proteins is highly variable, and the notion of clearly defined classes appears to have a weaker support than that for secondary structure elements.[15,18] Starting from this observation, we have recently developed a number of accurate regression-based methods for real-valued RSA prediction that outperform classification-based approaches.[18,19]

Here, we use such obtained real-valued RSA predictions to further improve secondary structure prediction. In order to test the hypothesis that RSA of an amino acid residue may be advantageous as a fingerprint of the overall topology of a protein in the context of a state-of-the-art secondary structure prediction, we first measure the effect of including the actual RSA derived from known protein structures in the feature space primarily based on the evolutionary profiles encoded by multiple alignments. Indeed, significant improvements (on the order of 3–4% in terms of per-residue classification accuracy) are obtained in the limit of perfect prediction (i.e., when the experimental RSA is included in the representation of a local environment of an amino acid residue). We then investigate whether predicted real-valued RSA, obtained using our novel approach to RSA prediction, might also be used to enhance secondary structure prediction.

The latter is not clear a priori, because the predicted RSA only approximates the actual and, in fact, highly variable (especially for exposed residues) RSA values observed in families of homologous proteins.[15,18,19] Moreover, both secondary structure and RSA prediction protocols discussed here incorporate the same evolutionary profiles in their input vectors that represent a local environment of an amino acid residue. Nevertheless, a systematic improvement of the classification accuracy between 0.4% and 1.4% is achieved on different control sets for 3-class secondary structure prediction when using the most accurate prediction of real-valued RSA available at present. Given the observed level of improvements in the limit of perfect RSA prediction, we expect that further progress in methods for predicting real-valued RSA may consequently yield further improvements in secondary structure prediction.

The structure of this article is as follows: In the next section, we describe the training and control sets, as well as alternative attributes (feature spaces) used to represent a residue (and its environment) for which we attempt to predict secondary structure. We also describe NN architectures and training protocols that we use to develop our system for secondary structure prediction. In the last sections we present results and discussion, followed by conclusions.

MATERIALS AND METHODS

Training and Control Sets

We used here the same training set that was used to develop in parallel different regression-based methods for accurate prediction of real-valued RSA.[18,19] Namely, the Protein Families (Pfam) database, version 6.6,[20] was used in order to guide the derivation of a representative and nonredundant (NR) training set of protein chains with known secondary structure states. In order to account for variable loops and sequence ends, entire PDB entries representative of a given Pfam family, rather than just the conserved fragments included in Pfam, were used. Specifically, an entire structure of an individual protein chain, as included in a PDB entry representative of a given Pfam family, was selected for each family represented by at least one 3D structure. Structures that could not be successfully parsed using the DSSP program,[21] which is used here (with default options) to assign amino acid residues to secondary structures, were excluded. Furthermore, by considering only unique PDB entries, a set of 860 PDB structures with about 210,000 residues was obtained. In this set, due to the fact that a PDB structure could be covered by multiple PFAM domains, there are 27 pairs of structures that share homologous fragments resulting in BLAST[22] sequence alignment matches with *E*-values lower than 0.001. As indicated by similar accuracy obtained on several independent test sets and in cross-validated training (see the Results section), the effect of this partial overlap concerning a small subset of structures in the training appears to be negligible. Since the DSSP program assigns each residue to 1 of 8 distinct classes, the following conversion from 8 to 3 classes was applied: {G, I, H} to H, {B, E} to E, and {T, S, and "other"} to C, where H denotes helix, E denotes β-strand, and C denotes coil.

We used an approach similar to that used by the EVA metaserver[4] in order to derive several independent control sets for assessing the accuracy of our predictions. Namely, NR structures submitted to the PDB database[7] were selected using initially 50% sequence identity threshold to remove redundant new entries. Further redundancies within the control sets were removed by applying BLAST sequence alignments to remove protein chains that resulted in matches with *E*-values lower than 0.001 with respect to sequences of PDB entries already included in any of the control sets. Next, structurally biased sequence alignments implemented in the LOOPP program[23] were used in order to remove proteins with homology (i.e., *Z*-scores higher than 6.5—a threshold that was found sufficient before[24]) to PDB structures included in our training set.

The resulting control sets with no homology to proteins included in the training will be referred to as S156 (156 structures submitted to PDB from January through March of 2002), S135 (135 structures submitted from April through

June), S163 (163 structures submitted from July through September), and S149 (149 structures submitted from October through December of 2002). Taken together, these control sets included 603 protein chains (only the first chain is considered in the case of PDB structures containing multiple protein chains) and a total of about 143,000 residues. In addition, we also used a set of 67 protein chains included in the EVA evaluation of secondary structure prediction servers in late 2003 and early 2004. This set will be referred to as S67. Since EVA metaserver collects results from other servers (including our own SABLE server), it allowed us to perform a direct comparison with other methods. The list of protein structures in the training and control sets can be downloaded from http://sable.cchmc.org.

