Hidden Neural Network Prediction of Cysteine Bonding States

Tara Chari

CS 4775

December 12, 2017

Introduction

Proteins are an extremely diverse group of macromolecules that allow for the multitude of biological processes that make up every living system. With functions ranging from reaction catalysis to antibody production, understanding how proteins are formed and how a protein's structure relates to function is key to studying the organization of and processes in all organisms. Given the vast natural diversity of life today, deciphering these structural properties of proteins offers its own complexities.

For each protein, there are several levels of structural development. The primary structure can be found from simple sequencing of proteins, resulting in a linear chain of the amino acid residues that form the basis of the protein. Secondary structure involves interactions between parts of the protein chain. The functional groups attached to each amino acid allow for a variety of interactions, which affect the folding properties of a protein (Cheriyedath, 2017). The tertiary structure refers to the 3D shape of the protein due to these interactions, and can lead to quaternary protein structure which involves interactions between multiple polypeptide chains (Cheriyedath, 2017). This paper focuses on using the primary structure to predict certain secondary structure properties.

Disulfide bonds are a part of a protein's secondary structure that form between cysteine residues. The sulfhydryl group on each cysteine can bond to each other, forming disulfide bridges (Yang et al., 2015). These bridges can form both within the same chain (intra-chain) and between multiple chains (inter-chain) as shown in Figure 1.

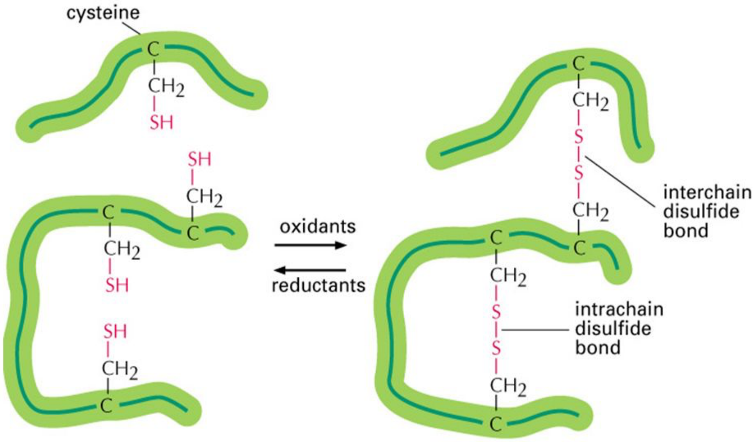

Figure 1. Illustration of inter/intra chain bonding between cysteine residues (from Molecular Biology of the Cell, 4th Edition)

These disulfide bonds help guide a protein's conformation and provide stability, by reducing the entropy of the chain(s) (from their unfolded states) (Harrison and Sternberg, 1994). Depending on the protein type, these bonds can be a part of necessary post-translational modifications and develop protein-protein interactions (Winther and Thorpe, 2014; Meitzler et al., 2013). Given the significance of disulfide bonds in the development and functionality of proteins, many computational tools have been developed to enable the prediction of these bonds.

This paper is specifically concerned with using a Hidden Neural Network (HNN) to predict the disulfide bonding states of cysteines, given a protein's amino acid sequence. Several studies have used neural nets (Fariselli et al., 1999; Martelli et al., 2002) to predict intra-chain bonding states of cysteines. The addition of a Hidden Markov Model (HMM) was shown to improve prediction accuracy on a per protein basis, from 57% to 84% (Martelli et al., 2002) on top of the neural net system. The combination of the Markov Model and the neural net is meant to encompass both local sequence information, including windows of residues around each cysteine, and global information, by taking into account all the cysteines in a protein. The goal of this paper, is to implement a similar version of this neural net and HMM combination, to provide a tool for improved prediction of cysteine bonding states.

