Protein Engineering vol.3 no.8 pp.667–672, 1990

Prediction of the disulfide-bonding state of cysteine in proteins

Steven M.Muskal[1], Stephen R.Holbrook[2] and Sung-Hou Kim[1,2,3]

[1]Department of Chemistry and [2]Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720, USA

[3]To whom correspondence should be addressed

The bonding states of cysteine play important functional and structural roles in proteins. In particular, disulfide bond formation is one of the most important factors influencing the three-dimensional fold of proteins. Proteins of known structure were used to teach computer-simulated neural networks rules for predicting the disulfide-bonding state of a cysteine given only its flanking amino acid sequence. Resulting networks make accurate predictions on sequences different from those used in training, suggesting that local sequence greatly influences cysteines in disulfide bond formation. The average prediction rate after seven independent network experiments is 81.4% for disulfide-bonded and 80.0% for non-disulfide-bonded scenarios. Predictive accuracy is related to the strength of network output activities. Network weights reveal interesting position-dependent amino acid preferences and provide a physical basis for understanding the correlation between the flanking sequence and a cysteine's disulfide-bonding state. Network predictions may be used to increase or decrease the stability of existing disulfide bonds or to aid the search for potential sites to introduce new disulfide bonds.

Key words: cysteine/disulfide bond/neural network/structure prediction

Introduction

Cysteine's thiol group is the most reactive of any amino acid side chain (Creighton, 1984). Existing as the free sulfhydryl, ionized as the thiolate ion or oxidized into disulfide, thioether or thioester bonds, cysteine plays an important functional and structural role in globular proteins. Functionally, cysteines fix hemes in cytochromes, bind metals in a variety of metalloproteins, and act as nucleophiles in thiol proteases. Structurally, cysteines form disulfide bonds that provide stability to snake venoms, peptide hormones, immunoglobulins, lysozymes and others (Schulz and Schirmer, 1984).

Because free thiols are unstable relative to S–S bridges in the presence of oxygen, cysteines are typically oxidized into disulfide bonds in proteins leaving the cell and reduced in proteins remaining inside the cell (Fahey *et al.*, 1977). Predictions of the disulfide-bonding state of cysteines based only on this criterion, however, result in failures for extracellular proteins containing free thiols, e.g. actinidin, immunoglobulin, papain and some virus coat proteins, and for cystine-containing intracellular proteins, e.g. trypsin inhibitor, thioredoxin and superoxide dismutase. Furthermore, to base positive disulfide-bond predictions on high cysteine content and even parity result in failures for ferrodoxins, metallothioneins and some cytochromes. Clearly, predictions based on these simple rules fail to capture the unique micro-environments a protein structure imposes on its cysteines to define their disulfide-bonding states.

Recently, computer-simulated neural networks have shown great promise in extracting structural features from sequence information (Bohr *et al.*, 1988; Qian and Sejnowski, 1988; Holley and Karplus, 1989; McGregor *et al.*, 1989; Holbrook *et al.*, 1990). In analogy to biological neuronal systems, computer-simulated neural networks consist of a large number of simple, highly interconnected computational units or nodes that operate in parallel. Integrating both 'excitatory' and 'inhibitory' input signals, each node generates an output based on some threshold value. When these functional units are organized into layers, a supervised network can be trained to map a set of input patterns to a set of output patterns (for a review of computational neural networks see Lippmann, 1987).

In this paper, we make use of computer-simulated neural networks for determining the effects of local sequence on the chemistry of cysteine. In particular, networks were trained to predict whether or not a cysteine participates in a disulfide bond, with the presumption that it is the local sequence that determines a cysteine's disulfide-bonding state. Once the networks learned the 'rules' from a large number of training sequences, they were tested on sets of sequences different from those used in training. The observation that trained networks made accurate predictions on the disulfide-bonding state of cysteines in the context of their flanking amino acids, supports the notion that locally surrounding amino acids greatly influence cysteines in forming disulfide bridges.

Materials and methods

Database

Cyst(e)ine-containing proteins of known tertiary structure were obtained from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977). Disulfide bond assignments were based on SSBOND records in this database and were confirmed with Kabsch and Sander's (1983) program DSSP. A set of 128 cysteine-containing protein structures (Table I) was selected to provide examples for network training and testing.

