PROTEINS: Structure, Function, and Genetics 47:142–153 (2002)

Prediction of Coordination Number and Relative Solvent Accessibility in Proteins

Gianluca Pollastri,[1] Pierre Baldi,[1,2*] Pietro Fariselli,[3] and Rita Casadio[3]
[1]*Department of Information and Computer Science, Institute for Genomics and Bioinformatics, University of California, Irvine, California*
[2]*Department of Biological Chemistry, College of Medicine, University of California, Irvine, California*
[3]*Department of Biology, CIRB Biocomputing Unit and Laboratory of Biophysics, University of Bologna, Bologna, Italy*

ABSTRACT **Knowing the coordination number and relative solvent accessibility of all the residues in a protein is crucial for deriving constraints useful in modeling protein folding and protein structure and in scoring remote homology searches. We develop ensembles of bidirectional recurrent neural network architectures to improve the state of the art in both contact and accessibility prediction, leveraging a large corpus of curated data together with evolutionary information. The ensembles are used to discriminate between two different states of residue contacts or relative solvent accessibility, higher or lower than a threshold determined by the average value of the residue distribution or the accessibility cutoff. For coordination numbers, the ensemble achieves performances ranging within 70.6–73.9% depending on the radius adopted to discriminate contacts (6Å–12Å). These performances represent gains of 16–20% over the baseline statistical predictor, always assigning an amino acid to the largest class, and are 4–7% better than any previous method. A combination of different radius predictors further improves performance. For accessibility thresholds in the relevant 15–30% range, the ensemble consistently achieves a performance above 77%, which is 10–16% above the baseline prediction and better than other existing predictors, by up to several percentage points. For both problems, we quantify the improvement due to evolutionary information in the form of PSI-BLAST-generated profiles over BLAST profiles. The prediction programs are implemented in the form of two web servers, CONpro and ACCpro, available at http://promoter.ics.uci.edu/BRNN-PRED/. Proteins 2002;47:142–153.**

Key words: protein structure prediction; protein contacts; contact map; contact number; recurrent neural networks; evolutionary information

INTRODUCTION

One approach toward predicting the structure of a protein is to predict a number of key attributes, in particular secondary structure, solvent accessibility, and coordination number. Deriving an accurate contact map from the primary sequence and these attributes is emerging as a promising strategy for solving the structure prediction problem.[1,2] For most of these attributes, machine learning methods in general, and more specifically neural network approaches, have proved particularly effective. For instance, the best secondary structure predictors today are neural network-based, with performance in the 75–80% range and these continue to improve.[2–4] In this work, we develop recurrent neural network methods for the improved prediction of coordination number and solvent accessibility.

Coordination Number

Knowing the correct positions of residue contacts in proteins has proved extremely useful in determining the three-dimensional (3D) structure of a given protein, as demonstrated in the CASP3 and CASP4 competitions [http://predictioncenter.llnl.gov/].[2,5] The number of stabilizing contacts that residues make in the protein-folded globule (see ref. 6, for a review) is a fundamental aspect of protein structure that is well worth predicting. In particular, this number can be used to "clean up" noisy contact map predictions, on the basis of primary sequence and secondary structure information. Furthermore, when a remote homology is searched, it benefits from deriving a surface potential from the distribution of contact numbers for each residue. This is computed by implementing an inverse of the Boltzmann rule,[7] or by using the notion of contacts among residues to improve existing threading algorithms.[8] In an off-lattice context, the number of contacts for each residue, or coordination number, is computed inside a spherical cutoff centered on each residue by counting the number of residues falling inside the sphere.[7]

During the past few years, researchers have made a number of attempts to predict contacts[9–11] and distances among residues in proteins,[12–14] with some degree of success. In ref. 15, a feed-forward neural network ap-

Grant sponsor: Laurel Wilkening Faculty Innovation; Grant sponsor: Sun Microsystems; Grant sponsor: Ministero della Università e della Ricerca Scientifica e Technologica (MURST); Italian Centro Nazionale delle Ricerche (CNR).

*Correspondence to: Pierre Baldi, Department of Information and Computer Science, Institute for Genomics and Bioinformatics, University of California, Irvine, CA 92697-3425. E-mail: Pfbaldi@ics.uci.edu

Received 6 July 2001; Accepted 16 November 2001

proach with a local window was developed to discriminate between two different states of residue contacts, characterized by a contact number higher or lower than the average value of the residue distribution. For a contact radius of 6.5 Å, this approach achieved a performance of 69% correct prediction, 12% above the level of the simple baseline classifier. By definition, the baseline classifier always selects the most frequent category for each amino acid independent of its environment.[16]

Solvent Accessibility

A second important feature of protein structural organization is the degree to which residues in the structure interact with the solvent molecules. Relative solvent accessibility classes are usually derived from the DSSP program[17] by normalizing it at the maximum value of exposed surface area obtainable for each residue. Different arbitrary threshold values of solvent accessibility are chosen to define binary categories (buried and exposed) or ternary categories (buried, partially exposed, or exposed).

Prediction of residue accessibility has been attempted with different methods based on neural networks with[18] or without[19] evolutionary information, Bayesian methods,[20] residue substitution matrices,[21] and information theory[22] (see also refs. 23–25). The baseline approach[26] classifies residues into a burial or nonburial category, using only the identity of the residue, independent of the surrounding context. Despite its simplicity, the baseline is as accurate as many previous more sophisticated methods. A comparison of all the available methods reported in ref. 16 showed that accuracy values level off ~69–71% when single protein sequences are used.

Recently, some groups[22,25,26] have revisited the problem of solvent accessibility prediction, using larger data sets. Naturally, the accuracy of the prediction depends on the threshold value of solvent accessibility. JNET[26] has been improved by means of new alignment methods, and the highest performance of 76.2% is achieved when the solvent accessibility threshold is 25%. Using approaches based on information theory and multiple sequence alignments, an accuracy of 71.5% with a threshold cutoff of 20% has been reported in ref. 25. A similar method, trained and scored using a program different from DSSP to compute solvent accessibility, has a reported accuracy of 74.4% when the relative solvent accessibility threshold is set at 25%.[22]

Based on the notion that less exposed residues are preferentially involved in hydrophobically driven chain compaction, solvent accessibility also has been routinely used to evaluate the number of residue contacts. To simulate the hydrophobic collapse in model proteins, the number of residue contacts is chosen as the inverse measure of residue solvent accessibility and, in the case of simple lattice protein models, is the only source of interaction.[27] In ref. 15, it was shown that, although a strong correlation between accessibility and contact number is commonly accepted, residue surface accessibility has a different distribution from the number of residue contacts, so that residue classification may be different, depending on which property is highlighted. This finding, which ought to be confirmed by a statistical correlation analysis, would support at least a partial separation between the problems of predicting coordination number and relative solvent accessibility.

Here we first extract a large curated data set of contact and accessibility information from the Protein Data Bank (PDB)[28] and generate a set of corresponding profiles using the BLAST,[29] and PSIBLAST[30] alignment/search programs. We compute detailed contact, accessibility, and secondary structure correlation statistics on this set and, in particular, examine the effect of the contact radius, ranging within 6–12 Å, as well as various accessibility thresholds. More importantly, we then develop a class of bidirectional recurrent neural network architectures, capable of partially capturing long-range information. In combination with the evolutionary profiles, these architectures are applied to the problem of predicting coordination number and relative solvent accessibility.

