PROTEINS: Structure, Function, and Genetics 36:340–346 (1999)

Role of Evolutionary Information in Predicting the Disulfide-Bonding State of Cysteine in Proteins

Piero Fariselli,[1] Paola Riccobelli,[1] and Rita Casadio[1,2]*
[1]*Laboratory of Biocomputing, Centro Interdipartimentale per le Ricerche Biotecnologiche (CIRB), Bologna, Italy*
[2]*Laboratory of Biophysics, Department of Biology, University of Bologna, Bologna, Italy*

ABSTRACT **A neural network-based predictor is trained to distinguish the bonding states of cysteine in proteins starting from the residue chain. Training is performed by using 2,452 cysteine-containing segments extracted from 641 nonhomologous proteins of well-resolved three-dimensional structure. After a cross-validation procedure, efficiency of the prediction scores were as high as 72% when the predictor is trained by using protein single sequences. The addition of evolutionary information in the form of multiple sequence alignment and a jury of neural networks increases the prediction efficiency up to 81%. Assessment of the goodness of the prediction with a reliability index indicates that more than 60% of the predictions have an accuracy level greater than 90%. A comparison with a statistical method previously described and tested on the same database shows that the neural network-based predictor is performing with the highest efficiency. Proteins 1999;36:340–346.**

Key words: multiple sequence alignment; neural networks; structure prediction; cysteine redox state prediction

INTRODUCTION

The tertiary folds of native proteins are defined by a large number of weak interactions such as hydrogen bonding, hydrophobic interactions, salt bridges, and weakly polar interactions. In addition to these noncovalent forces, certain proteins are also stabilized covalently by disulfide bridges formed by uniquely paired cysteine residues in the folded state. Reduction of disulfide bridges triggers functionally relevant conformational changes.[1]

The contribution of the disulfide bridge to the thermodynamic stability of proteins has been described as being due to a reduction of the conformational entropy of the unfolded polypeptide chain causing a destabilization of the unfolded state relative to the native state (for review, see Ref. 2), and it can be both experimentally[3,4] and theoretically estimated.[5] Several analyses of the characteristics of disulfide bridges in proteins have been performed, including structural and sequence features and classification of connectivity (see Ref. 6 and references therein). The disposition of cysteine residues relative to each other and relative to protein secondary structure is important in the classification of the structure of small disulfide-rich irregular proteins.[7]

Few studies have addressed so far the important problem of predicting the bonding state of cysteine in a protein chain. The correct prediction of this state can help in predicting ab initio the three-dimensional structure of proteins by adding structural constraints. The relevance of the flanking residues in predicting a cysteine bonding state has been demonstrated by using statistical[8] and neural network-based methods[9] with a much smaller database than we used in the present study.

We train and test with a cross-validation procedure a neural network system on a database of 2,452 cysteine-containing segments (34% of which contains half cystines) to distinguish between bonded and unbonded cysteine. The effect of evolutionary information not taken into account before is also investigated.

METHODS

The Database

Two thousand four hundred fifty-two segments containing cysteines (free and disulfide bonded [half cystines]) were taken from the crystallographic data of the Brookhaven Protein Data Bank. Disulfide bond assignment was based on the Define Secondary Structure of Proteins (DSSP) program.[10] Nonhomologous proteins (with an identity value $<25\%$) were selected by using the PDB_select_oct_97 algorithm (http://www.embl-heidelberg.de). Segments that in some residue position lack the corresponding atomic coordinates in the PDB file and segments whose cysteines are interchain disulfide bonded are not included in the database. After this filtering procedure, the total number of examples out of 641 proteins was 2,452, 842 of which were in the disulfide-bonded state and 1,610 of which were in the nondisulfide-bonded state. The PDB codes of the proteins whose cysteine-containing segments are included in the database are listed in Appendix A.