Identical cysteine-containing sequences were removed from the pool of examples leaving 689 examples, 379 of which were in the disulfide-bonded state and 310 of which were in the non-disulfide-bonded state. The 689 examples were used to create eight training and eight testing sets. One of the training-testing sets was used to find the optimum number of flanking residues (windowsize) for prediction purposes and contained a random selection of 30 examples used for the testing set, leaving the remaining 659 examples for the training set. The other seven training-testing sets were used for independent training and testing experiments and contained a random selection of 20 examples for each testing set, leaving the remaining 669 examples for each training set.

Table I. Proteins correspondıng to the following protein databank codes were used in network traınıng and testing

1ABP	1ACX	1AZU	1BP2	1CAC	1CC5	1CCR	1CHG	1CHO	1CRN	1CTX	1CY3
1CYC	1ETU	1F19	1FB4	1FBJ	1FC1	1FDH	1FDX	1FX1	1FXB	1GCR	1GD1
1GF1	1GF2	1GP1	1GPD	1HBS	1HDS	1HIP	1HKG	1HMG	1HMZ	1HOE	1HVP
1IG2	1LZ1	1MCP	1MEV	1NTP	1NXB	1P2P	1PAZ	1PFC	1PHH	1PP2	1PRC
1PSG	1PYP	1RBB	1RDG	1REI	1RHD	1SGT	1SN3	1TGN	1TON	1TRM	1WSY
1XY1	2ABX	2ACT	2ALP	2APP	2AZA	2CAB	2CCY	2CDV	2CGA	2CPP	2CYP
2DHB	2GN5	2HFL	2KAI	2LDX	2LHB	2MHR	2MT2	2OVO	2PAB	2PCY	2PRK
2RHE	2SGA	2SOD	2SSI	2STV	2TAA	2TBV	2TMV	2YHX	3ADK	3APR	3BCL
3C2C	3CPV	3EBX	3EST	3FAB	3FXC	3GAP	3GRS	3INS	3PGK	3PTB	3RP2
3SGB	3WGA	451C	4ADH	4APE	4DFR	4FD1	4FXN	4LDH	4MDH	4PFK	4RHV
4SBV	5CPA	5RXN	5TNC	6PAD	6API	7ATC	8CAT				

Network design and training procedure

The networks were of the feedforward type containing no hidden nodes (perceptrons). Because every sequence presented to the networks contained a centered cysteine, the input layer encoded a window of amino acid sequence surrounding, but not including, the central cysteine, as shown in Figure 1. A weighted connection existed between each node in the input layer and each node in the output layer. Each output node had an additional weight termed the bias. Each output node takes a weighted sum of its inputs,

$$X_o = \sum_i (W_{o,i} \times I_i) + B_o \qquad (1)$$

where $W_{o,i}$ is the weight to output node o from input node i, I_i is the value at the input node i (either 0.0 or 1.0) and B_o is the output bias. The actual activities A_o appearing at the output nodes are a function of a constant activation threshold T and defined by the following activation function:

$$\left(\begin{array}{l} X_o \geq T,\ A_o = X_o \\ X_o < T,\ A_o = 0.0 \end{array} \right) \qquad (2)$$

The weights were adjusted by back-propagation (Rumelhart *et al.*, 1986a,b) so as to minimize the difference between the actual output A_o of the network and the desired output D_o. The weight change is defined by

$$\Delta W_{o,i} = \eta \times v_o \times X_o \qquad (3)$$

where

$$v_o = (D_o - A_o) \qquad (4)$$

and η is the constant learning rate. If the weight change is averaged over all the training examples, equation (3) will minimize the total error

$$E = \sum_p \sum_o (A_{o,p} - D_{o,p})^2 \qquad (5)$$

across all examples p in a gradient descent fashion.

Each window position encoded a group of 21 nodes, one per amino acid and one to provide a null input used when the window overlapped one of the termini or a break in the chain. The output layer consisted of two nodes, one for a disulfide-bonded state and one for a non-disulfide-bonded state.

