Disulphide Bridge Prediction using Fuzzy Support Vector Machines

Jayavardhana Rama G. L.†, A. Shilton‡, Michael M. Parker‡, Palaniswami M.†
†Department of Electrical and Electronics Engineering
The University of Melbourne, Parkville, Victoria - 3010
‡St. Vincent’s Institute of Medical Research
Fitzroy, Victoria - 3065

Abstract—One of the major contributors to the native form of protein is cystines forming covalent bonds in oxidized state. The Prediction of such bridges from the sequence is a very challenging task given that the number of bridges will rise exponentially as the number of cystines increases. We propose a novel technique for disulphide bridge prediction based on Fuzzy Support Vector Machines. We call the system DiSzy. In our investigation, we look at disulphide bond connectivity given two Cystines with and without a priori knowledge of the bonding state. We make use of a new encoding scheme based on physico-chemical properties and statistical features such as the probability of occurrence of each amino acid in different secondary structure states along with psi-blast profiles. The performance is compared with normal support vector machines. We evaluate our method and compare it with the existing method using SPX dataset.

I. INTRODUCTION

The results of the Human Genome project have left a significant gap between the availability of the protein sequence and its corresponding structure [28]. The dependence on experimental methods may not yield protein structures fast enough to keep up with the requirement of today’s drug design industry. X-ray crystallography, which is the most popular technique for protein structure determination, makes use of tertiary structure in the molecular modelling stage [1], [18]. With the availability of abundant data, it has been shown that it is possible to predict the structure using machine learning techniques. As the prediction of tertiary structure from protein sequence is a very difficult task, the problem is usually subdivided into secondary structure prediction and super secondary structure prediction leading to tertiary structure. At the secondary structure level Solvent Accessibility and Disulphide Bridge prediction has been shown to significantly effect further processing.

There are twenty amino acids which are the primary building blocks of proteins. The chemical properties of these amino acids are mainly responsible for the unique fold each protein attains in its native state. Cystine is a unique amino acid which largely contributes to the final fold, function and evolution of proteins. These are often the most conserved regions in proteins. The proteins that contain these cystines are often subject to post-translational covalent modifications. Cystines can occur in two basic forms - Oxidized and Reduced State. The cystines in oxidized state often form covalent bond between each other in what is called a Disulphide Bridge.

With this background, we take a look at predicting disulphide bridges given a protein sequence alone. The problem is complex and can be further subdivided into several simpler ones. First of all, given a cystine we have to find out whether it is bondable or not. Secondly, we have to find whether two cystines are bonded when their bonding status is known (Disulphide Bridge Classification) and finally, without any priori knowledge, if it is possible to detect all the disulphide bridges in a given sequence (Disulphide Bridge Pattern Prediction). Several papers are available to predict the bonding state of Cystines. Fariselli et. al. [9] propose combining a neural network based predictor along with evolutionary information and report the highest accuracy of 81%. In 2000, Fiser and Simon [11] proposed a method based on multiple sequence alignment and reported an accuracy of 82% using Jack Knife test on a slightly larger dataset. Martelli et. al. [20] proposed a Hidden Neural Networks, a combination of hidden markov Model and Neural Network and reported 84% accuracy for a new larger data set. In the same year Frasconi et. al. [12] proposed a dual layered support vector machine for prediction of bonding state and reported a healthy accuracy of 84% and led the way to combined prediction of classification of sequences based on cystine content and finding the bonding state. As it can be seen, substantial work has been done in predicting bonding state. However, we feel little work has been done in prediction of the entire disulphide pattern. Vullo and Frasconi [29] used recursive neural networks and evolutionary information to predict bonding pattern with the knowledge of the Cystine bonding states. They tested their method on a small set of data from Swiss Prot (SP39). Prior to this, Fariselli and Casadio [7] equated the problem of connectivity prediction to graph matching. They furthered their work by making use of a Neural network [8] and reported a decent recognition accuracy. More recently in 2005, Ferre and Clote [10] emphasized the importance of secondary structure information and proposed a diereside neural network for predicting the bridges. Apart from these some substantial work has been done by Cheng et. al. [4], [3]. They have investigated the discussed questions in detail and raised two more questions, given a sequence, namely: is it possible to find whether it has disulphide bond or not and secondly how many disulphide bonds are contained in the sequence containing the bonds. They used a Support Vector Machines (SVM) to answer the first question and recursive
a neural network to answer the rest of the questions. They also curated a new dataset as an extension to SP39 and SP41 datasets, which they called the SPX dataset. Cheng et al. reported the highest accuracy so far in spite of larger curated dataset. We attempt to extend their work by improving the highest reported accuracy with the help of Fuzzy Support Vector Machines (FSVM).

