Journal Title: Proteins
ISSN: 0887-3585
Article Author(s): Rost B
Article Title: Conservation and prediction of solvent accessibility in protein families.
Year: 1994 Nov
Volume: 20
Issue: 3
Pages: 216-26
PMID: 7892171
User's Comments: In color, if available

PROTEINS: Structure, Function, and Genetics 20:216–226 (1994)

RESEARCH ARTICLES

Conservation and Prediction of Solvent Accessibility in Protein Families

Burkhard Rost and Chris Sander
Protein Design Group, European Molecular Biology Laboratory, 69012 Heidelberg, Germany

ABSTRACT **Currently, the prediction of three-dimensional (3D) protein structure from sequence alone is an exceedingly difficult task. As an intermediate step, a much simpler task has been pursued extensively: predicting 1D strings of secondary structure. Here, we present an analysis of another 1D projection from 3D structure: the relative solvent accessibility of each residue. We show that solvent accessibility is less conserved in 3D homologues than is secondary structure, and hence is predicted less accurately from automatic homology modeling; the correlation coefficient of relative solvent accessibility between 3D homologues is only 0.77, and the average accuracy of predictions based on sequence alignments is only 0.68. The latter number provides an effective upper limit on the accuracy of predicting accessibility from sequence when homology modeling is not possible. We introduce a neural network system that predicts relative solvent accessibility (projected onto ten discrete states) using evolutionary profiles of amino acid substitutions derived from multiple sequence alignments. Evaluated in a cross-validation test on 238 unique proteins, the correlation between predicted and observed relative accessibility is 0.54. Interpreted in terms of a three-state (buried, intermediate, exposed) description of relative accessibility, the fraction of correctly predicted residue states is about 58%. In absolute terms this accuracy appears poor, but given the relatively low conservation of accessibility in 3D families, the network system is not far from its likely optimal performance. The most reliably predicted fraction of the residues (50%) is predicted as accurately as by automatic homology modeling. Prediction is best for buried residues, e.g., 86% of the completely buried sites are correctly predicted as having 0% relative accessibility.**

Key words: evolutionary information, multiple alignments, neural networks, protein structure prediction

INTRODUCTION

Sequence–Structure Gap

One important task of theoretical biology is to reduce the sequence–structure gap: some 34,000 protein sequences are in Swissprot (release 27.0[1]), but less than about 2,000 three-dimensional (3D) structures are in the Protein Data Bank (PDB).[2] The number of unique proteins of known 3D structure is about 300.[3,4] Large-scale gene sequencing projects are increasing the gap rapidly.[5] Thus an important task is to predict 3D structure from the one-dimensional (1D) string of amino acids. Our optimism for the successful prediction of structure from sequence is based on the well-established credo that protein sequence uniquely determines protein structure.[6,7]

To What Extent Do the Current Theoretical Tools Narrow the Sequence-Structure Gap?

The bad news is that prediction in 3D is currently only possible for proteins with significant sequence identity (> 25%[8,9]) to proteins of known 3D structure. The good news is that by homology modeling,[10–20] structures can be predicted accurately for some 9,000 proteins.[21] For some of the remaining 25,000 known sequences, potentials of mean force may allow for homology modeling by threading (sequence–structure alignment), i.e., in the absence of significant sequence identity.[22–33] However, for most of the remaining 20,000 or so sequences, 3D prediction is not yet feasible. Thus, a compromise is advisable: the simplification of the prediction problem. An extreme compromise is to project 3D structure onto 1D, e.g., a string of secondary structure assignments.[34] Secondary structure (in three states) is conserved at the 90% level within families of proteins with homologous 3D structure.[35,36] This allows three-state prediction at better than 72% in overall

Received April 1, 1994; revision accepted June 23, 1994.
Address reprint requests to Burkhard Rost or Chris Sander, Protein Design Group, EMBL, Meyerhofstrasse 1, 69012 Heidelberg, Germany.

accuracy.[37] However, secondary structure captures only some aspects of 3D structure. A different aspect is contained in another 1D feature: the degree to which a residue in a protein is accessible to solvent. But is solvent accessibility conserved in 3D families? Can it be successfully predicted? Previously, a neural network prediction method evaluated on five proteins reported a two-state (buried/exposed) accuracy level of some 72%.[38] Will this number be confirmed in a thorough analysis on a large set of unique proteins of known structures?

Static Residue Solvent Accessibility

The concept of solvent accessibility was introduced by Lee and Richards.[39] Studies of solvent accessibility in proteins have led to many new insights into protein structure.[40–52] Knowledge of solvent accessibility (and hydrophobicity) has proved useful for identifying protein function, sequence motifs,[53–55] and domains,[56] and for formulating hypotheses about antigenic determinants,[57] site-directed mutagenesis,[58] humanization of antibodies,[59] and on the correctness of designed or experimentally determined protein structures.[60–62] Furthermore, knowledge of solvent accessibility has assisted alignments in regions of remote sequence identity.[33,63–68] Often a very crude approximation for residue accessibility has been used: a projection onto two states, i.e., buried-exposed.[45,69,51] Such a projection of surface area onto two states reduces the information available about solvent accessibility by about one-half.[70] Additionally, the problem arises of how to define the threshold to distinguish between the two states.

Aims of This Work

Here, we have chosen to describe relative solvent exposure not in two but in ten states (with a more fine-grained distinction between states in the protein core). First, we present the results of a study on the conservation of accessibility in pairs of proteins of homologous 3D structure. (A similar study with a different focus appears in Flores et al.[71]) Second, we describe a neural network prediction system that uses profiles of amino acid substitutions from multiple alignments as input. The system predicts relative solvent accessibility in ten states. This prediction method is evaluated in cross-validation tests on 238 proteins. Performance accuracy is compared with homology modeling and with random prediction.

METHODS

Measuring Solvent Accessibility and Conservation

Residue relative solvent accessibility

How can the solvent accessibility of a residue embedded in a 3D structure be cast into a simple number? One simple way is to count the number of water molecules in direct contact with the residue, as estimated by the program DSSP for the first hydration shell.[39,72] For comparison between amino acids of different sizes, the relative solvent accessibility is a useful quantity (defined in Table I). Here, the relative solvent accessibility is projected onto ten discrete states: $RelAcc_{10} = 0–9$, such that the square of $RelAcc_{10}$ gives the percentage of relative accessibility (definition in Table I).

Measures for conservation and prediction accuracy

How can the degree of conservation of solvent accessibility within a structure family be measured? A straightforward measure is the correlation between two sets of accessibility numbers (*CorAcc* defined in Table I). It is common practice to project the accessibility onto two (buried/exposed),[45] or three (buried/intermediate/exposed) states.[50,73] However, it is not clear a priori where to set the thresholds to distinguish between these states. All measures used for the evaluation of accuracy are defined in Table I.

Database for the Analyses

Prediction of solvent accessibility

For evaluation, we used a set of 126 unique proteins (dubbed the cross-validation set) of known 3D structure and less than 25% pairwise sequence identity (given in Table I of Rost and Sander[37]). As in our earlier studies on secondary structure prediction,[74,75] we performed a sevenfold cross-validation test on this set, i.e., the method is trained on 106 proteins, and tested on 18. This procedure was repeated seven times with different divisions of the data set until all proteins were tested exactly once. As a further test of the method, we applied it to a set of 112 proteins of recently determined structure (dubbed the pre-release set) with no significant sequence identity to any of the proteins in the cross-validation set (all PDB proteins given in Table II of Rost and Sander[37]).