Grant sponsor: Italian Centro Nazionale per le Ricerche (target project Biotechnology); Grant sponsor: Ministero della Università e della Ricerca Scientifica e Tecnologica (project "Biocatalisi e Bioconversioni").
*Correspondence to: Rita Casadio, Department of Biology, University of Bologna,Via Irnerio 42, 40126 Bologna, Italy.
E-mail: casadio@ kaiser.alma.unibo.it
Received 7 January 1999; Accepted 8 April 1999

The Neural Network-Based Predictor

Standard feed-forward neural networks are implemented with a back-propagation algorithm as learning procedure.[11] The network architecture consists of a perceptron without hidden layers, with two output nodes (discriminating the disulfide and free cysteine propensities, respectively). A cysteine residue, flanked by symmetrical segments of different length (from three to eight residues) is classified as disulfide bridge-forming or free, depending on the relative values of the network output neurons. To compensate for the disproportion between disulfide bond forming and free cysteines during the training phase, learning was accomplished by means of a procedure including a balancing probability factor to reduce the number of back propagation cycles for the most abundant class.[12] This was performed to minimize overprediction of free cysteines. Because of the limited number of examples presently available, an early learning stopping procedure is used to train the networks.[12]

Eight different input codings to the networks (N) are considered. One is based on single-sequence input (NSS), and the remaining seven are based on multiple-sequence profile (NMS). In the former case, only the cysteine flanking residues are taken into consideration by removing the cysteine from the center of the input windows, of variable length from 7 to 17 residues. This procedure is similar to that previously adopted,[9] and it is used to simplify the computation of the network junctions. Indeed, being cysteine always present in the central position of the segment, it does not carry any information. With single-sequence input, each residue is encoded as a vector of 21 elements, with all elements set to 0 but one, set to 1, whose position in the vector identifies the particular residue type. Twenty elements encode for the 20 amino acids, and the last one provides a signal when the input window overlaps either the C or N terminus of the protein.

When evolutionary information is presented to the network, coding is performed by using a sequence profile for the cysteine- containing segments taken from the HSSP files of the corresponding proteins.[13] Also in this case, each residue is encoded by a 21-element vector as in the former case, with the difference that each of the first 20 elements represents the frequency of residue in the sequence alignment. When a multiple- sequence profile is used, the central cysteine is taken into account. This is based on the observation that cysteines can be more or less conserved in the profile.

Alternative input codings based on multiple-sequence profile (Network Multiple Sequence, NMS) are:

- NMS+C (Charge): this network adds to the multiple-sequence profile explicit information regarding the charges in the neighboring sequence environment in the form of two more input neurons for each residue in the window. These neurons depending on the amino acidic charge are set respectively to 1 and 0 for positive, 0 and 1 for negative, and 0 and 0 for noncharged residues.
- NMS+H (Hydrophobicity): this network uses as input a matrix based on the hydrophobicity profile. Elements of the matrix are the values contained in the multiple sequence profile derived from the HSSP files[13] and multiplied by the hydrophobicity value of each residue. For this, we used different scales, but no significant differences have been observed and the data presented refer to the Rose's hydrophobicity scale.[14]
- NMS+WE (conservation Weight and relative Entropy): in this network two more neurons are added for each residue in the input window. One accounts for the conservation weight, and the other represents the relative entropy of each position in the multiple-sequence profile as computed by MaxHom and present in the HSSP files.[13]
- NMS+WEC: this network combines the input described for NMS+C and NMS+WE.
- NMS+WEH: this network combines the input described for NMS+H and NMS+WE.
- NMS+WECH: this network combines the input described for the three networks NMS+C, NMS+H and NMS+WE.

Statistical evaluation of the predictor efficiency is scored by computing the network accuracy (Q3), the correlation coefficient (C), and the probability of correct predictions (Pc) both for the disulfide bridge-forming and free cysteines (Pc(SS) and Pc(SH), respectively) (for the definition of the statistical indices see Appendix B and also Fariselli et al.[12]).

The predictor is validated with a cross-validation procedure, which is performed by splitting the whole set of segments of the database into 20 subsets containing an approximate equal number of examples (with the same proportion of disulfide bridge-forming and free cysteines). One subset at the time is removed from the training set and used as testing set. Identity between the segments of the training and testing sets is carefully kept $\leq$30%. For the evaluation of the statistical indices, the predictions of the 20 different networks are summed up, and the standard deviation for the network accuracy (Q3) is computed, assuming a binomial distribution of the assignments.

RESULTS AND DISCUSSION

The Predictor at Work

In Table I the results obtained with the neural network-based predictor using single sequence as input are listed depending on the window length presented to the network. It is evident that the discriminating capability of the predictor between the two different bonding states of cysteine is as high as 72% with a 13-residue-long window. This is also confirmed by the values of the other statistical indices computed to evaluate the network performance.