Methods

The overall structure of the HNN system, utilizes a neural net predictor based on residue windows surrounding each cysteine, and uses the outputs of the net as emission probabilities in the HMM. The bonding states of each cysteine are defined as free or disulfide bonded. The data used to test and train the models in this paper is from the 2002 Martelli study. This dataset contained 4136 cysteine-containing segments, and each residue is defined as free or disulfide bonded using the crystallographic data from the Brookhaven Protein Data Bank (Martelli et al., 2002). The bonding assignments were determined by the Define Secondary Structure of Proteins (DSSP) program (Kabsch and Sander, 1983; Martelli et al., 2002). The proteins used were also selected based on their non-homologous nature (with an identity value < 25% and without chain breaks) using the PAPIA system (Noguchi et al., 2001; Martelli et al., 2002). Each residue was labeled with a 1, -1, or 2 with 1 being intra-chain bound, -1 being free, and 2 being inter-chain bound. Inter-chain bound residues are considered free in this study, which only made up 0.8% of the segments used. The total number of proteins in the dataset is 969 with 4136 cysteine-containing segments, where 1446 were disulfide-bonded and 2690 were free/non-bonded. From the PDB id's given for each protein, multiple sequence alignment profiles were generated using BLAST. These alignments are then used as input into the neural net. The PDB id's were also divided into 20 subsets for the 20-fold cross validation performed in the Martelli study. These cross-validation sets for training and testing were also used in the models developed for this paper.

Neural Net

The development of the neural net was based on the details given in the Martelli study and the Fariselli study (whose neural net formed the basis for the Martelli study). To form the neural net, the python package Theanets was used. As input into the neural net, each amino acid residue in a protein is treated as a length 20 vector (given 20 amino acids) (Farselli et al., 1999). Each position in the vector can represent the identity of the amino acid or the amino acid's frequency given the obtained BLAST profiles. It has been shown that using the evolutionary information provided by these profiles leads to a more accurate neural network (Fariselli et al., 1999), thus the input vectors used for each amino acid represent the frequencies rather than the identity. The input vectors lead to a neural net with 540 inputs, given 27 residue long windows around each cysteine. The outline of this general neural net is shown in Figure 2.

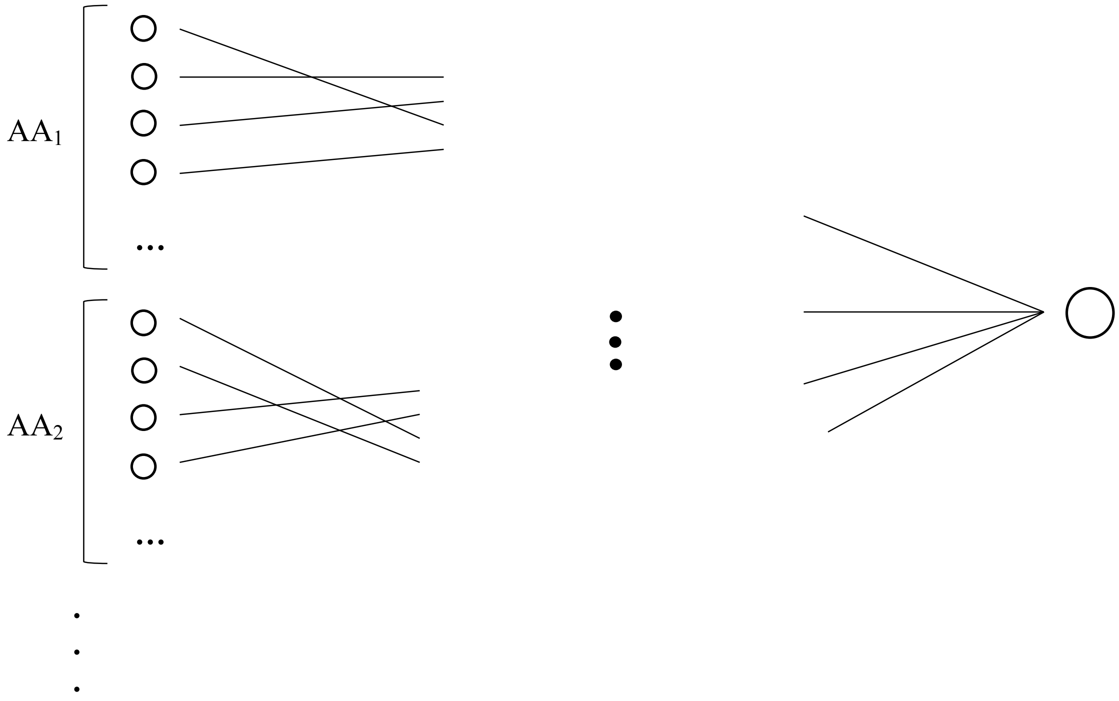

Figure 2. Input amino acid (AA) neurons, with possible hidden layer(s), with one output neuron