MATERIALS AND METHODS

Data Preparation

Coordination number

As is always the case in machine-learning approaches, the starting point is the construction of a well-curated data set. The data set used in this study was extracted from the PDB_select list[31] of June 2000. The list of structures and additional information can be obtained from the following ftp site: ftp://ftp.embl-heidelberg.de/pub/databases. To avoid biases, the set is redundancy-reduced, with an identity threshold based on the distance derived in ref. 32, which corresponds to a sequence identity of roughly 22% for long alignments, and higher for shorter ones. This set is further reduced by excluding those chains whose backbone is interrupted. We run Kabsch and Sander's DSSP program[17] on all the PDB files in the PDB-select list; excluding those on which DSSP crashed because of such factors as missing entries, erroneous entries, or format errors. The final set consists of 1,086 protein chains containing a total of 166,750 residues.

We compute the number of inter-residue contacts for each residue in the data set by defining a spherical protein volume centered on the C_{alpha} atom, with a given radius R Å, and counting the number of additional C_{alpha} atoms contained in the sphere. Thus, by this definition, a residue is in contact with its immediate primary sequence neighbors, but not with itself. For a given radius R, we compute the average number of contacts for each amino acid over the entire set (Table I). Each residue in a chain is then assigned to class 0 if the number of neighbors within the radius R is lower than the average, and to class 1 if higher than the average. The process was repeated for radii of 6, 8, 10, and 12 Å. For each radius, the range, average, and per amino acid distribution of the number of contacts are displayed in Table I and Figures 1 and 2.

In order to perform threefold cross-validation experiments, the data are then split evenly into three subsets, each containing 362 proteins (Table II). In all three subsets, the two classes are distributed almost evenly. Class 0 is slightly more numerous than class 1 for all four

TABLE I. Average and Range of Number of Contacts for Each Radius Across All Amino Acids

	6 Å	8 Å	10 Å	12 Å
Avg[a]	5.33	9.55	16.93	27.20
Min[b]	4.21 (*P*)	8.36 (*E*)	14.41 (*E*)	22.97 (*E*)
Max[c]	6.08 (*C*)	11.50 (*C*)	20.27 (*C*)	32.08 (*I*)

[a]Average over all amino acids.
[b]Lowest average number over all 20 amino acids with corresponding amino acid in parentheses.
[c]Highest average over all 20 amino acids with corresponding amino acid in parentheses.

radii, ranging from a minimum of 50.91 for 10 Å to a maximum of 52.12 for 8 Å over the total set. This effect is to be expected, as the possible contact values below the average have a more restricted range than the values above the average. The total number of amino acids in each cross-validation experiment is approximately 165,000: 110,000 used as a training set and 55,000 as a test set (Table III).

Solvent accessibility

For solvent accessibility, we use the same data set as for the number of contacts, except that 78 additional sequences have to be removed, leaving a total of 1,008 sequences. The removed sequences correspond to PDB files containing residues that are not completely resolved (i.e., with only C_{alpha} or C_{beta} atoms) or nonstandard amino acids. We build predictors for the two-state relative solvent accessibility. Accessibility values are computed again using the DSSP program. To predict the relative solvent accessibility RA(i) of each residue i, we calculate RA(i) = 100 * ACC(i)/MAXA(i), where ACC(i) is the solvent accessibility of residue i, as computed by the DSSP program (in Å^2), and MAXA(i) is the maximal accessibility of amino acid type i.[17]

For each relative accessibility percentage R, Figure 3 displays the percentage of amino acids that are more buried than R. As expected, most amino acids tend to be buried: roughly 50% are less than 25% exposed. Thus, when choosing a threshold for the classification, values around 25% are the most informative.

To perform threefold cross-validation experiments, the data are split in the same fashion as for coordination number, although in this case some sequences are missing. Twenty different classification schemes are extracted, from 0–95% exposure, with incremental steps of 5%. Table IV displays the number of amino acids for each classification threshold and for each of the three subsets, as well as for the entire set.

Profiles

It is well known that evolutionary information in the form of multiple alignments and profiles significantly improves the accuracy of, for instance, secondary structure prediction methods.[3,26,33–35] This is so because the secondary structure of a family is more conserved than the primary amino acid sequence. Similar effects have been reported for the prediction of contact number and relative solvent accessibility. For instance, in the case of contact number, an improvement of 3% has been reported in ref 15, using profiles over individual sequences. For relative solvent accessibility, a corresponding increase of 5% has been described both with neural networks[33] and Bayesian methods.[20] The work in ref. 3, has shown that, at least in the case of secondary structure, carefully generated PSI-BLAST profiles can give better results than BLAST profiles. We derive both BLAST and PSI-BLAST profiles and compare their effects on prediction performance.

BLAST

A first set of input profiles is constructed by running the BLAST program,[28] with standard default parameters (E = 10.0, BLOSUM62 matrix), against the nonredundant (NR) database. The version used was available online in October 1999 and contained approximately 420,000 protein sequences. For redundancy reduction, instead of using a hard threshold that requires an arbitrary choice, we use a graduated weighting scheme by assigning to each sequence a weight that measures how different the sequence is from the profile. Highly redundant sequences are assigned a lower weight. For any given sequence, the theoretical weight of information is given by the sum over all columns in the profile of the Kullback–Liebler distance between the delta distribution associated with the composition of the sequence in the column and the corresponding profile distribution.[4] Formally, the weight of sequence s is then

$$W(s) = -\sum_{c} logP[s(c)] \quad (1)$$

where $P[s(c)]$ is the probability of letter s in profile column c. In summary, every sequence in a given alignment is assigned a weight proportional to the Shannon information the sequence carries with respect to the unweighted profile. A weighted profile matrix is then compiled and used as input for the system (see also ref. 35).

PSI-BLAST

A second set of profiles is generated by PSI-BLAST.[29] All proteins are aligned against the NR database. Alignments are generated by the following four-step protocol.[36] First, filter and remove all database sequences with COILS to mark coiled-coil regions[37] and SEG to mark regions of low complexity.[38] Second, align the query protein against this filtered database with an E-value threshold for the iteration of 10^{-10} (PSI-BLAST h threshold) and a final threshold of $E \leq 10^{-3}$ to accept hits. The number of iterations is restricted to three to avoid drift.[3,36] Third, align the query against the unfiltered SWISS-PROT + TrEMBL + PDB using the previously found, position-specific profile. Finally, use the same weighting scheme as in the case of BLAST profiles to balance the profile and remove redundancy.

Recurrent Neural Network Architectures

Feed-forward neural networks have been one of the major machine-learning tools used in protein structure

