PROTEINS: Structure, Function, and Bioinformatics 56:753–767 (2004)

Accurate Prediction of Solvent Accessibility Using Neural Networks–Based Regression

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

ABSTRACT **Accurate prediction of relative solvent accessibilities (RSAs) of amino acid residues in proteins may be used to facilitate protein structure prediction and functional annotation. Toward that goal we developed a novel method for improved prediction of RSAs. Contrary to other machine learning–based methods from the literature, we do not impose a classification problem with arbitrary boundaries between the classes. Instead, we seek a continuous approximation of the real-value RSA using nonlinear regression, with several feed forward and recurrent neural networks, which are then combined into a consensus predictor. A set of 860 protein structures derived from the PFAM database was used for training, whereas validation of the results was carefully performed on several nonredundant control sets comprising a total of 603 structures derived from new Protein Data Bank structures and had no homology to proteins included in the training. Two classes of alternative predictors were developed for comparison with the regression-based approach: one based on the standard classification approach and the other based on a semicontinuous approximation with the so-called thermometer encoding. Furthermore, a weighted approximation, with errors being scaled by the observed levels of variability in RSA for equivalent residues in families of homologous structures, was applied in order to improve the results. The effects of including evolutionary profiles and the growth of sequence databases were assessed. In accord with the observed levels of variability in RSA for different ranges of RSA values, the regression accuracy is higher for buried than for exposed residues, with overall 15.3–15.8% mean absolute errors and correlation coefficients between the predicted and experimental values of 0.64–0.67 on different control sets. The new method outperforms classification-based algorithms when the real value predictions are projected onto two-class classification problems with several commonly used thresholds to separate exposed and buried residues. For example, classification accuracy of about 77% is consistently achieved on all control sets with a threshold of 25% RSA. A web server that enables RSA prediction using the new method and provides customizable graphical representation of the results is available at http://sable.cchmc.org. Proteins 2004;56:753–767.**

Key words: relative solvent accessibility; neural networks; regression; approximation; classification; protein structure prediction; solvent exposure; SABLE

INTRODUCTION

Many approaches for protein structure prediction rely on intermediate predictions of key attributes of amino acid residues in a protein, such as secondary structure, number of contacts, or solvent accessibility. In particular, improved secondary structure prediction methods, which achieved per residue accuracy of more than 75% for classification into three states (helix, beta strand, coil),[1,2] contributed significantly to the improved performance of fold recognition and de novo protein structure prediction methods.[3–6] The relative success of secondary structure predictions stems mainly from the use of evolutionary information and from the application of advanced machine learning techniques to solve the underlying classification problem.[7,8]

Similar to secondary structure prediction, accurate estimates of the extent of residue solvent exposure, as measured by the relative solvent accessibility (RSA) or accessible surface area, are likely to further enhance the level of success in fold recognition and de novo protein folding. Contrary to secondary structures, however, residue solvent accessibility is a real-value number, and there are no clearly defined, distinct classes of residues. Moreover, RSA appears to be less conserved than secondary structures in families of homologous structures. For example, the correlation coefficient of RSA between equivalent residues in homologous structures has been estimated to be equal to 0.77.[9] While these estimates are strongly dependent on the level of homology that one considers (see Results section), predicting solvent accessibility based on family profiles has proved to be a more difficult problem than secondary structure (SS) prediction.[2,9]

In the past, many different methods have been devised for predicting solvent exposure from a protein–amino acid

Grant sponsor: Cincinnati Children's Hospital Research Foundation. Grant sponsor: National Institutes of Health; Grant number: AI055338.

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

Received 3 January 2004; Accepted 20 February 2004

Published online 20 May 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20176

sequence. For example, neural networks (NN),[9] Bayesian,[10] substitution matrix–based, and simple baseline approaches[11] were proposed. The common denominator for most existing RSA prediction methods is that they cast the problem as a classification problem. Since NN proved to be particularly successful in the realm of SS classification, a similar approach was also adapted for the RSA prediction problem.[9] Following the SS prediction schemes, discrete classes are imposed on real-value solvent accessibilities, and the resulting classification problem is solved using NN or other machine learning techniques.

Recently, several groups attempted to further improve RSA prediction, using both feed forward[12,13] and recurrent NN,[14] support vector machines (SVMs),[15] or information theory approaches.[16] According to published accuracy estimates and our own tests, the recurrent NN-based ACCpro method appears to be the most accurate of these methods, achieving (in cross-validation) a classification accuracy of about 77% for the two-class problem with a threshold of 25% RSA.[14]

However, there are inherent problems with imposing an arbitrary threshold that separates buried from exposed residues, owing to the extent of variations in terms of RSA for residues that occupy equivalent positions in homologous structures. For example, the average difference (i.e., mean absolute error) in RSA for pairs of residues that are aligned according to PFAM alignments, with at least one of the residues having RSA in the range of 20–30%, is equal to 16.8% (see Results section). Therefore, imposing different classification for residues with RSAs of 24% and 26%, for example, while using the 25% threshold, is nonphysical and makes the training of such a classifier a difficult task. Moreover, the interpretation of the results is cumbersome, since multiple classifiers with different thresholds may be necessary to assess the actual level of solvent exposure and distinguish residues with RSAs of 25% from those with RSAs of 75%, for instance.

A novel method that provides real-value RSA prediction, based on simple binary encoding of amino acid sequences and feed forward NN with two output nodes and subsequent transformation of their excitations into a real-value prediction, was published recently.[17] While this new method goes beyond the classification protocols, the reported accuracy appears to be limited, with mean absolute errors between 18.0% and 19.5% RSA and correlation coefficients of up to 0.5.[17]

In this work, we develop a new method for accurate prediction of RSA using NN-based nonlinear regression instead of a classification approach. Rather than imposing arbitrary boundaries between the classes, we seek a continuous approximation of the real-value RSA. Several NNs, including feed forward and Elman recurrent networks, with a single logistic output node that approximates the experimental real-value RSA, were trained and combined into a consensus regression–based predictor. The expected level of variation in homologous structures in terms of RSA for residues with a different degree of exposure was assessed and used to weight individual contributions to the error function to be minimized. A metaclassifier is developed to provide reliability scores for real-value regression-based RSA predictions. We also develop an alternative, semicontinuous model with the so-called thermometer encoding of the output and several standard classifiers for comparison. The accuracy of all the methods is estimated and compared using EVAluation of automatic protein structure prediction (EVA)-like methodology[18] for evaluation of the accuracy of SS prediction methods.

MATERIALS AND METHODS

Relative Solvent Accessibility

The solvent-accessible surface area of an amino acid residue indicates its level of burial (or solvent exposure) in a protein structure and is often expressed in terms of RSA. The RSA of an amino acid residue i, which will be denoted as RSA_i throughout this article, is defined as the ratio of the solvent-exposed surface area of that residue observed in a given structure, denoted as SA_i, and the maximum obtainable value of the solvent-exposed surface area for this amino acid, denoted as MSA_i:

$$RSA_i = 100 \cdot \frac{SA_i}{MSA_i}[\%]. \quad (1)$$

Thus, RSA_i adopts values between 0% and 100%, with 0% corresponding to a fully buried and 100% to a fully accessible residue, respectively.

We used the Dictionary of Protein Secondary Structure (DSSP)[19] program to compute residue solvent-accessible surface areas, SA_i, for known protein structures. The maximum obtainable values of the solvent-exposed surface area are taken from Chothia[20] and correspond to surface-exposed area of the central residue observed in tripeptides in extended conformation. For comparison with other methods, we also used alternative values adopted by Rost and Sander.[9] Note that different normalization may lead to different classification when an arbitrary threshold is used. The regression approach, on the other hand, is expected to be less sensitive to the choice of the actual normalization scheme.

It should also be noted that estimating accuracy of approximation and classification approaches requires different error measures. In case of classification, with discrete classes imposed on the actual values of RSA, the natural and commonly used error measure is simply the classification accuracy. However, other error measures need to be applied in case of continuous approximation.