The network in this study, as well as those in the Martelli and Fariselli studies, are feed-forward networks that use back-propagation to compare output values to their correct values and re-adjust the weights between neurons in different layers. In order to develop the weights connecting the branches of the network, the training of the algorithms used root mean squared error to recalculate the weights. The error functions used in the Martelli and Fariselli studies are not discussed in their respective papers. Instead, the error function was determined based on the most common function used in Theanets regression networks. The training sets were based on the designated training sets given in the Martelli datasets. For training the network, 90% of the training data was used as training input while the other 10% was used as a validation set. The neural network developed used regression to determine the output value. The relative value of this output is used to designate a cysteine residue as 1 or -1 (bonded or free) (Fariselli et al., 1999). To determine the number of hidden layers and neurons inbetween the input and output neurons, several possibilities were tested including no hidden layer as this had shown promising results in the Fariselli study.

The network developed from the training data was used to predict the designated test data (designated in the Martelli dataset). A 10-fold cross validation was also done with the cross-validation sets as defined in the Martelli dataset.

HMM

The purpose of the Martelli study was to enhance the neural network for cysteine bonding prediction, by passing the information obtained from the network through an HMM. The HMM implemented in this paper, based on the work detailed in the Martelli paper, is a vector-based HMM that uses the neural network's output as emission probabilities. The state path diagram, shown in Figure 3, includes two bonded and not bonded states. This is because four is the minimum number of states required to assure prediction of even numbers of paired/bonded cysteines (to maintain realistic predictions).

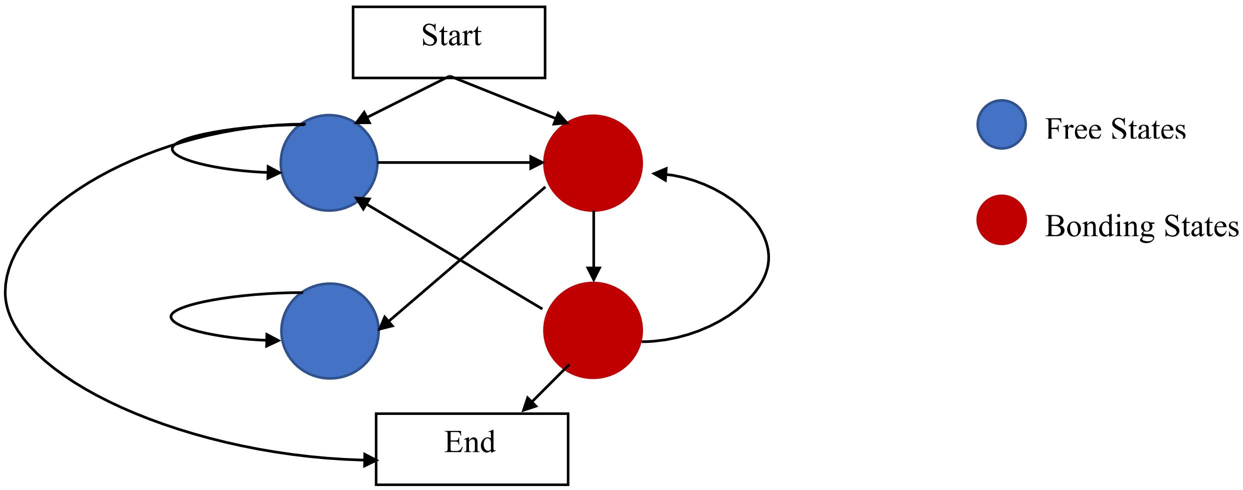

Figure 3. State path diagram of cysteine free/bonded states in given protein sequence

Once the state path probabilities have been obtained for a given protein sequence, Viterbi decoding is used to predict the most probable bonding state of each cysteine in the protein.