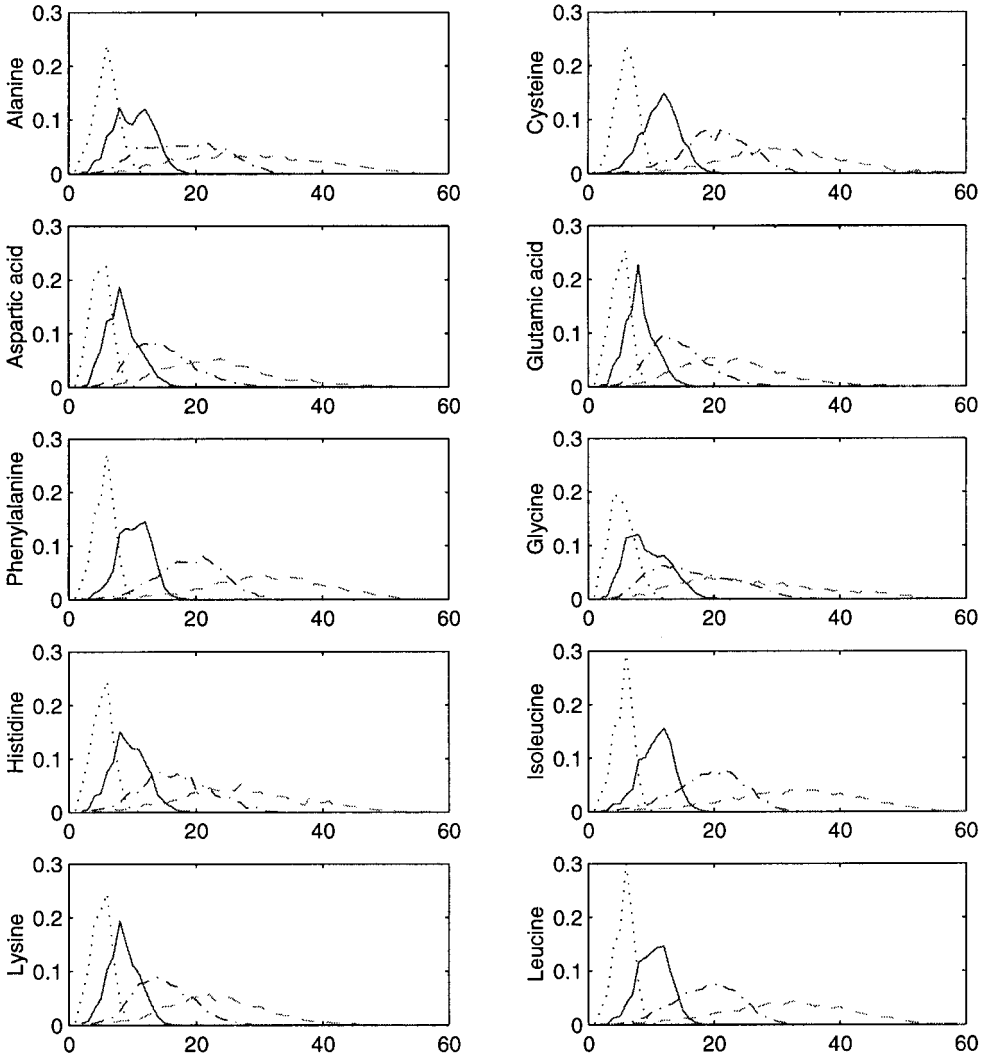

Fig. 1. Distribution of number of contacts for amino acids A–L in alphabetical order: dotted blue, 6 Å; solid green, 8 Å; dashdot red, 10 Å; dash light blue, 12 Å. *x*-axis = number of contacts, *y*-axis = probability.

prediction problems that range from the prediction of secondary structure to the number of contacts. The major weakness of feed-forward neural networks, however, is the use of a local input window of fixed size, which cannot provide any access to long-range information. Networks for contact prediction, for instance, have windows of size 1–15. Larger windows usually do not work, in part because the corresponding increase in the number of parameters leads to overfitting. Increase in the number of parameters, however, is not necessarily the main obstacle per se because data are becoming abundant and techniques such as weight sharing can be used to mitigate the risk of overfitting. The main problem is that long-range signals are very weak compared to the additional "noise" introduced by a larger window. Thus, larger windows tend to dilute sparse information present in the input that is relevant for the prediction.

The methods we use are designed to attempt to overcome the limitations of simple feed-forward networks have been described in refs. 35, 39, and 40 and consist of bidirectional recurrent neural networks (BRNNs). Letting t denote position within a protein sequence, the overall model for binary classification outputs for each t, a number $O_t(0 \leq O_t \leq 1)$ representing the membership probability of the residue at position t in the class. In the coordination or accessibility prediction applications, the output consists of a single logistic output unit that estimates the probability that the coordination number (respectively, solvent accessibility) is higher or lower than the average, or accessibility, cutoff in the center of the corresponding input window.

The output prediction has the functional form

$$O_t = \eta(F_t, B_t, I_t) \tag{2}$$

and depends on the forward (upstream) context F_t, the backward (downstream context) B_t, and the input I_t at time t. The vector $I_t \in IR^k$ encodes the external input at time t. In the most simple case, where the input is limited to a single amino acid, $k = 20$ by using orthogonal encoding. Larger input windows extending over several

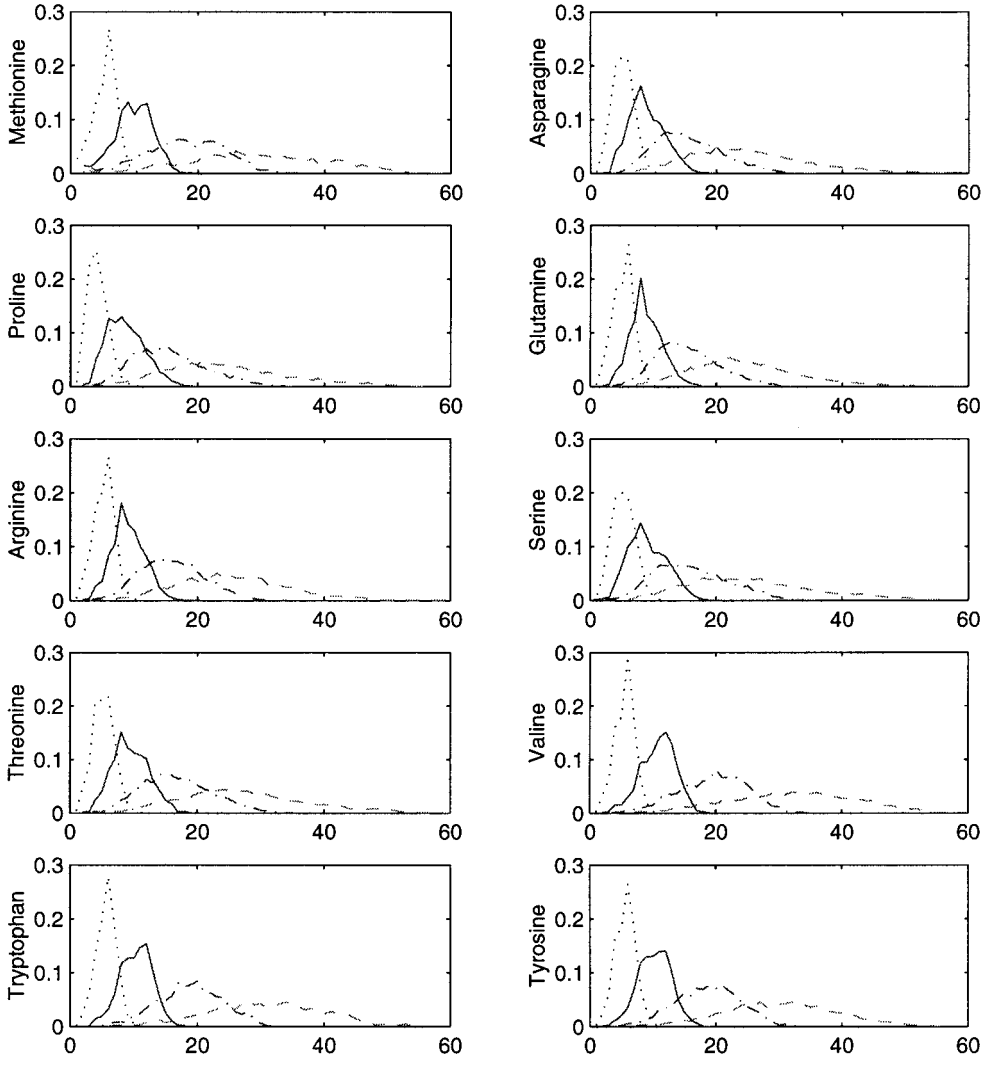

Fig. 2. Distribution of number of contacts for amino acids *M–Y* in alphabetical order: dotted blue, 6 Å; solid green, 8 Å; dashdot red, 10 Å; dash light blue, 12 Å. *x*-axis = number of contacts, *y*-axis = probability.