Here, we applied several different measures, including the root-mean-square error (RMSE) and weighted RMSE (wRMSE), defined as follows:

$$\mathrm{wRMSE} = \sqrt{\frac{1}{N}\sum_i \alpha(o_i)(y_i - o_i)^2}, \quad (2)$$

where y_i denotes the predicted and o_i the observed (experimental) value of RSA for residue i, respectively. The weights $\alpha(o_i)$ depend on the observed value of RSA and are used to scale the errors relative to expected level of RSA variation for equivalent residues in families of homologous

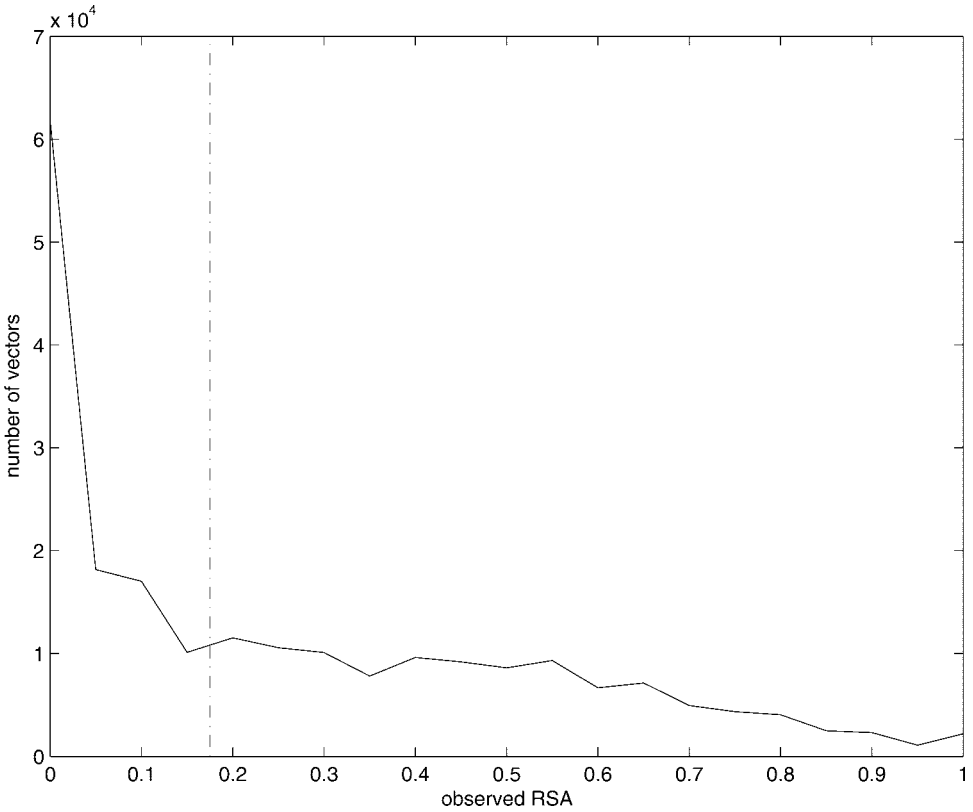

Fig. 1. The distribution of the observed values of the RSA, scaled to [0,1] interval, in our training set of 860 protein structures. Note the significant fraction of fully buried residues (RSA = 0%) relative to exposed residues. The vertical dashed line indicates the threshold (RSA = ca. 17% or 0.17 on the scale used in the figure), which results in a balanced definition of two classes: buried and exposed residues, with an equal number of residues in each class.

structures (the values used here are defined in the section on training protocols). The wRMSE reduces to RMSE when all the weights $\alpha(o_i)$ are equal to one. We also applied two other standard measures, namely, the mean absolute error (MAE):

$$\mathrm{MAE} = \frac{1}{N}\sum_i |y_i - o_i|, \qquad (3)$$

and the correlation coefficient between the predicted and observed values of RSA, as defined in Rost and Sander.[9]

For comparison with classification approaches, we also considered projections of the real-value RSA into discrete classes, measuring the error by using the standard classification accuracy. In case of a two-class problem, with a threshold separating buried (B) and exposed (E) residues, the two-state accuracy per residue, Q_2, defined as percentage of correctly predicted residues in two states, B and E, as well as Matthews correlation coefficients[21] (MCCs; note that for two-class problems the correlation coefficient and MCC are equivalent) were used.

Training and Control Sets

In order to derive a representative and nonredundant training set we used the PFAM database, version 6.6.[22] consisting of 3071 protein families (domains), of which about 45% had known three-dimensional (3D) structures. By excluding problematic structures (e.g., structures that were not parsed successfully by the DSSP program), a subset of 860 families represented by at least one 3D structure was obtained. For the training, a randomly chosen single structure (of an individual protein chain) per family was used, resulting in a set of 209,685 residues (Fig. 1). Multiple structures in each family, if available, were used to assess the level of variability in solvent-accessible surface area for pairs of residues that can be aligned and are in this sense structurally equivalent. The accuracy of RSA assignment based on homologous structures was estimated using this approach (see the section on training protocols).

Several control sets were derived following an approach used before by the EVA metaserver in order to assess the accuracy of SS prediction methods.[18] First, nonredundant new structures submitted to the Protein Data Bank (PDB) database[23] were selected using a filter available from the PDB server, with 50% sequence identity as the threshold to remove redundant new entries. Further redundancies in terms of more distant homology were removed by applying BLAST[24] sequence alignments to prune entries that resulted in matches with E-values lower than 0.001 with respect to sequences already included in any of the

control sets. Next, structurally biased sequence alignments, as implemented in the LOOPP (Learning, Observing and Outputting Protein Patterns) program,[25] were used in order to exclude proteins homologous to structures included in our training set. Specifically, matches resulting in Z scores higher than 6.5 for structurally biased sequence alignment, a threshold that was found sufficient before,[26] were excluded.

The resulting control sets of structures (assembled in specific months of 2002) with no homology to proteins included in the training will be referred to as S156 (156 structures submitted to the PDB from January through March 2002), S135 (135 structures submitted from April through June 2002), S163 (163 structures submitted from July through September 2002), and S149 (149 structures submitted from October through December 2002), respectively. Taken together, these control sets contain 603 protein chains (only the first chain for each structure is considered for protein complexes) and 143,348 residues. The average length (taking into account residues with unresolved coordinates) of proteins included in control sets is 238 as opposed to 268 in the case of proteins included in the training. The list of protein structures in the training and all control sets can be downloaded from http://sable.cchmc.org.

Feature Space

It has been demonstrated before that evolutionary information encoded in the form of a family profile, for example, as the position-specific scoring matrix (PSSM) generated iteratively by using the PSI-BLAST program,[24] improves significantly (by up to 10% for the three-state prediction) the accuracy of SS prediction.[2,7,8] On the other hand, some studies suggested that the extent of improvement due to the inclusion of evolutionary information is not as significant in the case of RSA.[2] Nevertheless, an improvement of up to 5% was achieved.[2,14]

Here, following these previous studies, we employed evolutionary information in the form of PSSMs and compare the results with the single sequence–based approach. In order to generate family profiles encoded in the form of a PSSM, we used the PSI-BLAST program (version 2.6 with default options unless specified otherwise; see also discussion in the Results section). Both, the SWISS-PROT database,[27] as of August 15, 2002 with 113,193 sequences, and the nr database[28] as of November 7, 2002, with 1,229,187 sequences, as well as a newer version as of August 8, 2003, with 1,486,372 sequences, were used to assess the effects of the growing size of sequence databases. Three iterations of PSI-BLAST were performed to generate family profiles, and no masking of low complexity regions or membrane domains was used.

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 for SS prediction. Since we subsequently use RSA predictions in order to improve SS prediction within the framework of multitask learning,[29] we did not attempt to optimize the length of the window independently for RSA. For the same reason (i.e., for the sake of consistency with our SS prediction system, in which we applied the same solution as an alternative to a special node indicating the edges of the alignment), windows for the first and last 5 residues were created by virtue of adding short artificial N- and C-termini peptide extensions to the query sequence. This resulted in the extended PSSMs of an appropriate length.

Since each residue in the window is initially represented by a column of the PSSM, the resulting input vectors have dimension 220. In addition to information derived from the PSI-BLAST PSSM, we also characterized each position in the window by the average hydrophobicity and volume of amino acids observed at that position in the multiple alignment, as well as the entropy at that position, adding an additional 33 features to the input vectors. Sixteen more features resulted from adding for the central residue and its two immediate neighbors a binary vector of length 5 that indicates the presence of amino acids belonging to 1 of the 5 groups with distinct SS propensities: {A, E, L}, {V, I}, {S, N}, {P}, {G}, and one more component indicating the presence of cysteine residues in the window. Thus, the input vectors consisted of 269 features. Some of the above arbitrary choices are discussed in more detail in the Results section.

Network Architectures

In this section we describe different types of networks used to obtain either a continuous approximation of the real-value RSA or various discrete classifications, as well as confidence level scores for regression-based prediction. For the regression-based continuous prediction we used a feed forward architecture shown in Figure 2 (without the loops indicated by dashed lines). The input layer consisting of 269 nodes, two hidden layers consisting of 30 nodes each, and a single logistic output node were used. Thus, the overall number of edges and, consequently, weights to be optimized was equal in this case to 9000, plus 61 additional (bias) parameters defining the sigmoidal activation functions for nodes in the hidden layers and the output node.

While the number of parameters to be optimized is large, it is still considerably smaller compared to methods that use a longer sliding window to characterize the structural environment around the residue of interest. We will refer to the above architecture as 269-30-30-1FFR, with FFR standing for feed forward regression network. Several of these networks were trained using different training algorithms, as described in the next section.

