Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Prediction of cystine connectivity using SVM

Jayavardhana Rama G. L.[1*], Alistair P. Shilton[1], Michael M. Parker[2], Marimuthu Palaniswami[1]

[1]Department of Electrical and Electronics Engineering, The University of Melbourne, Parkville, Victoria – 3010;

[2]St. Vincent's Institute of Medical Research, Fitzroy, Victoria – 3065;

Jayavardhana Rama G. L.* - Email: jrgl@ee.unimelb.edu.au; * Corresponding author

received November 10, 2005; revised December 6, 2005; accepted December 7, 2005; published online December 7, 2005

Abstract:

One of the major contributors to protein structures is the formation of disulphide bonds between selected pairs of cysteines at oxidized state. Prediction of such disulphide bridges from sequence is challenging given that the possible combination of cysteine pairs as the number of cysteines increases in a protein. Here, we describe a SVM (support vector machine) model for the prediction of cystine connectivity in a protein sequence with and without *a priori* knowledge on their bonding state. We make use of a new encoding scheme based on physico-chemical properties and statistical features (probability of occurrence of each amino acid residue in different secondary structure states along with PSI-blast profiles). We evaluate our method in SPX (an extended dataset of SP39 (swiss-prot 39) and SP41 (swiss-prot 41) with known disulphide information from PDB) dataset and compare our results with the recursive neural network model described for the same dataset.

Keywords: disulphide bridges; prediction; protein fold; SVM model; SPX dataset

Background:

The completion of the human genome project shows a significant gap between the protein sequence and known structure space. Determination of protein structures using conventional X-ray crystallography and NMR (nuclear magnetic resonance) techniques is not adequate to cover the sequence space in the context of drug discovery. Hence, protein structure prediction using computational methods is becoming critical. However, prediction of protein tertiary structure from sequence is non-trivial and is generally achieved by dividing the problem into finite levels of secondary structures and super secondary structures.

The native protein fold is dependent on the physical-chemical properties of the amino acid residues in the sequence. Disulphide bonds between cysteines are important features in the formation of several protein folds. It is shown that cysteines are highly conserved in a protein family and they exit in either oxidized or reduced states. **[1-3]** The cystines in oxidized state form covalent bond between each other and are referred as disulphide bridges. A schematic representation of conotoxin (PDB (protein databank) ID 1AS5) showing disulphide bonds is given in Figure 1. Information about the location of disulphide bridges find application in the understanding of protein folding **[1]** and have a role in thermodynamic stability of proteins. **[2]** Hence, studies on disulphide bridges have become very important.

Fariselli *et al.,* **[2]** proposed a disulphide prediction model combining a neural network based predictor and evolutionary data with an accuracy of 81%. In 2000, Fiser and Simon **[3]** proposed a method based on multiple sequence alignment and reported an accuracy of 82% using Jack Knife test on a larger dataset of 81 proteins. Martelli *et al.*, **[4]** proposed a Hidden Neural Network method (a combination of Hidden Markov Model and Neural Network) with an accuracy of 84% for a larger data set of 969 non-homologous proteins.

Vullo and Frasconi **[5]** used recursive neural networks and evolutionary data to predict bonding patterns using known information on cystine bonding states. The method was tested using a small dataset derived from Swiss-Prot release 39 (SP39) and an accuracy of 48% was reported. Prior to this, Fariselli and Casadio **[6]** linked connectivity prediction to graph matching. They also showed better connectivity prediction by combining with neural network models.

Recently, Ferre and Clote **[7]** emphasized the importance of secondary structure and solvent accessibility information in the development of a di-residue neural network model for predicting disulphide bridges. Cheng and colleagues discussed ways to find and count (using recursive neural network) disulphide bridges in a given sequence and tested the model performance in SPX (an extended dataset of SP39 and SP41 with known disulphide information from PDB). **[8]** Here, we describe a SVM (support vector machine) model for predicting cysteine bonding state as an extension of the work by Cheng and colleagues. **[8]** In this method, we predict disulphide bond connectivity given two cysteines with and without *a priori* knowledge on their bonding state using the SPX dataset.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net