TABLE I. Performance of the Predictor Using Single-Sequence-Based Input

Window length	Q3 (%)[a]	C	Pc (SS)	Pc (SH)
7	69.3 (0.9)	.33	.63	.72
9	70.4 (0.9)	.38	.64	.73
11	71.4 (0.9)	.40	.66	.73
13	71.8 (0.9)	.41	.67	.74
15	71.1 (0.9)	.40	.66	.73
17	70.0 (0.9)	.35	.60	.75

[a]Q3 (%) = percentage accuracy of the prediction; the standard deviation (within brackets) is computed assuming a binomial distribution of the assignments. C = correlation coefficient; Pc = probability of correct predictions (for the definitions of the statistical indices, see Appendix A).

TABLE II. Performance of the Predictor Using Multiple-Sequence Alignment-Based Input

Window length	Q3 (%)[a]	C	Pc (SS)	Pc (SH)
7	74.9 (0.9)	.43	.65	.79
9	76.5 (0.9)	.47	.67	.81
11	77.7 (0.8)	.50	.68	.82
13	77.5 (0.8)	.50	.68	.82
15	77.9 (0.8)	.50	.69	.82
17	78.2 (0.8)	.51	.70	.82

[a]See Table I for definitions.

TABLE III. Predictive Performance of Different Networks Trained With Multiple-Sequence-Based Input

Method[a]	Q3 (%)	C	Pc (SS)	Pc (SH)
NMS + C	78.3 (0.8)	0.51	0.71	0.81
NMS + H	78.3 (0.8)	0.51	0.70	0.82
NMS + WE	79.8 (0.8)	0.55	0.71	0.84
NMS + WEC	79.9 (0.8)	0.56	0.70	0.84
NMS + WEH	80.2 (0.8)	0.55	0.71	0.84
NMS + WECH	80.1 (0.8)	0.55	0.71	0.84
JURY	81.0 (0.8)	0.57	0.72	0.85

[a]MS = neural network with multiple sequence input. C = with residue charge information. H = with hydrophobicity profile. WE = with conservation weight and relative entropy as taken from the HSSP files.[13] JURY = the jury among the six trained networks of this table. For notation, see Table I.

The results confirm the observation that locally surrounding amino acids greatly influence cysteines in forming disulfide bridges. With the neural network-based predictor the local environment-dependent features are best discovered when cysteine is centered in a 13-residue-long segment.

Different types of network architectures were also tested, including networks with a hidden layer comprising from 2 to 6 neurons. This did not improve the predictor efficiency (data not shown) compared with that obtained with the perceptron without hidden layers. Indeed, the generalization capability of the network was progressively decreased by the increasing number of hidden neurons in the hidden layers, as previously noticed for small training sets as the one used in this study. An exploration of the effect of the number of examples presented to the perceptron without hidden layers (with a 13-residue-long window) (data not shown) indicates that when the training set is 50% and 75% reduced, the efficiency scored as network accuracy is 68% and 67%, respectively. However, the correlation coefficient values are drastically reduced to 0.34 (for the 50% reduced training set) and to 0.27 (for the 65% reduced training set), indicating that the network tends to perform similarly to a random predictor (C = 0). These results indicate that the size of the database used in the present work (and presently available) can be considered a lower limit to start with for predicting the bonding state of cysteine in proteins with neural networks.

It has been clearly shown that the evolutionary information embodied in a sequence profile can significantly improve protein structure predictions. This has been demonstrated for the prediction of secondary structures,[15] of solvent accessible surface,[16] and of transmembrane helices in membrane proteins.[17,18] In this work, evolutionary information is provided to the networks to test whether the discriminating capability between disulfide bond-forming and free cysteines is also affected by residue conservation in the local environment.

Using multiple-sequence alignment as input to the networks, the efficiency of the predictor improves by 6% (Table II). This is confirmed also by the increase of the values of the correlation coefficient and of the probability of correct predictions. Moreover, a new interesting result is that the performance of the prediction is independent of the window size ranging from 11 to 17 residues. It can be concluded that with the number of examples in the database, the addition of evolutionary information saturates the network performance already with an 11- residue-long window and that increasing the window size neither deteriorates nor diminishes the generalization capability of the system.