In addition to standard multilayered feed forward networks, we also used recurrent Elman networks that have been applied to sequential data, such as time series analysis.[30] 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] The overall architecture of the Elman networks used here was

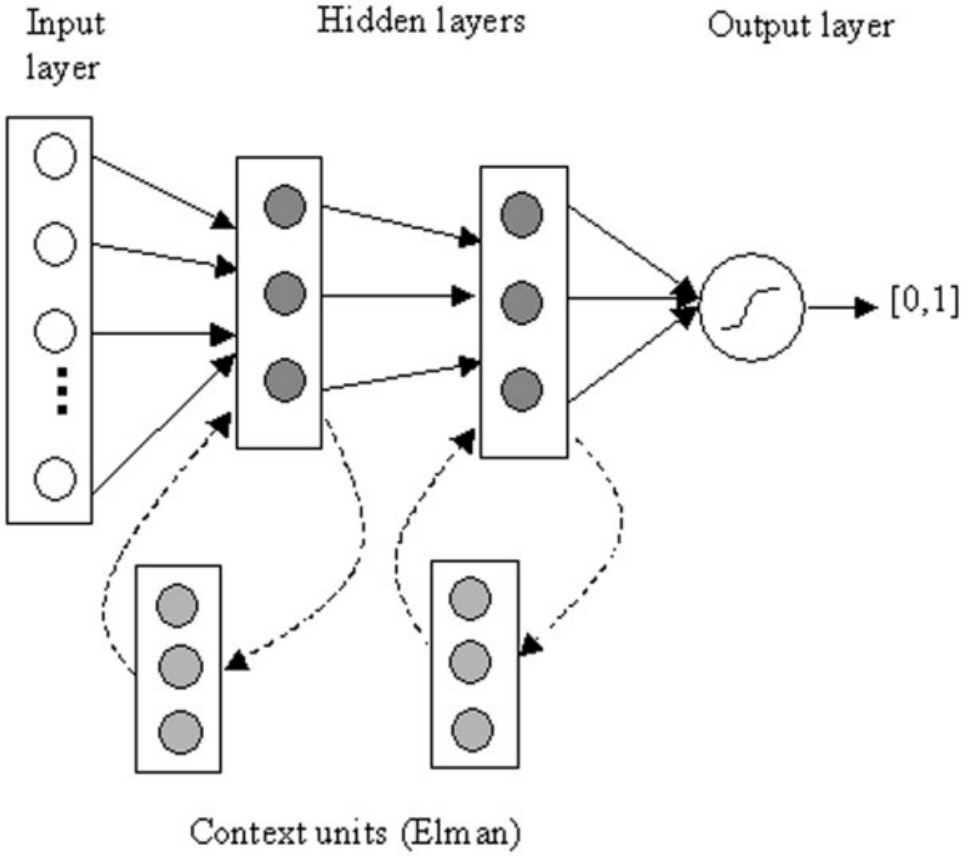

Fig. 2. Schematic representation of a multilayer feed forward and Elman network architectures, with a single real-value (logistic) output node used here for regression-based prediction of RSA. Subsequent layers are fully interconnected (i.e., each node of a given layer is connected to all nodes in the next layer). In the case of the recurrent Elman networks, the hidden layers are connected by additional (passive) feedback loops, with the context units that provide the context of prior internal states of the hidden layers.

similar to that of 269-30-30-1FFR. However, nodes in the hidden layers were connected by feedback loops, with all nodes in the same hidden layer, as shown in Figure 2. 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. We found that Elman-type networks typically require a smaller number of iterations (epochs) to achieve accuracy comparable to that of standard feed forward networks for both SS and relative solvent accessibility prediction. We will refer to this type of network as 269-30-30-1ER, with ER standing for Elman regression.

Another architecture that we considered in this work uses the so-called thermometer encoding[31] for the output node in order to obtain a semicontinuous approximation of the RSA. Specifically, we used 20 output nodes to represent 20 classes of residues with their RSA in the intervals: [0,0], (0,5], (5,10], (10,15], … (90,100] % RSA. However, contrary to the standard binary encoding used for classification problems, the real-value RSA was transformed into a binary vector of length 20, with *all* the bits (nodes) corresponding to lower RSA values being activated (rather than just a single node corresponding to a particular RSA). We found that, in practice, a reversed encoding, with buried residues represented as vectors of ones (as opposed to vectors of zeros) worked better. For example, an RSA of 8% is transformed into the following binary vector: $[1,1,\ldots,1,1,0,0,]^T$, whereas an RSA of 16% is transformed into the vector $[1,1,\ldots,1,1,0,0,0,0]^T$, and so on. When making predictions, the center of the last interval with excitation higher than 0.5 is assigned. A schematic representation of this network, which will be referred to as 269-30-30-20T, is included in Figure 3.

The network architectures discussed so far involve two hidden layers, following the setup that we used for SS prediction, for which it turned out to be advantageous. In order to measure the influence of the second hidden layer, we also trained a number of regression networks with only one hidden layer consisting of 33 nodes (in order to obtain a comparable overall number of parameters to be optimized). These networks were trained and combined into a consensus predictor, as discussed in the following sections, for the networks with two hidden layers. The results in terms of regression and classification accuracy, while marginally lower in case of the one-layer network, were not significantly different. In the remaining part of the

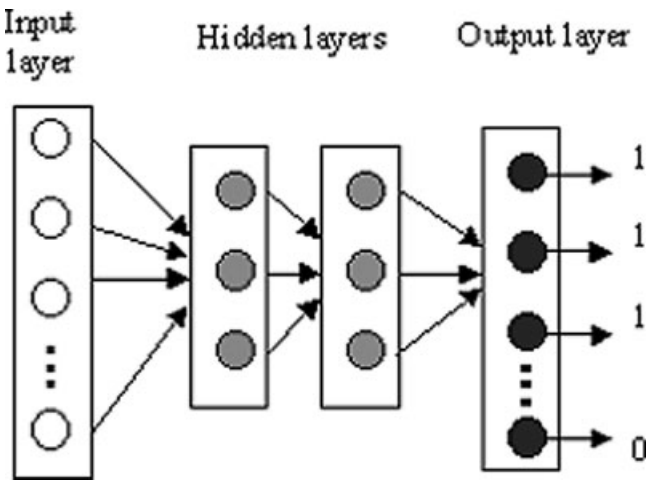

Fig. 3. Schematic representation of a multilayered feed forward network with a thermometer encoding in the output layer.

article, we only present the results for the two-layer regression and thermometer networks.

For classification problems, we used networks with the input and hidden layers, as described above. However, two binary output nodes were added in case of two-class problems. This kind of network will be referred to as 269-30-30-2. In addition, we also considered a simpler type of architecture for auxiliary classification networks that provide confidence scores for regression-based predictions. Four different feed forward networks, with an expanded input that includes 9 additional nodes representing RSA prediction for the residue of interest obtained by using 9 different networks and a single hidden layer, were used for that purpose. They will be referred to as 278-50-2, 278-30-2 (two different networks), and 278-20-2, respectively.

Training Protocols

The standard sum of squared error (SSE) cost function that we seek to minimize in the training is defined as follows:

$$\mathrm{SSE}(z) = \sum_i (y_i(z) - o_i)^2, \tag{4}$$

where $y_i(z)$ is the predicted value for the ith input vector given the parameters of the network (weights and biases) z, and o_i represents the observed real-value RSAs that are imposed in the training. For convenience, the RSA values were restricted to the set {0,1} for each of the binary classification nodes.

The minimization of the standard SSE function is strongly influenced by large errors that concern especially exposed residues. In Figure 4 and Table I, we summarize the observed level of variation in RSA for "equivalent residues" in families of homologous structures. Residues are regarded as equivalent here when they are aligned according to the PFAM database or, alternatively, using pairwise BLAST alignments for known family members derived from PFAM.

As can be seen from Figure 4, the average RMSE for pairs of equivalent residues is much higher for exposed residues. Therefore, in analogy to RMSE defined in Eq. (2), we introduced the corresponding weighted SSE cost function in order to account for naturally occurring variability in terms of RSA in families of homologous structures:

$$\mathrm{wSSE}(z) = \sum_i \alpha(o_i)(y_i(z) - o_i)^2. \tag{5}$$

The weights $\alpha(o_i)$ are defined using a coarse graining, as described below. Let B_i denote the interval (or bin) that contains the observed value of RSA for the ith residue (i.e., $o_i \in B_i$). Using this notation, $\alpha(o_i)$ may be defined as $\alpha(o_i) = 1/\langle (o_j - o_k)^2 \rangle_{o_j \in B_i}$, with the average in the denominator computed over all pairs of residues (j,k) that contain at least one residue with the observed value of RSA o_j in the interval B_i (we used the following bins: [0,10], (10,20], (30,40], etc.). The actual values of the weights were derived using PFAM alignments to identify pairs of equivalent residues, and they were used to scale down errors for exposed residues, relative to buried residues (see Table I).