Conservation of solvent accessibility in structure families

The analysis of the conservation of solvent accessibility within 3D families was based on the following 80 pairs of homologous 3D structures: 1aapA:1bpt, 1aapA:1dtx, 1bbpA:2apd, 1bbt2:2plv2, 1ccr:5cyt, 1ccr:1ycc, 1ccr:1yea, 1cd8:2mcgl, 1cd8:1reiB, 1cd8:2fbjL, 1cdtA:1ctx, 1clm:4cln, 1clm:5tnc, 1clm:4tnc, 1clm:1pal, 1clm:1cdp, 1cmbA:1cmaA, 1fas:2abxB, 1hgeB:2hmgB, 1hilA:1reiB, 1hilA:1mcpL, 1hilA:2fbjL, 1hilA:2fb4, 1hilA:2mcg, 1hilA:3fab, 1hsbA:2hlaA, 1hsbA:3hlaA, 1hsbA:1hsaA, 1mamH:1mcpH, 1mamH:1mcpH, 1mamH:2fbjH, 1mbd:2mb5, 1mbd:1pmbB, 1mbd:2mm1, 1mbd:1mbs, 1mbd:2dhbB, 1mbd:1nih, 1mbd:2dhbB, 1mbd:1nihB, 1mbd:2lhb, 1nxb:6ebxB, 1nxb:2abxB, 1nxb:1ctx, 1nxb:1cdtB, 1pbxA:2dhbA, 1pbxA:1nihC,

TABLE I. Definition of Solvent Accessibility States

Solvent accessibility

Acc = solvent accessibility of a residue (given in Å^2) calculated from coordinates using DSSP,[72] $W \approx Acc/10$, approximates the number of water molecules around the residue

Relative solvent accessibility

RelAcc = *Acc/MaxAcc*, with maximal accessibility (measured in Å^2) for the amino acids given by the table following (amino acids in one-letter code; B stands for D or N; Z for E or Q, and X for an undetermined amino acid).[50,73]

AA	A	B	C	D	E	F	G	H	I	K	L	M
MaxAcc	106	160	135	163	194	197	84	184	169	205	164	188
AA	N	P	Q	R	S	T	V	W	X	Y	Z	
MaxAcc	157	136	198	248	130	142	142	227	180	222	196	

Ten-state model for accessibility

$RelAcc_{10} = INTEGER\ \sqrt{100 \times RelAcc}$, i.e., the projection of relative accessibility onto ten states. The square root is chosen to describe buried residues in more detail than in exposed ones.

Two-state (binary) model for accessibility (B/E)

	Buried (B)	Exposed (E)
Thresholds to distinguish two states	*RelAcc* < 16%	*RelAcc* ≥ 16%

Three-state (ternary) model for accessibility (B/I/E)

	Buried (B)	Intermediate (I)	Exposed (I)
Thresholds to distinguish two states	*RelAcc* < 9%	RelAcc = 9–36%	*RelAcc* ≥ 36%

Measures for evaluation of conservation and accuracy of prediction

$$CorAcc\ \frac{\langle xy\rangle - \langle x\rangle\langle y\rangle}{\sqrt{\langle x^2\rangle - \langle x\rangle^2}\sqrt{\langle y^2\rangle - \langle y\rangle^2}}$$

Correlation of accessibility, with x and y being the relative accessibilities for a pair of homologous proteins (for analyzing accessibility conservation in 3D families), or the predicted and observed relative accessibilities for a single protein (for analyzing prediction accuracy)

Q_2 percentage of conserved (or correctly predicted) residues in two states (B, E) defined by thresholds given above

Q_3 percentage of conserved (or correctly predicted) residues in three states (B, I, E) defined by thresholds given above

Q_{nX} for n states: percentage of conserved (or correctly predicted) residues in state X

$Q_{3X}^{\%obs}$ percentage of conserved (or correctly predicted) residues as above; %obs indicates that for prediction the percentages are normalized by the number of residues observed to be in state X

$Q_{3X}^{\%pred}$ probability for a correct prediction, i.e., the number of residues predicted correctly in state X (× 100) divided by the number of all residues predicted to be in state X

1pbxA:2dhbA, 1pbxA:1nihC, 1pbxA:1fdhA, 1pbxA:21hb, 1plc:1pcy, 1plc:7pcy, 1plc:2mcg1, 1plc:1paz, 1prcM:2rcrL, 1r1a2:2plv2, 1rla2:2mev2, 1troA:2wrpR, 1vaaB:1hsaB, 2aviA:1stp, 2ltnA:1lte, 2ltnA:2cnaA, 2ltnb:1lte, 2ltnB:4cna, 2plv1:1r1a1, 2plv3:1r1a1, 2scpA:4tnc, 2scpA:5tnc, 2sicI:2ssi, 2tbvA:4sbvB, 3cd4A:1reiB, 3cd4A:2fbjL, 4bp2:4p2p, 4bp2:1ppa, 4bp2:1pp2R, 4gpd1:1ggaO, 4gpd1:1gd1O, 4sbvA:2tbvA, 5p21:1q21, and 7timA:lypiA (notation: PDB identifier, chain).

Homology Prediction

Currently, homology modeling is, in practice, the most accurate method for predicting many aspects of protein structure. Thus, an analysis of the conservation of solvent accessibility within a family of homologous 3D structures gives us an effective upper limit for the accuracy of predicting solvent accessibility. Two values are of interest: first, the accuracy in predicting solvent accessibility by automatic homology modeling based on sequence alignments; and second, the conservation of solvent accessibility between structurally aligned 3D homologues. Sequence alignments were compiled by a dynamic programming algorithm.[9,21] Structural alignments for the same proteins were made by optimized comparisons of distance matrices.[76] For a thorough analysis of any prediction method, it is also important to define the

