PROTEINS: Structure, Function, and Genetics 42:452–459 (2001)

Prediction of Protein Surface Accessibility with Information Theory

Hossein Naderi-Manesh,[1,2*] Mehdi Sadeghi,[1] Shahriar Arab,[2] and Ali A. Moosavi Movahedi[1]
[1]*Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran*
[2]*Department of Biophysics, Faculty of Science, Tarbiat Modarres University, Tehran, Iran*

ABSTRACT **A new, simple method based on information theory is introduced to predict the solvent accessibility of amino acid residues in various states defined by their different thresholds. Prediction is achieved by the application of information obtained from a single amino acid position or pair-information for a window of seventeen amino acids around the desired residue. Results obtained by pairwise information values are better than results from single amino acids. This reinforces the effect of the local environment on the accessibility of amino acid residues. The prediction accuracy of this method in a jackknife test system for two and three states is better than 70 and 60%, respectively. A comparison of the results with those reported by others involving the same data set also testifies to a better prediction accuracy in our case. Proteins 2001;42:452–459.** © 2001 Wiley-Liss, Inc.

Key words: protein structure prediction; solvent accessibility; hydropathy scale; local environment; pairwise information

INTRODUCTION

A knowledge of the information contained in a known protein structure is valuable both to the individual understanding of its function and to the general principles determining protein folding.

It is widely believed that the amino acid (AA) sequence of a protein contains sufficient information to determine its three-dimensional structure.[1] However, the specific mechanisms underlying protein folding still elude our understanding,[2] and a multitude of methods are available or are being developed to improve our ability to predict a protein structure from its AA sequence.

There is an enormous gap between the number of protein structures that have been resolved so far and the huge number of proteins that have been sequenced.[3–5] Consequently, the prediction of a protein structure from its AA sequence is of great interest.

An accurate prediction of the three-dimensional structures of proteins is currently possible for those that enjoy a significant sequence similarity to proteins of known three-dimensional structures.[6] For the remaining sequences, simplified approaches to the problems are inevitably attempted. An extreme form in this case is the prediction of the protein structure in one dimension, such as a characterization of AA residues to adopt one of the secondary structure conformations.[7–9] Another possibility is the characterization of AA surface accessibility, that is, the degree to which a residue in a protein is accessible to a solvent.[10] It has already been shown that in proteins, the hydrophobic free energies are directly related to the accessible surface area of both polar and nonpolar groups.[11] In the final folded structure of a protein, the hydrophilic side-chains have access to the aqueous solvent, but the contact between the hydrophobic side-chains and the solvent is minimized.[12–15] The studies of solvent accessibility in proteins have led to numerous insights into protein structures. Additionally, the prediction of residue accessibility can aid in elucidating the relationship between AA sequence and structure.[16–21]

Residue solvent accessibility often is divided into two states[22,23] (buried and exposed) or even three states[15,24] (buried, intermediate, and exposed), depending on the chosen percentage of the solvent-accessibility threshold. The prediction of solvent accessibility has been performed in a variety of ways, such as sequence alignment, neural network, and statistical analysis of AA composition.[25–28] In this study, we used information theory formalism to calculate the propensity of each residue in the various states of accessibility by considering self- and pair-information as was used in the GOR method for the prediction of secondary structures.[29,30] The predicted accessibility state was the state with the highest positional information value. Single and pair-information values for each possible pair from -8 to $+8$ positions were taken from the database. The relative solvent accessibilities in the various states were predicted with the different thresholds for each state, and the performance accuracy was compared to the previously reported results for the same data set. Similar predictions were made for three or more accessibility states, and a new hydropathy scale based on the characteristic of residues in the two-state model was developed. The prediction of solvent accessibility could be valuable for some applications, such as sequence-motif identification,[31] sequence alignment,[32,33] hydrophobic

Grant sponsors: Research Council of the University of Tarbiat Modarres; Research Council of the University of Tehran; Tarbiat Modarres Molecular Modeling Center; and IBB Bioinformatic Center.