The weighted cost function defined above was used to train 9 neural networks with the overall 269-30-30-1 architecture: 6 feed forward networks (269-30-30-1FFR), with 3 of them trained by the standard backpropagation (BP) algorithm, and 3 by using the resilient propagation (RP) algorithm,[32] as well as 3 networks with Elman architecture (269-30-30-1ER), with 2 of them trained by using RP and 1 by BP. These different networks achieved different local minima and were combined into a consensus predictor by using the arithmetic average of individual approximations. This new predictor will be referred to as SABLE weighted approximation (SABLE-wa). We also trained the above 9 networks using the standard error function of Eq. (4) and combined them into a consensus predictor, which will be referred to as SABLE approximation (SABLE-a). The consensus-based predictions achieved an increase in accuracy with respect to best individual networks (which were usually the Elman-type networks) between 1% and 2%.

In addition to regression networks, we also trained (using the RP algorithm) three Elman-type networks with the thermometer encoding (269-30-30-20T), which were combined into a semicontinuous consensus predictor that will be referred to as SABLE thermometer (SABLE-t). For comparison, we also developed consensus classifiers for two-class problems with several different thresholds that combined 4 different 269-30-30-2FF networks, 2 trained using BP and 2 trained using RP. These methods will be referred to as SABLE 2-class (SABLE-2c) and a number indicating the threshold used.

Our training set consisted of 209,685 vectors representing individual residues. This set was further split in the training into 2 subsets: an actual training set of 190,000 vectors, and a validation set containing the remaining vectors. The state of the network (adaptive parameters) was saved after each 10 epochs during learning, and the generalization was assessed using the validation subset. For further evaluation, we used these networks that

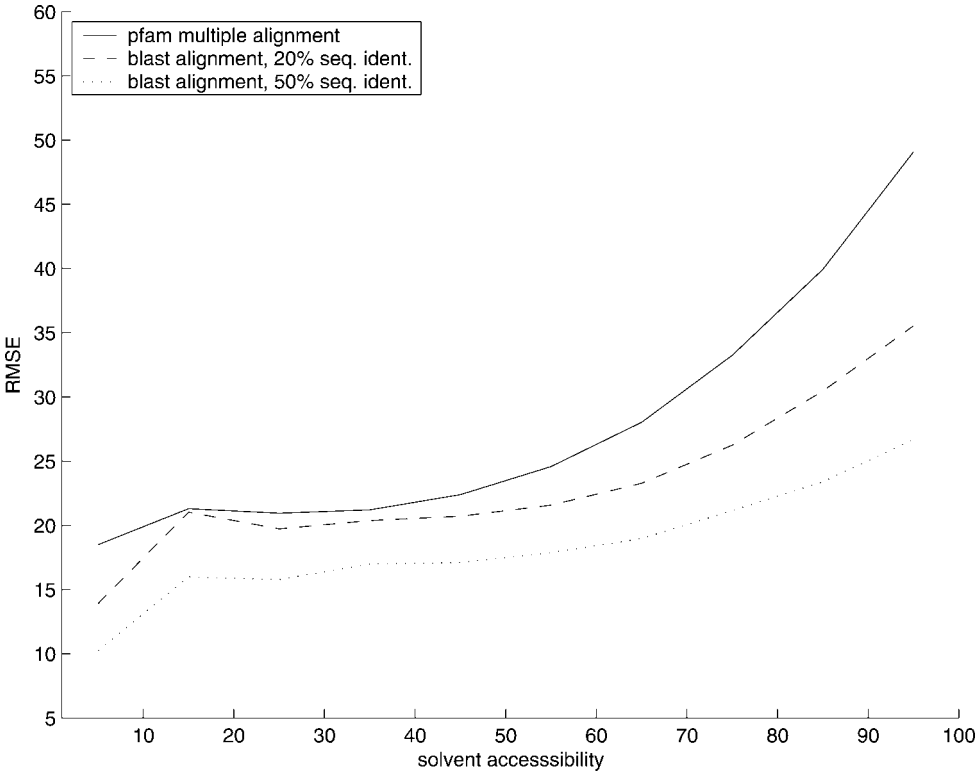

Fig. 4. Average root-mean-square error between RSA of "structurally equivalent residues" in different regions of solvent accessibility. For the sake of this analysis, pairs of structurally equivalent residues were defined by using original PFAM as well as BLAST pairwise alignments for a set of 809 protein families, with at least two known structures included in the PFAM database. The average RMSE value is computed in bins of length 10 and shown at the center of each bin. Note that the level of RSA deviations for structurally equiivalent residues depends strongly on the level of homology, as illustrated by differences between the 3 curves included: PFAM alignments include distant homologs, whereas BLAST pairwise alignments are restricted to homologs with at least 20% or 50% sequence identity, respectively. Note also the plateau between RSAs of 15–40%, which indicates a class of residues with an intermediate level of solvent exposure for which the extent of variation in terms of RSA in structural homologs is least dependent on the level of homology.

TABLE I. Level of Variation in RSA for Pairs of Equivalent Residues in Protein Families Identified by PFAM Multiple Alignments

RSA interval	⟨*MAE*⟩ [%]	⟨*RMSE*⟩ [%]	α (o_i)
[0,10]	8.1	18.5	5.4
(10,20]	14.1	21.3	4.7
(20,30]	15.7	20.9	4.8
(30,40]	16.8	21.2	4.7
(40,50]	17.8	22.4	4.5
(50,60]	19.0	24.5	4.1
(60,70]	21.4	28.0	3.6
(70,80]	25.2	33.3	3.0
(80,90]	31.2	40.0	2.5
(90,100]	39.1	49.0	2.0

See also Figure 4. MAEs, RMSEs, and the resulting (scaled) weights for the weighted SSE function of Eq. (5) in the respective intervals of RSA values (on the scale from 0% to 100%) are reported in columns 2, 3 and 4, respectively. Relatively large discrepancies between MAE and RMSE for buried residues result from a higher level of conservation of RSA for such residues and the binomial shape of the distribution of differences (absolute errors).

achieved the best accuracy on our validation set. For all the networks discussed here. the best generalization was obtained after around 150 epochs. Following standard practice, the input vectors were presented to the network in a random order. The Stuttgart Neural Network Simulator (SNNS) package with default settings for BP and RP algorithms[33] was used to train all the neural networks discussed in this work.

Assessing Statistical Confidence of RSA Predictions

Applying NNs to classification problems has the advantage of offering a convenient definition of reliability indices for prediction in terms of a normalized activation of the output nodes. In case of regression-based solvent accessibility prediction, however, this obvious prescription for obtaining estimates of statistical significance cannot be applied. Therefore, in order to derive reliability indices we used an auxiliary two-class classification problem, where the first class includes correct predictions, and the second, incorrect predictions, respectively.

Specifically, a prediction was classified as correct if the difference between the predicted and observed value of RSA was smaller than 1.5 times the average difference in RSA for pairs of equivalent residues in protein families, as defined for each range (bin) of RSA in Table I. For example, if the observed RSA was 1% and the predicted value was

TABLE II. Classification Accuracy Using Projected Real-Value RSA Predictions

	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
SABLE-a	75.8	75.9	76.3	76.8	76.8	77.1	77.5	78.0	79.2	80.6
SABLE-t	78.7	77.4	77.1	76.5	76.6	76.7	77.1	77.8	78.9	80.4
SABLE-wa	77.1	77.5	77.7	77.7	77.3	77.2	77.4	77.8	78.6	80.1

Results in terms of per residue accuracy, Q_2, for a series of two-class problems on the S163 set, defined by different thresholds (%) that separate buried and exposed residues, are reported.

14%, then the prediction was classified as incorrect, since the (scaled) average difference expected for the bin [0,10] was equal to 12.2%. We would like to stress that while guided by the observed level of variability in terms of RSA in homologous structure, this particular choice of tolerable errors in RSA prediction (as well as the previous choice of weights for the weighted SSE function) remains arbitrary.

Four NNs with architectures 278-50-2, 278-30-2 (two networks), and 278-20-2, respectively, were trained using the same training set as before and the standard BP algorithm, and subsequently combined into a consensus classifier for such defined (meta) classification problem. As for standard classification problems, the output vector was binary, $\vec{P} = [P_{\text{corr}}, P_{\text{inc}}]^T$, with $[1,0]^T$ representing correct prediction and $[0,1]^T$ representing incorrect prediction. The RSA values obtained by using 9 individual approximation networks, trained as described in the previous section, were included in the input vectors. The normalized consensus score for making a correct prediction, $P^{\text{con}}_{\text{corr}}$, was calculated as follows:

$$P^{\text{con}}_{\text{corr}} = \frac{\sum_k p^k_{\text{corr}}}{\sum_k (p^k_{\text{corr}} + p^k_{\text{inc}})}. \quad (6)$$

Here, k runs over different networks in the committee, whereas P^k_{corr} and P^k_{inc} denote the activation of the output node for classes "correct" and "incorrect" in the kth network, respectively.

RESULTS

Limits of Homology-Based Prediction

We start our discussion of the results with an analysis of limitations that are to be expected for RSA prediction methods based on evolutionary profiles of protein families (which is the case here). In their pioneering work, Rost and Sander[9] concluded that RSA is less conserved in protein families, relative to SSs. The correlation coefficient of RSA between homologous structures was estimated to be equal to 0.77 when using structure alignments and 0.68 when using less accurate sequence alignments (for the same set of 80 pairs of homologous structures). Due to lack of suitable alternatives, their analysis was biased toward globins that were represented by 19 out of 80 pairs of structures.