TABLE II. Threefold Cross-validation Subset Statistics With Number of Amino Acids in Each Class

	Class	6 Å	8 Å	10 Å	12 Å
Total set	0	85119	86906	84886	86401
n166750 AA	1	81631	79844	81864	80349
Subset 1	0	28415	29344	28675	29357
55859 AA	1	27444	26515	27184	26502
Subset 2	0	28072	28008	27430	27860
54355 AA	1	26283	26347	26925	26495
Subset 3	0	28632	29554	28781	29184
56536 AA	1	27904	26982	27755	27352

TABLE III. Typical Training Set Statistics Taken from the First Set

Sets	Class	6 Å	8 Å	10 Å	12 Å
Train	0	56704	57562	56211	57044
Train	1	54187	53329	54680	53847
Test	0	28415	29344	28675	29357
Test	1	27444	26515	27184	26502

amino acids are also possible. The function η is realized by a neural network $\mathcal{N}_\eta$ (see center and top connections in Fig. 4). The performance of the model can be assessed using the relative entropy between the estimated and the target distribution.

The novelty of the model is in the contextual information contained in the vectors $F_t \in IR^n$ and especially in $B_t \in IR^m$. These satisfy the recurrent bidirectional equations:

$$F_t = \phi(\mathrm{F}_{t-1},\mathrm{I}_t)$$
$$B_t = \beta(\mathrm{B}_{t+1},\mathrm{I}_t) \quad (3)$$

where $\phi(\cdot)$ and $\beta(\cdot)$ are learnable nonlinear state transition functions, implemented by two *NN*s, $\mathcal{N}_\phi$, and $\mathcal{N}_\beta$ (left and right subnetworks in Fig. 4). The boundary conditions

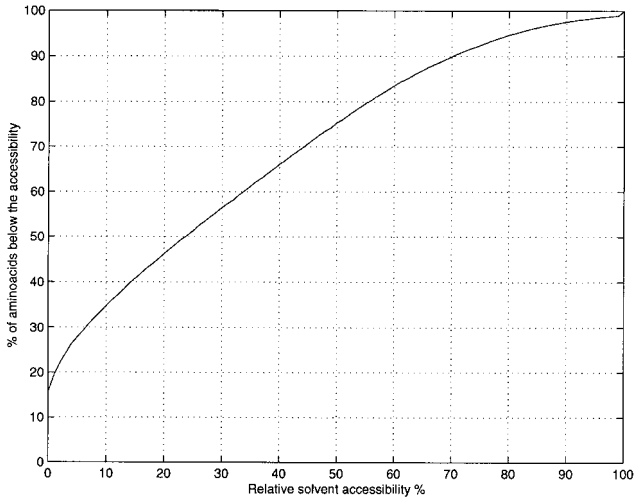

Fig. 3. Solvent accessibility distribution.

for F_t and B_t are set to 0, i.e., $F_0 = B_{T+1} = 0$, where T is the length of the protein being examined. Intuitively, we can think of F_t and B_t as "wheels" that can be rolled along the protein. To predict the class at position t, we roll the wheels in opposite directions from the N- and C-terminus up to position t and then combine what is read on the wheels with I_t to calculate the proper output using η.

All the weights of the BRNN architecture, including the weights in the recurrent wheels, can be trained in a supervised fashion from examples by a generalized form of gradient descent or backpropagation through time, by unfolding the wheels in time, or rather space. Architectural variants can be obtained by changing the size of the input windows, the size of the window of hidden states considered to determine the output, the number of hidden layers, the number of hidden units in each layer and so forth. The following discussion uses several notations:

Ct = size of semi-window of context states considered by the output network
NFB = number of output units in the left (forward) and right (backward) context networks (wheels)
NHO = number of hidden units in the output network
NHC = number of hidden units in the context networks

The BRNN networks are trained by back-propagation on the relative entropy error between the output and target probability distributions. In a typical case, we use a hybrid between online and batch training, with 300 batch blocks (2–3 proteins each) per training set. Thus, weights are updated 300 times per epoch after each block. The learning rate per block is initially set at about 2.7×10^{-4}, corresponding to the number of blocks divided by (10) times the number of residues ($0.1 \times 300/110{,}000$), and is progressively decreased. The training set is also shuffled at each epoch, so that the error is not decreasing monotonically. There is no momentum term or weight decay. When the error does not decrease for 50 consecutive epochs, the learning rate is divided by 2, and training is restarted from the lowest error model. Training stops after 8 or more reductions, corresponding to a learning rate that is 256 times smaller than the initial one, which usually happens after 1,500–2,500 epochs.

BRNNs have been used for secondary structure prediction and to develop the SSpro web server [http://promoter.ics.uci.edu/BRNN-PRED/]. They have also been used for the prediction of amino acid partners in β-sheets.[42] In ref. 35, evidence is provided that these architectures extend the range over which information can be effectively captured with respect to feed-forward neural networks, up to an effective window size of about 30 amino acids in the case of secondary structure prediction.

RESULTS AND DISCUSSION

Correlations

We used several different encoding schemes to compute correlations between different structural features:

1. Real numbers (i.e., relative accessibility, and the number of contacts for each cutoff).
2. Two states (1 and 0) for each descriptor; for instance,
   a. 1 if number of contacts is greater than average
   b. 0 otherwise
   or
   c. 1 if relative accessibility is greater than 16%
   d. 0 otherwise
3. Three states (−1 0 1) for each descriptor; for instance,
   a. 1 if contact is greater than average plus 1
   b. 1 if contact smaller than average minus 1
   c. 0 otherwise
   or
   d. 1 if relative accessibility >50%
   e. 1 if relative accessibility <9%
   f. 0 otherwise
4. Secondary structure
   a. 1 if residue in H, 0 otherwise (H)
   b. 1 if residue in E, 0 otherwise (E)
   c. 1 if residue in H, −1 if residue in E, 0 otherwise (HE)

The correlations between the contact numbers in the four different radius categories are shown in Table V, together with the correlations between contact numbers and relative solvent accessibility. As expected, correlations between contact numbers are high, especially between 8, 10, and 12 Å categories, while the 6-Å category is less correlated to the others. The correlation between the 6 Å and 12 Å numbers is only 0.46. Likewise, correlations between contact numbers and accessibility exist but are negative and far from perfect. They tend to decrease with smaller radius values: the correlation is −0.71 at 12 Å, but only −0.52 at 6 Å.

Similar results are obtained with two- and three-state correlation values (not shown). These results confirm that the 6.0-Å coordination cutoff captures a different picture of the local environment with respect to all other cutoffs. This suggests that the behavior at 6 Å is biased by the sequence neighboring contacts (helix or turns), while larger cutoffs also involve contacts with residues that are linearly distant along the primary sequence (e.g., β-structures).

TABLE IV. Numbers of Amino Acids in Each Accessibility Class for All 20 Thresholds, for Each of the Three Test Sets, and for the Total Set

Threshold	Set0-cl0	Set0-cl1	Set1-cl0	Set1-cl1	Set2-cl0	Set2-cl1	All-cl0	All-cl1
0	7935	43677	7648	42186	8367	43342	23950	129205
5	14169	37443	13486	36348	14906	36803	42561	110594
10	17843	33769	16925	32909	18537	33172	53305	99850
15	20996	30616	19882	29952	21586	30123	62464	90691
20	23814	27798	22587	27247	24389	27320	70790	82365
25	26450	25162	25093	24741	26987	24722	78530	74625
30	29072	22540	27701	22133	29685	22024	86458	66697
35	31652	19960	30100	19734	32084	19625	93836	59319
40	34155	17457	32502	17332	34544	17165	101201	51954
45	36565	15047	34860	14974	36909	14800	108334	44821
50	38958	12654	37200	12634	39199	12510	115357	37798
55	41176	10436	39455	10379	41338	10371	121969	31186
60	43188	8424	41461	8373	43357	8352	128006	25149
65	44918	6694	43204	6630	45070	6639	133192	19963
70	46442	5170	44672	5162	46587	5122	137701	15454
75	47751	3861	45951	3883	47888	3821	141590	11565
80	48900	2712	47084	2750	49076	2633	145060	8095
85	49729	1883	47933	1901	49909	1800	147571	5584
90	50408	1204	48562	1272	50528	1181	149498	3657
95	50848	764	48992	842	50949	760	150789	2366