Representation of an Amino Acid Residue and Its Environment

Following typical machine learning protocols, various classifiers discussed in this work assume that an amino acid residue to be classified is represented by a vector in a certain feature space defined by a set of attributes. In particular, all the methods discussed here assume that the local structural environment and evolutionary context of each residue is characterized by a sliding window of 11 amino acids, with the residue of interest at position 6. The window of length 11 proved to be sufficient in our tests to achieve accuracy essentially identical to those with longer windows. Moreover, a longer window would imply a larger number of parameters to be optimized, increasing the risk of overfitting.

In analogy to Petersen et al.,[25] we use an "expanded" output (i.e., predictions for 3 central residues at positions 5, 6, and 7 in the window are obtained simultaneously and then averaged over overlapping windows) (see next section). Contrary to previous studies, we do not distinguish explicitly terminal windows. Instead, for the first and last 4 residues, we create windows of length 11 by adding artificial N- and C-termini peptide extensions. Such extended sequences were then used as query sequences to generate multiple alignments. For example, if the original sequence was *M***A***VPAGL*…, it would be extended by adding the *EEDL* prefix, and the modified sequence *EEDL-M***A***VPAGL*… would be used to generate the multiple alignment and the resulting sliding window representation for the second (***A***) and subsequent residues (the first, and similarly, the last residues are only represented in terms of the expanded output for triplets of adjacent residues). Several different extensions, derived from terminal fragments of known proteins, had been tested and compared. We observed that such extensions help to improve slightly the accuracy for nonterminal residues in terms of the segment overlap (SOV) measure[26] (up to 0.5%), while providing similar accuracy in terms of the per-residue classification accuracy (Q_3) for the 3-state classification problem.[26]

Evolutionary information encoded in the form of family profiles derived from multiple alignments has been shown before to improve significantly the accuracy of secondary structure prediction.[5,27] Following in the footsteps of these previous efforts, each residue in the sliding window is represented here by the corresponding column of the position-specific scoring matrix (PSSM) generated iteratively by using the PSI-BLAST program.[22] Specifically, 3 PSI-BLAST iterations with the default options and without prefiltering of low complexity and membrane fragments (contrary to results in a previous work,[5] such prefiltering resulted in a slight decrease in accuracy) were performed, using the nr database[28] as of August 2003 with about 1,486,000 sequences. In addition to such obtained vectors (of dimension $11 \times 20 = 220$), each position in the window is characterized by the entropy at that position, which indicates explicitly the conservation of amino acids at that position, and an additional component indicating the presence of cystein residues in the whole window, adding, together, 12 features.

Furthermore, a majority of classifiers described in the next section were trained using an extended feature space with an additional 37 attributes: the average hydrophobicity and volume of amino acids observed at each position in the window (and the corresponding multiple alignment), thus adding 22 features, and 15 more attributes describing 3 central residues in terms of binary vectors of length 5, which indicate the presence of amino acids belonging to 1 of the 5 groups with distinct secondary structure propensities: {A, E, L}, {V, I}, {S, N}, {P}, {G}. The resulting total dimension of the input vectors was equal to 269. These features were found to improve somewhat the classification accuracy (on average by about 0.3%) on our validation sets.

The above-described representation of a local environment of an amino acid, with a sliding window of 11 residues and 269-dimensional input vectors was also used to train in parallel a novel NN-based regression system for predicting real-valued RSA.[18] Such independently predicted RSA values, or alternatively, their 2-class projections, with a threshold of 25% RSA separating the class of "buried" from the class of "exposed" residues, were subsequently used as additional features describing local environments of amino acid residues for the secondary structure prediction. However, the predicted RSA was not added to the original 269-dimensional input vectors, but rather to the input of the second-level ("structural") network that makes final predictions based on the results of the first-level networks trained in the original 269-dimensional feature space, as described in the next section. An attempt to include the predicted RSA also in the first-level networks did not result in improvements in secondary structure prediction.