We now revisit this question and report the results of our extended analysis using a set of 809 protein families included in the PFAM database and containing at least two known structures to enable comparison of RSA in homologous structures. The average number of alternative structures in these families was equal to 5.8 (with the standard deviation of 9.0 and 5 families including more than 50 structures per family). As can be seen from Figure 4 and Table I, which summarize our findings, the deviations between RSA of equivalent residues (i.e., those that can be aligned using a certain type of alignment) generally increase with the level of surface exposure, although the actual values vary significantly depending on the level of homology considered.

TABLE III. Classification Accuracy With 3 Different Databases Applied to Generate PSI-BLAST Multiple Alignments, Obtained by Projection of the Weighted Approximation (SABLE-wa) Onto the Two-Class Classification Problem, With the Threshold of 25% RSA

Database	S163	S156	S149	S135
1,486 kS nr	77.3%	76.5%	76.6%	77.3%
1,229 kS nr	77.0%	76.4%	76.2%	77.4%
113 kS SWISS-PROT	76.2%	75.0%	74.9%	76.4%

For example, using the original PFAM alignments, one obtains the correlation coefficient of 0.57, as opposed to 0.82, when more distant homologs are excluded by using 50% sequence identity threshold. Strictly speaking, in the latter case, the BLAST program is used to realign pairs of homologous sequences included in the PFAM database, and all pairs with less than 50% sequence identity are excluded from the analysis (the reason for that being that our predictions of RSA were based on BLAST alignments). For distant homologs, the level of RSA conservation is much lower even for a significant fraction of (the most abundant) nearly buried residues, resulting in lower overall correlation coefficients.

The above observation suggests that the overall performance of any method based on evolutionary profiles might suffer when very remote homologs are included, for example, in the PSI-BLAST multiple alignment. On the other hand, however, family profiles as opposed to single sequence–based methods improve the results by about 5%. An additional improvement of about 1–2% is observed when larger sequence databases and thus, in principle, better profiles are used (see Tables II and III). Thus, one might expect, in accord with previous analysis showing sensitivity of the PHDacc method to the choice of the E-value threshold in PsiBLAST,[2] a trade-off between the noise introduced by distant homologs and the requirement of an informative evolutionary profile. Nevertheless, our attempts to utilize different thresholds for PSI-BLAST iterations failed to yield significant improvements in terms

TABLE IV. Performance of Regression-Based Real-Value Prediction Models (SABLE-a, SABLE-wa) and Semicontinuous Thermometer Model (SABLE-t) on 4 Different Control Sets, as Measured by Correlation Coefficients (CCs) Between Predicted and Experimental RSA Values, MAEs, and RMSEs (the Latter Two in the Units of % RSA)

	S163			S156			S135			S149		
	CC	MAE	RMSE	CC	MAE	RMSE	CC	MAE	RMSE	CC	MAE	RMSE
SABLE-a	0.65	15.6	20.8	0.64	15.9	21.0	0.66	15.3	20.5	0.64	16.0	21.0
SABLE-t	0.65	15.6	22.5	0.63	15.8	22.6	0.65	15.5	22.3	0.63	16.0	22.9
SABLE-wa	0.66	15.5	21.2	0.64	15.7	21.3	0.67	15.3	20.9	0.65	15.8	21.4

TABLE V. Comparisons of the accuracy (% RSA) of 3 Prediction Models: Average RMSE (Top Lines) and Corresponding Standard Deviations (Bottom Lines) for Different RSA Ranges Using the S163 Control Set

	0–5	5–10	10–15	15–20	20–25	25–30	30–35	35–40	40–45	45–50
SABLE-a	15.6	17.9	17.0	16.5	15.2	14.9	14.9	15.0	16.0	17.5
	4.6	5.5	4.8	4.3	3.3	2.8	2.6	2.7	3.3	4.1
SABLE-t	10.4	14.0	14.8	16.0	16.6	17.5	18.6	19.8	21.4	23.3
	3.7	4.5	3.8	3.4	3.0	3.1	3.5	4.4	5.4	6.6
SABLE-wa	12.9	14.5	13.9	13.6	13.0	13.3	14.1	15.1	16.9	18.8
	3.3	3.9	3.4	2.9	2.2	1.9	2.1	2.7	3.6	4.5

Note, that the thermometer encoding (SABLE-t) results in more accurate predictions for buried residues (the first bin) and the simple approximation (SABLE-a) are most accurate for exposed residues. The weighted approximation (SABLE-wa) achieves best results in the range, 10–35% RSA.

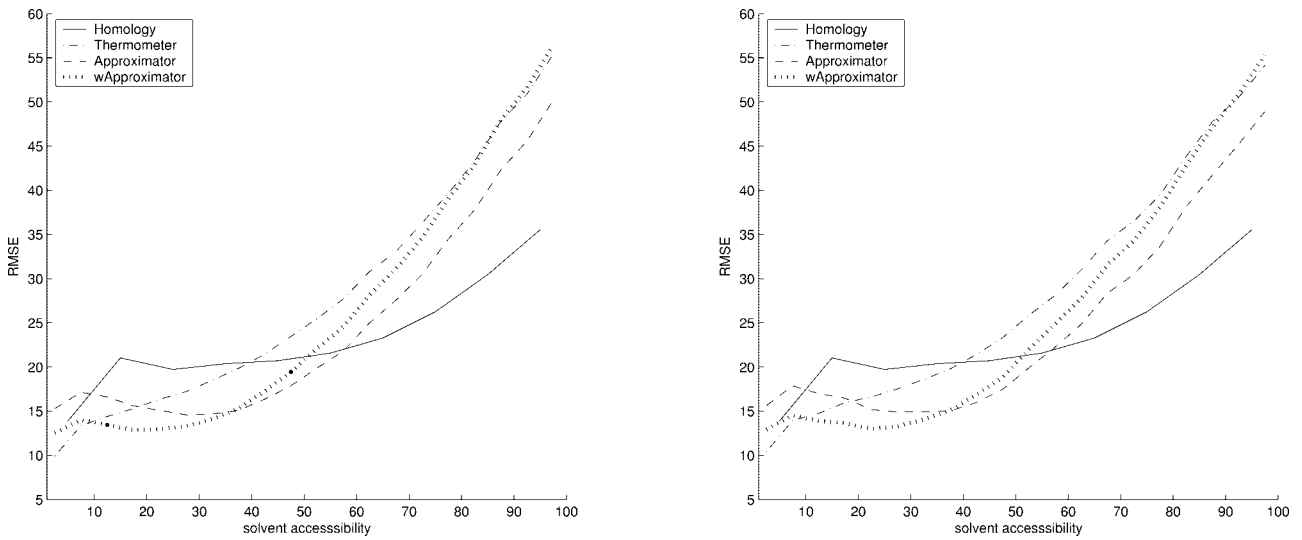

Fig. 5. RMSEs for RSA prediction in different ranges of the RSA in the training (**left panel**) and on the S163 control set (**right panel**), using several methods: homology-based prediction (using BLAST alignments for known structural homologs—see Fig. 2), continuous (weighted) approximation, and semicontinuous prediction using thermometer encoding.

of accuracy, indicating perhaps that a family-dependent level of homology would need to be considered, as opposed to a single threshold for all families.

Performance of Regression-Based Methods

In this section, we present and analyze the results of continuous predictors based on the simple (SABLE-a) and weighted approximation (SABLE-wa), as well as the semicontinuous predictions obtained by using the thermometer encoding (SABLE-t). Both the accuracy of real-value predictions and classification accuracy for projections into two-class problems with different thresholds are assessed. The results for all the control sets are summarized in Table IV. Further dissection of the results for the S163 control set is provided in Tables II and V, and for both the training set and S163 in Figures 5 and 6. The effects due to the size of the sequence database are assessed in Table III.