Before each simulation, network weights were assigned random values between −0.3 and 0.3. A training cycle then consisted

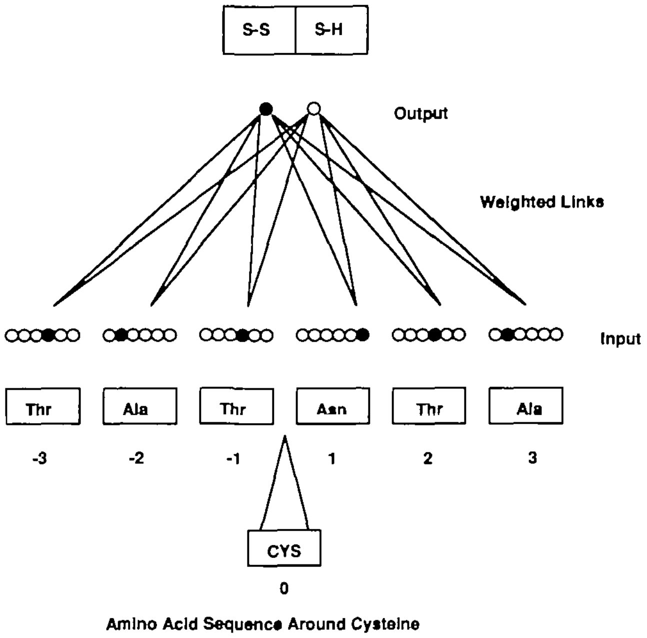

Fig. 1. A diagram of network architecture. For clarity, only six window positions (three amino acids to the N-terminal and three amino acids to the C-terminal side of an assumed centered cysteine) and six nodes per window position are illustrated. Within a window position, an amino acid is represented by giving a value of 1.0 to its node while setting all other nodes in that window position to 0.0. Input values are propagated through weighted links to produce activities at the two output nodes, S−S and S−H. The output node with the highest activity is the network's decision.

Table II. Dependence of training and testing success on window size

Window	% Train	C_{SS}-bond	% Test	C_{SS}-bond
−1:1	65.7	0.30	60.0	0.22
−2:2	72.8	0.45	66.7	0.34
−3:3	79.1	0.57	73.3	0.51
−4:4	83.9	0.67	73.3	0.48
−5:5	85.7	0.71	80.0	0.61
−6:6	88 2	0.76	80.0	0.60
−7:7	91.4	0.82	80.0	0.61

Dependence of network performance on the size of the window around a centered cysteine. Thirty examples (15 examples of sequences surrounding disulfide-bonded cysteines; 15 examples of sequences surrounding non-disulfide-bonded cysteines) were randomly selected from the pool of 689 examples to create a testing set, leaving the remaining 659 examples for a training set. Window −*x*:*x* stands for a window that has *x* amino acids on N-terminal and *x* amino acids on the C-terminal side of a central cysteine. % Train, % Test, and C_{SS}-bond are % correct prediction on training set, % correct prediction on the testing set, and Mathews' (1975) correlation coefficient respectively.

of presenting every example in the training set to the network, after which the weights were updated according to an average weight change as determined by the delta rule in equation (3) (Rumelhart *et al.*, 1986a). Once the total error over all examples in the training set converged to a minimum,. the weights were fixed and the network was evaluated on the training and testing sets. Network performance was defined by the percentage correct prediction at the output node with the highest activity. As there were approximately equal numbers of examples of disulfide-bonded and non-disulfide-bonded states in the training and testing sets, anything >50% prediction would be a non-random decision. The correlation coefficient by Mathews (1975) was another useful measure in determining the degree of randomness in the network's decisions.