We would like to add that we made another attempt to improve secondary structure prediction by adding information about the underlying exon–intron structure of genes coding for proteins of interest, using the Exon–Intron database (EID) derived from the GenBank 132.[29] Specifically, for each residue in the sliding window, we added a binary feature that indicates if an exon–intron boundary is observed with a sufficiently high frequency (with an arbitrarily chosen threshold) at that position in a given

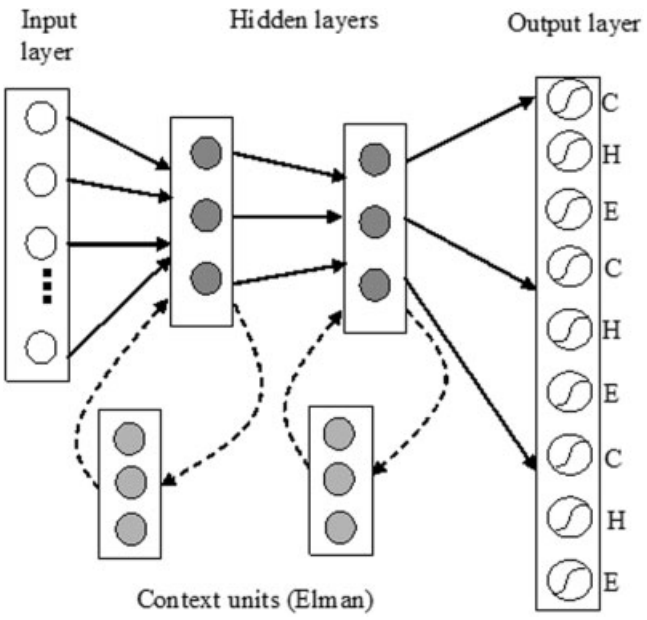

Fig. 1. Architecture of Elman-type recursive networks used in this work, with 2 hidden layers and the corresponding "context" layers, as well as an "expanded" output providing predictions for 3 central residues in the sliding window of 11 residues.

protein family. Unfortunately, even though we observed that splice junctions appear to be more frequent in the loops (as opposed to secondary structure elements), this information did not improve secondary structure prediction, suggesting that evolutionary profiles encoded by the PSSMs capture this information to some degree at least.

Neural Networks and Their Architectures

Following the protocol of Rost and Sander,[27] applied later by others (e.g., Jones[5]), our secondary structure prediction consists of 2 stages: (1) the initial ("sequence-to-structure") prediction that uses information derived from the amino acid sequence and input vectors in the feature space described in the previous section; and (2) the final ("structure-to-structure") prediction, which is based on the outcome of the first-level prediction and allows one to correlate better predictions for neighboring residues.

We used similar, Elman-type networks[30] for predictions at these 2 levels. Their architecture is schematically presented in Figure 1.

The advantage of Elman-type networks is that the representations of the problem developed in the hidden layers are fed back to themselves, providing dynamic memory of prior internal states of the network.[30] Specifically, each of the 30 nodes in both hidden layers of networks used in this study were connected by feedback loops with all the nodes in the same hidden layer through the so-called "context units." Weights associated with these additional edges were not optimized, as they simply add a scaled excitation pattern in the hidden layer obtained after presenting to the network previous training examples. Given either 232- or 269-dimensional input vectors for first-layer networks, this resulted in 8199 or 9309 adaptive parameters (weights and biases of the activation functions), respectively (the second-layer networks are discussed in detail below).

We compared Elman-type networks with standard feed-forward networks with the same number of nodes in the hidden layers and the same number of adaptive parameters. Both types of networks were trained using several standard learning algorithms, as discussed in the next section, and the results on our validation sets were on average about 1% better in terms of per-residue classification accuracy for Elman-type networks. For comparison, we estimated that the improvement due to the use of networks with 2 (as opposed to 1) hidden layers was on the order of 0.3%.

Another difference with respect to "standard" protocols is that we used an "expanded" output (a similar approach was in fact used before by another group[25]). The 3 classes of secondary structures are typically coded by binary vectors of length 3, resulting in 3 nodes in the output layer. Here, secondary structures of 3 central residues in the window were predicted at the same time (at both "sequence-to-structure" and "structure-to-structure" level networks); consequently, there were 9 nodes in the output layer. For comparison, we trained a system with just 3 output layer nodes and found that expanding the output layer to 3 residues and averaging the overlapping predictions from adjacent (1 to the left and 1 to the right) windows, improved slightly the accuracy (on the average by about 0.2% on the validation sets).

The input space for the second-level ("structure-to-structure") networks was based on the results from the first-step prediction. Again, the sliding window of 11 residues was used. Because of the expanded output, there are 3 alternative predictions for each residue, coming from 3 adjacent windows: 1 window centered at a given residue, 1 window centered at the left neighbor, and 1 window centered at the right neighbor. Thus, each residue is represented by 9 numbers (probabilities of that residue belonging to one of 3 classes according to 3 alternative first-level predictions from adjacent windows), resulting in a feature space of dimension 99 and 4209 parameters of the network to be optimized. Several networks trained using such a representation of an amino acid residue at the second level were combined into a consensus predictor (see next section) that will be referred to as Sable1 (or prediction without RSA).