*Correspondence to: Hossein Naderi-Manesh, P.O. Box 14115-111, Department of Biophysics/Biochemistry, TMU Tehran, Iran. E-mail: naderman@modares.ac.ir

Received 1 June 2000; Accepted 13 October 2000

TABLE I. Maximum Surface Accessibility (Max Acc) of the AAs (Å) in Extended β Conformation

AA	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile
Max Acc	188	312	258	234	188	293	233	112	252	257

AA	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Max Acc	243	290	280	265	231	193	217	303	274	228

TABLE II. Protein Data Bank (PDB) Code of the Protein Data Set

119L_	1BNCA	1DKTB	1GGGA	1KNYA	1OFGA	1PUD_	1SVPA	1VLS_	2CHSA	2SNS_
153L_	1BTMA	1DKZA	1GND_	1KPTA	1ONRA	1PYTA	1TADC	1WBA_	2CTC_	2TYSA
1ABA_	1BTN_	1DOSA	1GOTB	1KTE_	1OPR_	1QAPA	1TDX_	1WHI_	2END_	3CHY_
1ABRB	1CEM_	1DXY_	1GPC_	1KUH_	1OSPO	1RA9_	1TFE_	1WHO_	2GDM_	3COX_
1AFRA	1CEO_	1ECEA	1GPL_	1LBA_	1PBC_	1RCF_	1TFR_	1WSYB	2HFT_	3GRS_
1AFWA	1CEWI	1ECPA	1GSA_	1LCL_	1PDA_	1REC_	1THV_	1XGSA	2HHMA	3MDDA
1AMM_	1CFYA	1EDE_	1GTMA	1LKI_	1PDO_	1RGS_	1THX_	1XNB_	2HPDA	3MINB
1AMP_	1CHD_	1EDG_	1HAVA	1LKKA	1PEA_	1RNL_	1TIB_	1XVAA	2I1B_	3NLL_
1AOCA	1CHKA	1EDT_	1HFC_	1LTSA	1PEX_	1RRO_	1TML_	1XYZA	2LIV_	3SDHA
1ATLA	1CHMA	1ERV_	1HGXA	1MAI_	1PGS_	1RSY_	1TUPC	1YASA	2MTAC	5P21_
1ATNA	1CMKE	1ESC_	1HLB_	1MAZ_	1PHE_	1RVAA	1TYS_	1YSC_	2NACA	5PTP_
1AXN_	1CNV_	1EXNB	1HSBA	1MBD_	1PHP_	1SBP_	1UBI_	1YTW_	2PGD_	6GSVA
1BBPA	1CSEE	1EZM_	1HTP_	1MKAA	1PIOA	1SFTB	1UBY_	256BA	2PHLA	6PFKA
1BDO_	1CSGA	1FDS_	1IDAA	1MLDA	1PLC_	1SIG_	1UDII	2ABK_	2PHY_	7RSA_
1BEO_	1CSN_	1FJMA	1IDO_	1MML_	1PMI_	1SLUA	1UXY_	2ARCA	2PIA_	8ATCB
1BFG_	1CYX_	1FTPA	1IFC_	1MOLA	1PNE_	1SMEA	1VCAA	2AYH_	2PSPA	
1BGC_	1DEAA	1FUA_	1IRK_	1NAR_	1POA_	1SMPI	1VHH_	2BBVC	2RN2_	
1BHMB	1DELA	1GAI_	1ITG_	1NBAB	1POC_	1SRA_	1VHRA	2CAE_	2RSPB	
1BIB_	1DFJI	1GCB_	1JKW_	1NOX_	1POT_	1STD_	1VID_	2CBA_	2SCPA	
1BMFG	1DHR_	1GDOA	1KNB_	1NOZA	1PPN_	1STFI	1VIN_	2CCYA	2SIL_	

patches,[34,35] transmembrane-region prediction,[14,17,36] antigenic determinants,[37,38] and protein design.[39,40]

THEORY AND METHODS

We used an information theory similar to the GOR method for the prediction of secondary structures with this distinction: the conformational states were considered relative solvent-accessibility states. This enabled us to determine the propensity of single-residue and pairwise-residue interactions to adopt a conformational state. Naturally, it is necessary to consider the information contained by the neighboring residues on the conformation of a given residue.

The definition of the information that y carries on the occurrence of event x is as follows:

$$I(S = x{:}\bar{x}) = \log \frac{p(S = x|R)}{1 - p(S = x|R)} \log \frac{p(S = x)}{1 - p(S = x)} \quad (1)$$

where $p(S = X)$ is the probability of the occurrence of an event and $p(S = X|R)$ is the conditional probability of $S = X$ if event R has occurred. The complementary event of $S = X$ is $S = \bar{X}$.

The event $S = X$ corresponds to accessibility states of a residue, and the discrimination factor is the sum of the single-residue information (self-information), which depends on only one residue in a local sequence. In a protein structure, the conformation of AA residues may depend on the whole sequence or at least the local sequence. It is, therefore, necessary to consider the information carried by the neighboring residues on the conformation of a given residue.