Alternative Input Codings

It is well known (and in it has confirmed also in the previous paragraph) that the form of the input coding plays a key role in the neural network performance.[15–18] In this respect we have tried to increase the input information content by using alternative codings. The first takes into account the charges in the cysteine environment, the second adds also the hydrophobicity profile, and the third uses two more input neurons for each residue belonging to the input window, representing the conservation weight and the relative entropy of the residue position in the multiple sequence alignment.[13] Furthermore, three other input codings combining those described above, are also implemented. In Table III the best results obtained with these alternative codings with input window lengths rang-

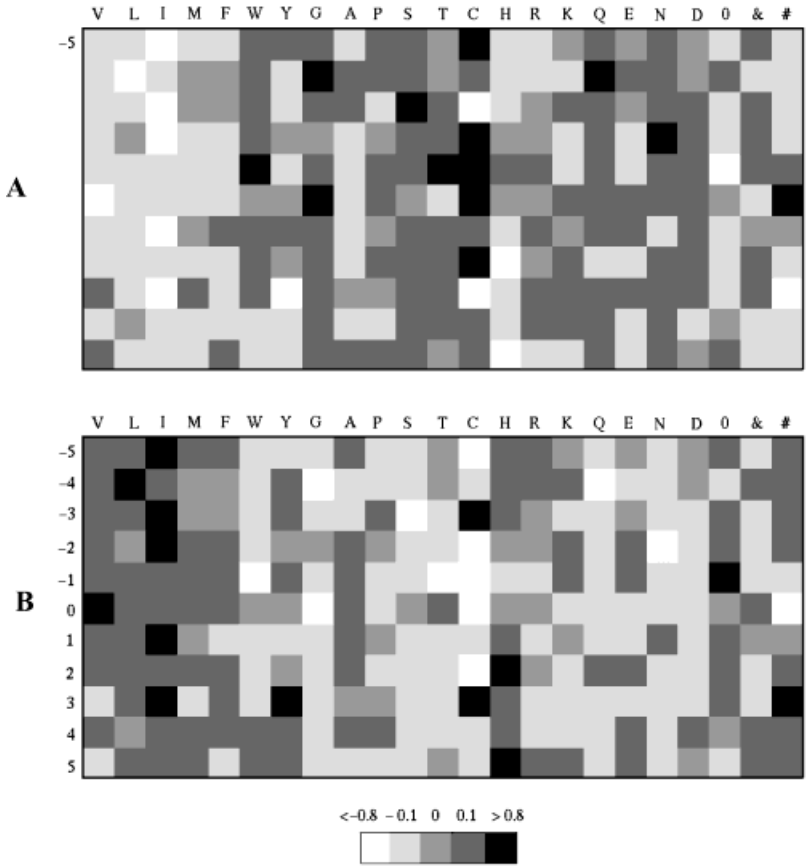

Fig. 1. Graphical representation of the values of the weight junctions averaged over the 20 networks used (values are scaled to a shade of black). Junctions are between the input window of 11 residues (shown along the vertical axis) and the network output. **A:** Propensity for the disulfide-bonding state; **B:** Propensity for the nonbonding state. Labels of the residues (single letter code) are placed horizontally; 0 represents the border condition; & and # represent the entropy and the conservation weight, respectively.

ing from 17 to 21 residues are listed. It is evident that neither the addition of information of the charge nor that of the hydrophobicity profile of the cysteine environment improves the network performance compared with that obtained by using as input the multiple-sequence profile alone (Table II). However, an improvement of 2% is obtained when the conservation weight and the relative entropy are explicitly taken into account. A graphical depiction of the weights from each amino acid at each window position averaged over the 20 different training sets, including the border condition, the entropy value, and the conservation weight is shown in Figure 1, both for the disulfide (A) and the nondisulfide nodes (B). In the plots, scaled to a shade of black, dark squares indicate positive weights (strong propensity for the bonded (A) and nonbonded (B) states), and light ones indicate negative weights (weak propensity for the bonded (A) and nonbonded (B) states). The inspection of the weight values under the conditions of maximal network performance highlights the following: (a) the presence of cysteine residues in the environment of the central cysteine strongly favors the disulfide bond formation, with the exception of positions ± 3. This is in agreement with the fact that metal binding cysteines are typically found in proteins in position i and i ± 3; (b) hydrophilic and/or charged residues in the environment are highly conducive toward disulfide bond formation compared with hydrophobic residues that are poorly conducive; (c) the entropy is lower for the environment of the nonbonded cysteines; and (d) the cysteine in the central position is highly conserved for the disulfide bond-forming cysteines.