As an alternative, we used an extended feature space, with real-valued RSA (either predicted using our NN-based regression system[18] or experimentally observed for the sake of comparison) for each of the residues in the sliding window added to the set of attributes described above, thus increasing the dimension of our input vectors to 110, and the number of parameters to be optimized to 4539. In order to assess the effects of projecting the real-valued RSA into discrete classes, we also trained several networks with RSA projected onto 2 classes, such that the additional node representing RSA of an amino acid residue is excited to 1 if the value of RSA is less or equal to 25%, and to 0 otherwise. A threshold of 25% RSA

is commonly used to separate buried and exposed residues in classification-based approaches to RSA prediction.[16,17] Predictions obtained with the information about predicted real valued RSA will be referred to as Sable2.

Training Protocols and Consensus Prediction

Our training set consisted of about 210,000 vectors representing individual residues and their local environment, primarily in the form of evolutionary profiles (using a sliding window of length 11, see the previous subsection). For each of the prediction systems (e.g., Sable1 vs Sable2) a number of individual networks was trained and combined into a consensus prediction. These individual networks were trained on different subsets consisting of 190,000 training vectors, with the remaining 20,000 vectors used as a validation set for a given network. Validation sets for individual networks were allowed to partially overlap; however, none of the networks combined into a final prediction system was trained on the same actual training set.

In order to develop and optimize the networks discussed here, we used the SNNS package.[31] The Resilient Propagation (RP) algorithm[32] with default parameters was used as the basic training protocol. Other algorithms, such as Quickprop[33] and the standard back-propagation (BP),[34] were applied as well. Nevertheless, the fastest convergence, smallest sum of squared errors (SSE), and the best generalization on the validation sets were obtained when using the RP algorithm.

Therefore, final prediction systems discussed in the remaining parts of the article are those obtained using the RP algorithm.

The training set was unbalanced in the sense that the number of examples for the class "β-strand" (E) was much smaller than for classes "helix" (H) or "coil" (C), reflecting the relative frequencies of secondary structure elements. Specifically, in our training set classes E, H, and C account for 21%, 37%, and 42% of the residues, respectively. Therefore, we used a modified ("weighted") sum of squared errors (*wSSE*) function which we sought to minimize in the training:

$$wSSE(z) = \sum_i [\alpha(C_i)\Delta_i(z)^2 + \alpha(C_{i-1})\Delta_{i-1}(z)^2 + \alpha(C_{i+1})\Delta_{i+1}(z)^2]. \quad (1)$$

The summation in the above formula runs over all training vectors i, with the exclusion of the first and last residue in each chain, which are only included via expanded output of their immediate neighbors. There are 3 quadratic terms in the error function (for the central residue and its 2 immediate neighbors) that explicitly reflect the expanded output implemented here, with the errors for individual residues obtained as a result of the summation over different classes j:

$$\Delta_i(z)^2 = \sum_{j=1}^{3} (\hat{y}_{ij}(z) - y_{ij})^2. \quad (2)$$

Here, $\hat{y}_{ij}(z)$ denote the predicted probabilities (assuming that the excitations of the output nodes are normalized) of each class given the parameters of the network, z, whereas y_{ij} represent the assigned truth (encoded as a binary output vector $\mathbf{y}_i$) for residue i, given its "true" class assignment, C_i, which is imposed in the training. The following values of the weighting parameters, $\alpha(C_i)$, which scale the importance of errors for different classes, were used: 0.3, 0.5, and 1.0 for residues in class C, H, and E, respectively. We would like to stress that while guided by the relative frequencies of secondary structure elements observed in our training set, this choice remains arbitrary. In particular, in order to help improve the prediction for β-strands, the relative weight for the class E is somewhat larger compared to weights resulting from the distribution of the 3 classes in the training.

Prediction schemes that are compared in the Results section differ in the setup of their second-layer networks. All of them, however, were based on a common first-layer consensus predictor. Specifically, 9 different networks (each trained on a different subset of the training set), with best generalization on the respective validation sets and with significant contributions to consensus classifiers, were combined into the final first-level prediction. Three of these first-level ("sequence-to-structure") networks were trained using 232-dimensional feature space, and the remaining 6 networks were trained in the extended 269-dimensional feature space, described earlier. Using different input spaces and different training subsets was meant to minimize correlations between individual networks. Adding more networks to the consensus did not lead to further significant improvements in the classification accuracy (the overall number of networks optimized in this study was much larger; see also the discussion in the Results section).