In a sequence environment of eight residues on either side of a central residue, the preference (informational content) I of a residue with sequence number j and AA type R_j for an accessibility state, for example, type $S\in$ (buried, intermediate, exposed) in a three-state model, is approximated as

$$I(S_j = x{:}\bar{x}; R_{j-m}, \ldots R_j, \ldots, R_{j+m}) \approx I(S_j = x{:}\bar{x}; R_j) + \sum_{j=-8}^{8} (S_j = x{:}\bar{x}; R_{j+m}|R_j) \quad (2)$$

That is called pair-information, the information carried by the residue at $j \pm m$ on the accessibility state of the residue at j on the basis of the type of residue at j and $j \pm m$. If there are enough observations, the frequency ratio is a good approximation for the probability required. For a few observations, an estimation based on Bayesian reasoning of the information parameters was used.

For the three-state prediction, $x \in (B,I,E)$, where B represents the buried residues, I represents the intermediate residues, and E represents the exposed residues. The first term of Equation 2 requires a contingency table with $3 \times 20 = 60$ entries, and for pair-information it needs $20 \times 20 \times 3 = 1{,}200$ entries. The data set used contains about

TABLE III. Hydropathy Scale Based on Self-Information Values in the Two-State Model[†]

	5%	9%	16%	20%	25%	36%	50%
Cys	116	137	169	182	194	224	329
Ile	107	106	104	106	102	83	28
Val	100	108	116	113	111	117	114
Leu	95	103	103	104	103	82	36
Phe	92	108	128	132	131	117	120
Met	78	73	77	82	90	83	62
Trp	59	69	102	118	116	130	179
Ala	58	51	41	32	24	5	−2
Thr	−7	−3	10	20	34	79	174
Gly	−11	−13	−18	−22	−28	−47	−66
Tyr	−11	11	36	44	43	27	−7
Ser	−34	−26	−31	−34	−36	−41	−52
His	−73	−55	−35	−25	−31	−50	−70
Pro	−79	−79	−81	−82	−85	−103	−132
Asn	−93	−84	−74	−73	−76	−77	−97
Asp	−97	−78	−47	−29	0	45	248
Glu	−131	−115	−90	−74	−57	−8	117
Gln	−139	−128	−104	−95	−87	−67	−37
Arg	−184	−144	−109	−95	−79	−57	−41
Lys	−244	−205	−148	−124	−96	−38	115

[†]With different thresholds of accessibility. For 5% accessibility, the scale has been ranked from more hydrophobic (positive value) to more hydrophilic (negative value). AA rankings are different in different accessibility cutoffs.

51,000 residues that corresponds to an average of 850 frequencies for single-residue information and 42 frequencies for pair-information.

The prediction quality was evaluated by the percentage of correctly predicted residues divided by the total number of residues in the data set. For example, for three states we have

$$Q\% = [(NB + NI + NE)/N_{tot}] \times 100$$

where $Q\%$ is the percentage of correctly predicted residues and NB, NI, and NE represent the number of residues correctly predicted as buried, intermediate, and exposed, respectively. The correlation coefficient between the observed (x) and predicted (y) solvent-accessibility states for a data set of N residues was calculated form

$$\text{Correlation coefficient} = \frac{\langle xy\rangle - \langle x\rangle\langle y\rangle}{\sqrt{\langle x^2\rangle - \langle x\rangle^2}\ \sqrt{\langle y^2\rangle - \langle y\rangle^2}}$$

Relative Solvent Accessibility of a Residue

Accessible surface areas for individual atoms of the proteins were calculated from atomic coordinates deposited in the protein data bank with the program devised by our group. For each atom, a sufficiently large number of approximately evenly distributed points were placed on the solvation sphere of radius $R_a + R_w$ centered at the atomic position, where R_a and R_w are the Van der Waals radii of atom A and the solvent probe, respectively.[10,41]

In the absence of hydrogen atoms, group radii were used.[42] Accessible surface areas of individual residues were calculated with the peptide Gly-R-Gly, which has an extended β conformation ($\phi = -140$, $\psi = 135$) and with a fully extended side-chain (Table I). The relative solvent accessibility of each residue in the folded protein was calculated by the surface accessibility being divided by the maximum accessibility of that residue. Here, the relative accessibility was divided over two to nine states with various thresholds.