The overall accuracy of each method, as measured by the correlation coefficients, as well as MAE and RMSE error measures, is provided in Table IV. In terms of correlation coefficients and MAE, SABLE-wa performs somewhat better than the two other methods, achieving correlation

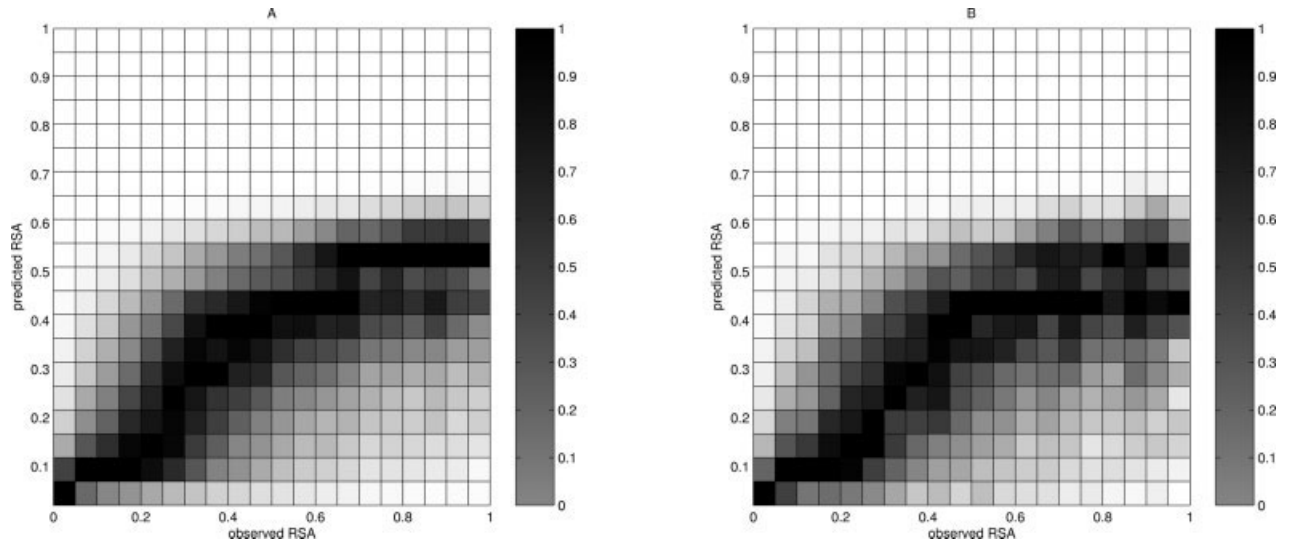

Fig. 6. Correlations between observed and predicted (using weighted approximation) values of RSA for different ranges of solvent exposure, scaled to [0,1] interval. The results for the training set (**A**) and the S163 control set (**B**) are shown. The density of vectors is normalized in each column independently, and the boxes with maximum density are marked in black to indicate intervals with the most frequent values of predicted RSA for a given interval of observed values. Note that while buried and partially exposed residues are predicted with relatively small errors and the densities approximately following the diagonal, surface-exposed residues with an RSA of 60% (0.6) or more (which are, however, much less frequent—see Fig. 1) are systematically predicted as more buried.

between 0.64 and 0.67, and MAE between 15.8% and 15.3% RSA. On the other hand, SABLE-a is somewhat better than SABLE-wa in terms of RMSE, with errors from 20.5% to 21.0% RSA. When correlation coefficients, as well as the MAE and RMSE error measures, are computed for each protein independently and then averaged (rather than computing an average over all residues in the control set), SABLE-wa predictions yield the average correlation coefficient of 0.65, average MAE of 17.4% RSA, and average RMSE of 22.9% RSA on the S163 control set, for instance. The standard deviations are in this case equal to 0.12, 5.6% and 5.9% RSA, respectively, reflecting relatively high variability in accuracy achieved for different proteins included in the S163 control set. Similar results were obtained for other control sets and for both SABLE-a and SABLE-t, indicating that the differences between the three methods in terms of the overall accuracy (which are on the order of 0.01 for the correlation coefficients) are not significant. Nevertheless, the three methods achieve significantly different accuracies for residues with different levels of solvent exposure, as discussed below.

As can be seen from Figure 5, the regression accuracy as measured by the RMSE is comparable to that of homology-based predictions for buried and partially exposed residues. In fact, the accuracy of all three systems is better than the homology-based prediction, as obtained by using BLAST alignments for pairs of homologous structures with at least 20% sequence identity (see Materials and Methods section on training protocols), in the range between 10% and 40% RSA (and the range between 10% and 50% for the two regression-based methods). The errors increase, relative to the homology based-prediction for more exposed residues, with weighted approximation being (as expected) least accurate in this region.

A more detailed comparison of the three prediction systems for the range between 0% and 50% RSA is also included in Table V. The weighted approximation (SABLE-wa) achieved best results in the range from 10% to 35% RSA, with differences between the means being at least as large as the standard deviation for SABLE-wa in the 20–25% RSA interval, for instance. Consequently, also the classification accuracy of SABLE-wa is higher for two-class problems obtained by projecting the real-value RSA into discrete bins, with threshold from 10% to 30% RSA (see Table II). Thresholds in this range coincide approximately with a balanced definition of the two classes and provide the most informative classifications (see next section).

It is interesting to note that the approximation methods reproduce the pattern observed before for homology-based predictions (see Fig. 4), with a plateau for residues of intermediate levels of solvent exposure in the range from 10% to 40% RSA, whereas the semicontinuous thermometer encoding resulted in monotonically decreasing accuracy for the whole range of RSA values. Since the input information is identical in each case, these differences result from alternative encoding of the output and error definitions. For example, the weighting of the error function in case of SABLE-wa leads to a better performance with respect to SABLE-a for buried and partially buried residues.

Correlations between predicted and observed RSA are further presented in Figure 6, for the training (panel A) and control set S163 (panel B), respectively. Each column in these "heat maps" is normalized independently by dividing the number of residues with a specific range of predicted RSA values by the maximum number of residues in that column. The highest density is represented by

TABLE VI. Errors in Regression-Based RSA Prediction Decrease With the Growing Reliability Scores (Probabilities) Derived Using an Auxiliary Classification Method (Described in the Text)

Probability	0.9	0.8	0.7	0.6	0.5	0.0
Fraction of residues	9.3%	18.1%	31.1%	47.2%	66.8%	100%
Correct—wide	94.0%	91.8%	89.4%	87.0%	84.0%	77.7%
(fraction of buried)	99.9%	96.7%	90.0%	81.8%	72.2%	62.5%
Correct—medium	89.8%	80.4%	70.0%	62.6%	56.1%	48.4%
(fraction of buried)	100%	97.9%	93.0%	85.8%	76.2%	65.2%
Correct—narrow	77.0%	60.0%	48.7%	42.4%	37.6%	32.4%
(fraction of buried)	100%	98.1%	93.0%	85.4%	75.4%	64.0%

The overall fraction of residues predicted with a reliability score (probability) higher than certain threshold and the percentage of correctly predicted residues, as well as the fraction of buried residues among the correctly predicted ones, are given for each threshold for the S163 control set. Three error bars are considered: wide, medium, and narrow, corresponding to 1.5, 0.75, and 0.5 times the average difference in RSA for pairs of equivalent residues (defined for each range of RSA values in Table I), respectively.

black squares. An ideal solution, with perfect correlation between the observed and predicted RSA values, would result in a sharp diagonal density. In our case, the density is reasonably diagonal (for both training and control sets) in the range of 0% to 50% RSA, indicating the range of relatively accurate real-value RSA predictions using the weighted approximation.

While simple approximation and thermometer encoding resulted in a slightly more diagonal density also for the exposed residues (data not shown), they too overestimated systematically the level of burial for surface exposed residues. Our attempts to introduce an a posteriori correction accounting for this systematic error failed to yield significantly improved results, because higher accuracy for relatively infrequent highly exposed residues was offset by increased errors for buried residues.

A lot of consideration has been given in the literature to improving the quality of multiple alignments in order to further improve the accuracy of SS and RSA prediction.[2,12,14] We estimated how the accuracy of our RSA prediction depends on (primarily the size of) sequence databases used to generate the PSI-BLAST PSSM. The SABLE-wa real-value RSA predictions were projected for that purpose onto the two-class classification problem with the threshold of 25% RSA (see Table III). Three databases were used: nr with 1,486,372 sequences (denoted as 1,486 kS nr), a somewhat older version of nr with 1,229,187 sequences (denoted as 1,229 kS nr), and the SWISS-PROT database with 113,193 sequences. Note, that while the results obtained using the SWISS-PROT database were worse by about 1.5% on 4 different test sets, the difference between 1,229 kS and 1,486 kS nr databases was not significant. Therefore, we hypothesize that no significant gains will be obtained due to the further growth of sequence databases.

In order to further assess the effects of multiple alignments and to simulate the worst-case prediction scenario, we estimated the accuracy of regression-based predictions assuming that no homologous sequences were available to build the multiple alignment. For this case, the BLOSUM62 matrix was applied to represent the amino acids instead of the PSI-BLAST PSSM. The two-class classification accuracy of the projected SABLE-wa method (without retraining) dropped to about 55% on our test sets, with the class threshold set to 25% RSA. For comparison, we also trained two regression-based systems, based on information derived from a single sequence and amino acid residues represented either by the respective column in the BLOSUM62 matrix (i.e., implicitly containing some family nonspecific evolutionary information) or by simple binary vectors. Such systems achieved similar accuracy of about 72% when projected into two-class problem with a 25% threshold and will be referred to as SABLE-wa BS62 and SABLE-wa binary, respectively (see discussion in the next section).

