FOR THE RECORD

Prediction of the disulfide-bonding state of cysteines in proteins at 88% accuracy

PIER LUIGI MARTELLI, PIERO FARISELLI, LUCA MALAGUTI, AND RITA CASADIO

Laboratory of Biocomputing, CIRB/Department of Biology, University of Bologna, 40126 Bologna, Italy

(RECEIVED June 7, 2002; FINAL REVISION August 6, 2002; ACCEPTED August 9, 2002)

Abstract

The task of predicting the cysteine-bonding state in proteins starting from the residue chain is addressed by implementing a new hybrid system that combines a neural network and a hidden Markov model (hidden neural network). Training is performed using 4136 cysteine-containing segments extracted from 969 non-homologous proteins of well-resolved three-dimensional structure. After a 20-fold cross-validation procedure, the efficiency of the prediction scores as high as 88% and 84%, when measured on cysteine and protein basis, respectively. These results outperform previously described methods for the same task.

Keywords: Hidden neural networks; hidden Markov models; neural networks; cysteine-bonding state

The formation of disulfide bonds between the correct pairs of cysteine residues is essential for the folding and stability of many proteins (Narayan et al. 2000; Wedemayer et al. 2000). Such bonds increase the conformational stability of the protein both by lowering the entropy of the folded state and by forming stabilizing interaction in the native state.

Several analyses of the characteristics of disulfide bond formation in proteins have been performed, including structural and sequence features and classification of connectivity (Harrison and Sternberg 1994; Wedemayer et al. 2000; Fariselli and Casadio 2001). The stability of the protein mainly increases by constraining the unfolded conformation, as many experimental and theoretical studies indicate (among others, see Skolnick et al. 1997; Abkevich and Shakhnovich 2000; Clarke et al. 2000; Welker et al. 2001).

The ever-increasing number of fully sequenced genomes requires powerful tools to perform protein sequence analysis at a genomic scale. However, only few studies have addressed the important problem of predicting the bonding state of cysteine in a protein chain.

Reprint requests to: Rita Casadio, Department of Biology, University of Bologna, Via Irnerio 42; 40126 Bologna, Italy; e-mail: casadio@alma.unibo.it; fax: 0039-051-242576.

Article and publication are at http://www.proteinscience.org/cgi/doi/10.1110/ps.0219602.

The relevance of the flanking residues in predicting a cysteine-bonding state has been shown using statistical methods (Fiser et al. 1992), neural networks (NNs; Muskal et al. 1990; Fariselli et al. 1999), and methods that combine local context and global information about protein sequences (Fiser and Simon 2000; Mucchielli-Giorgi et al. 2002).

In the present paper, we present an approach based on hidden neural networks (HNNs) that combines NNs and hidden Markov models (HMMs) and that outperforms all the existing methods, scoring as high as 88% and 84%, on a cysteine and protein basis, respectively.

Results and Discussion

Basically, the system we implemented takes advantage both of local and global characteristics of the protein chains. The local information is extracted by a feed-forward NN, with a sliding window centered on the cysteine residue at hand. NN captures local features of the residue context conducive to the bonding or nonbonding state (Fariselli et al. 1999).

It was previously shown that a perceptron without hidden layers could extract general characteristics of the local contexts conducive to disulfide bond formation (Fariselli et al. 1999). More specifically, when a segment centered in the cysteine to be predicted and including 11 residues of the protein chain was considered, the following features could be captured: (1) The presence of cysteine residues in the

environment of the central cysteine strongly favors the disulfide bond formation, with the exception of position 3 [this is in agreement with the fact that in proteins, metal-binding cysteines are typically found in positions i and i+3], and (2) hydrophilic and/or charged residues in the environment are highly conducive toward disulfide-bond formation compared with that of hydrophobic residues, which are poorly conducive (Fariselli et al. 1999).