The second-level ("structure-to-structure") classifiers were also based on a consensus of 9 different second-level networks, trained using predictions of their first-level counterparts. Each of the final prediction systems (e.g., Sable1 or Sable2) was built as a combination of predictions from these individual networks, using the following rule to assign $\hat{C}_i$ as the class predicted for the ith residue by the consensus classifier:

$$\hat{C}_i = \max_j \left(\sum_{k=1}^{9} \hat{y}_{ij}(z_k) \right), \quad (3)$$

where $\hat{y}_{ij}(z_k)$ denotes the probability of assigning residue i to class j according to the kth individual network (z_k denotes the adaptive parameters for network k). The probabilities $\hat{y}_{ij}(z_k)$ were derived as normalized excitations of the corresponding output nodes and, similar to the first-level networks, averaged over 3 adjacent sliding windows to take into account the expanded output for 3 central residues in the window.

RESULTS

In this section we discuss the results obtained using new systems developed in this study for secondary structure

TABLE I. Comparison of Classification Accuracy Obtained Without and With Information About RSA Derived From Known Protein Structures (Limit of Perfect RSA Prediction)

Test sets Method	S163 Q3/SOV/MCC	S156 Q3/SOV/MCC	S149 Q3/SOV/MCC	S135 Q3/SOV/MCC
Without RSA	77.9/74.6/0.66	76.6/72.9/0.64	76.6/73.7/0.64	77.8/75.1/0.66
Real-valued RSA	81.0/78.4/0.70	80.3/77.6/0.69	80.5/78.0/0.69	80.9/78.2/0.70
Discretized RSA	79.3/77.4/0.68	78.5/75.8/0.66	78.2/75.0/0.66	79.4/76.9/0.68

Measured on Several control sets using the percentage of correctly classified (in 3 classes) residues, the SOV[3] and MCC. The last row includes results obtained with projection of the observed RSAs into 2 discrete classes.

TABLE II. Classification Accuracy of Secondary Structure Prediction for Networks Trained Using Predicted Real-Valued (Sable2) and Discretized RSA

Test set accuracy measure	S163 Q3/SOV/MCC	S156 Q3/SOV/MCC	S149 Q3/SOV/MCC	S135 Q3/SOV/MCC
Sable1	77.9/74.6/0.66	76.6/72.9/0.64	76.6/73.7/0.64	77.8/75.1/0.66
Sable2	78.4/75.1/0.66	77.5/73.3/0.65	77.0/73.4/0.64	78.3/75.3/0.66
Discretized RSA	78.2/74.8/0.66	76.8/72.5/0.64	76.7/73.6/0.64	78.2/75.5/0.66

prediction. The accuracy is compared primarily in terms of the standard per-residue classification accuracy, which is defined as the percentage of residues correctly classified into one of the 3 classes (the Q_3 measure), as well as the SOV measure that captures the extent of overlap between the predicted and observed secondary structure elements, as defined formally in Zemla et al.[26] In addition we used the Matthews correlation coefficient (MCC) defined for each 2-class problem (e.g., MCC_H for H vs other classes' problems),[35] with the overall MCC coefficient for the 3-class problem defined as the geometric average of the MCC coefficients for the 3 different 2-class problems.

We first compared the accuracy of secondary structure prediction obtained without the information about RSA with the classifiers trained using the actual (experimentally observed) real-valued RSAs, as well as their 2-class projections. As can be seen from Table I, which summarizes the results of these 3 classifiers in terms of Q_3, SOV, and MCC measures on 4 control sets, the actual RSA values, if known, could improve the accuracy of the secondary structure prediction between 3% and 5% in terms of both Q_3 and SOV measures. Improvements are observed for all classes, as indicated by MCCs for 2-class problems on the S163 control set: MCC_H (H vs non-H problem) increases from 0.73 to 0.77, MCC_E from 0.65 to 0.7, and MCC_C from 0.6 to 0.64 (similar results were obtained for other control sets as well). Interestingly, roughly only half of these improvements are observed when the real valued RSA is projected into 2 discrete classes, providing additional support for regression-based rather than classification-based approach for RSA prediction (see also Adamczak et al.[18]).

Thus, solvent accessibility of an amino acid residue can be used as an effective local fingerprint of the overall topology and protein structure packing, allowing one (at least in the limit of perfect prediction) to improve considerably the accuracy of secondary structure prediction. We would like to comment that such a conclusion a priori was not obvious, especially in light of the observed high level of variability of solvent exposure for structurally equivalent residues in families of homologous proteins.[15,18]

TABLE III. Comparison of Sable1 and Sable2 Results With Two State-of-the-Art Methods

	Q3	SOV
Sable1	75.4	75.8
Sable2	76.8	75.2
PROFsec	74.7	73.5
SAM-T99sec	76.9	74.9

Ranked among four best secondary structure prediction servers by the EVA metaserver,[4] using a set of 67 proteins derived recently from new submission to PDB.