Table III. Performance of independent training and testing sessions

Set	% Train		% Test	
	S−S	S−H	S−S	S−H
1	89.7	83.3	80.0	80.0
2	89.4	82.3	80.0	80.0
3	89.7	83.3	90.0	70.0
4	90.2	83.0	70.0	90.0
5	90.5	83.0	70.0	100.0
6	90.5	84.3	90.0	70.0
7	90.0	82.7	90.0	70.0
Av	90 0	83.1	81.4	80.0

Network performance for each set was evaluated by testing on a random subset of 20 examples (10 examples of sequences surrounding disulfide-bonded cysteines; 10 examples of sequences surrounding non-disulfide-bonded cysteines) taken from the pool of 689 examples after training on the remaining 669 examples. Each experiment was conducted independently on networks with a window −5:5 (five amino acids to the left and five to the right of a central cysteine). For the training and testing sets, S−S and S−H stand for % correct prediction for disulfide-bonding state and % correct prediction for non-disulfide-bonding state respectively.

Results and discussion

To determine the influence of flanking sequence on a centered cysteine in predicting its disulfide-bonding state, we first asked how an increasing window of sequence would affect the network's predictive performance. As seen in Table II, the network's performance on both the training and testing sets increases with increasing window size. The phenomenon of increasing training performance despite stabilized testing performance is attributed to memorization, an effect that results in networks containing a large number of weights relative to training examples. It should be noted that after window −7:7 (14 flanking amino acid positions, 21 nodes per amino acid position, two output nodes and two output node biases correspond to 14 × 21 × 2 + 2 = 590 weights), the number of weights begins to exceed the number of training examples. As this occurs, the capacity that the network has for memorization increases considerably, and hence experiments were stopped after a window size of 14. Table II indicates that the windows −5:5 or −6:6 are optimal for predictive purposes, where the percentage correct prediction and correlation coefficient on the testing set reach a maximum and memorization between the training and testing sets is minimized. Furthermore, Table II shows that trained networks made accurate predictions on examples never seen before, thus supporting the hypothesis that a cysteine's propensity and/or aversion for disulfide bond formation depends to a great extent on its neighbors in sequence.

After window size experiments were completed, seven independent training and testing experiments were conducted so as to determine an average performance that was not dependent on any particular training and testing set. Table III indicates that a network can be trained to predict disulfide-bonded scenarios 81.4% correctly and non-disulfide-bonded scenarios 80.0% correctly. It should be noted that the trained networks made accurate predictions on sequences from both extracellular and intracellular proteins. In fact, for the extracellular proteins

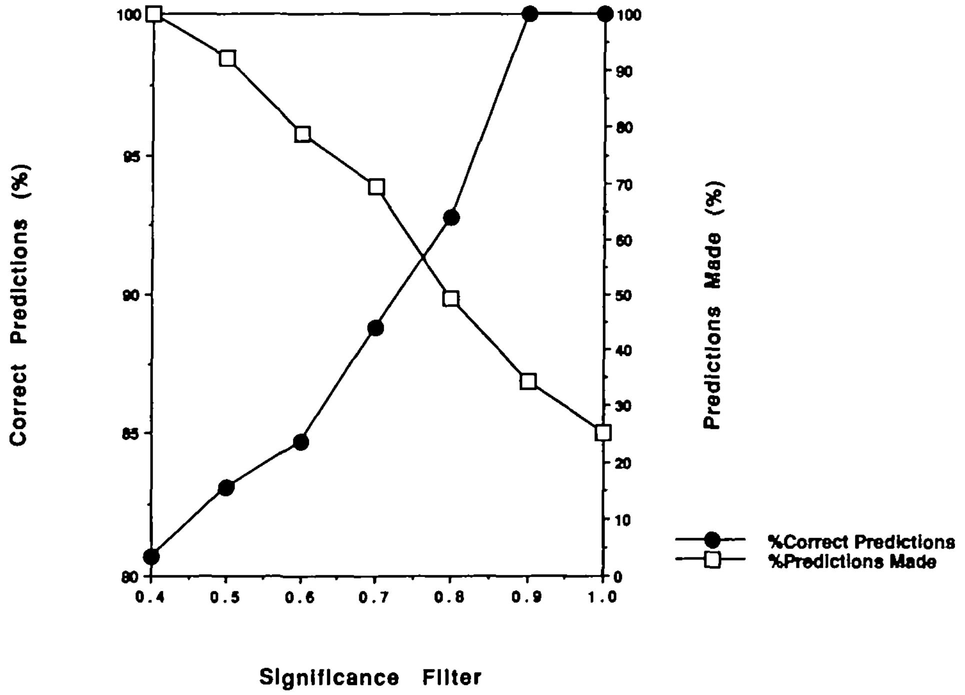

Fig. 2. Dependence of predictive accuracy on the strength of output node activities. The significance filter is placed over the output nodes so that only activities greater than the filter can pass through for prediction. Data were average from the seven testing sets in Table III.