In the present paper, the network includes an input window comprising a larger number of residues (27 instead of 11). This was performed after a careful search in the parameter space. Also, the number of proteins included in the data set is 1.5-fold higher (and that of cysteine comprising segments 1.7-fold higher than that used before; Fariselli et al. 1999). A statistics of the characteristics of the segment composition conducive to correct bond formation gives, however, results similar to those described before (Fariselli et al. 1999). This again confirms that the local context of the central cysteine is determining the correct bonding state and that a NN is sufficient to capture all the relevant interaction within the input window conducive to the bonding or non-bonding state.

However, the network is unable to capture global information. A network predicts one central cysteine at the time, and this is performed without keeping records of the different predictions associate to a given sequence. In other words, when a cysteine is predicted in a chain, its prediction does not take into account whether other cysteines in the chain are present and what their predicted bonding state is.

For modeling this global information, we use a four-state HMM that ensures that the number of cysteines predicted in the bonding state is even in each chain. This constrains the hybrid system to predict an even number of cysteines in the bonding state in each given chain, independently of the number of cysteines in the protein; NN outputs are then used as emission probabilities of the HMM (Fig. 1).

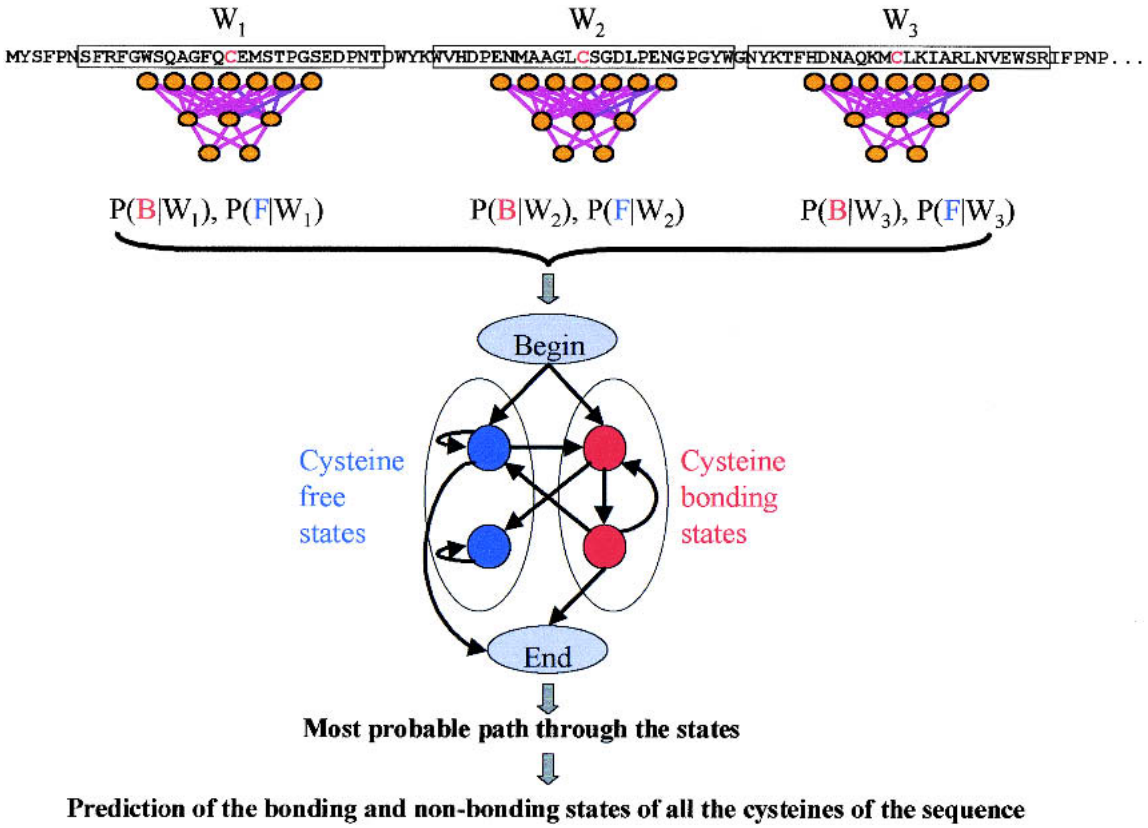


Fig. 1. Scheme of the hybrid system comprising a feed-forward neural network (NN) and a hidden Markov model (hidden neural network). The first predictor consists of a NN that uses evolutionary information as input and is trained on a 27-long window centered on the cysteine residues of the protein chain. The local context information is conducive to the bonding or nonbonding state of the cysteine residue (Fariselli et al. 1999). Then the NN outputs are used as emission probabilities of a four-state hidden Markov model, which constrains the number of cysteines predicted in the disulfide-bonding state per chain to be even. The arrows indicate allowed transitions; red and blue balls, the bonding and nonbonding cysteine states, respectively. The path can end only from an even state, which guarantees that only correct even predictions are assigned when considering intrachain disulfide bonds. Interchain disulfide-bridged cysteines are considered as free cysteines, because they represent only 0.8% of the training set of the cysteine-containing segments. $P(B|W_i)$ and $P(F|W_i)$ are the probabilities for the bonding (B) and free (nonbonding, F) states, respectively, at a given local context (W).