Of course, in real applications, the actual solvent accessibility is not known, and its predicted value has to be used instead. Comparison of 2 systems included in Table I, one with real valued and another with projected experimental RSAs, suggests that the extent of improvements in secondary structure prediction will strongly depend on the accuracy of RSA prediction. In particular, it may require a real valued RSA prediction rather than a simple classification of amino acid residues as buried or exposed.

In order to test that hypothesis, we used our novel regression-based real-valued RSA prediction protocol (in the original work[18] referred to as Sable-wa—weighted approximation), which outperformed in rigorous tests classification-based approaches and achieved the correlation coefficient between predicted and observed RSA values of about 0.66 on control sets used also in this work.[18] Such predicted real valued RSAs, as well as their 2-class projections for comparison, were added to other attributes when training our new secondary structure prediction methods, as described in the previous section. The results are summarized in Tables II and III.

The results for 4 different control sets, comprising together 603 proteins without homology to proteins used in

the training, are included in Table II. Overall, the improvements with respect to the Sable1 method that does not utilize information about RSA, while being systematic, are rather small: between 0.4% for S149 set and 0.9% for S156 set in terms of the per-residue classification accuracy. However, as can be seen from Table II, the results are indeed somewhat better when real-valued RSA predictions are used (Sable2) as opposed to 2-class projections (discretized RSA). In terms of specific secondary structure elements, we observed a slightly bigger improvement for the class H (MCC_H increasing from 0.73 to 0.74 when the predicted RSA is used) than for the other classes for which the increase in terms of MCC is less than 0.01 on the S163 set.

The level of improvements observed when applying real-valued RSA prediction is similar to that obtained by applying 2-level prediction instead of using just first-layer ("sequence-to-structure") networks: In our case, the improvement due to the second-layer ("structure-to-structure") was between 0.6% and 0.9% on the same control sets. On the other hand, the effects of the committee of 9 networks used in our approach appear to be more significant: The improvement due to consensus was between 0.9% and 1.3% compared to the result of the best network in the committee.

Although the level of improvement observed in Table II is comparable to that due to other commonly used ways of enhancing the final accuracy of prediction (such as the 2-step strategy mentioned above), the question remains whether the observed differences are statistically significant.

In order to address this question we used the fact that large populations of networks with and without information about RSA were trained in this study. Since these networks were trained on 9 different subsets of the training set and tested on complementary validation subsets (see previous section on the amino acid residue and its environment), one can compare average results on these validation sets for both type of networks.

We would like to comment, however, that our protocol is not equivalent to a standard cross-validation procedure, because training subsets for different networks were allowed to overlap to some extent. Specifically, 27 networks trained without the information about RSA (with 9 of them included later in the Sable1 consensus classifier) and 27 networks trained with real-valued RSA predictions (with 9 of them included later in the Sable2 consensus classifier) were compared. Each of the training sets was used to train 3 alternative networks of both types, with the same architecture but with different random assignment of initial values for adaptive parameters.

The average per-residue classification accuracy (Q_3 measure) on the validation sets was 76.9 $\pm$ 0.4% for networks trained without RSA and 77.3 $\pm$ 0.3% for networks trained with predicted RSA. Thus, the difference between the means is comparable to standard deviations for these 2 samples. According to the standard t-test (with the assumption of equal variances), the 2 means are indeed different, with a p-value of 0.0001, suggesting that the observed differences are statistically significant.

Table III compares the performance of the new method with other state-of-the-art secondary structure prediction methods. Since we could not rule out correlations between our test sets and training sets used to develop other methods discussed here, we decided to use instead a NR set of 67 proteins (the S67 set) without homology to previously known structures. The S67 set was derived from recently submitted new protein structures by the EVA metaserver for continuous evaluation of registered secondary structure prediction servers (starting from the moment when our own server was added to the evaluation performed by EVA). The additional advantage of such an approach was that the results of other methods on this set were readily available.