Instead of interpreting the protein sequences/cysteines as a sequence of characters, the model instead uses a sequence of vectors (Martelli et al., 2002). To define these sequence vectors, *L* is the number of cysteines in the protein and *A* is the size of the alphabet over which the vectors are built (meaning *A* = 2 since there are bonding and nonbonding cysteine states). This means that sequence vector notation can be written as:

$$s = s^1 s^2 \ldots\ldots s^L = \big(s^1(1), s^1(2)\big)\big(s^2(1), s^2(2)\big) \ldots \big(s^L(1), s^L(2)\big) \tag{1}$$

The components of each vector s^t (regardless of position *t*) are positive, and sum to a constant value.

As previously shown in Figure 3, the HMM used is a Markov model with four states. Each of these states is connected by transition probabilities, represented as a_{ij} (from state i to j). The derivations of the emission and transition probabilities followed the format defined in the Martelli study. The probability density function for the emission of these vectors (from each state) includes the number of parameters (*A*) that are unique to each state (*k*) which are written as $e_k(c)$ where *c* = 1, 2, …*A* (or simply 1, 2):

$$P(s^t|\pi^t = k) = \left(\frac{1}{Z}\right)\sum_c s^t(c)e_k(c) \tag{2}$$

Here π^t is the *t*th state in the given path. *Z* is the normalizing factor which ensures that $\sum_c e_k(c) = 1$. Compared to the usual emission probability $e_k(s^t)$ used in an HMM (Durbin et al., 1998), this HMM substitutes the emissions probability with

$$(\frac{1}{Z}) \sum_c e_k(c) s^t(c) \tag{3}$$

The neural network is used to directly obtain the vector s^t defined in (1). Using the neural net output, s^t is defined as:

$$s^t = (NN(B,W), NN(F,W)) \tag{4}$$

where *W* encompasses the local context/window of the particular cysteine, and *NN* stands for neural net. *NN(B,W)* and *NN(F,W)* are then the estimated probabilities of the cysteine being in a bonded (*B*) or free (*F*) state, respectively. The hope of this hybrid system was to increase the predictive capabilities of the local information used in the neural net with the global information of the HMM.

The training of this HMM was also based on the generic form of the Expectation-Maximization algorithm (Durbin et al., 1998). The transition and emission probabilities were updated using:

$$a_{ij} = A_{ij} / \sum_{j=1}^{N} A_{ij} \tag{5}$$
$$e_k(c) = C_k(c) / \sum_c C_k(c) \tag{6}$$

where:

$$A_{ij} = \sum_{d \epsilon D} 1/(Z^L P(d|M)) \sum_{t=1}^{L} f_i(t-1) \\ \times a_{ij} b_j(t) \sum_c e_j(c) s^t(c) \tag{7}$$

$$C_k(c) = \sum_{d \epsilon D} 1/(Z^L P(d|M)) \sum_{t=1}^{L} f_i(t) \\ \times b_j(t) \sum_c e_j(c) s^t(c) \tag{8}$$

and *d* represents the sequences in the training set *D*, *M* is the Markov model, *f* represents the forward probabilities, and *b* represents the backward probabilities. The updating of the parameters during the HMM training was stopped once there was a small/no change in the (log) likelihood. This was around a difference of less than 0.001. Starting values for the emission and transition probabilities were randomly generated (except for the transitions designated as zero).

Using one protein at a time, each cysteine's bonding state was predicted using the Viterbi path through the protein sequence. The final predictions from both the neural net and the total HNN were compared to the results of the Martelli study, and its predecessors.

Results

The main metric used and reported in the Martelli and Fariselli papers, to compare the predictive capabilities of the models developed, was residue accuracy:

$$Q2 = P/N \tag{9}$$

where P represents the total number of correctly predicted cysteines and N is the total number of residues. Another measure for comparison was accuracy per protein:

$$Q2_{prot} = P_p/N_p \tag{10}$$

where P_p represents the number of proteins with all their cysteines correctly predicted, and N_p is the total number of proteins.