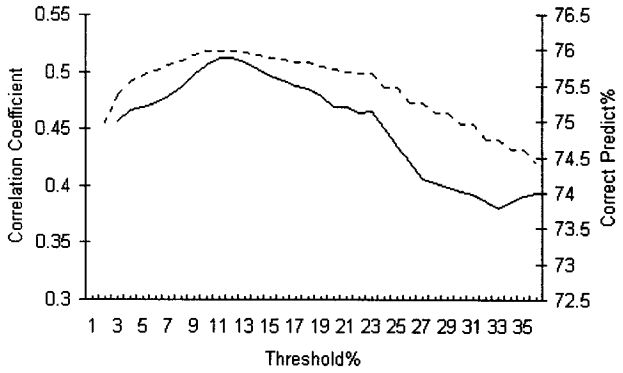

Fig. 1. Dependence of the two-state prediction accuracy and the correlation coefficient on the solvent-accessibility threshold. The results were obtained with information theory over the 215-protein data set. The solid line represents the correct prediction percentage, and the dashed represents the correlation coefficient.

Data Base and Prediction Procedure

A set of 215 protein chains of known three-dimensional structures determined by X-ray crystallography with no more than 25% pairwise-sequence identity, no sequence with length less than fifty residues, and crystallographic resolution >2.5 Å was used (Table II),[43] and a jackknife test was performed on this set. In this method, each protein in the data set was selected as a test protein and was removed from the data set. Informational parameters used in predicting solvent-accessibility states were calculated for the remaining proteins in the data set. This procedure was repeated until all the proteins were tested exactly once. For comparison with the results of other methods, the data set of 126 protein selected by Rost and Sander[44] was used, and the aforementioned procedure was applied.

RESULTS AND DISCUSSION

Hydropathy Scale

Many different scales of hydrophobicity have been determined for AAs.[12–15,22,45] Solution scales are based on the degree of the AA partition coefficient between water and a noninteracting, isotropic phase to calculate the free energy of transfer. Other scales are derived via statistical analysis of the observed distribution of the residues between the solvent-accessible surface and the buried interior in proteins of known structures. These scales in general are qualitative descriptions of the hydropathic behaviors of AAs, and in statistical scales, qualitative disagreements arise because of criterion differences for the determination of residue buriedness.

TABLE IV. Prediction of Solvent Accessibility in Various States[†]

		Pair-information		Self-information	
No. of states	State threshold	Accuracy	Correlation	Accuracy	Correlation
Two states					
	4	75.1	0.49	67.5	0.35
	9	75.9	0.51	70.0	0.38
	16	75.5	0.50	68.5	0.37
	25	74.4	0.47	63.2	0.33
	36	74.1	0.41	58.0	0.22
	49	79.9	0.36	57.1	0.14
	64	97.2	0.46	70.2	0.04
	81	80.5	0.05	63.4	0.01
Three states					
	4;25	49.3	0.39	48.9	0.32
	4;36	57.9	0.43	42.3	0.30
	9;16	62.3	0.42	62.4	0.40
	9;36	57.4	0.41	43.7	0.32
	9;64	74.1	0.47	44.4	−0.27
	16;64	73.7	0.47	35.2	−0.21
Four states					
	9;16;36	45.2	0.32	40.6	0.35
	9;36;49	41.2	0.25	23.0	0.03
	4;16;36	46.4	0.36	36.8	0.35
	4;16;49	51.8	0.37	34.4	0.00
	4;25;49	47.1	0.34	27.4	0.08
Seven states					
	4;9;16;25;36;49	23.7	0.15	16.1	0.10
Nine states					
	4;9;16;25;36;49;64;81	15.3	0.09	6.8	−0.19

[†]The prediction accuracy and correlation coefficient results are based on the use of self- and pair-information values obtained from the data set in two-, three-, four-, seven-, and nine-state models with various accessibility thresholds defined for each state.

Self-information values were calculated with Equation 1 and are listed in Table III for the two-state accessibility models with different thresholds of accessibility for buried and exposed states. Information values show different tendencies for different residues in the core or surface of globular proteins. The order of residues from the most hydrophobic (positive values) to the most hydrophilic (negative values) residues does not agree in all respects and varies with the determination of the accessibility state threshold for the classification of residues in the buried or exposed states.

Prediction of the Data Set

A jackknife test was applied for the prediction. After the test protein was removed, the parameters were recalculated for the remaining data set. This procedure was repeated until the entire data set had been predicted.

A problematic factor in this regard is the choice of solvent-accessibility cutoffs. The obtained results would change with changes in the cutoff levels. Figure 1 shows the effects of the solvent-accessibility threshold on the prediction accuracy and the correlation coefficient for a two-state prediction. As shown, the prediction accuracy and correlation coefficient are threshold-dependent. Therefore, the thresholds for various accessibility state models were selected on the basis of these factors and the distribution of different residues into states.