We included in our comparison the results of the 4 best servers according to the current (as of March 2004) ranking by the EVA metaserver, namely, PsiPRED,[5] SAM-T99sec,[6] Scratch3,[9] and PROFsec[4] servers. The results for all 67 proteins were available at the EVA metaserver for 3 methods: Sable1, SAM-T99sec, and PROFsec. These results are summarized in Table III. The results of Sable2 were available for only a subset of S67 set, and they were recomputed locally (also to avoid biases due to initial problems with our server that resulted in a number of incorrectly submitted Sable2 predictions). Following the EVA methodology,[4] the classification accuracy is computed slightly differently in this case (i.e., Q_3 and SOV measures are first computed for each protein separately and then averaged over all proteins included in the test). This way of computing accuracy is more sensitive to outliers and reveals better potential failures of prediction protocols for specific classes of proteins (e.g., short proteins).

As can be seen from Table III, there is an improvement of about 1.4% on the S67 set in terms of the Q_3 measure; on the other hand, the results of Sable2 are worse than those for Sable1 by about 0.6% in terms of the SOV measure. These trends likely result from different biases in the S67 set as opposed to other control sets considered here. In particular, the EVA evaluation set includes a significant number of short proteins: The average length of proteins in the S67 set was 106 as opposed to 268 amino acids in case of the other control sets used here.

In addition, pairwise comparisons with PsiPRED[5] and Scratch3[9] servers were performed on smaller sets of proteins for which prediction of both methods and the corresponding Sable predictions were included in the EVA evaluation. We found that the Sable2 method was worse than PsiPRED by about 1% (on a set of 57 proteins) and better than Scratch3, also by about 1% (on a set of 61 proteins) in terms of Q_3 measure. Thus, while further tests may be required to better assess the relative performance (the reader is referred to the future results of the EVA metaserver evaluation[4] in that regard), the new method appears to be competitive with other state-of-the-art protocols for secondary structure prediction. Therefore, the new method could potentially be used as part of consensus-

based metaclassifiers, which have been shown to improve the accuracy with respect to best individual methods.[11,36]

CONCLUSIONS

In this work, we developed a new method for secondary structure prediction that utilizes predicted solvent exposure of an amino acid residue as an additional attribute describing its environment. Such a "local" descriptor may capture some characteristics of the overall packing of a protein and enhance secondary structure prediction by taking into account the effects of long-range interactions on local secondary structure propensities. In order to test the above hypothesis, we investigated whether an improvement in accuracy of secondary structure prediction can be achieved by using both: the actual RSA, derived from known structures (perfect prediction limit), as well as RSA predicted using our accurate regression-based protocol for real-valued RSA prediction, with the overall correlation coefficient between predicted and observed RSAs of about 0.66 obtained in rigorous tests.[18]

Our secondary structure prediction protocol follows in the footsteps of previous studies and utilizes evolutionary information encoded by family profiles, as well as advanced machine learning techniques.[5,9,27,37] Using a training set consisting of 860 NR protein chains, we first trained a number of Elman-type recurrent NNs that do not use the information about RSA of an amino acid residue. When combined into a consensus predictor, this system achieved 3-state classification accuracy of 76.6–77.9% on different control sets, derived using EVA-like methodology and comprising 603 NR proteins without homology to proteins included in our training set. We next trained similar systems with actual or predicted RSAs included as additional features.

When the actual RSAs derived from known structures were used, an improvement between 3% and 4% was observed, with per-residue classification accuracy between 80.3% and 81.0% on the same control sets. These numbers represent simply the upper bounds in the limit of perfect RSA prediction, whereas, in practice, RSA is predicted with a certain level of errors. Moreover, due to the variability in RSA of structurally equivalent residues in protein families, and due to conformational flexibility of protein chains in solution, which is likely to be higher than that of secondary structures,[15,18,38] caution should be exerted when drawing conclusions based on the suggested here limit. Nevertheless, these results support the idea of using RSA as an additional feature capable of capturing to some extent nonlocal effects. Interestingly, when the experimentally observed real-valued RSAs were discretized into 2 classes with the threshold of 25% RSA, the accuracy dropped between 1.5% and 2.3% on different control sets. This observation provides an additional support for casting the RSA prediction problem as a regression rather than classification problem.[18]

On the other hand, an improvement of up to 1.4% was achieved when the predicted RSA obtained by applying our most accurate NN-based regression method[18] was used. Thus, incorporating the best available real-valued RSA prediction introduces a significant amount of noise with respect to real RSA. Nevertheless, the improvements are systematic and statistically significant as shown by the *t*-test analysis for 2 sets of networks, each consisting of 27 networks trained with and without predicted RSA, respectively. Moreover, on a control set of 67 proteins recently submitted to PDB and derived by the EVA metaserver,[4] the new method achieved accuracy of 76.8% and 75.2% in terms of averaged per protein Q_3 and SOV measures, respectively. These results appear to be competitive with other state-of-the-art protocols for secondary structure prediction, opening a way for the new method to be included in consensus-based metaclassifiers.[36] While the extent of improvement in the context of other protocols for secondary structure prediction remains to be seen, we propose that accurate real-valued RSA prediction is likely to further enhance secondary structure prediction.