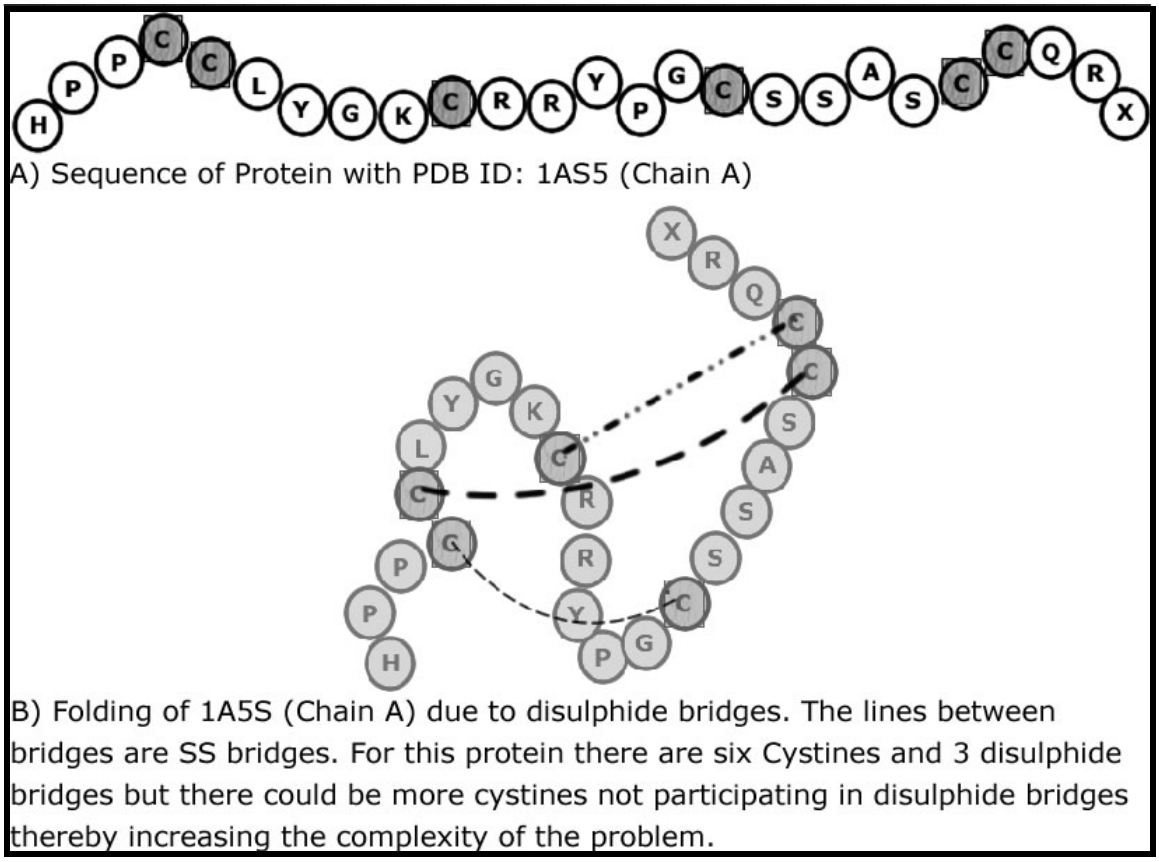

Figure 1: A schematic representation of CONOTOXIN (PDB (protein databank) ID 1AS5) showing disulphide bonds.

Methodology:

Support Vector Machines:

SVM (Support Vector Machine) is a class of tool used in classification and regression as described elsewhere by Vapnik. **[9]** When used as a binary classifier, an SVM will construct a hyperplane which acts as the decision surface between the two classes. This is achieved by maximizing the margin of separation between the hyperplane and those points nearest to it. The idea is further extended for data that is not linearly separable by first mapping it to a possibly higher dimension feature space. The SVM formulation is desirable due to its mathematical tractability and good generalization properties.