The results are computed with a 20-fold cross-validation procedure over a test set comprising 969 proteins from the Protein Data Base (PDB), with 4136 cysteine-containing segments, 1446 of which are in the disulfide-bonded state and 2690 of which are in the non–disulfide-bonded state. The protein chains have low sequence identity (<25%).

When only the NN-based predictor is adopted, the average accuracy per cysteine residue is ~80% (similar to the accuracy obtained with NN using a smaller set of proteins; Fariselli et al. 1999), and that per protein is 57% (Fig. 2, blue bars in the bar plot). Remarkably, when the hybrid system (HNN) is tested on the same protein set, accuracy per cysteine residue increases up to 88%, and that per protein improves by at least 27% points (Fig. 2, red bars). Concomitantly, the cross-correlation coefficient increases from 56% to 73%. The improvement obtained with the hybrid method compared with NN is seemingly caused by the introduction of global "rules" defined by the regular grammar implemented in the HMM. This second step captures the number of cysteines in a chain and also keeps track of the bonding states of all the cysteines in the same chain.

Even though it is difficult to compare methods tested on different databases, it can be claimed that the accuracy obtained with HNN is greater than that previously described and obtained with other methods, incorporating also global protein rules (Fisher and Simon 2000; Mucchielli-Giorgi et al. 2002). The method implemented by Fiser and Simon (2000) is based on a simple majority rule and reaches accuracy of 82% when predicting the bonding states of cysteines on a small set of proteins comprising 81 chains; that of Mucchielli-Giorgi (2002) uses global protein descriptors and scores as high as 84% for the same task on 869 chains. The higher accuracy (88%) obtained with HNN on 969 chains is probably owing to the higher flexibility of our system in capturing features of the sequences essential for the prediction of the cysteine-bonding state.

In conclusion, in the present paper, we show that a hybrid system combining local with global information outperforms previously developed methods to predict the cysteine-bonding states, confirming that for the problem at hand, a crucial step forward can be made only when global features of the protein chains are taken into consideration.