In the development of the neural net, it was determined that the highest accuracy (per cysteine residue) was obtained using a neural net with no hidden layers (only the input and output layers). The addition of a two neuron hidden layer between the input and output layers, as used in the Martelli model, lowered protein accuracy by 10%. This finding, regarding the benefit of a lack of hidden layers, was reported by the Fariselli paper. The Martelli study also did not discuss if other hidden layers were tested in their study, and if varying those layers had significant effects. The residue window size was maintained at a length of 27 residues (as defined in the Martelli paper), to keep the window constant while varying the hidden layer details.

Table 1 below reports the accuracy scores, both per cysteine and per protein for the neural net (no hidden layer) and HNN models developed in this paper. The average accuracies (from the 10-fold cross validation) are reported, as well as the highest accuracies obtained. The accuracy values reported in the Martelli study are also noted as comparison.

Table 1. Comparison of re-implementation and original neural net/HNN accuracies

	Re-implemented Neural Net	Re-implemented HNN	Martelli et al Neural Net	Martelli et al HNN
Protein Accuracy ($Q2_{prot}$)	Max: 91% Avg: 85%	Max: 85% Avg: 80%	Max: n/a Avg: 57%	Max: 84% Avg: n/a
Cysteine Accuracy (Q2)	Max: 91% Avg: 88%	Max: 85% Avg: 81%	Max: n/a Avg: 80%	Max: 88% Avg: n/a

Discussion

Though the goal of the Martelli study, through the combination of the neural net and the HMM, was to enhance the locally-focused abilities of the neural net, the Theanets based neural net developed in this paper seemed to exceed the predictive capabilities of both the neural net and HNN developed in the Martelli study (on the same test sets). Interestingly, the addition of the HMM in this paper did not lead to an HNN with an increased accuracy (per cysteine or per protein).

These findings suggest that there are quite significant differences in the development of the neural net in this paper versus in the Martelli study. Due to the time difference since the publishing of the Martelli model and the lack of details into the parameters of the neural net (especially in relation to the back-propagation/error functions), it is possible that a neural net is now capable of the same predictive/learning capabilities as an HNN. The inability for the addition of an HMM to significantly increase the accuracy of bonding state prediction could be the result of the inherent nature of the given protein sequences, which may not, alone, provide enough information to predict bonding at a higher accuracy. The randomized starting emission/transition probabilities may also be a player in this inability, but several rounds of randomly generated probabilities did not produce significantly different results (in the HNN's accuracy).

In the future, it may be helpful (in improving the neural net's accuracy) to include other information as input. This could be information like subcellular localization, correlated mutations, and other non-traditional sequence-derived features (Yang et al., 2015).

References

1. Durbin,R., Eddy,S., Krogh,A. and Mitchinson,G. (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge Univ. Press, Cambridge.

2. Fariselli, P., Riccobelli, P. and Casadio, R. (1999), Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. Proteins, 36: 340–346. doi:10.1002/(SICI)1097-0134(19990815)36:3<340::AID-PROT8>3.0.CO;2-D

3. Harrison,P.M. and Sternberg,M.J. (1994) Analysis and classification of disulphide connectivity in proteins: the entropic effect of cross-linkage. J. Mol. Biol., 244, 448–463

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

5. Martelli, P. L., Fariselli, P., Malaguti, L. and Casadio, R. (2002), Prediction of the disulfide-bonding state of cysteines in proteins at 88% accuracy. Protein Science, 11: 2735–2739. doi:10.1110/ps.0219602

6. Meitzler,J.L. et al. (2013) Conserved cysteine residues provide a protein-protein interaction surface in dual oxidase (DUOX) proteins. J. Biol. Chem., 288, 7147–7157

7. 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.

8. Smith, Y. (2017, May). *Protein Structure and Function*. Retrieved from (https://www.news-medical.net/life-sciences/Protein-Structure-and-Function.aspx

9. Winther,J.R. and Thorpe,C. (2014) Quantification of thiols and disul- fides. Biochimica et Biophysica Acta (BBA)-General Subjects, 1840, 838–846.

10. Yang, J., He, B., Jang, R., Zhang, Y., Shen, H. (2015) Accurate disulfide-bonding network predictions improve ab initio structure prediction of cysteine-rich proteins. *Bioinformatics*. 31(23), 3773-3781. https://pdfs.semanticscholar.org/b633/3d3a9f72af2009c7c4ff99b5b722f51b49e9.pdf