The data to be classified is formally written as

$$\Theta = \{(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)\}$$
$$x_i \in \Re^m$$
$$y_i \in \{-1, 1\} \quad \text{(Equation 1)}$$

The nonlinear feature map $\phi(x) : x \subset \Re^m \rightarrow \Re^d$ ($m << d$) is never explicitly used in the calculation. Vapnik **[9]** suggests the form of the hyperplane $f(x) \in F$ to be chosen from a family of functions with sufficient capacity. In particular, F contains functions for the linearly and non-linearly separable hyperplane having the following forms:

$$f(x) = \sum_{i=1}^{n} w_i x_i + b \quad \text{(Equation 2)}$$

$$f(x) = \sum_{i=1}^{n} w_i \phi_i(x) + b \quad \text{(Equation 3)}$$

Now for separation in feature space, we would like to obtain the hyperplane with the following properties:

$$f(x) = \sum_{i=1}^{n} w_i \phi_i(x) + b$$
$$f(x) > 0 \forall i : y_i = +1$$
$$f(x) < 0 \forall i : y_i = -1$$
(Equation 4)

The conditions in equation Equation 4 can be described by a strict linear discriminant function, so that for each element pair in Θ we require:

$$y_i \left(\sum_{i=1}^{n} w_i \phi_i (x) + b \right) \geq 1$$ (Equation 5)

The distance from the hyper-plane to points lying closest to it is given geometrically as $\frac{1}{\|w\|}$. The soft-margin minimization problem relaxes the strict discriminant in equation 5 by introducing slack variables, ξ_i and is formulated as:

$$\min_{w,\xi} \Im(w,\xi) = \frac{1}{2}\sum_{i=1}^{n} w_i^2 + C\sum_{i=1}^{n} \xi_i$$
$$\text{s.t} \quad \begin{cases} y_i \left(\sum_{i=1}^{n} w_i \phi_i (x) + b \right) \geq 1 + \xi_i \\ \xi_i > 0 \\ \forall i = 1..n \end{cases}$$
(Equation 6)

The constant C is selected so as to compromise between the minimization of training error and prevention of over-fitting. Applying Lagrangian Theory, the following dual problem in terms of Lagrange multipliers α_i is usually solved

$$\min_{\alpha \in D} \Im(\alpha) = -\sum_{i=1}^{n} \alpha_i + \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$
$$D = \left\{ \alpha \middle| 0 \leq \alpha_i \leq C, \sum_{i=1}^{n} \alpha_i y_i = 0 \right\}$$
(Equation 7)

The explicit use of the nonlinear function $\phi(.)$, has been circumvented by the use of a kernel function, defined formally as the dot products of the nonlinear functions

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$$ (Equation 8)

Kernels can be chosen according to Mercer's theorem. In all our experiments we use polynomial kernel with degree d = 2 given by

$$K(x_i, x_j) = (1 + x_i \cdot x_j)^d$$ (Equation 9)

This was chosen based on preliminary experiments involving fewer protein chains. The SVM classifier is given by:

$$f(x) = sign\left(\sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b \right)$$ (Equation 10)

Disulphide bonding patterns in proteins:

The human alkaline phosphatase (PDB ID: 1EW2) have 5 cysteines with 2 disulphide bonds formed between 2nd - 3rd and 4th -5th cysteines in the order of the sequence. It should be noted that the 1st cysteine is not involved in any disulphide bond formation. This describes the nature and selectivity of disulphide bond formation in human alkaline phosphatase and gives information on the bonding states of the cysteines in the sequence. However, disulphide bonds are formed in various combinations in different proteins. Therefore, it is of potential interest to predict the nature of disulphide bonds from sequence for which structure is unknown. Nonetheless, this task is non-trivial and predictions of disulphide bonds are generally preformed with and without prior knowledge on cysteine bonding states in a sequence of interest. If we have to predict the disulphide bonding patterns in human alkaline phosphatase assuming the structure is un known, then it can performed either with or without a prior knowledge on the bonding state of cysteines. Prediction of disulphide bonding patterns with prior knowledge on the bonding state (6 different possible combinations) is relatively simpler to that without any prior knowledge on the bonding state of the cysteines (10 different possible combinations) in human alkaline phosphatase.

Dataset:

The SPX dataset was created by Cheng *et al.*, **[8]** was used in this study. The dataset contains non-homologous (at a sequence similarity cut-off of < 25%) sequences (containing information on intra-chain disulphide bonds) from PDB.