As previously shown by other authors,[15,19] a jury of networks improves the prediction accuracy. This is so also when a jury of the six different networks previously described is used and the performance increases to 81%.

Our predictive method is also evaluated by measuring the reliability of the prediction. This can be estimated by computing the reliability index that relates the absolute value of the difference between the two output values of the network with the efficiency of the prediction (Q3).[15–18] In the case of the jury, the values are averaged among the different concurring networks. In Figure 2, the accuracy of the prediction, together with the percentage of the examples in the database whose prediction is characterized by a certain value of the reliability index, are plotted as a function of the value of the reliability index itself. When the jury of networks based on evolutionary information is used, more than 60% of the cysteine containing segments of the database is predicted with an accuracy of 90%.

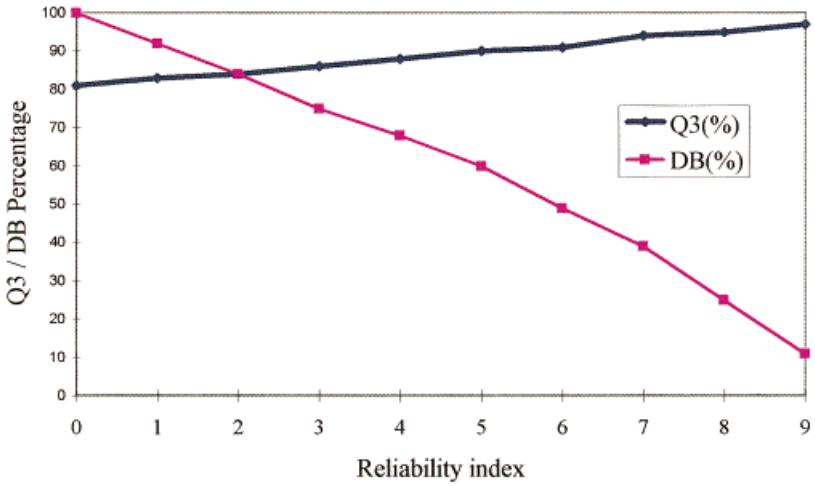

Fig. 2. The reliability of the prediction of the neural network based predictor using evolutionary information as input. DB(%) = percentage of segments in the database. Q3(%) = percentage efficiency of the predictions.

TABLE IV. Predictive Performance of Different Methods on the Data Base

Method	Q3 (%)	C	Pc (SS)	Pc (SH)
MR	68.1 (0.9)	.36	.50	.81
NSS	71.8 (0.9)	.41	.67	.78
JURY	81.0 (0.8)	.57	.72	.85

MR = a statistical method implemented as described in Ref. 8 and using our database. NSS = neural network with single-sequence input. JURY = the network jury of Table III. For notation, see Table I.

Comparison With Other Methods

To compare our predictor with a statistical method previously described,[8] we implemented the same method using our database. This is promoted by the fact that a direct comparison of our results with those reported in Reference 8 is hampered by the different and smaller database previously used. The statistical method is based on the compilation of two matrices representing the statistical frequencies for the residues along the cysteine flanking regions for both disulfide bond-forming and free cysteines. These frequency matrices are used to compute another matrix (MR), whose elements are taken as the ratio of each position of the frequency matrices of the disulfide bond-forming and free cysteines.[8] Provided that the segment whose central cysteine state is to be predicted is not included in the compilation of the frequency matrix, it is possible to evaluate the prediction of the cysteine-bonding state by computing the product of each element in MR associated to each residue in the segment.[8] If the product is >1, the cysteine in the segment is predicted to form a disulfide bridge; otherwise, the cysteine in the segment is assigned to the free cysteine class.

After a cross-validation on the database, the results obtained with the statistical method are compared with those obtained with the neural network approach (Table IV). For comparison, the efficiency obtained by using single sequence and the jury based on multiple-sequence inputs are also shown. It is evident that the neural network-based method scores higher than the statistical one.