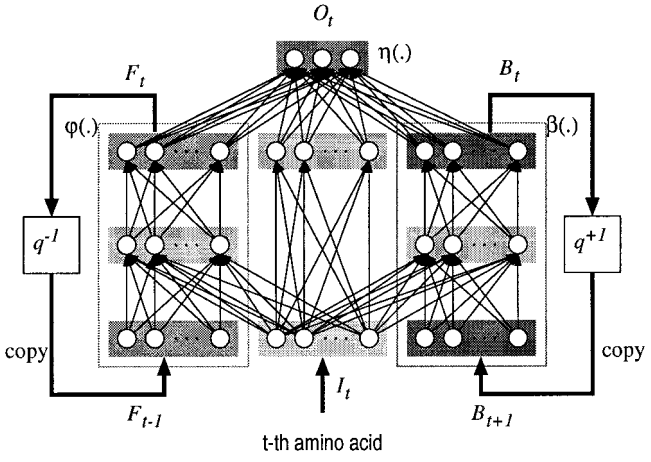

Fig. 4. BRNN architecture with a left (forward) and right (backward) context associated with two recurrent networks (wheels).

TABLE V. Correlations Between Contact Numbers for Different Radius Values and Between Contact Numbers and Relative Solvent Accessibility

	6 Å	8 Å	10 Å	12 Å	−0.52
6 Å	1.0	0.63	0.53	0.46	−0.52
8 Å		1.0	0.86	0.79	−0.70
10 Å			1.0	0.92	−0.72
12 Å				1.0	−0.71
ACC					1.0

TABLE VI. Simple, Two-State (2), and Three-State (3) Correlations Between Number of Contacts for Different Radius Values and Relative Accessibility With Secondary Structure Classes *H*, *E*, and *HE**

	H	*E*	*HE*
ACC	−0.10	−0.24	0.07
ACC (2)	−0.07	−0.20	0.07
ACC (3)	−0.08	−0.24	0.08
6 Å	0.42	0.07	0.23
6 Å (2)	0.40	0.01	0.25
6 Å (3)	0.32	0.02	0.20
8 Å	0.06	0.33	−0.15
8 Å (2)	0.05	0.24	−0.10
8 Å (3)	0.04	0.29	−0.14
10 Å	0.04	0.30	−0.14
10 Å(2)	0.02	0.24	−0.12
10 Å (3)	0.02	0.26	−0.13
12 AÅ	0.05	0.32	−0.15
12 Å(2)	0.05	0.24	−0.11
12 Å (3)	0.05	0.25	−0.11

*See text for details.

This hypothesis is supported by the correlations between accessibility or contact number and secondary structure, as reported in Table VI. Overall these correlations are quite weak, but the correlation between residue contact number and helical structure at 6 Å (0.42) is higher compared with all other correlations. In contrast, residue contact number within the 8–12 Å range is far more correlated with extended (*E*) structures than helical structures. This provides further evidence that larger cutoffs are more suitable to capture contacts associated with large sequence separations. Note also that overall the correlations decrease when going from real-valued to two-state encoding. This indicates that even though a two-state (or three-state) classification is quite useful in real applications, the threshold definitions are, of course, somewhat arbitrary.

Prediction of Coordination Number

Preliminary tests were conducted with a number of different BRNN architectures. We finally focused on seven BRNNs with the same structure as those used in the early

TABLE VII. Total Number of Weights and Size Parameters of the Seven BRNN Models*

Model No.	Weights	Ct[a]	NFB[b]	NHO[c]	NHC[d]
0	2241	3	8	11	9
1	1959	2	9	11	8
2	3009	3	12	11	9
3	2615	3	12	9	9
4	4232	3	15	12	13
5	4896	3	17	12	15
6	5430	3	17	14	15

[a]Size of semi-window of context states considered by the output network.
[b]Number of output units in the left and right context networks.
[c]Number of hidden units in the output network.
[d]Number of hidden units in the context networks.

TABLE VIII. Percentage of Threefold Cross-validation Results Obtained With Several BRNNs and the Corresponding Ensemble on the Test Set, for PSI-BLAST-Based Input Profiles*

Model	6 Å	8 Å	10 Å	12 Å
0	71.59	69.29	71.04	73.00
1	72.03	69.45	70.96	72.42
2	71.04	68.91	70.58	72.71
3	71.39	69.28	70.84	72.68
4	69.99	67.80	69.79	72.54
5	69.77	67.72	69.54	71.93
6	69.95	67.49	70.16	71.69
Ens	73.02	70.57	72.00	73.93
Comb	73.24	70.95	72.13	74.09
Filter	73.13	70.56	72.02	73.92

*Performance results expressed in percentages of correct prediction ($Q2$).
[a]Ensemble of models in a given radius category.
[b]Combination of four ensembles associated with different radius categories.
[c]Plain filter applied to four ensembles associated with different radius categories.

TABLE IX. Threefold Cross-validation Results Obtained With the Four Ensembles*

		6 Å	8 Å	10 Å	12 Å
$Q2$[a]	PSI	73.02	70.57	72.00	73.93
	BLAST	72.54	70.09	71.20	73.03
	Diff	0.48	0.48	0.80	0.90
Corr[b]	PSI	0.462	0.410	0.440	0.478
	BLAST	0.452	0.400	0.424	0.460
	Diff	0.010	0.010	0.016	0.017

*Gains of the PSI-BLAST-based ensembles over the BLAST-based ensembles.
[a]Percentage of correctly assigned residues.
[b]Matthews' correlation coefficients.
[c]PSI-BLAST-based profiles.
[d]BLAST-based profiles.
[e]Difference PSI-BLAST.

TABLE X. Comparison With Baseline Predictor*

Q2[a]	6 Å	8 Å	10 Å	12 Å
Ens[b]	73.02	70.57	72.00	73.93
Comb[c]	73.24	70.95	72.13	74.09
Base[d]	57.01	54.11	52.86	52.66
Diff[e]	16.23	16.84	19.27	21.43

*PSI-BLAST-based profiles.
[a]Percentage measure.
[b]Ensemble of models in a given radius category.
[c]Combination of ensembles across categories.
[d]Baseline predictor that selects the largest class for each amino acid.
[e]Difference in $Q2$ between Comb and Base.

version of the SSpro software[35] for protein secondary structure prediction. The basic parameters of each architecture are given in Table VII. The number of parameters in each architecture ranges within 1,959–5,430. We used a network of 16 Sun Microsystems UltraSparc workstations for training and testing, roughly equivalent to 2 years of a single CPU, excluding the preliminary experiments. The seven architectures are combined by simple averaging of the outputs into an ensemble predictor. For a given radius category, each ensemble is the average of 21 predictors (7 networks $\times$ 3 cross-validation subsets).

Several indices can be used to score the efficiency[41] of the algorithm. Here, we use $Q2$, the number of correctly predicted residues divided by the total number of residues, and the Matthews' correlation coefficient. The results of threefold cross-validation, corresponding to $3 \times 4 \times 7 = 84$ tests, for each one of the 7 BRNNs and for the ensemble, are summarized in Table VIII for the test sets.