In this paper, we propose a novel technique based on Fuzzy Support Vector Machines for Disulphide Bridge Prediction. We call the system DIzzy. In our investigation we look at disulphide bond connectivity given two cystines with and without a priori knowledge about the bonding status. Hence we address the two major problems of bridge classification and bridge prediction. As the disulphide bridge is affected by chemical and physical characteristics, we make use of a novel encoding scheme. We first investigate the performance of normal Support Vector Machines (SVM) and then compare it with the performance after fuzzifying the feature space.

II. DISULPHIDE BRIDGES

The native form of protein is largely the result of to the chemical properties of the amino acids. Formation of disulphide bonds between cystines is a unique feature responsible for the final fold, function and evolution of proteins. As a result they are the most conserved regions in the proteins. Cystines can occur in basically two forms - Oxidized and Reduced State. The cystines in oxidized state often form covalent bond between each other which is termed as Disulphide Bridge. The chemical reaction forming the bridge is given by the following equation:

$$2CH_2SH + \frac{1}{2}O_2 \rightarrow CH_2SSCH_2 + H_2O$$

Information about the location of the disulphide bridges will reduce the search space in the protein folding problem [23], [16]. The presence of disulphide bridges results in thermodynamic stability in any given proteins [2] and the reduction of the same leads to functional differences [9]. Hence the study of disulphide bridge pattern has become very important. Figure 1 shows the sequence of protein 1AS5 (Chain A) and the folding nature due to disulphide bonding (2).

III. SUPPORT VECTOR MACHINES

Support Vector Machines (SVM) are a new class of tools used in classification and regression introduced by Vapnik [5]. 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

$$\Theta = \{ (x_1, y_1), (x_2, y_2), \ldots (x_n, y_n) \}$$

$$x_i \in \mathbb{R}^m$$

$$y_i \in \{-1, 1\}$$

The nonlinear feature map $\phi(x): x \in \mathbb{R}^m \rightarrow \mathbb{R}^d (m \ll d)$ is never explicitly used in the calculation. Vapnik [5] suggests that the form of the hyperplane, $f(x) \in \mathbb{F}$ be chosen from a family of functions with sufficient capacity. In particular, $\mathbb{F}$ contains functions for the linearly and non-linearly separable hyperplane having the following forms:

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

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

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

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

$$f(x) > 0 \ \forall \ i : y_i = +1$$

$$f(x) < 0 \ \forall \ i : y_i = -1$$

The conditions in eq. 4 can be described by a strict linear discriminant function, so that for each element pair in $\Theta$ we require:

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

The distance from the hyperplane to those points lying closest to it is $\frac{1}{\|w\|}$ by geometry. The soft-margin minimization

Fig. 1. Sequence of Protein with PDB ID: 1AS5 (Chain A)