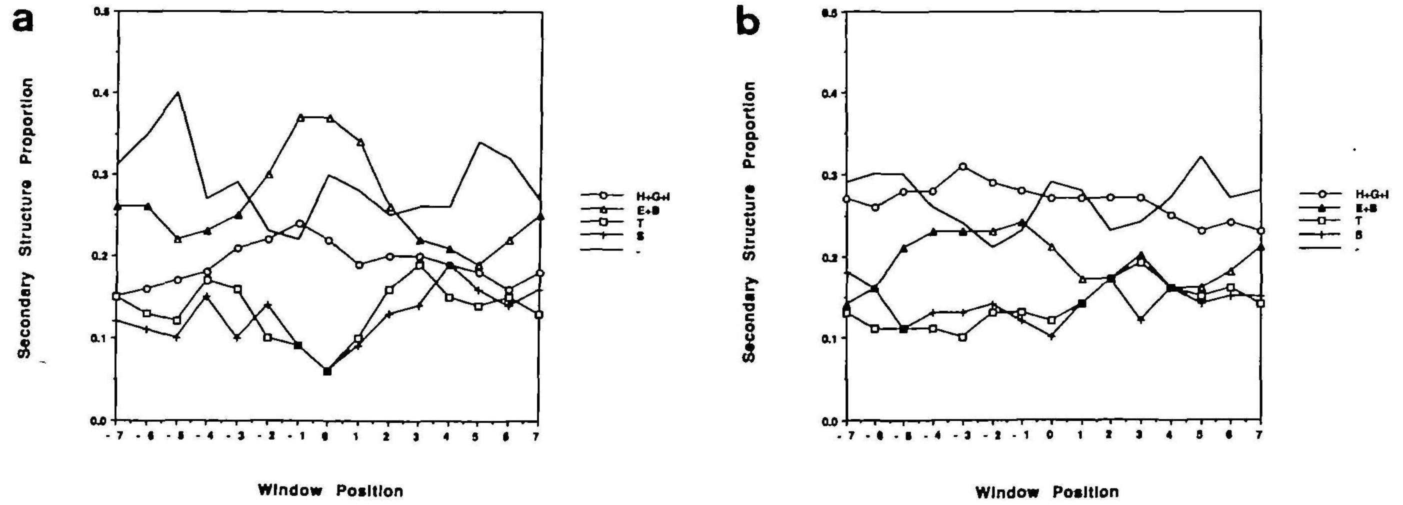

Fig. 3. Secondary structure surrounding disulfide-bonded (a) and non-disulfide-bonded cysteines (b) used in this study. Secondary structure proportion is calculated by summing number of individual secondary structure types and dividing by the total number of secondary structure types occurring in that window position. Secondary structure assignments were made by the method of Kabsch and Sander (1983). H, alpha helix; G, 3_{10} helix; I, pi helix; E, extended strand; B, isolated β-bridge; T, turn; S, bend; –, other.

actinidin, immunoglobulin and papain, the cysteines not involved in disulfide bonds were correctly predicted as such. Likewise, for the intracellular cystine-containing proteins, such as trypsin inhibitor and superoxide dismutase, every cysteine's state was correctly predicted.

Furthermore, if one accepts only those predictions with high output activities, the network's performance will greatly increase. As seen in Figure 2, predictive performance will increase while the number of predictions made will decrease with an increasing significance filter. The significance filter is a means of considering only those predictions with network activities greater than the value of the filter. Because stronger output activities result in more accurate predictions, Figure 2 can be used to describe a confidence level based on the magnitude of the network's output activity.