Unfortunately, a direct comparison of our predictor with that previously described, also based on neural networks and segment single sequence, is not possible because of a lack of cross- validation in the testing procedure adopted by the authors.[9]

CONCLUSIONS

In this study we show that a neural network-based predictor is capable of discriminating the disulfide bridge-forming potential of cysteine residues by weighting the effect of the local environment with and without evolutionary information. The results indicate that the jury of neural networks using as input sequence profile perform well (with an efficiency equal to 81%, which is 16% higher than that obtained with a random predictor) and that the accuracy of the prediction for the 60% of the database used is extremely good (90%). The neural network system scores higher than a statistical method implemented and tested with a cross-validation procedure on the same database. The neural network predictor here described can therefore provide a useful tool for protein modeling and protein engineering.

The software is available upon request from the authors.

ACKNOWLEDGMENTS

This work was supported partially by a grant for a target project in Biotechnology from the Italian Centro Nazionale per le Ricerche (C.N.R.) and by a grant from Ministero della Università e della Ricerca Scientifica e Tecnologica (MURST) delivered to the project "Biocatalisi and Bioconversioni."

APPENDIX A. The Database of Proteins Containing Cysteines

153l_	1bgw_	1cpt_	1edg_	1gof_	1iso_	1mdaH	1ounA	1rcy_	1thjA	1xvaA	2gstA	2vil_
1aa6_	1bhmA	1crkA	1edhA	1gotB	1itg_	1mhcA	1ovaA	1rec_	1thtA	1xyzA	2hbg_	3bcl_
1aa8A	1bia_	1crl_	1efnB	1gowA	1ivd_	1mhyG	1oxa_	1regX	1thv_	1yasA	2hft_	3chy_
1aayA	1bip_	1csh_	1eft_	1gplA	1jacA	1mkaA	1oyc_	1reqA	1thx_	1yppA	2hhmA	3cla_
1ab8A	1bmfA	1csn_	1emk_	1gpb_	1japA	1mla_	1pax_	1rfbA	1tiv_	1ytw_	2hlcA	3cox_
1ad2_	1bmfD	1ctn_	1eny_	1gpc_	1jcv_	1mml_	1pbe_	1rga_	1tlk_	1yua_	2hmzA	3dfr_
1ad3A	1bmfG	1ctt_	1epaB	1gph1	1jer_	1mpp_	1pbn_	1rgs_	1tml_	1zia_	2hpdA	3grs_
1adeA	1bmtA	1cxsA	1eriA	1gpl_	1jon_	1mrj_	1pbwA	1rie_	1tnrA	1znbA	2hpeA	3minA
1aep_	1bncA	1cydA	1erw_	1gpmA	1jpc_	1msc_	1pco_	1rmd_	1tpg_	1zxq_	2hts_	3minB
1aerA	1bndA	1cyw_	1esc_	1gpr_	1jswA	1msfC	1pda_	1rnl_	1trkA	1zymA	2kauC	3pga1
1afb1	1bp1_	1daaA	1esfB	1grj_	1jud_	1msk_	1pdnC	1rpa_	1ttbA	2aaa_	2lbp_	3pmgA
1afrA	1bp2_	1dar_	1esl_	1gsa_	1jul_	1mspA	1pea_	1rrgA	1tupB	2abd_	2madL	3pte_
1afwA	1bplA	1dbqA	1etpA	1gtmA	1jvr_	1mtyB	1pedA	1rsy_	1tys_	2abh_	2mev1	3sdhA
1agrE	1bpyA	1deaA	1eur_	1gtqA	1kapP	1mtyD	1pfkA	1ryc_	1tyv_	2abk_	2mnr_	3tgl_
1aihA	1briC	1def_	1exnA	1gtrA	1kaz_	1mup_	1pgs_	1ryt_	1u9aA	2acq_	2mprA	4aahA