Use of Pair-Information

The extensive data set chosen allowed us to use the pair-information parameters (Equation 2) if the number of observations was sufficient to give a good estimation of the information values. Therefore, we calculated pair-information parameters and performed a prediction of the data set. A prediction was also made with self-information values (Equation 1). To obtain more detailed information, solvent accessibility was classified into two sets of nine states, each with various cutoffs. The results are shown in Table IV for the whole data set. As expected, the results obtained by pair-information are better than the self-information values. This shows that residue periodicity and pair-interaction can affect the accessibility of the AA residues. Tables V and VI show the accuracy of the prediction for each AA in various states. The buried state is better predicted for hydrophobic AAs, and for hydrophilic residues, the exposed state shows better results. However, the overall predictions over different states are the same.

Table VII shows a comparison of the results obtained by the application of the information theory procedure to the same data set of 126 proteins listed by Rost and Sander.[26] The percentage of correctly predicted residues in two-state and three-state models with the same solvent-accessibility thresholds obtained by information theory is compared with the results of a neural network method based on

TABLE V. AA Accuracy in Different Thresholds in the Two-State Model[†]

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Total
Threshold: 4%																					
No. AA in st 1	1,790	146	328	344	436	179	316	1,126	197	1,640	2,205	80	529	1,010	388	715	698	288	544	1,806	14,765
% pred	88.2	44.5	52.1	40.1	93.6	50.3	39.9	71.2	72.1	95.7	94.1	46.3	90.4	91.4	54.6	60.8	62.3	84.4	66.9	94.2	81.2
No. AA in st 2	2,275	2,288	2,030	2,824	344	1,766	2,953	2,790	933	1,369	2,137	3,078	607	1,077	1,896	2,386	2,211	438	1,438	1,697	36,537
% pred	46.5	97.2	90.1	93.7	44.5	94.2	94.4	68.1	87.8	22.2	25.1	99.0	51.2	36.2	85.6	77.3	76.1	63.9	70.0	27.0	72.7
Threshold: 9%																					
No. AA in st 1	2,207	308	547	579	552	310	582	1,501	340	2,014	2,834	170	656	1,364	569	1,060	989	405	887	2,267	20,141
% pred	88.3	31.5	48.6	32.5	96.2	41.9	41.9	66.7	63.5	95.6	95.0	24.7	90.2	95.2	47.6	62.3	62.1	85.9	76.1	94.7	78.9
No. AA in st 2	1,858	2,126	1,811	2,589	228	1,635	2,687	2,415	790	995	1,508	2,988	480	723	1,715	2,041	1,920	321	1,095	1,236	31,161
% pred	49.7	96.0	88.7	93.9	42.5	93.7	91.7	69.4	82.7	22.7	22.7	99.1	50.8	30.3	86.2	75.7	75.1	63.2	57.4	23.9	73.8
Threshold: 16%																					
No. AA in st 1	2,638	611	842	963	643	534	997	1,938	494	2,350	3,314	410	772	1,633	813	1,411	1,351	537	1,243	2,656	26,150
% pred	87.0	24.5	46.9	32.3	95.5	39.3	41.7	66.0	64.0	95.4	95.4	12.7	88.6	96.1	44.5	60.2	62.3	93.3	83.3	94.8	75.7
No. AA in st 2	1,427	1,823	1,516	2,205	137	1,411	2,272	1,978	636	659	1,028	2,748	364	454	1,471	1,690	1,558	189	739	847	25,152
% pred	51.6	93.9	87.9	92.0	48.2	91.2	89.7	70.4	81.0	24.4	25.1	99.3	51.4	23.8	85.5	75.9	73.2	58.7	47.2	25.4	78.7
Threshold: 25%																					
No. AA in st 1	3,027	1,061	1,261	1,514	710	849	1,697	2,477	687	2,627	3,747	963	899	1,832	1,135	1,873	1,778	634	1,568	2,968	33,307
% pred	84.5	30.4	48.7	39.0	97.2	37.7	51.1	65.4	68.9	95.7	94.8	12.6	91.4	96.7	41.8	63.6	62.2	96.1	88.6	93.9	72.4
No. AA in st 2	1,038	1,373	1,097	1,654	70	1,096	1,572	1,439	443	382	595	2,195	237	255	1,149	1,228	1,131	92	414	535	17,995
% pred	58.7	93.4	85.4	90.2	68.6	91.4	82.8	71.6	77.2	31.2	29.7	97.3	64.6	32.9	84.9	73.0	72.1	62.0	44.9	29.3	76.7
Threshold: 36%																					
No. AA in st 1	3,485	1,633	1,792	2,258	753	1,321	2,648	3,043	865	2,811	4,042	1,868	1,011	1,954	1,532	2,429	2,325	679	1,776	3,237	41,462
% pred	82.9	40.8	61.8	49.0	99.5	44.4	71.9	61.8	71.9	94.0	93.3	21.9	89.3	95.2	43.4	66.9	66.9	97.6	89.5	93.1	72.9
No. AA in st 2	580	801	566	910	27	624	621	873	265	198	300	1,290	125	133	752	672	584	47	206	266	9,840
% pred	64.8	89.6	80.7	83.8	92.6	86.4	71.8	76.5	82.3	47.0	41.3	94.0	72.0	54.9	87.8	75.9	71.1	91.5	59.7	46.2	78.0

[†]The percentage correctly predicted for each AA (% pred), the number of AAs (No. AA) in each state (st), and the total prediction result in the two-state model with various accessibility thresholds.