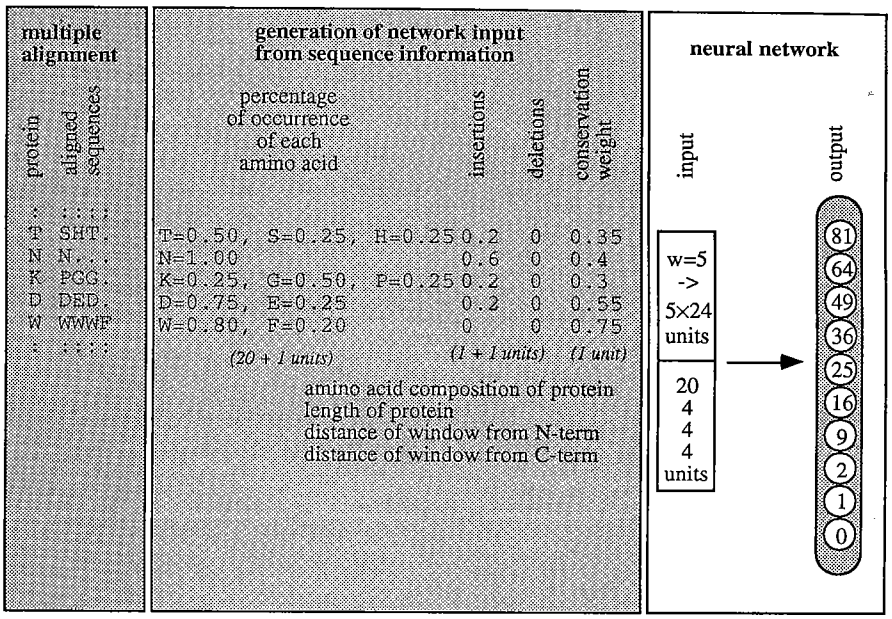

Fig. 1. Profiles of amino acid substitutions were compiled from multiple alignments. The conservation weights were computed as in the HSSP alignment database.[9,21] Input to the network consisted of profiles of amino acid frequencies for $w = 9$ or 13 consecutive residues (here we show only five). Each of the w residues in the input window was coded by 20 units, corresponding to the 20 different amino acids. To allow the first and last $w - 1$ residues in a sequence to occur in the central position in the gliding window, an additional "spacer" was used, i.e., an input unit coding for anything other than amino acids. The amino acid composition was coded by 20 units, one for each amino acid. The length of the protein was coded by four units: input to unit $i = 1$, if $N \geq 2^{i-1} \times 60$, and $= (N - 2^{i-2} \times 60)/2^{i-1} \times 60$, for $i = 1 - 4$. The distance of the first residue in the window to the N-terminal, and of the last window residue to the C-terminal end of the protein was coded by four units: input to unit $i = 1$, if $D \geq 2^{i-1} \times 10$, and $= (D - 2^{i-2} \times 10)/2^{i-1} \times 10$, for $i = 1 - 4$, where D gives the number of residues in the protein before the beginning or after the end of the input window. The output of the network was the relative accessibility of the central residue in ten states ($RelAcc_{10}$, Table I). The numbers shown in the output units define the lower limits of relative accessibility for each unit. The relative accessibility of the central residue (here K) was predicted to be the unit with maximal value. For example, suppose that unit six has the largest output sum. Then this residue would be predicted to have an accessibility between 25% and 36%. If this were a lysine, then it would be predicted to be in contact with between $0.25 \times 205/10 \approx 5$ and $0.36 \times 205/10 \approx 7$ water molecules (Table I).

lower limit of prediction accuracy. The worst one can do is to predict accessibility at random.

Predicting Accessibility by Neural Networks

We used a system of neural networks for predicting solvent accessibility. This system is similar to that we previously used for predicting secondary structure.[74,75,37] Thus, we only briefly describe the system; the differences from the previous system are described in detail.

Principle architecture of the network

The input to the network was a window of $w = 9$ or 13 consecutive residues; the window was shifted through each protein sequence (Fig. 1). The output was the relative accessibility of the central residue. Two different projections of relative accessibility were compared: first, a three-state description for which three output units code for buried (B), intermediate (I), and exposed (E) residues; and second, a ten-state description for which 10 output units were used (coding as in Fig. 1).

Profiles of evolutionary conservation as input

Using a profile of amino acid substitutions (as given in a data base of multiple sequence alignments[9,21]) rather than single sequences led to improvement in secondary structure prediction by some 10 percentage points.[37] Here, the same idea was applied to predicting accessibility: instead of single sequences we used as input to the network the profiles for a window of w consecutive residues in a protein (Fig. 1).

Additional input information

When using profiles from multiple alignments, the following information was used as additional input (Fig. 1): 1) for each residue, a weight was defined that reflects the conservation of this residue in the alignment[77]; 2) at each alignment position, i.e., for each residue, one unit was used for the percentage of insertions in the alignment, and one for the percentage of deletions[37]; 3) the relative content of each amino acid in the entire protein[37]; 4) the length of the protein N; and 5) the distance of the window to the N- and C-terminal end of the protein.

Training the neural networks

The networks were trained by the usual *conjugate gradient descent* algorithm (also known as *back-propagation*[78]; for details, see Rost and Sander[74]). Training was stopped at $CorAcc > 0.56$ for the training set. At each step of the minimization, the examples presented were chosen at random from the training set.

Final prediction

We added two further computational steps to obtain the final network prediction (called *PHDacc*). First, a jury decision was performed, i.e., an arithmetic average over the output from different networks. Different networks were generated using different input information, and two different sizes for the input window ($w = 9$ and $w = 13$). Second, a filter was applied. The formation of clusters of adjacent units with high-output values suggested a nearest neighbor average, i.e., to filter the output by:

$$o_i = \frac{1}{3}\{o_{i-1} + o_i + o_{i+1}\}, \qquad i = 1, \ldots, 10 \quad (1)$$

where o_i is the output of network unit i. The final prediction was assigned to the output unit with the largest average.

RESULTS

Conservation of Solvent Accessibility in 3D Homologues

Solvent accessibility is less conserved than secondary structure

Comparing residue accessibility numbers for 80 pairs of homologous 3D structures yielded a correlation coefficient of 0.77 (Table II). In three states the conservation is about 74% residues in identical accessibility states (Table II); this is significantly lower than the conservation of secondary structure in three states (Fig. 6).

Accuracy of homology prediction sharply decreases with sequence identity

The accuracy of predicting solvent accessibility by homology modeling (Fig. 3) was far more sensitive to sequence identity than the accuracy of predicting secondary structure by homology modeling.[36] Surprisingly, for protein pairs of low sequence identity, buried residues were found to be less conserved than exposed ones (data not shown). By contrast, secondary structure is better conserved in protein cores.[36]

Prediction of Solvent Exposure by Neural Networks

Previous results hold for a larger data set

In a pilot study on prediction of surface exposure, Holbrook et al.[38] selected a very small test set of five proteins. In general, this test set is too small to estimate prediction accuracy.[34] The authors reported a correlation between observed and predicted accessibility of 0.44. Did this result hold up in our study with a data set 30 times larger? For direct comparison, we trained and used a network similar to that of Holbrook et al.[38] with three output states (B/I/E), and tested on the cross-validation set of 126 proteins. The average correlation predicted-observed was 0.36. Projected onto a binary or ternary description of accessibility, the results were similar to the results of Holbrook et al.[38] (Table II). Does this imply that the authors were fortunate in choosing the set of five proteins? They were, as the variation of prediction accuracy between different proteins was significant (Fig. 4).

Evolutionary information improves the prediction markedly

Evaluated on 126 proteins with the three output-states network, the gain in accuracy from using evolutionary information was about 3 percentage points in both binary and ternary accuracy. The correlation improved more markedly: from 0.36 to 0.45 (Table II).

Improvement of prediction by ten-state output

Given the relatively low prediction accuracy, is it possible that the neural network system can predict in more detail? Networks with ten output units using single sequences as input in our hands resulted in the same correlation coefficient as three-state networks using profiles (Table II). Using multiple alignments for ten-state networks raised the correlation above 0.5. The best network used additional information as input: conservation weight, number of insertions and deletions in the alignment, amino acid composition in the protein, length of the protein, and distance of the input window to the N- and C-terminals of the protein. Such a network resulted in a correlation of 0.536. Any additional information used tended to improve prediction of more exposed residues at the expense of buried ones (Table II). Note: a surprising result was that for the ten-state prediction based on profiles, the test set performance was always best for the highest training set performance, i.e., we did not observe overtraining.