Fig. 2. Folding of 1AS5 (Chain A) due to disulphide bridges. The lines between bridges are SS bridges. For this protein there are six Cystines and 3 disulphide bridges but there could be more cystines not participating in disulphide bridges thereby increasing the complexity of the problem.
problem relaxes the strict discriminant in (5) by introducing slack variables, $\xi_i$ and is formulated as:

$$\min_{\mathbf{w}, \xi} \mathcal{J}(\mathbf{w}, \xi) = \frac{1}{2} \sum_{i=1}^{n} w_i^2 + C \sum_{i=1}^{n} \xi_i$$

s.t. $$\begin{align*}
y_i \left( \sum_{i=1}^{n} w_i \phi_i(\mathbf{x}) + b \right) &\geq 1 + \xi_i \\
\xi_i &> 0 \\
\forall i &= 1..n
\end{align*}$$

(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 $\alpha_i$ is usually solved

$$\min_{\mathbf{D}} \mathcal{J}(\alpha) = -\sum_{i=1}^{n} \alpha_i + \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\mathbf{D} = \left\{ \alpha | 0 \leq \alpha \leq C, \sum_{i=1}^{n} \alpha_i y_i = 0 \right\}$$

(7)

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

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$$

(8)

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

$$K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i \cdot \mathbf{x}_j)^d$$

(9)

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

$$f(\mathbf{x}) = \text{sign} \left( \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \right)$$

(10)

IV. FUZZY SUPPORT VECTOR MACHINES

In most real world problems we find that a given situation can fall into different categories with different levels of confidence. Fuzzy logic is a tool that embeds such human knowledge into workable algorithms [27]. When we incorporate such fuzziness into our SVM input the resulting machine is called a Fuzzy Support Vector Machine (FSVM). Based on the earlier work by Lin and Wang [19], we reformulate SVM to FSVM and apply them to disulphide bridge classification and prediction.

In FSVM, the data to be classified can be formally written as

$$\Theta = \{(x_1, y_1, s_1), (x_2, y_2, s_2), \ldots, (x_n, y_n, s_n)\}$$

$$x_i \in \mathbb{R}^m, y_i \in \{-1, 1\}, s_i \in S$$

(11)

where $S$ is a set of fuzzy membership values derived membership function (i.e. attitude of $x_i$ towards one class), $\sigma \leq s_i \leq 1$ and $\sigma > 0$. From the standard SVM formulation, $\xi_i$ is the measure of classification error. In the FSVM case, $s_i \xi_i$ is the new (weighted) classification error. Hence the FSVM problem primal is:

$$\min_{\mathbf{w}, \xi} \mathcal{J}(\mathbf{w}, \xi) = \frac{1}{2} \sum_{i=1}^{n} w_i^2 + C \sum_{i=1}^{n} s_i \xi_i$$

s.t. $$\begin{align*}
y_i \left( \sum_{i=1}^{n} w_i \phi_i(\mathbf{x}) + b \right) &\geq 1 + \xi_i \\
\xi_i &> 0 \\
\forall i &= 1..n
\end{align*}$$

(12)

where $C$ is the regularization parameter as before. Hence we say that an input point is less important by making $s_i$ small, the effect of the error is thereby reduced allowing compromise in the treatment of corresponding point $x_i$.

This optimization problem is solved by constructing the Lagrangian

$$L(\mathbf{w}, b, \xi, \alpha, \beta) = \frac{1}{2} \mathbf{w} \cdot \mathbf{w} + C \sum_{i=1}^{n} s_i \xi_i$$

(13)

$$-\sum_{i=1}^{n} \alpha_i \left( y_i (\mathbf{w} \cdot \mathbf{z}_i) + b \right) - 1 + \xi_i \right) - \sum_{i=1}^{n} \beta_i \xi_i$$

and finding the saddle point of $L(\mathbf{w}, b, \xi, \alpha, \beta)$. The parameters must satisfy:

$$\frac{\partial L(\mathbf{w}, b, \xi, \alpha, \beta)}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i \mathbf{z}_i = 0$$

$$\frac{\partial L(\mathbf{w}, b, \xi, \alpha, \beta)}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\frac{\partial L(\mathbf{w}, b, \xi, \alpha, \beta)}{\partial \xi_i} = s_i C - \alpha_i - \beta_i = 0$$

(14)