Figure 3(a and b) shows the secondary structure proportion as a function of window position for disulfide- and non-disulfide-bonded cysteines. It can be seen here that the sequences surrounding and including half-cystines seem to prefer the extended conformation of β-sheets over that of turns and bends, whereas those sequences containing non-disulfide-bonded cysteines show little, if any, secondary structure preference. The secondary structural preferences of half-cystines perhaps enable the high prediction rate of a cysteine's disulfide-bonding state. It should be noted that in Figure 3(a), beyond ±5 residues from the central half-cystine (coinciding with the selected network window size) the preferences for any secondary structure disappear.

In contrast to networks containing hidden nodes, perceptron weights can be easily interpreted. Figure 4 is a graphical depiction of the weights averaged from the seven network experiments. The weights from each amino acid at each window position to the non-disulfide node (S–H) were subtracted from those respective weights to the disulfide node (S–S) and scaled to a shade of black. The dark squares indicate large weights (strong SS-bonding propensity) while the light squares indicate low weights (weak SS-bonding propensity). Note that cysteines at positions ±3 are not very conducive towards disulfide bond formation. This can be explained by the frequent occurrence of Cys–x–x–Cys in heme and metal-binding proteins. Conversely, cysteines at position ±1 increase the propensity considerably. This can be explained by the frequent occurrence of Cys–Cys in extracellular proteins, where the cysteines can form a basis for linking three chain segments in close proximity (Brown, 1976). Figure 4 also shows a positive disulfide bond propensity of closely centered β-sheet forming residues such as Ile and Tyr.

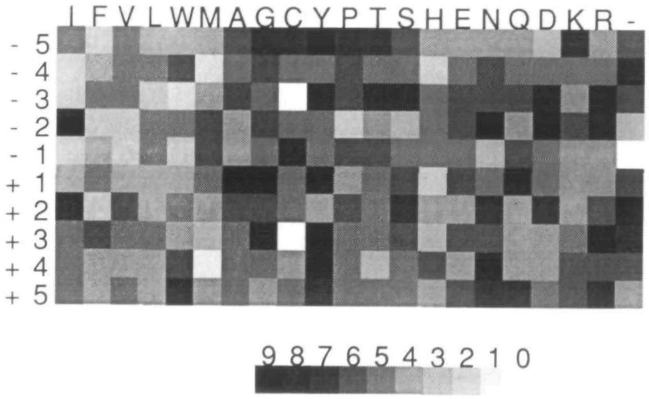

Fig. 4. Weights averaged over the seven network experiments in Table III. Weights for an amino acid in a window position to the S–H output node were subtracted from those to the S–S output node. Black shades indicate high and white shades indicate low S–S propensity. Displayed vertically is window positioning –5:5 (five positions to the N-terminal and five positions to the C-terminal side of an assumed centered cysteine). Displayed horizontally are amino acids ordered in decreasing hydrophobicity according to Eisenberg (1984).

Thornton (1981) has shown that residues linking two closely spaced bonded half-cystines often possess positive β-turn potential. In Figure 4, cysteines at position ±5 show strong to medium disulfide bond influence as do high β-turn potential residues such as Asp, Asn, Ser, Pro and Gly (Wilmot and Thornton, 1988) at positions ±4, ±3, ±2 and ±1. This suggests that given cysteines at positions 0 and ±5 separated by residues with strong β-turn potential, not only is the cysteine at position 0 likely to be involved in a disulfide bond, but it is likely to be bonded to the cysteine at ±5.

The contribution an individual amino acid may have towards

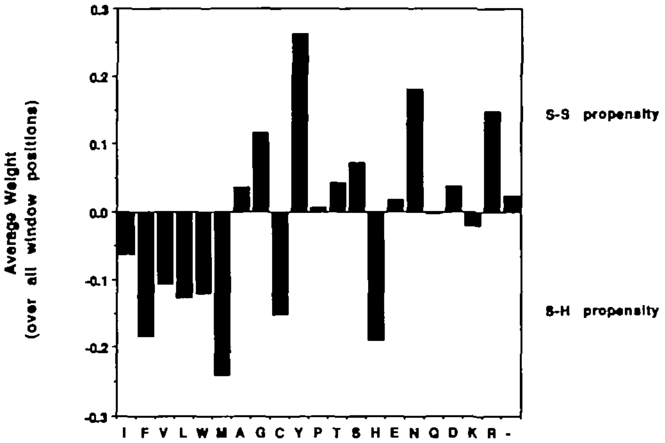

Fig. 5. Weights from the seven network experiments in Table III were averaged for each amino acid across each window position and displayed in bar form. Bars represent the weights to the S−H node subtracted from the weights to the S−S node. The amino acids are ordered in decreasing hydrophobicity according to Eisenberg (1984).