The choice of the window size has marginal influence

The size of the input window w has a very small influence on prediction accuracy. We found a difference of only about 0.1 in the correlation coefficient between using single residues for the prediction and using a window of consecutive residues (data not shown). One-third of the improvement from a "window" of $w = 1$ (single residue) to $w = 9$–17 was gained by using triplets ($w = 3$), and another third by using quintets ($w = 5$).

TABLE II. Prediction of Solvent Accessibility by Different Methods*

	Two states	Three states							Ten states	
			Q_{3B}		Q_{3I}		Q_{3E}			
	Q_2	Q_3	%o	%p	%o	%p	%o	%p	Q_{10}	*CorAcc*
Reference Methods†										
Random prediction	**52.0**	**33.9**	**30**		**30**		**38**		**10.9**	**0.005**
Prediction by sequence alignment	**83.8**	**71.6**	**76**	**78**	**60**	**60**	**76**	**74**	**37.1**	**0.676**
Prediction by subset of 22 monomers	86.6	73.9	82	71	61	58	77	85	39.2	0.731
Prediction by structural alignment	**84.8**	**73.6**	**78**	**78**	**63**	**61**	**77**	**79**	**39.1**	**0.765**
Prediction by subset of 22 monomers	87.2	75.5	82	69	64	61	78	88	40.6	0.799
Three-state networks: B,I,E.‡										
Percent of residues observed in B, I, E			32		53		14			
Single sequences	71.4	55.1	29		77		30			0.356
Multiple alignment profiles	74.6	58.0	39		73		40			0.450
Previous methods**										
Wako and Blundell[89](13 families)	76.5									
Holbrook et al.[38](5 proteins)	72.0	52.0	51		44		62			
Ten State Networks†										
Percent of residues observed in B, I, E			32		29		37			
Single sequences	70.0	52.4	71	52	21	36	60	59	21.6	0.432
Multiple alignment profiles + conservation weights	74.2	57.1	78	57	6	42	77	58	23.9	0.522
Profiles + conservation weights + amino acid composition	74.2	56.8	77	57	9	41	75	58	23.8	0.515
Profiles + conservation weights + composition + protein length + distance to N- and C-terminals	75.0	57.2	75	58	6	41	80	57	24.0	0.530
Profiles + conservation weights + indels	74.1	57.0	77	57	7	42	78	57	23.9	0.520
All = profiles + weights + indels + composition + length + distance	74.8	57.5	76	58	9	41	78	58	24.3	0.533
Our best networks tested on different data sets†										
***PHDacc* 126 = cross-validation set**	**75.0**	**57.9**	**76**	**60**	**12**	**43**	**81**	**58**	**24.4**	**0.541**
***PHDacc* 112 = pre-release set**	**74.7**	**57.9**	**77**	**59**	**12**	**44**	**75**	**58**	**25.3**	**0.544**
PHDacc 99 monomers (of 238)	77.7	60.5	77	59	13	45	81	64	25.8	0.589
PHDacc 13 from Wako and Blundell[89]	79.2	60.8	77	57	12	46	86	66	26.2	0.609
PHDacc 5 from Holbrook et al.[38]	75.7	58.4	76	63	10	43	79	56	26.5	0.549

*The random and homology predictions are based on 80 pairs of homologous proteins of known 3D structure; the network predictions are based on the cross-validation set of 126 proteins (except for the last rows). The quantities Q and *CorAcc* are defined in Table I (%o and %p refer to $Q_{3X}^{\%obs}$ and $Q_{3X}^{\%pred}$). To examine the performance for proteins with only one chain, 22 proteins were extracted from the set of 80 homologous pairs, and 99 of the 238 (126 + 112) proteins were used for the prediction analysis. All numbers have been averaged over all residues. Values for prediction by homology modeling based on automatic sequence alignments are averaged over the fragments aligned by the dynamic programming method; the results for prediction by structural alignment are averaged over fragments aligned by comparison of distance matrices. Thus, values for structural alignments do not include non-homologous loop regions, while those for sequence alignments do. The correlation coefficients were based on a projection of the largest of the three or ten output units onto one real value for relative accessibility. Thus the comparison between ten- and three-output state neural networks can be based on such a coefficient.

Methods. *Holbrook et al.*[38]: Neural network prediction of relative accessibility in three states based on single sequences; results as published (evaluated on five proteins); *Wako and Blundell*[89]: Profile-based prediction of relative accessibility in two states; results as published (evaluated on 13 protein families); *PHDacc n:* Results of jury average over eight networks (five as shown with an input window of $w = 9$ consecutive residues, and three over networks using $w = 13$ consecutive residues as input), after compiling the neighbor averages (eq. 1); n indicates the number of proteins used for evaluation; results for the following data sets are given—*PHDacc* 126, 126 proteins of cross-validation set; *PHDacc* 112, 112 proteins in the pre-release set; *PHDacc* 99 monomers, 99 monomers in the cross-validation and in the pre-release set (total of 238 proteins); *PHDacc* 5: Five proteins used by Holbrook et al.[38]; *PHDacc* 13, 13 proteins, one representative from each of the 13 families used by Wako and Blundell.[89] Thresholds used to define discrete accessibility states were as follows:

†Two states, 16% (B/E); three states, 9% (B/I) and 36% (I/E).

‡Two states, 9% (B/E); three states, 9% (B/I) and 64% (I/E).

**Two states, 20% (B/E); three states, 5% (B/I) and 40% (I/E).

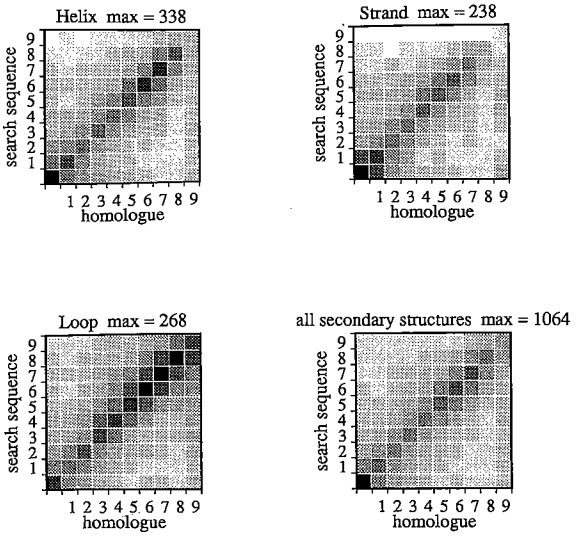

Fig. 2. Correlation between solvent accessibility in 3D homologues. The vertical dimension corresponds to one protein, and the horizontal dimension to its homologue. The scale used is $RelAcc_{10}$ (Table I), i.e., the squares of the axis labels are the percentage of relative accessibility. Counts are coded by different shading, with the maximal counts (black) given on top of each of the four graphs (separate statistics for residues in the secondary structure states helix, strand, loop and for all residues). The graphs were generated using the program CARTE.[88] The plot confirms that solvent accessibility is largely conserved in 3D families. In general, conservation is best for buried residues; this is particularly true for strands. Helical residues with intermediate accessibility (25–64% relative accessibility) are also well conserved. This supports the observation that helices often provide a scaffold between protein interior and solvent. Residues in irregular secondary structure in exposed positions are more conserved in accessibility than those in buried positions. Note that the plots are not symmetric, since the classification of secondary structure is based on the search sequence, i.e., the results for, e.g., helix hold for all residues that are observed in helices in the search sequences.