1air_	1broA	1dehA	1fbaA	1gym_	1kcw_	1mut_	1phg_	1sacA	1uae_	2admA	2mtaC	4enl_
1aisB	1btv_	1dekA	1fbr_	1ha1_	1kinA	1mxa_	1phr_	1sbp_	1uby_	2ak3A	2myr_	4fgf_
1ako_	1bucA	1dhpA	1fbtA	1han_	1kit_	1nal1	1pkm_	1schA	1ucwA	2amg_	2nacA	4rhn_
1akz_	1burT	1dhr_	1fc2D	1har_	1klo_	1nbaA	1pkp_	1scmB	1ulp_	2arcA	2ncm_	4rhv3
1alkA	1bvp1	1din_	1fcdC	1havA	1knb_	1ndh_	1pls_	1scuA	1umuA	2ayh_	2olbA	4sbvA
1alo_	1bw4_	1div_	1fd2_	1hbq_	1knyA	1neu_	1pmi_	1scuB	1uxy_	2azaA	2omf_	4ts1A
1amm_	1byb_	1dja_	1fgkB	1hce_	1kobA	1nfkA	1pms_	1seiA	1vcc_	2bbi_	2ora_	4xiaA
1amp_	1cauA	1djxA	1fib_	1hcnA	1kuh_	1nfn_	1pnkA	1sesA	1vdc_	2bbvC	2pcdM	5fbpA
1anu_	1cauB	1dkgA	1fieA	1hcnB	1kveB	1nfp_	1pnkB	1sfe_	1vhh_	2bgu_	2pf1_	5nul_
1anv_	1ccr_	1dkzA	1fil_	1hcp_	1kxu_	1nhkL	1poc_	1sftA	1vhiA	2bltA	2pgd_	5rubA
1aocA	1cdq_	1dlc_	1fim_	1hcz_	1l68_	1nhp_	1pot_	1sgt_	1vhrA	2bnh_	2phy_	5timA
1aozA	1cdy_	1dlhA	1fjmA	1hdgO	1lba_	1nif_	1poxA	1shcA	1vid_	2bpa1	2pia_	7rsa_
1apyA	1celA	1dlhB	1fkj_	1hfh_	1lbd_	1nipA	1ppfE	1sig_	1vin_	2bpa2	2pldA	8abp_
1arb_	1cem_	1dnpA	1fkx_	1hgxA	1lbiA	1nkl_	1pprM	1sltB	1vls_	2btfA	2polA	8acn_
1ars_	1ceo_	1doi_	1fnc_	1hjrA	1lbu_	1nox_	1prcC	1sly_	1vmoA	2cae_	2por_	8atcA
1arv_	1cewI	1dorA	1fnf_	1hlb_	1lckA	1npoA	1prcH	1smd_	1vnc_	2cas_	2prk_	8fabB
1ash_	1cex_	1dosA	1froA	1hmy_	1lcl_	1nsj_	1prcM	1smeA	1vokA	2cba_	2pspA	8tlnE
1aszA	1cfb_	1dpb_	1fruA	1hngA	1lcpA	1nsyA	1preA	1smnA	1vom_	2chr_	2qilA	9pap_
1atiB	1chd_	1dpe_	1frvA	1hslA	1ldm_	1nulA	1prr_	1snc_	1vpsA	2cmd_	2reb_	9wgaA
1atlA	1chkA	1dpgA	1frvB	1htp_	1lenA	1nzyA	1pscA	1sra_	1vpt_	2cpl_	2rslB	
1atpE	1chmA	1drw_	1fua_	1httA	1lfaA	1oacA	1ptvA	1sriA	1vscA	2csmA	2rveA	
1axn_	1cid_	1dsn_	1fvkA	1hxn_	1lgr_	1obpA	1ptx_	1svb_	1vsd_	2ctc_	2sas_	
1ayl_	1cksB	1dts_	1gai_	1hxpA	1lid_	1obwA	1pud_	1svpA	1vsgA	2dkb_	2scpA	
1babB	1clc_	1dupA	1gal_	1i1b_	1lis_	1occA	1pueE	1tag_	1wad_	2dri_	2sil_	
1bbpA	1cmbA	1dxy_	1garA	1iae_	1lki_	1occB	1put_	1tam_	1wba_	2drpA	2stv_	
1bbt1	1cmvA	1dynA	1gca_	1ignA	1lnh_	1occC	1pvc1	1taq_	1wdcC	2dtr_	2tbd_	
1bbt2	1cnsA	1dyr_	1gcb_	1ihfB	1lrv_	1occD	1pvdA	1tca_	1whi_	2ebn_	2tct_	
1bbt3	1cnt1	1eal_	1gdoA	1ilk_	1ltdA	1octC	1pyaB	1tcmA	1who_	2end_	2tgi_	
1bcfA	1cnv_	1ebpA	1gds_	1iml_	1lucA	1ois_	1qapA	1tcoB	1xel_	2er7E	2tmdA	
1bco_	1cof_	1eca_	1gen_	1inp_	1lxa_	1omp_	1qba_	1tcrA	1xer_	2fal_	2tmvP	
1bdmB	1colA	1eceA	1ghr_	1iol_	1lylA	1onc_	1qorA	1tdtA	1xgsA	2fcr_	2tplA	
1bec_	1cpcA	1ecl_	1gky_	1iow_	1lzr_	1oroB	1qpg_	1tf4A	1xikA	2fha_	2trcP	
1beo_	1cpcB	1ecpA	1glcG	1irl_	1masA	1ospL	1raiD	1tfe_	1xjo_	2gdm_	2tysA	
1berA	1cpo_	1ecrA	1gln_	1irsA	1maz_	1ospO	1rbu_	1tfr_	1xnb_	2gmfA	2tysB	
1bglA	1cpq_	1ede_	1gnd_	1iscA	1mbd_	1otgA	1rcb_	1tgxA	1xsm_	2gsq_	2ull_	