Materials and methods

The database

We took 4136 segments containing cysteines (free and disulfide bonded [half cystines]) from the crystallographic data of the Brookhaven Protein Data Bank. Disulfide-bond assignment was

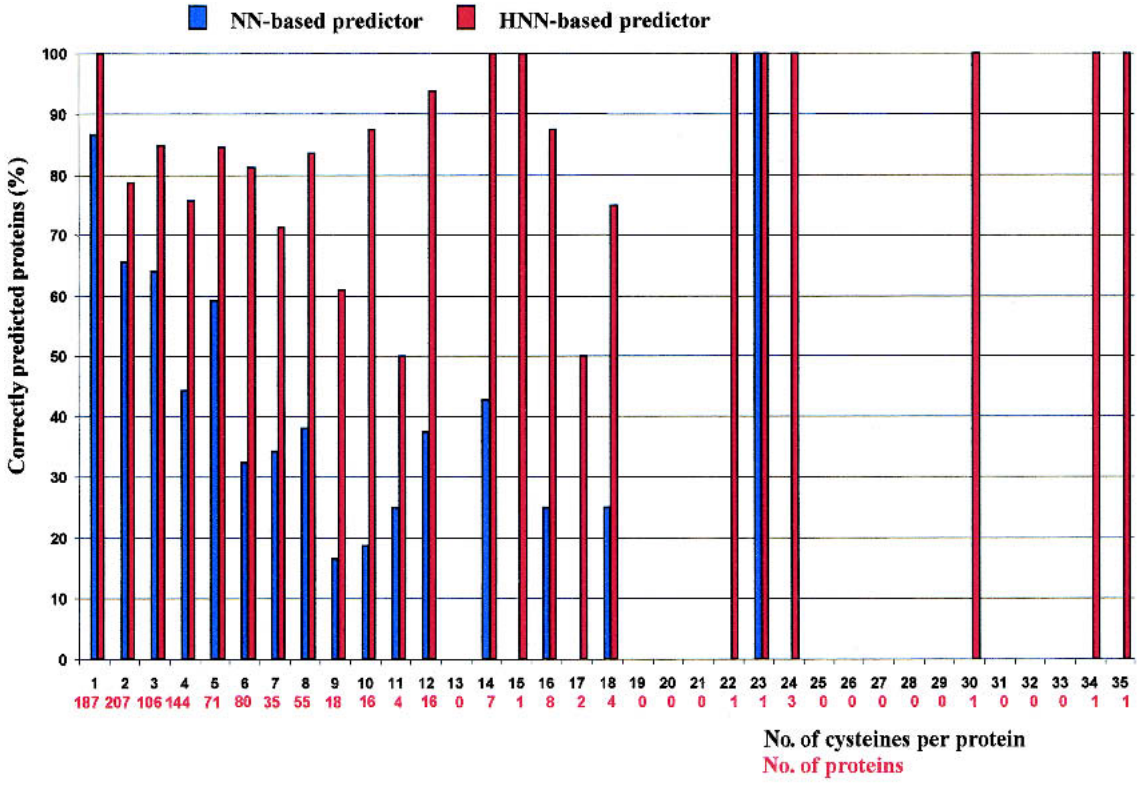


Fig. 2. Bar plots of the predictions. The percentage of proteins correctly predicted (one chain is accepted only if the bonding and nonbonding states of all the cysteines in the chain are correctly predicted) is shown as a function of the number of cysteines in the chain. Furthermore, the number of proteins with a given number of cysteines in the chain is also shown along the *X*-axis. Red and blue bars indicate the predictions of the hybrid system and of the neural network, respectively.

based on the Define Secondary Structure of Proteins (DSSP) program (Kabsch and Sander 1983).

Nonhomologous proteins (with an identity value <25% and without chain breaks) were selected using the PAPIA system (Noguchi et al. 2001). Segments with cysteines that are interchain disulfide-bonded are included as "free" cysteines in the database (34 segments extracted from 27 monomeric chains and amounting to 0.8% of the database of segments). After this filtering procedure, the total number of proteins is 969 (2.8% contain segments corresponding to interchain disulfide bonds, which are considered free cysteine), with 4136 cysteine-containing segments—1446 of which were in the disulfide-bonded state, and 2690 of which were in the non–disulfide-bonded state. For each protein in our database, a profile based on a multiple sequence alignment was created using the BLAST program on the nonredundant data set of sequences. The obtained profiles are used for creating the NN input.