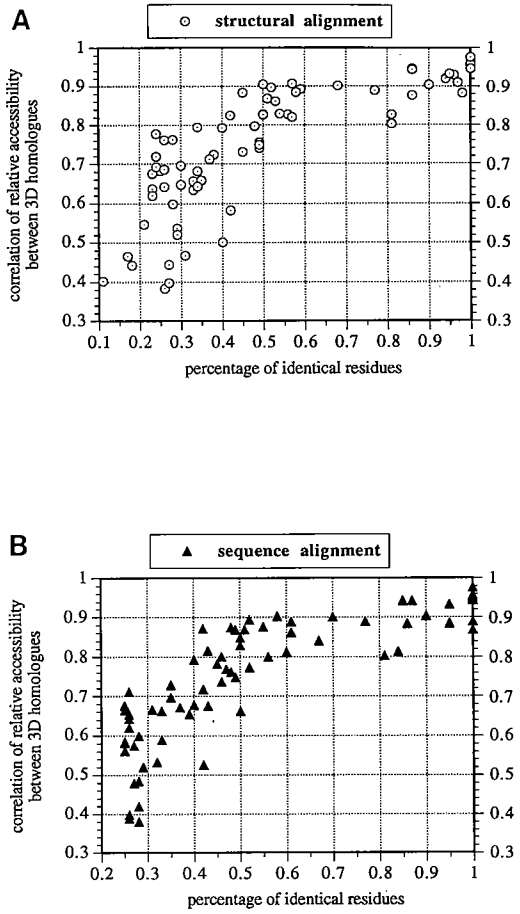

Fig. 3. Conservation of relative solvent accessibility (averaged over protein pairs) vs. the percentage of identical residues between the sequences of the pair. Alignments of homologues were done by (**A**) structural or by (**B**) sequence alignments. It is well established that sequence identity > 25–30% (for alignment length > 80 residues) is usually sufficient to conclude that two sequences have the same 3D structure.[8,9] Particularly for sequence alignments, there is a clear decrease in conservation of relative accessibility at about 30–40% identity. The reason for the decrease is mostly that for low levels of sequence identity, it becomes increasingly difficult to align the sequences (without structural information). Structural alignments (**A**) have a better conservation for low levels of sequence identity than sequence alignments (**B**). The clear decrease in conservation for sequence alignments indicates the limitations in gaining prediction accuracy by using more distant family members for the network input (Fig. 1).

Jury and filtering slightly improves and balances prediction

The final jury decision had a slightly inferior correlation coefficient than the best single network (but a better three-state accuracy than the best network); (Table II). The result of the neighbor average filter [Eq. (1)] was a slight improvement in prediction accuracy. The averaging filter had an effect on the prediction very similar to the (more complicated) use of a second level of neural networks (as used for secondary structure prediction[74]; data not shown). Particularly striking was the poor performance of the ten-state network for the intermediate state (I). However, this was not a general result for neural network prediction: by using a different training algorithm, some 40% of the residues observed in the intermediate state could be predicted correctly, at the expense of a reduced accuracy in predicting buried and exposed residues (data not shown).

Half the residues are predicted as accurately as by homology modeling

The reliability index [*RI*, defined in Fig. 5, Eq. (2)] correlated well with prediction accuracy. About half of all residues were in reliability classes with an accuracy level comparable to that of homology modeling (Fig. 5): for residues predicted with $RI \geq 4$, the correlation between observed and predicted relative accessibility was as large as 0.69.

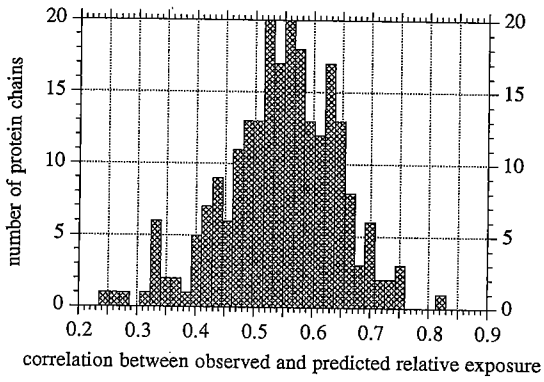

Fig. 4. Distribution of the correlation between predicted and experimentally observed relative accessibility averaged over all residues in all 238 (126 + 112) protein chains examined in this study. This distribution is approximately Gaussian with a mean of 0.54 and a standard deviation of 0.12. Shown is the distribution of the final filtered jury prediction (called *PHDacc*). Three negative outliers are not shown: 2MEV_4 (−0.08) 4RHV_4 (0.03), and 2GN5 (0.16; note that this structure is likely to be wrong[61]) (PDB + chain identifier). Note that the prediction accuracy was significantly higher for monomeric proteins (Table II).

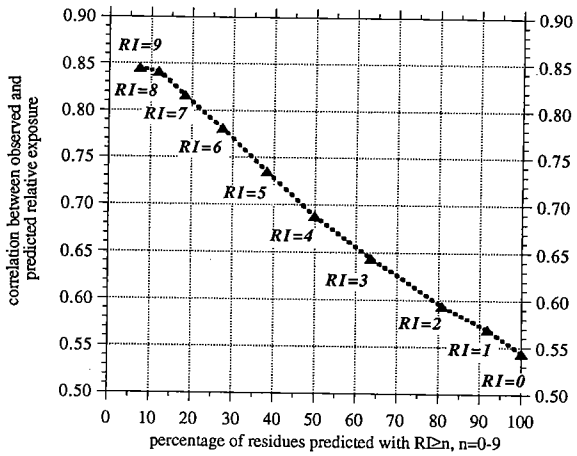

Fig. 5. Correlation between predicted and observed relative solvent accessibility versus the fraction of residues with a reliability index RI > *n*, where $n = 0, \ldots, 9$. The results have been averaged over all residues in all 238 proteins (cross validation + pre-release set), e.g., about 50% of all residues were predicted with RI ≥ 4. For these residues, the correlation between predicted and observed accessibility was 0.69, i.e., comparable to automatic homology modeling. The reliability index, RI, is defined as:

$$RI = \text{INTEGER}\,(30 \cdot \{o_{imax} - o_{2nd}\}) \qquad (2)$$

where imax is the index of the unit with the largest output and

$$o_{2nd} = \max\{o_j\}, \quad \text{with } |j - imax| > 2, \quad \text{and } j = 0\text{–}9.$$

In other words, RI is the difference between the largest output and the second largest unit at least two positions away from the largest unit. The empirical factor of 30 that RI lies between 0 and 9.

Prediction performance confirmed by test on pre-release set

To make doubly sure that the prediction results were not biased by the data set chosen, we performed another test: we trained a network on all 126 proteins of the cross-validation set and tested on the 112 proteins of the PDB pre-release set. The values of prediction accuracy evaluated based on the cross-validation set were largely confirmed (Table II). This implies that the correlation between predicted and observed solvent accessibility is likely to remain at about 0.54 for the next 240 or so protein structures that are determined, using the currently trained networks.