TABLE VI. AA Accuracy in Different Thresholds in the Three-State Model[†]

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Total
Threshold: 9;16%																					
No. AA in st 1	2,207	308	547	579	552	310	582	1,501	340	2,014	2,834	170	656	1,364	569	1,060	989	405	887	2,267	20,141
% pred	58.4	10.7	18.1	12.6	46.6	13.9	19.4	33.7	21.2	65.7	72.7	2.9	36.4	63.0	14.4	26.5	23.7	51.1	41.1	64.8	47.7
No. AA in st 2	431	303	295	384	91	224	415	437	154	336	480	240	116	269	244	351	362	132	356	389	6,009
% pred	38.3	51.2	51.2	46.6	76.9	58.9	47.2	45.8	74.7	44.6	34.6	34.2	62.1	55.4	54.9	53.3	58.8	77.3	71.9	40.4	50.4
No. AA in st 3	1,427	1,823	1,516	2,205	137	1,411	2,272	1,978	636	659	1,028	2,748	364	454	1,471	1,690	1,558	189	739	847	25,152
% pred	64.3	87.5	83.8	85.3	73.0	87.7	83.7	74.3	79.9	47.3	39.1	95.2	76.9	43.8	84.8	77.7	75.0	67.7	48.3	45.2	76.7
Threshold: 9;36%																					
No. AA in st 1	2,207	308	547	579	552	310	582	1,501	340	2,014	2,834	170	656	1,364	569	1,060	989	405	887	2,267	20,141
% pred	60.8	11.0	21.0	15.2	93.5	13.2	26.3	35.8	33.8	74.7	78.3	5.9	61.0	73.5	15.1	35.1	34.5	79.5	53.6	69.6	55.9
No. AA in st 2	1,278	1,325	1,245	1,679	201	1,011	2,066	1,542	525	797	1,208	1,698	355	590	963	1,369	1,336	274	889	970	21,321
% pred	32.1	61.8	68.8	62.5	54.2	60.8	81.2	38.3	67.0	24.7	22.0	52.5	49.3	35.6	47.2	54.8	57.9	70.1	62.2	22.0	52.3
No. AA in st 3	580	801	566	910	27	624	621	873	265	198	300	1,290	125	133	752	672	584	47	206	266	9,840
% pred	71.9	73.4	68.9	69.3	92.6	73.7	49.6	78.2	77.4	65.7	55.0	81.0	84.8	69.9	81.9	71.4	64.6	91.5	60.2	64.7	71.7
Threshold: 4;36%																					
No. AA in st 1	1,790	146	328	344	436	179	316	1,126	197	1,640	2,205	80	529	1,010	388	715	698	288	544	1,806	14,765
% pred	61.8	19.2	26.2	19.5	90.8	21.2	20.9	37.0	37.6	77.5	80.0	11.3	61.2	75.0	20.4	34.5	36.0	79.2	44.5	74.8	59.6
No. AA in st 2	1,695	1,487	1,464	1,914	317	1,142	2,332	1,917	668	1,171	1,837	1,788	482	944	1,144	1,714	1,627	391	1,232	1,431	26,697
% pred	30.0	58.8	69.4	60.7	52.7	59.4	80.7	38.4	70.4	25.1	26.6	47.7	51.0	43.1	42.7	55.6	57.8	68.3	73.1	29.4	51.5
No. AA in st 3	580	801	566	910	27	624	621	873	265	198	300	1,290	125	133	752	672	584	47	206	266	9,840
% pred	71.0	75.0	69.6	73.3	96.3	77.1	54.9	78.2	78.5	62.1	48.3	84.2	83.2	63.2	84.0	73.2	66.6	93.6	55.3	57.5	73.0
Threshold: 4;25%																					
No. AA in st 1	1,790	146	328	344	436	179	316	1,126	197	1,640	2,205	80	529	1,010	388	715	698	288	544	1,806	14,765
% pred	66.0	17.8	21.3	18.6	80.7	15.1	19.9	42.5	35.5	78.7	79.8	8.8	56.1	70.6	17.8	33.4	32.4	66.3	38.1	72.3	58.5
No. AA in st 2	1,237	915	933	1,170	274	670	1,381	1,351	490	987	1,542	883	370	822	747	1,158	1,080	346	1,024	1762	18,542
% pred	28.6	47.3	56.3	47.1	58.0	52.2	59.5	36.4	70.6	32.9	32.0	30.4	56.8	52.8	42.7	51.2	51.6	73.4	77.8	36.7	47.0
No. AA in st 3	1,038	1,373	1,097	1,654	70	1,096	1,572	1,439	443	382	595	2,195	237	255	1,149	1,228	1,131	92	414	535	17,995
% pred	67.2	83.2	76.2	80.8	70.0	85.2	72.5	72.3	71.6	38.0	33.3	92.6	76.4	40.8	81.9	71.7	69.2	68.5	39.4	38.1	73.3