APPENDIX B

In this study the efficiency of the predictors is scored by using three different indices. The network accuracy is defined as:

$$Q3 = P/N \tag{1A}$$

where P is the total number of correct predictions and N is the total number of possible predictions.

Because only two classes (the bonding and the nonbonding state of cysteine) are discriminated, the correlation coefficient C is single valued and defined (for the bonding or nonbonding state) as:

$$C = (p * n - u * o)/((p + u)(p + o)(n + u)(n + o))^{1/2} \tag{2A}$$

where p and n are the total number of correct predictions and that of correctly rejected assignments, respectively for one state; u and o are the numbers of under and over predictions for same state.

The probability of correct predictions Pc is evaluated as:

$$\mathrm{Pc}(s) = p(s)/(p(s) + o(s)) \tag{3A}$$

where for the state s (bonding and non bonding) $p(s)$ and $o(s)$ are the numbers of correct and over predictions, respectively.

REFERENCES

1. Creighton T. Proteins: structures and molecular properties. New York: WH Freeman, 1996.

2. Betz SF. Disulfide bonds and the stability of globular proteins. Protein Sci 1993;2:1551–1558.
3. Freire E. Structural thermodynamics: prediction of protein stability and protein binding energy. Arch Biochem Biophys 1993;303: 181–184.
4. Privalov PL, Gill SJ. Stability of protein structure and hydrophobic interaction. Adv Prot Chem 1988;39:191–324.
5. Casadio R, Compiani M, Fariselli P, Vivarelli F. Predicting free energy contributions to the conformational stability of folded proteins from the residue sequence with radial basis function networks. Ismb 1995;3:81–88.
6. Harrison PM, Sternberg MJE. Analysis and classification of disulfide connectivity in proteins. J Mol Biol 1994;244:448–463.
7. Harrison PM, Sternberg MJE. The disulfide β-cross: from cystine geometry and clustering to classification of small disulfide-rich protein folds J Mol Biol 1996;26:603–623.
8. Fiser A, Cserzo M, Tudos E, Simon I. Different sequence environment of cysteines and half cystines in proteins. FEBS Lett 1992;302:117–120.
9. Muskal SM, Holbrook RS, Kim SH. Prediction of the disulfide-bonding state of cysteine in proteins. Protein Eng 1990;3:667–672.
10. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
11. Rumelhart DE, Hinton GE, Williams RJ. Learning representation by back-propagation error. Nature 1986;323: 533–537.
12. Fariselli P, Compiani M, Casadio R. Predicting secondary structures of membrane proteins with neural networks Eur Biophys J 1993:22:41–51.
13. Schneider R, Sander C. Database of homology-derived structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
14. Rose GD, Gesolowittz AR, Lesser GJ, Lee RH, Zehfus MH. Hydrophobicity of amino acid residues in globular proteins. Science 1985;229:834–838.
15. Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 1994;19:55–72.
16. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
17. Rost B, Casadio R, Fariselli P, Sander C. Prediction of helical transmembrane segments at 95% accuracy. Protein Sci 1995;4: 521–533.
18. Rost B, Casadio R, Fariselli P. Topology prediction for helical transmembrane proteins at 86% accuracy. Protein Sci 1996;5: 1704–1718.
19. Krogh A, Sollich P. Statistical mechanics of ensemble learning. Phys Rev 1997;E55:811–825.