disulfide bond formation, irrespective of window position, can be seen in Figure 5. Here, the weights for a particular amino acid across all window positions to the S−H node were averaged and subtracted from those to the S−S node and depicted in bar form. One clear pattern is that the residues contributing towards S−S bond prediction are polar and/or charged while those against S−S bond prediction are primarily hydrophobic. Perhaps during the folding process, a locally hydrophilic environment helps to maintain a cysteine more accessible, thus increasing the chances of disulfide bond formation; whereas a locally hydrophobic environment helps to bury a cysteine, thus decreasing the chances of disulfide bond formation.

The most striking features in Figure 5 exist between similar amino acids. Tyr, for example, is highly conducive towards disulfide bond formation, yet Phe and Trp disfavor formation quite strongly. Reid *et al.* (1985) discuss an electrostatic interaction between the edge of aromatic rings and sulfur atoms. This interaction is found to be more frequent between aromatics and half-cystines than with aromatic and free cysteines. Figure 5 suggests that Tyr will favor disulfide bond formation over the other aromatics simply because Phe and Trp lack hydrophilic character. Likewise, Arg's greater polarity suggests S−S formation more strongly than does Lys. Less obvious, however, is the strong S−S propensity of Asn relative to Gln. One explanation is that Asn's smaller size may better enable the close approach of a potential half-cystine. Consistent with this, the S−S propensity of Gly, Asp and Ser exceeds that of their slightly larger counterparts Ala, Glu and Thr respectively. These differences in S−S propensity between otherwise very similar amino acids (Dayhoff, 1972) may make feasible the stabilization and/or destabilization of disulfide bonds through the site-directed mutagenesis of sequences surrounding half-cystines.

For proteins containing a single disulfide bond with more than two cysteines, it would be of great use to know which two cysteines are disulfide bonded. In such proteins, an accurate prediction of each cysteine's disulfide-bonding state would produce a correct partner assignment. In the cases of cytochrome c5 (1CC5: 2 S−S state; 2 S−H state), superoxide dismutase (2SOD: 2 S−S state; 1 S−H state) and mengo-encephalomyocarditis virus coat protein (1MEV: 2 S−S state; 8 S−H state), the weights averaged from the seven independent network experiments correctly predict the half-cystines as S−S and the free thiols as S−H, thus making accurate disulfide bond assignments. However, for azurin (2AZA: 2 S−S state; 1 S−H state) and glutathione reductase (3GRS: 2 S−S state; 7 S−H state), only one of the two half-cystines was correctly predicted for both.

Similarly, for proteins containing a single free thiol amidst disulfide bonds, the identification of the free sulfhydryl would be quite useful. For actinidin (2ACT: 6 S−S state; 1 S−H state), papain (6PAD: 6 S−S state; 1 S−H state) and lambda immunoglobulin Fab (3FAB: 10 S−S state; 1 S−H state), the bonding states of all cysteines are correctly predicted, thus identifying the free sulfhydryl. However, for proteinase k (2PRK: 4 S−S state; 1 S−H state), beside the correct S−H prediction, one other cysteine had an incorrect high S−H prediction, thus making no identification possible.

A few successful attempts at assigning half-cystines to their partners were made by choosing those pairs with the most similar network activities and the smallest residue separation. These successes [alpha-lytic protease (2ALP: 6 S−S state; 0 S−H state), aspartic proteinase (3APR: 4 S−S state; 0 S−H state) and actinoxanthin (1ACX: 4 S−S state; 0 S−H state)], however, are proteins with half-cystine pairs not enclosing other half-cystines.