Next, we estimated how the accuracy of our RSA predictions (using weighted approximation, SABLE-wa) correlates with the reliability indices obtained using the meta-classification system described in the Materials and Methods section on assessing statistical confidence of RSA predictions. The results for the S163 test set are summarized in Table VI. Real-value RSA predictions were classified as correct if the difference between the predicted and observed value of RSA was either smaller than the 1.5 times the average difference (correct—wide), 0.75 times the average difference (correct—medium) or half of the average difference (correct—narrow) in RSA for pairs of equivalent residues, as defined for each range of RSA values in Table I. For example, if the observed RSA is 1% and the predicted value is 7.5%, then prediction was classified as correct when using the wide and medium error bars, but incorrect when using the narrow error bar, which is equal in this case to about 4% RSA (half of the expected error bar of 8.1% RSA for residues in this range of RSA—see Material and Methods section on assessing statistical confidence of RSA predictions).

As can be seen from Table VI, the fraction of residues predicted within the expected error bars for a given range of RSA values increases with the growing reliability scores. On the other hand, the fraction of buried residues

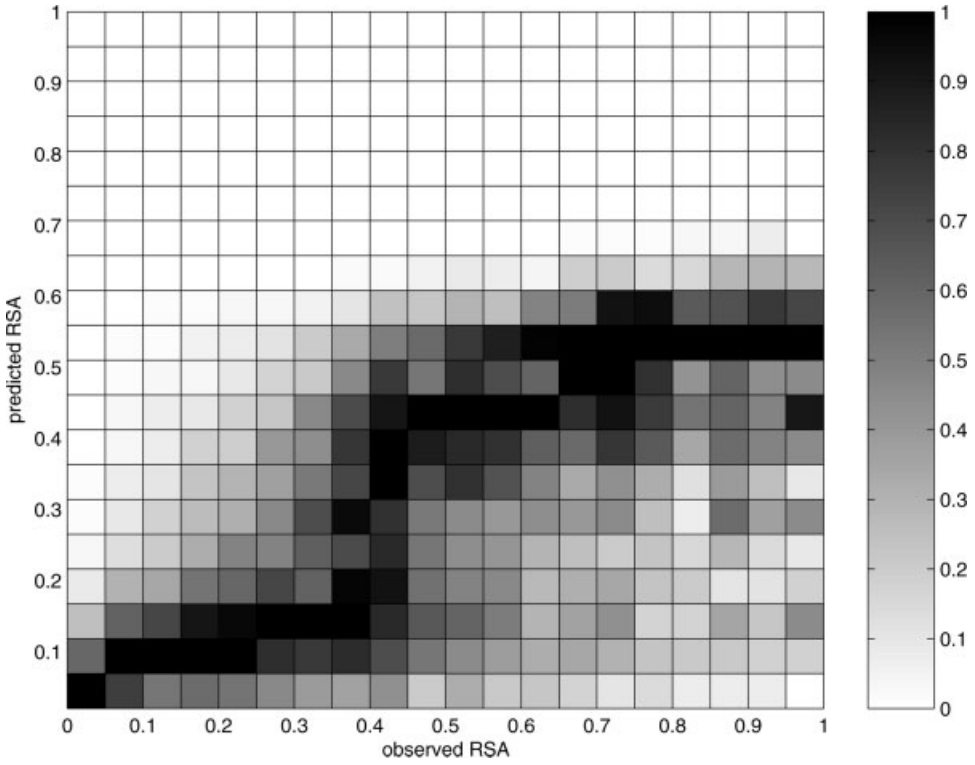

Fig. 7. Correlations between observed and predicted values of RSA for different ranges of solvent exposure: Same as panel B in Figure 6, except that only predictions with probability higher than 0.5 are taken into account. Note that most of such predictions concentrate reasonably close to the diagonal for a relatively wide range of RSA values.

in the subpopulation of residues that were predicted within the respective error bars increases quickly with the reliability scores as well. For example, even for the wide error bars, essentially all residues predicted with probability higher than 0.9 and within the error bars are buried. Since for buried residues the error bars used here are equal to 1.5, 0.75, and 0.5 times 8.1% RSA, respectively, one may conclude that approximately 94%, 90%, and 77% of residues predicted with probability higher than 0.9 is predicted with errors smaller than 12%, 6%, and 4% RSA, respectively.

Thus, buried residues appear to be easier to predict accurately. In fact, buried residues constitute the majority of residues that are predicted within the error bars even without using reliability scores (see last column in Table VI). Nevertheless, a growing fraction of nonburied residues with predicted RSA within the respective error bars at a given level of exposure can be identified with lower probability thresholds (see also Fig. 7). As a consequence of the above trends, a significant increase in classification accuracy was also observed when the results were projected into two-class problems. For example, 81.4%, 89.6%, and 96.6% correct classification was achieved when considering residues with reliability scores higher than 0.5, 0.7, and 0.9, respectively, for the two-class problem with a class threshold of 25% RSA. For the problem with 15% RSA threshold, the corresponding numbers were 78.3%, 83.3%, and 93.2% correct classification.

In light of the above, we conclude that the reliability scores can be used to identify subsets of residues predicted with lower errors, both in terms of real-value predictions and classification. It is also worth noting that all the reliability scores (probabilities) of correct prediction dropped consistently below 0.5 when the simple BLOSUM62 representation was used instead of the PSSM (see earlier discussion). Thus, a low overall probability of the prediction can be used to indicate when the method fails due to the lack of sufficient evolutionary information.

Comparison With Other Methods

Several new methods for predicting RSA have been proposed recently.[13–16] These efforts follow earlier attempts[9,12] and cast the RSA prediction as a classification problem, which is typically solved using machine learning techniques, such as NNs or SVMs. Consequently, the performance of such methods is evaluated in terms of classification accuracy. For example, the Q_2 measure is typically used for two-class problems when an arbitrary threshold (e.g., RSA of 25%) is imposed to separate the buried and exposed residues and define the B and E classes, respectively. Therefore, a direct comparison of our regression-based approach with such classification-based methods was not possible.

In order to address the above dilemma, we compared the performance of our continuous approximation in terms of projections into two-class classification problems, with

different thresholds defining the class of buried and the class of exposed residues. We present here a detailed comparison for two thresholds, namely, 15% and 25% RSA, that define an approximately balanced division into the two classes and coincide with thresholds used before in the literature. Moreover, according to the analysis of Pollastri et al.,[14] and also in accord with our own findings (see Figs. 1 and 4), other thresholds, such as 5% or 50% for instance, are much less informative. For example, the classification accuracy for systems trained with 5% or 50% thresholds was only slightly better than the baseline prediction, in which one simply assumes that all the residues belong to the class represented by the largest number of examples in the training.

We included in our analysis three relatively successful methods from the literature, namely ACCpro,[14] Jnet,[12] and NETASA,[13] for which Web servers are available, facilitating a more detailed comparison. We developed a metaserver that was used to submit sequences of all the proteins included in our four control sets to ACCpro, Jnet, and NETASA servers (using several available thresholds for classification). We did not attempt to remove structures that might be homologous to proteins included in the respective training sets of the ACCpro, Jnet, and NETASA methods. Therefore, the following results (which are discussed in detail below) should be regarded as an upper bound for the actual accuracy.

According to published accuracy estimates and our own tests, the most accurate of these methods appears to be the recurrent NN-based ACCpro method, which achieved in three-fold cross-validation about 77% correct classification for two-class problem with a threshold of 25% RSA.[14] The ACCpro method was trained using PSI-BLAST PSSMs for several alternative definitions of the two-class problem, with thresholds 0%, 5%, 10%, and so on, using a set of 1008 protein chains for cross-validated training. As a reference, the PHDacc method achieved a cross-validated accuracy of about 75% for the two-state problem with 25% threshold.[9]

The Jnet prediction method is based on feed-forward NN and combines into a consensus prediction classifier trained with Psi-BLAST PSSM and HMMER profile Hidden Markov Model (HMM) representations of evolutionary family profiles.[12] Jnet achieved classification accuracy of 76.2% in 7-fold cross-validation on the training set of 480 proteins for the two-class problem with a threshold of 25% (the Jnet server also provides results for thresholds 0% and 5%).[12] Contrary to Jnet or ACCpro, the NETASA method, which is also based on a feed-forward NN, does not utilize evolutionary profiles of protein families for predicting RSA. Instead, a simple binary coding of amino acids is used.[13] Therefore, NETASA, with a reported accuracy of 71.1% for the threshold 25%, estimated on a small control set of 71 proteins,[13] may be used to provide a reference for methods that are based on family profiles.

We would like to stress that any objective comparison of the Web servers for solvent accessibility prediction has to be done with care, as different groups use different definitions of RSA. For example, while the DSSP program was applied to derive solvent-accessible areas for ACCpro and Jnet methods, NETASA used an alternative program, which is not readily available. In addition, each of these methods utilizes a different normalization scheme to derive relative solvent accessibilities, defined in Eq. (1). Taking this into account in our evaluation of the ACCpro and Jnet servers, we defined the true state of each residue by applying the DSSP program and the maximum accessibilities used in the original papers.[12,14] However, such a direct comparison was not possible for the NETASA server, for which we applied the DSSP program and the maximum accessibilities used in this work. Thus, any results obtained this way only approximate the actual performance of NETASA.