Feature parameters:

We used five parameters for each cysteine based on physico-chemical properties and probability of occurrence in secondary structures (alpha helix, beta strand, coil), Chou-Fasman conformational parameters **[10]** (3 in number), Kyte-Dolittle hydrophobicity scale **[11]** and Grantham polarity **[12]** (1 in number each) were chosen as features. The Chou-Fasman parameter for helix (α) is given by $P_{\alpha i} = f_{\alpha i} / \langle f_\alpha \rangle$, where, $\langle f_\alpha \rangle$ = (number of residues in helix / total number of residues) and i is the set

of amino acids residues. Similar conformational parameters for strand $P_{\beta i}$ and coil $P_{\gamma i}$ were calculated. Kyte-Dolittle hydrophobicity values and Grantham Polarity values were taken from the Protscale website. **[17]** We chose the above parameters after preliminary experimentation with a small dataset (30 protein chains) at different hydrophobic and polarity scales.

Use of homologous sequence information:

Recent CAFASP and CASP results showed that the use of homologous sequences can improve secondary structure prediction, solvent accessibility calculations and cystine connectivity identification. This attempts to capture the evolutionary information for sequences and is generated by developing matrices from sequence profiling. The PSSM (position specific scoring matrix) is generated by calculating position-specific scores for each position from sequence profiles and the scores are a measure of residue variability or similarity in the profile. **[13]** The PSSM generated by PSI-BLAST (http://www.ncbi.nlm.nih.gov/) from a non-redundant (NR) dataset of protein sequences was used in this analysis with an E-value (expect value) of 0.001 at 3 iterations. A window of length w was considered for every cysteine under consideration at the center of the window and this is used as a feature for the classifier. In PSSMs, there are $w*L$ elements and L is the protein length. In this study, we used L = 5 after several trails. The PSSM values vary approximately between -10 and +10. However, SVM require values between 0 and +1. Therefore, we normalized the PSSM values using the following function as described elsewhere. **[14]**

$$g(x) = \begin{cases} 0.0 & x <= -5 \\ 0.5 + 0.1x & -5 < x < 5 \\ 1.0 & x >= 5 \end{cases}$$

In this formulation x is the value in the PSSM matrix. Instead of taking just 20 values per residue as a feature vector, we considered a window of length w and all the values within the window were considered in feature definition. **[13]** We were able to incorporate the gradual variation required for the classifier to make a better decision by selecting a window w = 5 for PSSM values. We included 5 X 20 PSSM values in addition to five physical-chemical features for every cysteine under consideration and the total feature length for every cysteine was 105. Hence, the final feature length for each cysteine pair is $((w*20)+5)*2$.

SVM parameters and performance measures:

We use SVM with $C = 10$ and a polynomial kernel with $D = 2$ in this analysis. We used the SVM implementation SVMHeavy developed based on incremental training of support vector machines as described elsewhere. **[14,16]** A five fold cross validation was performed for each experiment reported in the study. We compared the performance of the model with the results of Cheng and colleagues using specificity, sensitivity and accuracies Q_c and Q_p. Specificity is the ability to reject false positive matches given by $TN/(FP+TN)$ and sensitivity is the ability to detect true positive matches given by $TP/(TP+FP)$ (TP = True Positive; FP = False Positive; TN = True Negative). Q_c defined per disulphide bond is given by (TP + TN)/(TP+TN+FP+FN) and Q_p is the accuracy defined per protein sequence.

Table 1A: Disulphide Bridge Prediction with *a priori* knowledge about bonding state

Number of Bridges	Specificity ┼	Sensitivity ┼	Specificity	Sensitivity
1	0.48	0.71	0.61	0.65
2	0.63	0.63	0.63	0.61
3	0.67	0.62	0.66	0.60
4	0.55	0.50	0.61	0.51
5	0.41	0.37	0.56	0.38
6	0.33	0.29	0.59	0.37
7	0.36	0.31	0.47	0.36
8	0.32	0.30	0.44	0.32
9	0.71	0.61	0.55	0.35
10	0.40	0.37	0.59	0.45
12	0.55	0.50	0.60	0.50

14	0.62	0.57	0.65	0.58
16	0.23	0.22	0.43	0.25
17	0.40	0.35	0.51	0.31
25	0.40	0.24	0.63	0.30
26	0.73	0.42	0.69	0.30
Overall	0.54	0.55	0.62	0.59

Table 1B: Disulphide Bridge Prediction without *a priori* knowledge about Bonding State