Overall, compact models tend to show better performance. Larger models perform worse because they overfit the training set. The effect of overfitting is considerable in the 6-Å and 8-Å categories, moderate for 10 Å and 12 Å. Although large models sometimes have significantly poorer performance, they still prove useful when combined in an ensemble. In each radius category, the ensemble represents a sizeable improvement over each individual architecture and performs considerably better than the simple baseline predictor that always assigns a residue to its most abundant class independent of its environment.[16] The gains over the baseline predictor range from 16.0% for the 6-Å ensemble to 21.3% for the 12-Å ensemble (Tables X and XI). Note that the error bar on the performance estimates at the level of single amino acids is 0.11%. The best previously known predictor,[15] trained and tested only on a 6.5-Å radius data set, achieved a performance of 69%, 12% better than the corresponding baseline predictor. In the 6-Å category, closest to the one used in ref. 15, the ensemble of BRNNs trained using PSI-BLAST profiles achieves a $Q2$ of 73.02%, a gain of more than 4%. At 12 Å, the ensemble of BRNNs achieves a performance of 73.93% correct prediction, with a correlation coefficient of 0.48. The ensemble of BRNNs trained on BLAST profiles show slightly poorer performances. Table IX shows how the PSI-BLAST profiles are responsible for $Q2$ gains of 0.5–0.9%. The error bar on the performance estimates at the level of single amino acids is 0.11%.

At least two reasons should be considered to explain performance differences across the four radius categories.

TABLE XI. Comparison With Baseline Predictor*

Corr	6 Å	8 Å	10 Å	12 Å
Ens	0.462	0.410	0.440	0.478
Comb	0.467	0.419	0.443	0.482
Base	0.195	0.119	0.063	0.051

*Same as Table X, but using correlation measure.

First, the performance of the baseline predictors decreases with radius size. This particularly affects the 6-Å predictor, whose base level is 3% higher than the others. Second, as the radius is increased, the total length of the chain becomes increasingly relevant. The average number of contacts in the 12-Å data set is comparable to the length of short proteins, making it less likely or even impossible sometimes to have residues belonging to class 1. For example, isoleucine requires 33 contacts to be classified as 1, which is, of course, impossible in proteins shorter than 34 residues, and unlikely for proteins that are just slightly longer.

It is natural to wonder whether performance could be further improved by combining predictors across the four radius categories. Thus, we can combine the previous ensembles using a small BRNN (a small feed-forward neural network gives similar results) with parameters Ct = 2, NFB = 3, NHO = 4, and NHC = 3. To avoid retraining on the same training set, we perform a twofold cross-validation on each of the three subsets of the previous cross-validation. The results (Comb) are reported in Tables VIII and X. Each number is the average of six different values, since each of the three subsets of the previous cross-validation experiment is split into two and the two resulting subsets are used alternatively as test and training sets in this experiment, yielding a total of $6 \times 4 = 24$ numbers. The improvements obtained by pooling different radius categories range from 0.13% for 10 Å to 0.38% for the 8-Å category.

To make sure that these improvements are attributable to the combination of diverse information and *not* to a filtering effect associated with the additional BRNN used in the combination, we also test the same BRNN architecture as a filter for each single-category predictor (Filter in Table VIII). The latter simple output filtering approach gives results that are extremely similar to the unfiltered case, with differences in the -0.01 or $+0.02$ percentage range, except for the 6-Å category, where a small improvement of 0.11% is observed. Thus, the small but significant improvements observed with Comb can be imputed to the combination of different information associated with the 6-Å, 8-Å, 10-Å, and 12-Å categories.

Prediction of Relative Solvent Accessibility

As in the contact case, it is possible to define a baseline statistical predictor that assigns an amino acid to the largest class for the given amino acid.[16] We do so in a threefold cross-validation context; i.e., the largest class for a given amino acid is determined on the training set and is not always the largest on the test set. This cannot be avoided, but it has very little or no impact. The results are displayed in Table XII.

TABLE XII. Percentage Performance of Baseline Accessibility Predictors for All Thresholds, Threefold Cross-validation*

Threshold	Base0	Base1	Base2	BaseAll
0	84.63	84.65	83.82	84.37
5	72.55	72.94	71.17	72.22
10	65.43	66.04	64.15	65.21
15	62.12	62.91	61.33	62.12
20	65.68	66.75	65.76	66.06
25	65.35	67.96	67.33	66.88
30	65.54	66.30	65.75	65.87
35	67.54	68.08	67.93	67.85
40	68.56	68.46	69.20	68.74
45	71.30	70.98	72.32	71.53
50	75.48	74.65	75.81	75.31
55	79.78	79.17	79.94	79.63
60	83.68	83.20	83.85	83.57
65	87.03	86.70	87.16	86.96
70	89.98	89.64	90.09	89.91
75	92.52	92.21	92.61	92.45
80	94.75	94.48	94.91	94.71
85	96.35	96.19	96.52	96.35
90	97.67	97.45	97.72	97.61
95	98.52	98.31	98.53	98.45

*BaseN, baseline results on the *N*th test set; BaseAll, average of baseline results.

The threefold cross-validation was carried using the same seven BRNN architectures used for the number of contacts (Table VII). Results of the threefold cross-validation for all models and thresholds (test sets), using PSI-BLAST-based input profiles, are summarized in Table XIII.

Performance of the ensemble on both training and test sets is displayed in Figure 5, together with the baseline prediction. For thresholds within the range of 15–30% exposed, neither class covers more than 60% of the set, and therefore the classification problem is more balanced, hence harder. It is in this balanced region that the ensemble outperforms the baseline predictor by more than 10%. For an exposure threshold of 25%, the two classes are almost perfectly balanced. In this case, we achieve 77.2% correct classification. In the balanced region, error bars are again of the order of 0.1% at the single amino acid level. The best improvement with respect to the baseline prediction is 16.2%, achieved for an exposure threshold of 15%. As in the case of the coordination number, PSI-BLAST profiles prove useful for the prediction of relative solvent accessibility. Table XIV shows how 0.6–0.8% gains over the BLAST-based ensemble are observed in the 15–30% threshold region.

The current ensemble outperforms other recently published solvent accessibility approaches. With a threshold of 20%, Li and Pan[25] achieved a performance of 71.5% using single sequences, claiming that multiple align ments in this case are less useful than in secondary structure prediction. Although there may be some truth to that claim, we still find profiles useful. For the same threshold,

TABLE XIII. Threefold Cross-validation Results for Each of the Seven BRNN Models and for the Ensemble, in the case of PSI-BLAST Profiles

Model	0	1	2	3	4	5	6	Ens[a]
0	86.12	86.08	86.18	86.22	86.14	86.21	86.12	86.49
5	80.63	80.59	80.68	80.74	80.59	80.98	80.57	81.20
10	78.62	78.66	78.72	78.71	78.60	78.88	78.46	79.26
15	77.51	77.61	77.65	77.53	77.41	77.76	77.33	78.34
20	76.80	76.83	76.85	76.80	76.72	77.04	76.52	77.49
25	76.47	76.49	76.53	76.57	76.27	76.64	76.23	77.18
30	76.29	76.30	76.32	76.33	76.11	76.58	76.03	77.01
35	76.27	76.35	76.34	76.35	76.24	76.63	76.17	77.03
40	76.80	76.86	76.78	76.83	76.71	77.01	76.70	77.53
45	77.85	77.84	77.83	77.83	77.77	78.04	77.75	78.44
50	79.62	79.55	79.58	79.62	79.54	79.72	79.53	80.10
55	82.09	81.93	82.05	82.14	82.07	82.07	82.10	81.92
60	84.88	84.81	84.87	84.90	84.86	84.91	84.79	84.42
65	87.71	87.65	87.69	87.73	87.68	87.70	87.72	87.80
70	90.40	90.40	90.39	90.43	90.38	90.42	90.34	90.45
75	92.80	92.82	92.81	92.82	92.77	92.83	92.77	92.85
80	94.97	94.99	94.98	94.98	94.95	94.98	94.97	95.02
85	96.52	96.52	96.52	96.52	96.51	96.50	96.50	96.27
90	97.69	97.66	97.66	97.65	97.69	97.65	97.67	97.66
95	98.46	98.46	98.46	98.46	98.46	98.45	98.45	98.45

[a]Ensemble of the 7 models.