[†]The percentage correctly predicted for each AA (% pred), the number of AAs in each state, and the total prediction result in the three-state model with various accessibility thresholds.

TABLE VII. Comparison of the Solvent Accessibility Predictions in the 126-Protein Data Set†

No. of states	Threshold (%)	Percentage correct IT[a]	BT[b]	NN[c]
Two states				
	9	78.2	72.8	71.4
	16	77.5	71.1	70
	23	77.4	70	
Three states				
	9;36	61.5	54.2	52.4

†The predictions of solvent accessibility in two and three states for the threshold reported on the 126-protein data set of Rost and Sander[43] are compared.
[a]Information theory described in this work (Equation 2).
[b]Bayesian probabilistic prediction of Thompson and Goldstein.[27]
[c]Neural network prediction of Rost and Sander.[25]

TABLE VIII. A Comparison of the Prediction Base From DSSP and Our Program for Surface Accessibility Calculations†

No. of states	Threshold (%)	Accuracy (%) Our program	DSSP
Two states			
	4	75.1	69.5
	9	75.9	70
	16	75.5	69.7
	25	74.4	69.1
	36	74.1	68.9
Three states			
	4;36	57.9	53.9
	9;16	62.3	58.1
Four states			
	9;16;36	45.2	39.3

†The prediction accuracy obtained with the DSSP program for surface accessibility calculations is compared.

single-sequence data by Rost and Sander. Furthermore, the results obtained by Thompson and Goldstein[28] via Bayesian statistics on this data set are compared. As shown in Table VII, the results achieved by information theory (IT) are superior to those obtained by neural network (NN) and Bayesian theory (BT) for the same data set with the same accessibility thresholds for two-state and three-state models. We used a homemade algorithm instead of DSSP (Definition of the Secondary Structure of Protein)[46] for accessible surface calculations, which could be part of our improvement over the neural network method. To check the effect of this algorithm, we used data obtained from DSSP; the results are shown in Table VIII. This comparison was performed on the 215-protein data set, and our method shows a 5% improvement in accuracy.