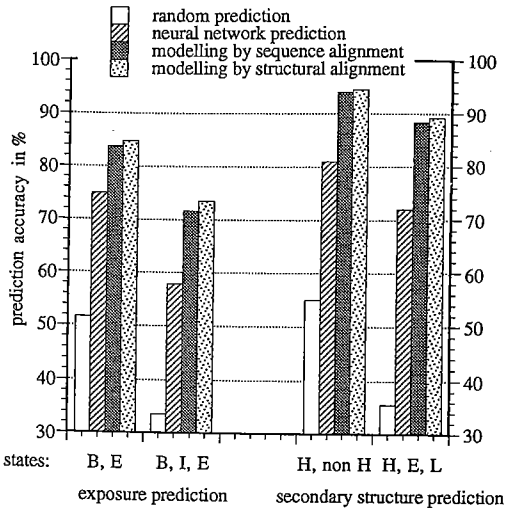

Fig. 6. Here we compare the accuracy of prediction of relative solvent accessibility with that of secondary structure. The prediction methods compared are: random prediction, cross-validated neural networks, and homology modeling (either by sequence or structural alignments). For solvent accessibility we show results for both two-state (B, E), and three-state (B, I, E) models; for secondary structure, we show results for two-state (helix, non-helix) and three-state (helix, strand, loop) descriptions.[36]

DISCUSSION

Can Prediction of Solvent Accessibility Be Useful for Predicting 3D Structure?

We do not yet know to what extent prediction of solvent accessibility is useful for prediction of detailed 3D structures. Here, we pursued a far less ambitious goal: to evaluate the accuracy of exposure prediction as such. Appropriate evaluation of the performance of any prediction requires two elements: first, an analysis of the best prediction in practice (e.g., homology modeling) and of the worst prediction (random); and second, an evaluation of the method on a unique test set of sufficient size (here 238 proteins).

How Well Is Solvent Accessibility Conserved?

Comparing relative accessibility between structurally aligned corresponding residues in 3D homologous structures yields a correlation coefficient of 0.77. Prediction of accessibility by automatic homology modeling (sequence alignment) results in a cor-

relation coefficient of about 0.68. The accessibility of completely buried residues was best conserved (Fig. 2). Evaluated in three states, relative accessibility was much less conserved than secondary structure. This low degree of conservation raises the question of whether or not there are descriptions of the relative position of a residue in a structure that are better conserved between 3D homologues.

How Good Is Prediction of Solvent Accessibility?

The final *PHDacc* prediction correlated with the experimental observation with a correlation coefficient of about 0.54. Expressed in units of the difference between automatic homology modeling and random prediction, the improvement of *PHDacc* over simple networks that do not use multiple alignments and predict accessibility only in three states was 26%: about 13% resulted from using multiple alignments profiles (residue frequencies) as input; 3% from using further information contained in the multiple alignment (conservation, number of insertions and deletions) and the protein (amino acid composition, length, position of predicted residue relative to protein ends); and 10% from using a ten-state instead of a three-state model for relative accessibility. Prediction was better for residues in regular secondary structure segments, and in general better for the extreme cases, i.e., completely buried and fully exposed residues, e.g., 86% of the completely buried sites were correctly predicted as having 0% relative accessibility. The network method was superior to previously published methods (Table II), but in absolute terms, prediction of relative solvent accessibility may appear to be rather inaccurate. However, its accuracy was relatively closer to prediction by automatic homology modeling than it was to prediction of secondary structure (Fig. 6).

Can Prediction Accuracy Be Improved?

Our belief is that improvements of the method could by made by searching for an alternative 1D feature of 3D structure that contains information similar to solvent accessibility but has been better conserved in evolution. Another idea is to combine predictions of accessibility with data from nuclear magnetic resonance experiments.[79]

Is Prediction of Accessibility Useful?

Possible applications for solvent accessibility predictions are as follows. First, an approach that has been pursued for a long time[80,31] is to start from fragments, such as secondary structure segments and to try to arrange these in 3D using predictions of accessibility. Second, the prediction could be used as an estimate for the number of protein–protein and protein–solvent contacts of a residue. Such an estimate could be useful for normalization of predicted contact maps,[81,82] i.e., descriptions of the 3D structure in 2D by a matrix of all inter-residue contacts in a protein. Third, a classification into families,[83] e.g., for threading procedures, could be assisted by predictions of accessibility. In particular, predictions of secondary structure and solvent accessibility could be aligned to known 3D structures to detect putative remote homologues, or at least to provide independent evidence for such detection by other techniques.[84] Fourth, predictions could be used as additional constraints in energy minimization of large molecules.[85–87] Fifth, predictions could be used to improve predictions of glycosylation sites (Brunak et al., in preparation). Sixth, solvent accessibility predictions could be helpful to predict epitopes (antigenic sites).

Is the Network Prediction Available?

Predictions of solvent accessibility (and secondary structure) are provided via by an automatic electronic mail server. For information, send the word *help* to the internet address *PredictProtein@EMBL-Heidelberg.DE* by electronic mail.

ACKNOWLEDGMENTS

We thank colleagues at EMBL: Séan O'Donoghue, for very detailed comments on the manuscript, proof-reading, and motivating discussions; Gerrit Vriend and Reinhard Schneider, for many valuable ideas and fruitful discussions; Reinhard Schneider and Simon Hubbard, for help with software tools; Liisa Holm, for the generation of 3D alignments; and Ulrike Göbel, for proof-reading. Furthermore, thanks to Tim Hubbard (MRC, Cambridge, U.K.), and the referees for valuable comments on the manuscript. Last but not least, many thanks to all those who publish experimental data on 3D protein structures and deposit the coordinates in public databases.

REFERENCES

1. Bairoch, A., Boeckmann, B. The SWISS-PROT protein sequence data bank. Nucleic Acids Res. 20:2019–2022, 1992.
2. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A computer based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
3. Hobohm, U., Scharf, M., Schneider, R., Sander, C. Selection of representative protein data sets. Protein Sci. 1:409–417, 1992.
4. Hobohm, U., Sander, C. Enlarged representative set of protein structures. Protein Sci. 3:522–524, 1994.
5. Oliver, S., der, A.Q.J.M.v., Agostioni-Carbone, M.L., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, J.P.G., Benit, P., Berben, G., Bergantino, E., Biteau, N., Bolle, P.A., Bolotin-Fukuhara, M., Brown, A., Brown, A.J.P., Buhler, J.M., Carcano, C., Carignani, G., Cederberg, H., Chanet, R., Contreras, R., Crouzet, M., Daignan-Fornier, B., Defoor, E., Delgado, M., Demolder, J., Doira, C., Dubois, E., Dujon, B., Dusterhoft, A., Erdmann, D., Esteban, M., Fabre, F., Fairhead, C., Faye, G., Feldmann, H., Fiers, W., Francingues-Gaillard, M.C., Franco, L., Frantali, L., Fukuhara, H., Fuller, L.J., Galland, P., Gent, M.E., Gigot, D., Gilliquet, V., Glansdorff, N., Goffeau, A., Grenson, M., Grisanti, P., Grivell,

L.A., Haan, M.d., Haasemann, M., Hatat, D., Hoenicka, J., Hegeemann, J., Herbert, C.J., Hilger, F., Hohmann, S., Hollenberg, C.P., Huse, K., Iborra, F., Indge, K.J., Isono, K., Jacq, C., Jacquet, M., James, C.M., Jauniaux, J.C., Jia, Y., Jimenez, A., Kelly, A., Kleinhans, U., Kreisl, P., Lanfranchi, G., Lewis, C., Linden, C.G.v.d., Lucchini, G., Lutzenkirchen, K., Maat, M.J., Mallet, L., Mannhaupt, G., Martegani, E., Mathieu, A., Maurer, C.T.C., McDonnell, D., McKee, R.A., Massenguy, F., Mewes, H.W., Molemans, F., Montague, M.A., Muzi Falconi, M., Navas, L., Newlon, C.S., Noone, D., Pallier, C., Panzeri, L., Pearson, B.M., Perea, J., et al. The complete DNA sequence of yeast chromosome III. Nature 357:38–46, 1992.