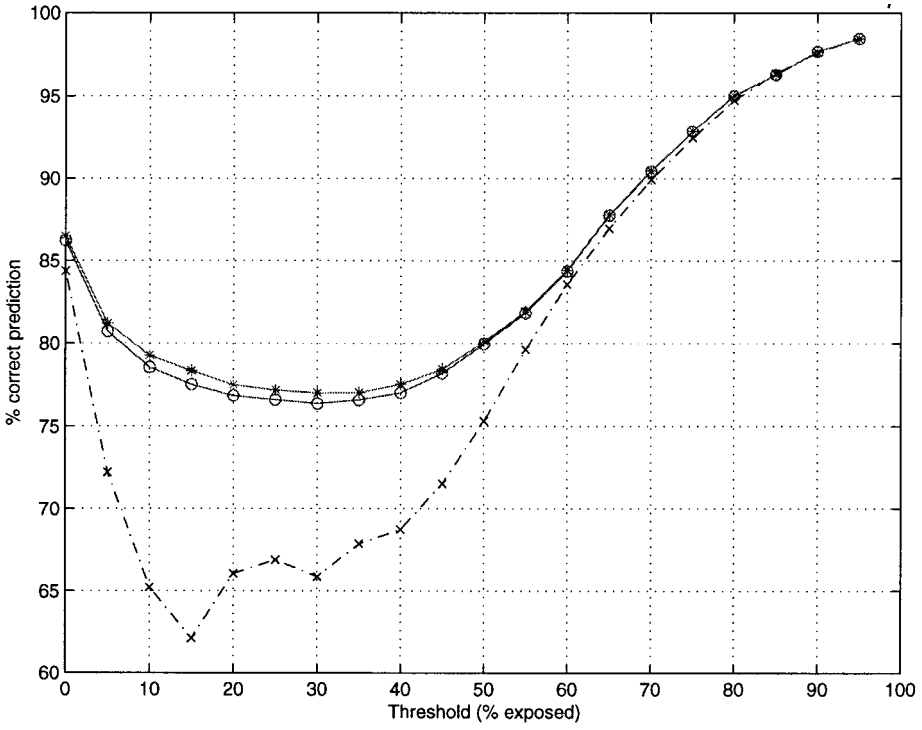

Fig. 5. Ensemble solvent accessibility prediction. Baseline predictor (blue crosses). BLASTbased ensemble (red circles), PSI-BLAST-based ensemble (magenta stars). There are 20 different thresholds.

our PSI-BLAST-based ensemble achieves 77.5% correct prediction. Another recent prediction server[23] claims 70.7% correct prediction at a 25% threshold, versus the 77.2% of our ensemble for the same threshold. The less recent PHDacc server[17] claims 74% with a threshold of 16%. For comparison, at a 15% threshold, the baseline method performs the worst, while our system achieves 78.3% accuracy. Closest to our performance is perhaps the system described in ref. 26, which achieves 76.2% at 25%, where we achieve 77.2%, 1% better. Thus, to the best of our knowledge, this is the top performance achieved so far by any method, although a completely fair comparison would require comparing all methods on exactly the same data. Such comparison may become possible in the near future

TABLE XIV. Threefold Cross-validation Results for Different Input Profiles, Compared With the Baseline Predictor.

	Base[a]	BLAST[b]	PSI[c]	PSI−Base	PSI−BLAST
0	84.37	86.20	86.49	2.12	0.29
5	72.22	80.71	81.20	8.98	0.50
10	65.21	78.55	79.26	14.05	0.71
15	62.12	77.51	78.34	16.22	0.82
20	66.06	76.85	77.49	11.42	0.64
25	66.88	76.59	77.18	10.30	0.59
30	65.87	76.36	77.01	11.15	0.65
35	67.85	76.58	77.03	9.18	0.45
40	68.74	77.03	77.53	8.79	0.50
45	71.53	78.10	78.44	6.00	0.24
50	75.31	79.96	80.10	4.79	0.15
55	79.63	81.84	81.92	2.28	0.08
60	83.57	84.37	84.42	0.84	0.05
65	86.96	87.76	87.80	0.83	0.04
70	89.91	90.43	90.45	0.54	0.01
75	92.45	92.84	92.85	0.40	0.01
80	94.71	94.99	95.02	0.30	0.03
85	96.35	96.27	96.27	−0.08	0.01
90	97.61	97.67	97.66	0.05	−0.01
95	98.45	98.46	98.45	0.00	0.00

[a]Baseline predictor.
[b]BLAST-based profiles.
[c]PSI-BLAST-based profiles.

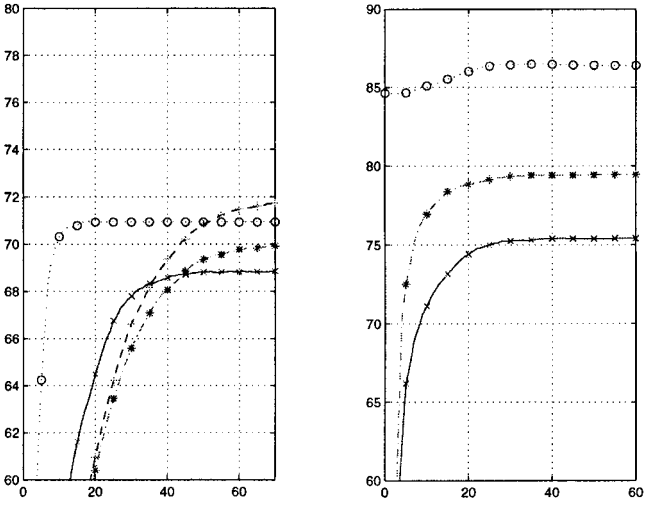

Fig. 6. Distant information exploited by the BRNN, corresponding to model 3 and to validation subset 1. The horizontal axis represents half the window size, i.e., the distance τ from a given position beyond which all entries are set to null values. The vertical axis represents the percentage of correct prediction. Left plot represents contact predictions at 6 Å (dotted line/balls), 8Å (solid line/*x*), 10 Å (dash-dot line/stars), and 12 Å (dashed line/+). Right plot represents accessibility predictions with thresholds 0 (dotted line/balls), 25 (solid line/*x*), and 50 (dash-dot line/stars).

through an automated web server similar to the EVA server (http://cubic.bioc.columbia.edu/eva/).

Long-Range Effects

We believe our improvements are due to both an increase in the size of the training sets and in the architectures we have developed, in particular their ability to capture long-range interactions that are beyond the reach of conventional feed-forward neural networks, with their relatively small and fixed input window sizes. In order to test the capabilities of our models to capture long-range information, we looked at the performance of a typical BRNN architecture (in case 3) when fed with a sequence where all inputs are replaced by 0 outside the range $[t - \tau, t + \tau]$, as in ref. 35. The experiment was repeated for different values of τ from 0 to 70, for both contact and accessibility (Fig. 6). For contacts in the 6-, 8-, 10-, and 12-Å categories, 0.1 below optimal performance is achieved for $\tau = 20$, 45, 62, and 75, corresponding to window sizes of 41, 91, 125, and 151 residues), respectively, in the 6-, 8-, 10-, and 12-Å categories. The signal of the protein terminus is in fact propagated beyond 70 amino acids in the 12-Å system by the BRNN architectures. This signal implicitly provides a sense of protein size during the classification process.

For accessibility, only minor changes are observed beyond $\tau = 30$, i.e., a window size of 61 residues. For instance, for $\tau = 35$ and an accessibility threshold of 25%, the performance of the model trained with incomplete data achieves a prediction only 0.1% below the performance of the same model trained with complete data.