During the training/testing phase, the database has been split in 20 subsets (almost equally sized and distributed) to perform a 20-fold cross-validation. The PDB codes of the proteins included in the data base, the 20-fold cross validation lists, and the training profiles are available at http://www.biocomp.unibo.it/piero/cyspred/cysdataset.tgz.

Measures of performance

The efficiency of the predictors is scored using the statistical indexes defined as follows.

The accuracy is

$$Q2 = P/N \tag{1}$$

where P is the total number of correctly predicted cysteines, and N is the total number of cysteines.

The correlation coefficient C is defined as

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

where, for each class s (free or bonded cysteines), $p(s)$ and $n(s)$ are, respectively, the total number of correct predictions and correctly rejected assignments; $u(s)$ and $o(s)$ are the numbers of under- and overpredictions.

Finally, the accuracy per protein is

$$Q2prot = Pp/Np \tag{3}$$

where Pp is the number of the proteins with cysteines that are all correctly predicted, and Np is the total number of proteins.

Neural networks

Standard feed-forward NNs are implemented with a back-propagation algorithm as learning procedure. The network architecture is similar to that previously used (Fariselli et al. 1999) and consists of a two-layer perceptron with two hidden neurons, one output node (discriminating the disulfide and free cysteine propensities, respectively), and an input layer that consists of 540 neurons (27-residue-long input window). Because of the limited number of examples presently available, an early learning stopping procedure is used to train the networks (Fariselli et al. 1999).

Hidden neural network

A vector-based HMM that can handle emission probability vectors is used on top of the NNs described above. The hybrid system is defined HNN, following the definition of Krogh and Riis (1999). A vector-based HMM, similar to that used in this paper, was recently developed (Martelli et al. 2002).

Briefly, if L is the number of cysteines in the protein and A is the size of the alphabet over which vectors are built (i.e., $A = 2$, bonding and nonbonding/free cysteine states), we refer to this sequence vector with the following notation:

$$\boldsymbol{s} = \boldsymbol{s^1 s^2} \ldots \boldsymbol{s^L} = (s^1(1),s^1(2))(s^2(1),s^2(2)) \ldots . (s^L(1),s^L(2)) \tag{4}$$

The components of each vector $\boldsymbol{s}^t$ are positive and sum to a constant value S (independent of the position t).

The HMM for the specific problem at hand is composed of a Markov model with four states connected by means of the transition probabilities a_{ij} (Fig. 1). The four states are the minimum number of states required to constrain to an even number the paired cysteines in a chain. The probability density function for the emission of a vector from each state is determined by a number A of parameters that are peculiar for each state k and are indicated with the symbols $e_k(c)$ (with $c = 1,2, \ldots ,A$):

$$P(\boldsymbol{s}^t \mid \pi^t = k) = (1/Z) \cdot \sum_c s^t(c) \cdot e_k(c), \tag{5}$$

where π^t is the t'th state in the path. Z is the normalizing factor with $\Sigma_c\ e_k(c) = 1$ (for further details, see Martelli et al. 2002).

The vector $\boldsymbol{s}^t$ is obtained directly from the NN outputs as

$$\boldsymbol{s}^t = (NN(B,W),NN(\mathrm{F},W)) \tag{6}$$

where W is the local context of the cysteine, and $NN(B,W)$ and $NN(F,W)$ are the NN estimated probabilities of being in the bonding (B) or nonbonding/free (F) state, respectively. In this way, the local context exploited by the NN is coupled with the global information captured by the hybrid system.

Training the HMM parameters is accomplished by using a modified expectation-maximization algorithm (Martelli et al. 2002). To keep the constraints derived by the selected HMM model (Fig. 1), the prediction of each cysteine is made using one protein at a time and by means of the Viterbi decoding (Durbin et al. 1998).