REFERENCES

1. Anfinsen CB. Principles that govern the folding of protein chains. Science 1973;181:223–230.
2. Lesk AM. Computational molecular biology. In: Kent A, Williams GJ, Hall CM, Kent R, editors. Encyclopedia of computer science and technology. Volume 31, Supplement 16. New York: Marcel Dekker; 1994. p 101–165.
3. Bernstein FC, Koetzle TF, Williams GJ, Meyer EFJ, Brice MD, Rodgers JR, Kenneard O, Shimanochi T, Tasumi M. The protein data bank: A computer based archival file for macromolecular structures. J Mol Biol 1977;112:532–542.
4. Bairoch A, Boeckmann B. The SWISS-PROT protein sequence data bank. Nucleic Acid Res 1992;20:2019–2022.
5. Oliver SG. The complete DNA sequence of yeast chromosome III. Nature 1992;357:38–64.
6. Mosimann S, Meleshko R, James MNG. A critical assessment of comparative molecular modeling of tertiary structures of proteins. Proteins 1995;23:301–317.
7. Rost B, Sander C, Schneider R. Redefining the goals of protein secondary structure prediction. J Mol Biol 1994;235:13–26.
8. Cohen BI, Cohen FE. Prediction of protein secondary and tertiary structure. New York: Academic; 1994. 430 pp.
9. Barton GJ. Protein secondary structure prediction. Curr Opin Struct Biol 1995;5:372–376.
10. Lee BK, Richards FM. The interpretation of protein structure: Estimation of static accessibility. J Mol Biol 1971;55:379–400.
11. Ooi T, Oobatake M, Nemethy G, Scheraga HA. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. Proc Natl Acad Sci U S A 1987;84:3086–3090.
12. Chothia C. The nature of accessibility and buried surface in proteins. J Mol Biol 1976;105:1–14.
13. Wolfender R, Anderson L, Cullis PM, Soulhgate CCB. Affinities of amino acid side chains for solvent water. Biochemistry 1981;20: 849–855.
14. Kyte J, Doo Little RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132.
15. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. Hydrophobicity of amino acid residues in globular proteins. Science 1985;229:834–838.
16. Richmond TJ. Solvent accessible surface area and excluded volume in proteins. J Mol Biol 1984;178:63–89.
17. Eisenberg D, Schwartz E, Komaromy M, Wall R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J Mol Biol 1984;179:125–142.
18. Rao MJK, Argos P. A conformational preference parameter to predict helices in integral membrane proteins. Biochim Biophys Acta 1986;889:197–214.
19. Hubbard TJ, Blundell TL. Comparison of solvent-inaccessible cores of homologous proteins: Definitions useful for protein modeling. Protein Eng 1987;1:159–171.
20. Degli Esposti M, Crimi M, Venturoli GA. A critical evaluation of the hydropathy profile of membrane proteins. Eur J Biochem 1990;190:207–219.
21. Jones DT, Thornton JM. Potential energy functions for threading. Curr Opin Struct Biol 1996;6:195–209.
22. Janin J. Surface and inside volume in globular proteins. Nature 1979;227:491–492.
23. Miller S, Janin J, Klesk AM, Chothia C. Interior and surface of monomeric proteins. J Mol Biol 1987;196:640–656.
24. Sander C, Scharf M, Schneider R. Design of protein structure. In: Rees AR, Sternberg MJE, Wetzel R, editors. Protein engineering. Oxford: IRL; 1992. p 82–115.
25. Holbrook SR, Muskal SM, Kim SH. Predicting surface exposure of amino acids from protein sequences. Protein Eng 1990;3:659–665.
26. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
27. Wako H, Blundell T. 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 1994;238:682–692.
28. Thompson MJ, Goldstein RA. Predicting solvent accessibility: Higher accuracy using Bayesian statistics and optimized residue substitution classes. Proteins 1996;25:38–47.
29. Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implication of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 1978;120: 97–120.
30. Gibrat JF, Garnier J, Robson B. Further developments of protein secondary structure prediction using information theory. J Mol Biol 1987;198:425–443.
31. Cornette JL, Cease KB, Margalit H, Spouge JL, Berzofsky JA, Delisi C. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. J Mol Biol 1987;195: 659–685.
32. Gaboriand C, Bissery V, Benchetrit T, Mornon JP. Hydrophobic

cluster analysis: An efficient new way to compare and analysis amino acid sequences. FEBS Lett 1987;224:149–155.
33. Lemesle-Varloot L, Henrissat B, Gaboriand C, Bissery V, Morgat JP. Hydrophobic cluster analysis: Procedures to derive structural and functional information from 2-D representation of protein sequences. Biochimie 1990;72:555–574.
34. Eisenhaber F, Argos P. Hydrophobic region on protein surface: Definition based on hydration shell structure and a quick method for their computation. Protein Eng 1996;9:1121–1133.
35. Lijnzaad P, Berendsen HJC, Argos P. A method for detecting hydrophobic patches on protein surface. Proteins 1996;26:192–203.
36. Engelman DM, Steitz TA, Goldman A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu Rev Biophys Biophys Chem 1986;15:321–353.
37. Both GW, Sleigh MJ. Complete nucleotide sequence of the haemagglutinin gene from a human influenza virus of the Hong Kong subtype. Nucleic Acids Res 1980;8:2561–2575.
38. Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci U S A 1981;78: 3824–3828.
39. Baumenn G, Frommel C, Sander C. Polarity as a criterion in design. Protein Eng 1989;2:329–343.
40. Sippl MJ. Recognition of errors in three-dimensional structures of proteins. Proteins 1993;17:355–362.
41. Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J Mol Biol 1973;79:351–371.
42. Pauling LC. The nature of the chemical bond. 3rd ed. New York: Cornell University Press; 1960. 644 pp.
43. Hobohm U, Sander C. Enlarged representative set of protein structure. Protein Sci 1994;3:522–524.
44. Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 1994;19: 55–72.
45. Nozaki Y, Tanford C. The solubility of amino acids and two glycin peptides in aqueous ethanol and dioxane solutions: Establishment of a hydrophobicity scale. J Biol Chem 1971;246:2211–2217.
46. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bond and geometrical features. Biopolymer 1983;22:2577–2637.