An additional difficulty in assessing accuracy of different servers stems from the fact that outdated sequence databases might have been used to generate multiple sequence alignments and evolutionary profiles. According to our own analysis, this effect should be small (between 1% and 2%— see Table III). Nevertheless, for further comparison, we developed our own classifiers for two-class problems (SABLE-2c) with the thresholds 15% and 25% RSA, as discussed in the Materials and Methods sections on network architectures and training protocols. These classifiers were based on evolutionary profiles as encoded by PSI-BLAST PSSMs and used the same representation of amino acid residues as our regression-based predictors.

As can be seen in Table VII, which summarizes the results of different two-class predictors, the classification accuracy achieved for both thresholds by the projected continuous approximation (SABLE-wa) was higher than accuracy of specifically trained two-class classifiers with these thresholds, for all but one control set. Although the increase in accuracy was not larger than 0.7% with respect to our own classifiers (SABLE-2c), it demonstrates that a single regression–based system may replace multiple classifiers and achieve classification accuracy of about 77% for the 25% RSA threshold and up to 78% for the 15% RSA threshold. Note also that the performance of the alternative approximations that were trained without family profiles (SABLE-wa BS62 and SABLE-wa binary, which are very much comparable in terms of their classification accuracy) was worse by 5–8% for classification problems considered here.

The ACCpro server achieved an accuracy of up to 71% (i.e., significantly lower than the original estimate of about 77% and lower than the accuracy of our own two-class systems). While the latter might be partially explained by differences in the sequence databases used and the extended representation of a sliding window used in this work (which incorporates several additional physical and chemical profiles—see the Materials and Methods section on feature space), the original estimates of the ACCpro accuracy based on cross-validation appear to be too optimistic. Nevertheless, we would like to stress that the ACCpro server outperformed other servers in our tests. The Jnet server achieved up to 69% correct classification for the 25% threshold, as opposed to original estimate of 76.2%. The NETASA server, on the other hand, also achieved an accuracy of around 69%, which is, however, only 2% below

TABLE VII. Performance of Two-Class Predictors, Measured in Terms of Classification Accuracy (%) and Matthews Correlation Coefficients, on 4 Test Sets with Thresholds 25% (A) and 15% (B) RSA

Method	S163	S156	S135	S149
A. ACCpro server 25%	70.4/0.41	69.8%/0.41	70.6%/0.42	71.1%/0.43
SABLE-wa BS62	71.7%/0.43	71.1%/0.42	72.2%/0.44	72.2%/0.44
SABLE-wa binary	71.4%/0.42	70.9%/0.41	71.9%/0.43	72.1%/0.44
SABLE-2c 25%	76.7%/0.53	75.8%/0.52	77.1%/0.54	76.4%/0.53
SABLE-wa	77.3%/0.54	76.5%/0.52	77.3%/0.54	76.6%/0.53
B. ACCpro server 15%	70.0%/0.39	69.5%/0.38	70.3%/0.39	70.7%/0.40
SABLE-wa BS62	70.1%/0.39	70.0%/0.38	70.9%/0.40	71.1%/0.41
SABLE-wa binary	70.4%/0.40	70.2%/0.39	70.8%/0.40	71.1%/0.40
SABLE-2c 15%	77.1%/0.54	76.6%/0.53	77.8%/0.56	76.9%/0.54
SABLE-wa	77.7%/0.55	76.8%/0.53	78.3%/0.56	77.3%/0.54

The advantage of continuous approximation is that a single system can be used (without retraining) for different thresholds, outperforming in fact two-class classifiers trained for specific thresholds considered here, as seen by comparing the projected weighted approximation (SABLE-wa) with profile-based two-class classifiers (ACCpro and SABLE-2c). The use of multiple alignments improves the accuracy between 5% and 8%, as seen by comparing the results of the profile-based SABLE-wa with single sequence-based approximations (SABLE-wa BS62 and SABLE-wa binary).

the original estimate (note that the number of parameters to be optimized is much smaller in case of NETASA, which reduces the risk of overfitting), even though our own definition of RSA was used in this case to define the true classes.

We stress that our validation scheme follows the EVA methodology for evaluating the accuracy of SS prediction methods by using new submissions to the PDB and avoiding homology with structures included in the training. This kind of evaluation is more likely to sample new folds and new types of structures with specific biases not included in the training compared to cross-validation that was used in the original papers to estimate the accuracy of the ACCpro and Jnet methods. Further continuous validation of RSA prediction methods using EVA-like methodology may be required to assess progress in this field.

CONCLUSIONS

Knowing approximately the location of locally ordered SS elements greatly facilitates the search for the correct structural template in fold recognition, as well as the search for the native conformation in de novo simulations, as illustrated by the subsequent critical assessment of protein structure prediction (CASP) competitions in blind structure prediction. A similar premise concerns the prediction of relative solvent accessibility, which indicates the level of exposure of an amino acid residue in a protein structure. In this work, we proposed a novel method for predicting RSA based on nonlinear regression rather than a classification approach.

For comparison, we also trained several NN-based classifiers with standard binary output, as well as networks utilizing the so-called "thermometer encoding." In addition, we also developed a novel metaclassifier approach in order to derive reliability scores for regression-based real-value predictions. The accuracy of all the methods was validated using EVA-like methodology for evaluation of the accuracy of SS prediction methods. Newly deposited PDB structures were used to create control sets that consist of nonredundant protein structures with no homology to proteins included in the training.

In our rigorous validation tests, the new method achieved accuracy close to that of RSA assignment based on homology for a wide range of relative solvent exposure, with the overall MAE (defined as the average absolute difference between the predicted and observed values of RSA) of about 15.5% and the correlation coefficient between the predicted and observed values of about 0.66. Another method providing real-value prediction of RSA was published recently[17] and achieved significantly lower accuracy, with 18.0–19.5% MAE and correlation coefficients of up to 0.5. Moreover, in our tests, the regression-based predictor outperforms classification-based approaches from the literature when real-value predictions are projected into discrete classes, with several most relevant thresholds separating the buried and exposed residues. Therefore, one simple regression-based predictor may replace multiple classifiers, providing accurate prediction of the real RSA values.

ACKNOWLEDGMENTS

We thank all the authors of solvent accessibility prediction servers for making their methods available online.

REFERENCES

1. Rost B. Protein secondary structure prediction continues to rise. J Struct Biol 2001;134:204–218.
2. Przybylski D, Rost B. Alignments grow, secondary structure prediction improves. Proteins 2002;46:197–205.
3. Venclovas C, Zemla A, Fidelis K, Moult J. Comparison of performance in successive CASP experiments. Proteins 2001;Suppl 5:163–170.
4. 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; Suppl 5:171–183.

5. Schonbrun J, Wedemeyer WJ, Baker D. Protein structure prediction in 2002. Curr Opin Struct Biol 2002;12:348–354.
6. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
7. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
8. Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 1994;19: 55–72.
9. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
10. Thompson MJ, Goldstein RA. Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. Proteins 1996;25:38–47.
11. Richardson CJ, Barlow DJ. The bottom line for prediction of residue solvent accessibility. Protein Eng 1999;12:1051–1054.
12. Cuff JA, Barton GJ. Application of enhanced multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 1999;40:502–511.
13. Ahmad S, Gromiha MM. NETASA: neural network based prediction of solvent accessibility. Bioinformatics 2002;18:819–824.
14. Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 2002;47:142–153.
15. Kim H, Park H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. Proteins 2004;54:557–562.
16. Naderi-Manesh H, Sadeghi M, Arab S, Moosavi Movahedi AA. Prediction of protein surface accessibility with information theory. Proteins 2001;42:452–459.
17. Ahmad S, Gromiha MM, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. Proteins 2003;50:629–635.
18. 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.
19. Kabsch W, Sander C. Dictionary of Protein Secondary Structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
20. Chothia C. The nature of accessible and buried surfaces in proteins. J Mol Biol 1976;105:1–14.
21. Matthews BW. Comparison of predicted and observed secondary structure of T4 ohage lysozyme. Biochim Biophys Acta 1975;405: 442–451.
22. 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.
23. 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.
24. 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://www.tc.cornell.edu/cbio/loopp
25. Meller J, Elber R. Linear optimization and a double statistical filter for protein threading protocols. Proteins 2001;45:241–261.
26. 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.
27. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31: 365–370.
28. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2003;31:23–27.
29. Adamczak R, Porollo A, Meller J. Improved secondary structure prediction with real value solvent accessibility approximation. Forthcoming.
30. Elman JL. Finding structure in time. Cognit Sci 1990;14:179–211.
31. Smith M. Neural networks for statistical modeling, Boston: International Thomson Computer Press; 1996.
32. Riedmiller M, Braun H. A direct adaptive method for faster backpropagation learning: the RPROP algorithm. Proc IEEE Int Conf Neural Networks 1993; pp. 123–134.
33. 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