Acknowledgments

This work was partially supported by a grant of the Ministero della Università e della Ricerca Scientifica e Tecnologica (MURST) for the project "Hydrolases from Thermophiles: Structure, Function, and Homologous and Heterologous Expression", as well as a grant for a target project in Biotechnology and a project on Molecular Genetics, both of the Italian Centro Nazionale delle Ricerche (CNR; to R.C). P.L.M. is the recipient of a fellowship from CNR devoted to a target project of Molecular Genetics (Law no. 449-1997).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

Abkevich, V.I. and Shakhnovich, E.I. 2000. What can disulfide bonds tell us about protein energetics, function and folding: Simulations and bioinformatics analysis. *J. Mol. Biol.* **300:** 975–985.

Clarke, J., Hounslow, A.M., Bond, C.J., Fersht, A.R., and Daggett, V. 2000. The effects of disulfide bonds on the denatured state of barnase. *Protein Sci.* **9:** 2394–2404.

Durbin, R., Eddy, S., Krogh, A. and Mitchinson, G. 1998. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, UK.

Fariselli, P. and Casadio, R. 2001. Prediction of disulfide connectivity in proteins. *Bioinformatics* **17:** 957–964.

Fariselli, P., Riccobelli, P., and Casadio, R. 1999. The role of evolutionary information in predicting the disulfide bonding state of cysteines in proteins. *Proteins* **36:** 340–346.

Fiser, A. and Simon, I. 2000. Predicting the oxidation state of cysteines by multiple sequence alignment. *Bioinformatics* **6:** 251–256.

Fiser, A., Cserzo, M., Tudos, E., and Simon, I. 1992. Different sequence environments of cysteines and half cystines in proteins: Application to predict disulfide forming residues. *FEBS Lett.* **302:** 117–120.

Harrison, P.M. and Sternberg, M.J.E. 1994. Analysis and classification of disulfide connectivity in proteins. *J. Mol. Biol.* **244:** 448–463.

Kabsch, W. and Sander, C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22:** 2577–2637.

Krogh, A. and Riis, S.K. 1999. Hidden neural networks. *Neural Comput.* **11:** 541–563.

Martelli, P.L., Fariselli, P., Krogh, A., and Casadio, R. 2002. A sequence-profile–based HMM for predicting and discriminating β-barrel membrane proteins. *Bioinformatics* **18 (Suppl.1):** S46–S53.

Mucchielli-Giorgi, M.H., Hazout, S., and Tuffery, P. 2002. Predicting the disulfide bonding state of cysteines using protein descriptors. *Proteins* **46:** 243–249.

Muskal, S.M., Holbrook, R.S., and Kim, S.H. 1990. Prediction of the disulfide-bonding state of cysteine in proteins. *Protein Eng.* **3:** 667–672.

Narayan, M., Welker, E., Wedemeyer, W.J., and Scheraga, H.A. 2000. Oxidative folding of proteins. *Acc. Chem. Res.* **33:** 805–812.

Noguchi, T., Matsuda, T.H., and Akiyama, Y. 2001. PDB-REPRDB: A database of representative protein chains from the Protein Data Bank (PDB). *Nucleic Acids Res.* **29:** 219–220.

Skolnick, J., Kolinski, A., and Ortiz, A.R. 1997. MONSSTER: A method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.* **265:** 217–241.

Wedemeyer, W.J., Welkler, E., Narayan, M., and Scheraga, H.A. 2000. Disulfide bonds and protein folding. *Biochemistry* **39:** 4207–4216.

Welker, E., Narayan, M., Wedemeyer, W.J., and Scheraga, H.A. 2001. Structural determinants of oxidative folding in proteins. *Proc. Natl. Acad. Sci.* **98:** 2312–2316.

Web reference

http://www.biocomp.unibo.it/piero/cyspred/cysdataset.tgz; PDB codes of the proteins included in the data base, the 20-fold cross validation lists, and the training profiles.