Number of Bridges	Accuracy at Bridge level ┼	Accuracy at Protein Level ┼	Accuracy at Bridge level	Accuracy at Protein Level
1	-	0.59	0.65	0.53
2	-	0.59	0.59	0.50
3	-	0.54	0.61	0.56
4	-	0.34	0.63	0.46
Overall	-	0.51	0.63	0.52

┼Chang *et al.,* **[8]**

Results and Discussion:

Prediction of disulphide bonds from sequence has a critical role to play in protein fold identification and folding simulation. A number of statistical models have been described using ANN (artificial neural network), HMM (hidden Markov model) and evolutionary algorithm for the prediction of disulphide bonding patterns in protein sequence. **[2-8]** However, a SVM model was not available for disulphide boding pattern prediction in protein sequences. Table 1A shows the performance of the described SVM model (with prior knowledge on disulphide bonding states). The results were compared with the recursive neural network model by Cheng and colleagues **[8]** in SPX dataset. We compared with the results of Cheng and colleagues **[8]** because the dataset used in the both studies were identical. The comparison shows that the SVM method (4% higher sensitivity and 8% higher specificity) performs better than the recursive neural network model for classification with *a priori* knowledge. Although, the method performs better than the recursive neural network model, variations in performance are noticed among different prediction runs.

Table 1B shows the performance of the described SVM model (without *a priori* knowledge on disulphide bonding states) and compares with the results of a recursive neural network by Cheng and colleagues **[8]** in SPX dataset. The results from SVM model were found to be similar to that of the recursive neural network presented by Cheng and colleagues. **[8]** We measured the performance using the overall accuracy for disulphide bridges and proteins. These results (Table 1) show the utilization of SVM models for the prediction of disulphide connectivity in proteins. In our opinion, the combination of SVM parameters and the encoding method chosen in model development played an important role in better performance even in small datasets.

Conclusion:

Disulphide bridge pattern identification for fold prediction from sequence is not trivial. In this paper, we have described a SVM model to predict disulphide bridges with and without *a priori* knowledge on their bonding states. The SVM method is found to perform better than a recursive neural network model described elsewhere. **[8]** In future investigation, we plan to extend our approach to classify sequences with and without disulphide bonds.

Acknowledgment:

The authors would like to thank Prof. David Jones for providing the *pfilt* program. We are also grateful to NCBI for *PSI-BLAST* program and Cheng and colleagues for making the *SPX* dataset available on the web. We also thank the anonymous reviewers for help in improving the manuscript.

References:

[1] A. Tony, *et al., European Journal of Biochemistry*, 267:566 (2000) [PMID: 10632727]

[2] P. Fariselli, *et al., Proteins: Structure, Function, and Genetics*, 36:340 (1999) [PMID: 10409827]

[3] A. Fiser & I. Simon, *Bioinformatics*, 16:251 (2000) [PMID: 10869018]

[4] P. L. Martelli, *et al., Protein Engineering*, 15:951 (2002)

[5] A. Vullo & P. Frasconi, *Bioinformatics*, 10:653 (2004) [PMID: 15033872]

[6] P. Fariselli & R. Casadio, *Bioinformatics*, 17:957 (2001) [PMID: 11673241]

[7] F. Ferre & P. Clote, *Bioinformatics*, 21:2336 (2005) [PMID: 15741247]

[8] J. Cheng, *et al., Proteins: Structure, Function, Bioinformatics*, 1 (2005)

[9] C. Cortes & V. Vapnik, *Machine Learning*, 20:273 (1995)

[10] P.Y. Chou & G.D. Fasman, *Biochemistry*, 13 :211 (1974)
[11] J. Kyte & R.F. Doolittle, *Journal of Molecular Biology*, 157:105 (1982) [PMID: 7108955]
[12] R. Grantham, *Science*, 185:862 (1974) [PMID: 4843792]
[13] D. T. Jones, *Journal of Molecular Biology*, 292:195 (1999) [PMID: 10493868]
[14] H. Kim & H. Park, *Protein Engineering*, 16:553 (2003) [PMID: 12968073]
[15] A. Shilton, *et al., IEEE Transactions on Neural Networks*, 16:114 (2005).
[16] http://www2.ee.mu.oz.au/pgrad/apsh/svm/
[17] http://au.expasy.org/tools/protscale.html

Edited by P. Kangueane
Citation: Jayavardhana Rama *et al.,* Bioinformation 1(2): 69-74 (2005)