A reasonable interpretation of these results is that the BRNN architectures can leverage information in a window of ≤60–70 residues in both kinds of prediction. In the case of contact prediction, however, the size of the protein is particularly important as well, so the BRNNs learn how to measure it by the distance to the N- and C-terminus. Obviously, this effect is increasingly important with the increase in the size of the sphere used to determine contacts, i.e., the increase in the number of neighbors needed to be above the average.

CONCLUSIONS

We have combined recursive neural network techniques and profiles to improve the state-of-the-art prediction of contact number and relative solvent accessibility prediction. The predictors achieve performances within the 71–74% range for contact numbers, depending on radius, and greater than 77% for accessibility in the most interesting range. In both cases, we have found evidence that more sensitive PSI-BLAST profiles provide a small but sizeable improvement over BLAST profiles. We have also collected contact and accessibility statistics and studied the effects of contact radius and relative accessibility threshold on prediction.

The predictors are implemented in the form of two Internet servers, CON-pro for contact number and ACCpro for relative solvent accessibility, accessible at http://promoter.ics.uci.edu/BRNN-PRED/. For coordination numbers, predictions are returned for 6, 8, 10, and 12 Å. For solvent accessibility, users can select the threshold, 25% being the default value. Predictions are emailed back to the users after a brief period, depending on server load.

ACCpro and CONpro are part of a broader suite of programs aimed at predicting protein 3D structure via contact map prediction, and contact map prediction via prediction of structural features. Prediction of structural

features, such as accessibility, can also be used as a filter for other tasks, for instance the study of contact sites involved in protein–protein interactions. Although perfect prediction of structural features should not be expected for a variety of reasons, including the fact that some proteins do not fold spontaneously, it is encouraging to see performance in this area improve year after year as a result of data expansion and algorithmic improvements. The performance levels now achieved by these methods, coupled with their speed, allows one to use them to sift through large sets of proteins and, for instance, considerably narrow down the number of targets that need to be tested by much more time-consuming computer or experimental methods.

ACKNOWLEDGMENTS

The work of P.B. and G.P. is supported by a Laurel Wilkening Faculty Innovation award and by a Sun Microsystems award to P.B. at UCI. The work of PF and RC is supported by a grant from the Ministero della Università e della Ricerca Scientifica e Tecnologica (MURST) for the project Structural Functional and Applicative Prospects of Proteins from Termophiles, and by a grant for a target project in Biotechnology of the Italian Centro Nazionale delle Ricerche (CNR).

REFERENCES

1. Baldi P, Pollastri G. Machine learning structural and functional proteomics. IEEE Intelligent Systems. Special Issue on Intelligent Systems in Biology, 2001 (in press).
2. Lesk AM, Lo Conte L, Hubbard TJP. Assessment of novel fold targets in CASP4: predictions of three-dimensional structures, secondary structures, function and genetics. Proteins 2001. (submitted).
3. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
4. Baldi P, Brunak S. Bioinformatics: the machine learning approach. 2nd ed. Cambridge, MA: MIT Press; 2001.
5. Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J. Ab initio folding of proteins using restraints derived from evolutionary information. Proteins 1999;3(suppl):177–185.
6. Dill K. Polymer principles and protein folding. Protein Sci 1999;8: 1166–1180.
7. Flokner H, Braxenthaler M, Lackner P, Jaitz M, Ortner M, Sippl MJ. Progress in fold recognition. Proteins 1995;3:376–386.
8. Olmea O, Rost B, Valencia A. Effective use of sequence correlation and conservation in fold recognition. J Mol Biol 1999;293:1221–1239.
9. Shindyalov IN, Kolchanov NA, Sander C. Can three-dimensional contacts of proteins be predicted by analysis of correlated mutations? Protein Eng 1994;7:349–358.
10. Olmea O, Valencia A. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. Fold Des 1997;2:S25–32.
11. Fariselli P, Casadio R. Neural network based predictor of residue contacts in proteins. Protein Eng 1999;12:15–21.
12. Aszodi A, Gradwell MJ, Taylor WR. Global fold determination from a small number of distance restraints. J Mol Biol 1995;251: 308–326.
13. Lund O, Frimand K, Gorodkin J, Bohr H, Bohr J, Hansen J, Brunak S. Protein distance constraints predicted by neural networks and probability density functions. Protein Eng 1997;10: 1241–1248.
14. Gorodkin J, Lund O, Andersen CA, Brunak S. Using sequence motifs for enhanced neural network prediction of protein distance constraints. In: Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB99), La Jolla, CA. Menlo Park, CA: AAAI Press, 1999;p 95–105.
15. Fariselli P, Casadio R. Prediction of the number of residue contacts in proteins. In: Proceedings of the 2000 Conference on Intelligent Systems for Molecular Biology (ISMB00), La Jolla, CA. Menlo Park, CA: AAAI Press; 2000;146–151.
16. Richardson CJ, Barlow DJ. The bottom line for prediction of residue solvent accessibility. Protein Eng 1999;12:1051–1054.
17. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
18. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
19. Holbrook SR, Muskal SM, Kim SH. Predicting surface exposure of amino acids from protein sequence. Protein Eng 1990;3:659–665.
21. Pascarella S, De Persio R, Bossa F, Argos P. Easy method to predict solvent accessibility from multiple protein sequence alignments. Proteins 1999;32:190–199.
22. Naderi-Manesh H, Sadeghi M, Arab S, Moosavi Movahedi AA. Prediction of protein surface accessibility with information theory. Proteins 2001;42:452–459.
23. Mucchielli-Giorgi MH, Hazout S, Tuffery P. PredAcc: prediction of solvent accessibility. Bioinformatics 1999;15:176–177.
24. Carugo O. Predicting residue solvent accessibility from protein sequence by considering the sequence environment. Protein Eng 2000;13:607–609.
25. Li X, Pan XM. New method for accurate prediction of solvent accessibility from protein sequence. Proteins 2001;42:1–5.
26. Cuff JA, Barton G/J. Application of multiple sequence alignments profiles to improve protein secondary structure prediction. Proteins 2000;40:502–511.
26. Thompson MJ, Goldstein RA. Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. Proteins 1996;25:38–47.
27. Sali A, Shakhnovich E, Karplus M. How does a protein fold? Nature 1994;369:248–251.
28. 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.
29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
30. 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.
31. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative data sets. Protein Sci 1992;1:409–417.
32. Abagyan RA, Batalov S. Do aligned sequences share the same fold? J Mol Biol 1997;273:355–368.
33. Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 1994;19: 55–72.
34. Barton GJ. Protein secondary structure prediction. Curr Opin Struct Biol, 1995;5:372–376.
35. Baldi P, Brunak S, Frasconi P, Pollastri G, Soda G. Exploiting the past and the future in protein secondary structure prediction. Bioinformatics 1999;15:937–946.
36. Przybylski D, Rost B. Alignments grow, secondary structure pre-diction improves. Proteins 2001 (submitted).
37. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. Science 1991;252:1162–1164.
38. Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. Methods Enzymol 1996;266:554–571.
39. Baldi P, Brunak S, Frasconi P, Pollastri G, Soda G. Bidirectional dynamics for protein secondary structure prediction. In: Sun R, Giles CL, editors. Sequence learning: paradigms, algorithms, and applications. New York: Springer-Verlag; 2000;99–120.
40. Pollastri G, Przybylski D, Rost B, Baldi P. Improving the prediction of protein secondary strucure in three and eight classes using recurrent neural networks and profiles. Proteins 2001 (in press).
41. Baldi P, Brunak S, Chauvin Y, Anderson CAF, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 2000;16:412–424.
42. Baldi P, Pollastri G, Andersen CAF, Brunak S. Matching protein β-sheet partners by feedforward and recurrent neural networks. In: Proceedings of the 2000 Conference on Intelligent Systems for Molecular Biology (ISMB00), La Jolla, CA. Menlo Park, CA: AAAI Press; 2000;25–36.