6. Anfinsen, C.B., Haber, E., Sela, M., White, F.H., Jr. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc. Natl. Acad. Sci. U.S.A. 47:1309–1314, 1961.
7. Anfinsen, C.B. Principles that govern the folding of protein chains. Science 181:223–230, 1973.
8. Chothia, C., Lesk, A.M. The relation between the divergence of sequence and structure in proteins. EMBO J. 5:823–826, 1986.
9. Schneider, R., Sander, C. Database of homology-derived structures and the structural meaning of sequence alignment. Proteins 9:56–68, 1991.
10. Greer, J. Model for haptoglobin heavy chain based upon structural homology. Proc. Natl. Acad. Sci. U.S.A. 77: 3393–3397, 1980.
11. Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E., Thornton, J.M. Knowledgebased prediction of protein structures and the design of novel molecules. Nature 326:347–352, 1987.
12. Blundell, T.L., Carney, D., Gardner, S., Hayes, F., Howlin, B., Hubbard, T.J.P., Overington, J., Singh, D.A., Sibanda, B.L., Sutcliffe, M. Knowledge-based protein modelling and design. Eur. J. Biochem. 172:513–520, 1988.
13. Summers, N.L., Karplus, M. Construction of side-chains in homology modelling. J. Mol. Biol. 210:785–811, 1989.
14. Taylor, W.R., Orengo, C.A. A holistic approach to protein structure alignment. Protein Eng. 2:505–519, 1989.
15. Overington, J., Johnson, M.S., Sali, A., Blundell, T.L. Tertiary structural constraints on protein evolutionary diversity: Templates, key residues and structure prediction. Proc. R. Soc. Lond. [Biol.] 241:132–145, 1990.
16. Greer, J. Comparative modeling of homologous proteins. Methods Enzymol. 202:239–252, 1991.
17. Vriend, G., Sander, C. Detection of common three-dimensional substructures in proteins. Proteins 11:52–58, 1991.
18. Levitt, M. Accurate modeling of protein conformation by automatic segment matching. J. Mol. Biol. 226:507–533, 1992.
19. Overington, J., Donnelly, D., Johnson, M.S., Sali, A., Blundell, T.L. Environment-specific amino acid substitution tables: Tertiary templates and prediction of protein folds. Protein Sci. 1:216–226, 1992.
20. Overington, J.P. Comparison of three-dimensional structures of homologous proteins. Curr. Opin. Struct. Biol. 2:394–401, 1992.
21. Sander, C., Schneider, R. The HSSP data base of protein structure-sequence alignment. Nucleic Acids Res. 21: 3105–3109, 1993.
22. Sippl, M.J. The calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures of globular proteins. J. Mol. Biol. 213:859–883, 1990.
23. Finkelstein, A.V., Reva, B.A. Search for the stable state of a short chain in a molecular field. Protein Eng. 5:617–624, 1992.
24. Sippl, M.J., Hendlich, M., Lackner, P. Assembly of polypeptide and protein backbone conformations from low energy ensembles of short fragments: Development of strategies and construction of models for myoglobin, lysozyme, and thymosin β_4. Protein Sci. 1:625–640, 1992.
25. Sippl, M.J., Weitckus, S. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. Proteins 13:258–271, 1992.
26. Bryant, S.H., Lawrence, C.E. An empirical energy function for threading protein sequence through the folding motif. Proteins 16:92–112, 1993.
27. Miyazawa, S., Jernigan, R.L. A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. Protein Eng. 6:267–278, 1993.
28. Nishikawa, K., Matsuo, Y. Development of pseudoenergy potentials for assessing protein 3D-1D compatibility and detecting weak homologies. Protein Eng. 6:811–820, 1993.
29. Ouzounis, C., Sander, C., Scharf, M., Schneider, R. Prediction of protein structure by evaluation of sequence-structure fitness: Aligning sequences to contact profiles derived from 3D structures. J. Mol. Biol. 232:805–825, 1993.
30. Stultz, C.M., White, J.V., Smith, T.F. Structural analysis based on state-space modeling. Protein Sci. 2:305–314, 1993.
31. Taylor, W.R. Protein fold refinement: Building models from idealized folds using motif constraints and multiple sequence data. Protein Eng. 6:593–604, 1993.
32. Wodak, S.J., Rooman, M.J. Generating and testing protein folds. Curr. Opin. Struct. Biol. 3:247–259, 1993.
33. Sippl, M.J., Jaritz, M. Predictive power of mean force pair potentials. In: H. Bohr "Protein Structure by Distance Analysis." Bohr, H., Brunak, S., eds. Washington DC: IOS Press, 1994:113–134.
34. Rost, B., Sander, C., Schneider, R. Progress in protein structure prediction? TIBS 18:120–123, 1993.
35. Russell, R.B., Barton, G.J. The limits of protein secondary structure prediction accuracy from multiple sequence alignment. J. Mol. Biol. 234:951–957, 1993.
36. Rost, B., Sander, C., Schneider, R. Redefining the goals of protein secondary structure prediction. J. Mol. Biol. 235: 13–26, 1994.
37. Rost, B., Sander, C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 19:55–72, 1994.
38. Holbrook, S.R., Muskal, S.M., Kim, S.-H. Predicting surface exposure of amino acids from protein sequence. Protein Eng. 3:659–665, 1990.
39. Lee, B.K., Richards, F.M. The interpretation of protein structures: Estimation of static accessibility. J. Mol. Biol. 55:379–400, 1971.
40. Chothia, C. The nature of the accessible and buried surfaces in proteins. J. Mol. Biol. 105:1–12, 1976.
41. Janin, J. Surface area of globular proteins. J. Mol. Biol. 105:13–14, 1976.
42. Richards, F.M. Areas, volumes, packing, and protein structure. Annu. Rev. Biophys. Bioeng. 6:151–176, 1977.
43. Richmond, T.J., Richards, F.M. Packing of α-helices: Geometrical constraints and contact areas. J. Mol. Biol. 119: 537–555, 1978.
44. Rose, G.D. Prediction of chain turns in globular proteins on a hydrophobic basis. Nature 272:586–590, 1978.
45. Janin, J. Surface and inside volumes in globular proteins. Nature 277:491–492, 1979.
46. Wodak, S., Janin, J. Analytical approximation to the accessible surface area of proteins. Proc. Natl. Acad. Sci. U.S.A. 77:1736–1740, 1980.
47. Kyte, J., Doolittle, R.F. A simple method for displaying the hydrophathic character of a protein. J. Mol. Biol. 157:105–132, 1982.
48. Sweet, R.M., Eisenberg, D. Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. J. Mol. Biol. 171:479–488, 1983.
49. Eisenberg, D., Schwartz, E., Komaromy, M., Wall, R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J. Mol. Biol. 179:125–142, 1984.
50. Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., Zehfus, M.H. Hydrophobicity of amino acid residues in globular proteins. Science 229:834–838, 1985.
51. Miller, S., Janin, J., Lesk, A.M., Chothia, C. Interior and surface of monomeric proteins. J. Mol. Biol. 196:641–656, 1987.
52. Ponder, J.W., Richards, F.M. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193:775–791, 1987.
53. Eisenberg, D., Weiss, R.M., Terwilliger, T.C. The hydrophobic moment detects periodicity in protein hydrophobicity. Proc. Natl. Acad. Sci. U.S.A. 81:140–144, 1984.
54. Cornette, J.L., Cease, K.B., Margalit, H., Spouge, J.L., Berzofsky, J.A., DeLisi, C. Hydrophobicity scales and com-