Conclusion

The results from the neural network analysis provide a convenient means of predicting possible sites for disulfide bond formation based only on amino acid sequence. Cysteines with high as well as those with low disulfide bond propensity can be predicted with a very high confidence level. Successful prediction of each cysteine's disulfide-bonding state in some cases can lead to immediate disulfide partner assignment and represents the most important step towards eventual assignment of all disulfide pairs. Such assignments would provide key distance constraints for predicting the tertiary structure of cyst(e)ine containing proteins.

Our method can be of practical utility in protein engineering. Amino acid sequences can be searched for residues which appear to be good candidates for mutation into cysteine, whose flanking sequence suggests a high likelihood of forming a disulfide bridge. Likewise, sequences surrounding a cysteine can be altered to change the propensity that cysteine has for disulfide bond formation. In this way, when combined with other structural information, one can design sequences to either increase or decrease the likelihood of S−S formation.

We plan to increase the size of the database by including those proteins with disulfide bond information in the Protein Information Resource (George *et al.*, 1986) because the crystal structures in the Protein Data Bank are only a small subset of naturally occurring proteins. By increasing the present database, more sequence space would be represented, thereby allowing the neural net to extract those features that are at present only weakly suggested. Further experiments will consider cysteine content, cysteine spacing, protein size and protein family with the goal of improving the prediction accuracy.

Acknowledgements

The authors are grateful to Marge Hutchinson for her assistance and advice, Fan Jiang for his graphics program and Mike Milburn for useful discussions. We acknowledge the support of the Health Effects Research Division, Health and Environmental Research, Office of Energy Research of the US Department of Energy. One of the authors (S.M.M.) was supported by a University of California Reagents fellowship.

References

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol.*, **112**, 535–542.

Bohr,H., Bohr,J., Brunak,S., Cotterill,R., Lautrup,B., Norskov,L., Olsen,O.H. and Petersen,S.B. (1988) *FEBS Lett.*, **241**, 223–228.

Brown,J.R. (1976) *Fed. Proc.*, **35**, 2141–2144.

Creighton,T. (1984) *Proteins: Structures and Molecular Properties.* W.H.Freeman, New York.

Dayhoff,M.O. (1972) *Atlas of Protein Sequence and Structure.* The National Biomedical Research Foundation, Maryland, Vol. 5.

Eisenberg,D. (1984) *Annu. Rev. Biochem.*, **53**, 595–623.

Fahey,R.C., Hunt,J.S. and Windham,G.C. (1977) *J. Mol Evol.*, **10**, 155–160.

George,D.G., Barker,W.C. and Hunt,L.T. (1986) *Nucleic Acids Res.*, **14**, 11–15.

Holbrook,S.R., Muskal,S.M. and Kim,S.-H. (1990) *Protein Engng.*, 3, 659–665.

Holley,H.L. and Karplus,M. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 152–156.

Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.

Lippmann,R. (1987) IEEE ASSP. April, 4–22.

Matthews,B.W. (1975) *Biochim. Biophys. Acta*, **405**, 442–451.

McGregor,M.J., Flores,P.T. and Sternberg,M.J. (1989) *Protein Engng*, **2**, 521–526.

Qian,N. and Sejnowski,T.J. (1988) *J. Mol. Biol.*, **202**, 865–884.

Reid,K.S.C., Lindley,P.F. and Thornton,J.M. (1985) *FEBS Lett.*, **190**, 209–213.

Rumelhart,D.E., Hinton,G.E. and Williams,R.J. (1986a) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* MIT Press, Cambridge, MA, Vol. 1.

Rumelhart,D.E., Hinton,G.E. and Williams,R.J. (1986b) *Nature*, **323**, 533–536.

Schulz,G.E. and Schirmer,R.H. (1984) *Principles of Protein Structure.* Springer Verlag, New York.

Thornton,J.M. (1981) *J. Mol. Biol.*, **151**, 261–287.

Wilmot,C.M. and Thornton,J.M. (1988) *J. Mol. Biol.*, **203**, 221–232.

Received on October 17, 1989; accepted on March 23, 1990