REFERENCES

1. Venclovas C, Zemla A, Fidelis K, Moult J. Comparison of performance in successive CASP experiments. Proteins 2001;5:163–170.
2. Fischer D, Elofsson A, Rychlewski L, Pazos F, Valencia A, Rost B, Ortiz AR, Dunbrack RL. CAFASP2: the second critical assessment of fully automated structure prediction methods. Proteins 2001;5: 171–183.
3. Schonbrun J, Wedemeyer WJ, Baker D. Protein structure prediction in 2002. Curr Opin Struct Biol 2002;12:348–354.
4. Eyrich VA, Marti-Renom MA, Madhusudhan MS, Fiser A, Pazos F, Valencia A, Sali A, Rost B. EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics 2001; 17:1242–1243.
5. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
6. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14,846–856.
7. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
8. Cuff JA, Barton GJ. Application of enhanced multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 1999;40:502–511.
9. Pollastri G, Przybylski D, Rost B, Baldi P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 2002;47:228–235.
10. Bystroff C, Thorsson V, Baker D. HMMSTR: a hidden Markov model or local sequence-structure correlations in proteins. J Mol Biol 2000;301:173–190.
11. Rost B. Protein secondary structure prediction continues to rise. J Struct Biol 2001;134:204–218.
12. Przybylski D, Rost B. Alignments grow, secondary structure prediction improves. Proteins 2002;46:197–205.
13. MacDonald JR, Johnson WC Jr. Environmental features are important in determining protein secondary structure. Protein Sci 2001;10:1172–1177.
14. Zhu ZY, Blundell TL. The use of amino acid patterns of classified helices and strands in secondary structure prediction. J Mol Biol 1996;260:261–276.
15. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
16. Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 2002;47:142–153.
17. Ahmad S, Gromiha MM. NETASA: neural network based prediction of solvent accessibility. Bioinformatics 2002;18:819–824.
18. Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks based regression. Proteins 2004;56:753–767.
19. Wagner M, Adamczak R, Porollo A, Meller J. Linear regression

models for solvent accessibility prediction in proteins. J Comp Biol 2005;12:355–369.
20. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL. The Pfam protein families database. Nucleic Acids Res 2002;30:276–280.
21. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
22. Altschul SF, Madden TL, Schaffer AA. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
23. Meller J, Elber R. LOOPP: Learning, Observing and Outputting Protein Patterns (LOOPP)—a program for protein recognition and design of folding potentials. Available online at http://cbsu.tc.cornell.edu/software/loopp
24. Meller J, Elber R. Linear optimization and a double statistical filter for protein threading protocols. Proteins 2001;45:241–261.
25. Petersen TN, Lundegaard C, Nielsen M, Bohr H, Bohr J, Brunak S, Gippert GP, Lund O. Prediction of secondary structure at 80% accuracy. Proteins 2000;41:17–20.
26. Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins 1999;34:220–223.
27. Rost B. Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
28. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2003;31:23–27.
29. Saxonov S, Daizadeh I, Fedorov A, Gilbert W. EID: The Exon–Intron Database: an exhaustive database of protein-coding intron-containing genes. Nucleic Acids Res 2000;28:185–190.
30. Elman JL. Finding structure in time. Cognit Sci 1990;14:179–211.
31. Zell A, Mamier G, Vogt M, et al. The SNNS users manual version 4.1. Available online at http://www-ra.informatik.uni-tuebingen.de/SNNS, 1995.
32. Riedmiller M, Braun H. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In: Proceedings of the IEEE International Conference on Neural Networks (ICNN 93); San Francisco, 1993.
33. Fahlman SE. Faster-learning variations on back-propagation: an empirical study. In Sejnowski TJ, Hinton GE, Touretzky DS, editors. 1988 Connectionist Models Summer School. San Mateo, CA: Morgan Kaufmann; 1988.
34. Rumelhart DE, McClelland JL. Parallel distributed processing. Vol. 1. Cambridge, MA: MIT Press; 1986.
35. Matthews BM. Comparison of predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:445–451.
36. Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. JPred: a consensus secondary structure prediction server. Bioinformatics 1998;14:892–893.
37. Baldi P, Brunak S, Frasconi P, Soda G, Pollastri G. Exploiting the past and the future in protein secondary structure prediction. Bioinformatics 1999;11:937–946.
38. Andersen CA, Palmer AG, Brunak S, Rost B. Continuum secondary structure captures protein flexibility. Structure (Camb) 2002;10:175–184.