putational techniques for detecting amphipathic structures in proteins. J. Mol. Biol. 195:659–685, 1987.
55. Viari, A., Soldano, H., Ollivier, E. A scale-independent signal processing method for sequence analysis. CABIOS 6:71–80, 1990.
56. Wodak, S.J., Janin, J. Location of structural domains in proteins. Biochemistry 20:6544–6552, 1981.
57. Hopp, T.P., Woods, K.R. Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. U.S.A. 78:3824–3828, 1981.
58. Vriend, G., Eijsink, V. Prediction and analysis of structure, stability and unfolding of thermolysin-like proteases. J. Comput. Aided Mol. Design 7:367–396, 1993.
59. Pedersen, J.T., Henry, A.H., Searle, S.J., Guild, B.C., Roguska, M., Rees, A.R. Comparison of surface accessible residues in human and murine immunoglobulin Fv domains. J. Mol. Biol. 235:959–973, 1994.
60. Baumann, G., Frömmel, C., Sander, C. Polarity as a criterion in protein design. Protein Eng. 2:329–334, 1989.
61. Sippl, M.J. Recognition of errors in three-dimensional structures of proteins. Proteins 17:355–362, 1993.
62. Sippl, M.J. Boltzmann's principle, knowledge based mean fields and protein folding. An approach to the computational determination of protein structures. J. Comput. Aided Mol. Design 7:473–501, 1993.
63. Gaboriaud, C., Bissery, V., Benchetrit, T., Mornon, J.P. Hydrophobic cluster analysis: An efficient new way to compare and analyse amino acid sequences. FEBS Lett. 224:149–155, 1987.
64. Lemesle-Varloot, L., Henrissat, B., Gaboriaud, C., Bissery, V., Morgat, A., Mornon, J.P. Hydrophobic cluster analysis: Procedures to derive structural and functional information from 2-D representation of protein sequences. Biochimie 72:555–574, 1990.
65. Bowie, J.U., Lüthy, R., Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 253:164–169, 1991.
66. Lüthy, R., McLachlan, A.D., Eisenberg, D. Secondary structure-based profiles: Use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. Proteins 10:229–239, 1991.
67. Casari, G., Sippl, M.J. Structure-derived hydrophobic potential. J. Mol. Biol. 224:725–732, 1992.
68. Lüthy, R., Bowie, J.U., Eisenberg, D. Assessment of protein models with three-dimensional profiles. Nature 356: 83–85, 1992.
69. Hubbard, T.J.P., Blundell, T.L. Comparison of solvent-inaccessible cores of homologous proteins: Definitions useful for protein modelling. Protein Eng. 1:159–171, 1987.
70. Lawrence, C., Auger, I., Mannella, C. Distribution of accessible surfaces of amino acids in globular proteins. Proteins 2:153–161, 1987.
71. Flores, T.P., Orengo, C.A., Moss, D.S., Thornton, J.M. Comparison of conformational characteristics in structurally similar protein pairs. Protein Sci. 2:1811–1826, 1993.
72. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. Biopolymers 22:2577–2637, 1983.
73. Sander, C., Scharf, M., Schneider, R. Design of protein structures. In: "Protein Engineering." Rees, A.R., Sternberg, M.J.E., Wetzel, R., eds. Oxford: IRL Press, 1992:89–115.
74. Rost, B., Sander, C. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol. 232:584–599, 1993.
75. Rost, B., Sander, C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc. Natl. Acad. Sci. U.S.A. 90:7558–7562, 1993.
76. Holm, L., Sander, C. Protein structure comparison by alignment of distance matrices. J. Mol. Biol. 233:123–138, 1993.
77. Rost, B., Sander, C. Secondary structure prediction of all-helical proteins in two states. Protein Eng. 6:831–836, 1993.
78. Rumelhart, D.E., Hinton, G.E., Williams, R.J. Learning representations by backpropagating error. Nature 323: 533–536, 1986.
79. Esposito, G., Lesk, A.M., Molinari, H., Motta, A., Niccolai, N., Pastore, A. Probing protein structure by solvent perturbation of NMR spectra: III. Combination of experiment and theory. In: "Protein Structure by Distance Analysis." Bohr, H., Brunak, S., eds. Washington DC: IOS Press, 1994:51–63.
80. Cohen, F.E., Sternberg, M.J.E., Taylor, W.R. Analysis and prediction of protein β-sheet structures by a combinatorial approach. Nature 285:378–382, 1980.
81. Goebel, U., Sander, C., Schneider, R., Valencia, A. Correlated mutations and residue contacts in proteins. Proteins 18:309–317, 1994.
82. Neher, E. How frequent are correlated changes in families of protein sequences? Proc. Natl. Acad. Sci. U.S.A. 91:98–102, 1994.
83. Wodak, S.J., Janin, J. Computer analysis of protein-protein interaction. J. Mol. Biol. 124:323–342, 1978.
84. Sippl, M.J., Weitckus, S., Flöckner, H. In search of protein folds. In: "The Protein Folding Problem and Tertiary Structure prediction." Merz, K.H., LeGrand, S., eds. Boston: Birkhäuser Boston Inc., 1994: in press.
85. Brünger, A.T., Clore, G.M., Gronenborn, A.M., Saffrich, R., Nilges, M. Assessing the quality of solution nuclear magnetic resonance structures by complete cross-validation. Science 261:328–331, 1993.
86. Brünger, A.T., Nilges, M. Computational challenges for macromolecular structure determination by x-ray crystallography and solution NMR-spectroscopy. Q. Rev. Biophys. 26:49–125, 1993.
87. O'Donoghue, S.L., Junius, F.K., King, G.F. Determination of the structure of symmetric coiled-coil proteins from NMR data: Application of the leucine zipper proteins Jun and GCN4. Protein Eng. 6:557–564, 1993.
88. Hubbard, S. "Carte-graphics for the Display of Contact Maps." Heidelberg: EMBL, 1994.
89. Wako, H., Blundell, T.L. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins I. Solvent accessibility classes. J. Mol. Biol. 238:682–692, 1994.