By applying the above conditions to Lagrangian (13) the dual form can be written as:

$$\min_{\alpha \in \mathbf{D}} \mathcal{J}(\alpha) = -\sum_{i=1}^{n} \alpha_i + \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\mathbf{D} = \left\{ \alpha | 0 \leq \alpha \leq C, \sum_{i=1}^{n} \alpha_i y_i = 0 \right\}$$

(15)

Fuzzy membership function

Choosing an appropriate fuzzy membership function can be challenging. From our experiments, we were able to make out that the outliers were the major cause of bad classification results when using the SVM method. Hence we used the class center method to reduce the effect of these outliers [19].

Denoting the mean of class +1 as $m_+$ and the mean of class -1 as $m_-$, the radii of the two classes are:

$$r_+ = \max_{(x_i, y_i=+1)} \| m_+ - x_i \|$$

$$r_- = \max_{(x_i, y_i=-1)} \| m_- - x_i \|$$

(16)

We derive the fuzzy membership function from the above collected values to be:

$$s_i = \begin{cases} 1 - \frac{\| m_- - x_i \|}{r_- + \delta} & \text{if } y_i = +1 \\ 1 - \frac{\| m_+ - x_i \|}{r_+ + \delta} & \text{if } y_i = -1 \end{cases}$$

(17)
V. Method and Performance Measures

We have used the SPX dataset for all our investigations. This was developed by Cheng et al. [4], [3] and contains non-homologous sequences containing intra-chain disulphide bonds from PDB with sequence similarity < 25%. The dataset contains 1018 chains totally. There are 5983 cystines out of which 5082 are bonded. This dataset is an extended version of SP39 and SP41 datasets and hence any result obtained on this can be extended to other sets. We used five parameters derived from physico-chemical properties and probability of occurrence of each cystine in different secondary structure states (Alpha Helix, Beta Strand, Coil). Chou-Fasman conformational parameters [21] (3 parameters), Kyte-Dolittle Hydrophobicity scale [13] and Grantham Polarity [22] (1 parameter each) were chosen as the features. The Chou-Fasman parameter for Helix(\(\alpha\)) is given by \(P_{\alpha}\) where \(f_{\alpha}\) = Number of Residues in Helix/Total Number of Residues and \(i\) is the set of amino acids residues. Similar conformational parameters for strand \(P_{\beta}\) and coil \(P_{\gamma}\) are calculated. Kyte-Dolittle hydrophobicity values and Grantham Polarity values were taken from the Protscal\(^1\) website. Although several scales of hydrophobicity and polarity are available, based on our earlier experiments [17] the above were chosen. A second set containing position specific scoring matrix (PSSMs) generated by PSI-BLAST [24] using non-redundant (NR) database was used. pfilt [6] was used to filter the low complexity regions, coiled-coil regions and transmembrane helices before subjecting to PSI-BLAST. We choose an E-value of 0.001 and 3 iterations for PSI-BLAST. We filter the low complexity regions, coiled-coil regions and transmembrane helices before subjecting to PSI-BLAST. We choose an E-value of 0.001 and 3 iterations for PSI-BLAST. We next consider the problem where the bonding status of the given Cystine is known. As the bonding status is known, the problem reduces to 2-class classification. We first apply Support Vector Machines with \(C = 10\) and a polynomial kernel with \(D = 2\). We then use Fuzzy Support Vector Machines with the same parameters. The SVM implementation we used was SVMHeavy \(^2\) [25], [26] based on Incremental training of support vector machines. The same software was extended to incorporate fuzziness in feature space.

\[
g(x) = \begin{cases} 
0.0 & x \leq -5 \\
0.5 + 0.1x & -5 < x < 5 \\
1.0 & x \geq 5
\end{cases}
\]  

where \(x\) is the value of the PSSM matrix. Instead of taking only 20 values per residue as a feature vector, we considered a window of length \(w\) and all the values within the window were considered [14]. The final feature length for each cystine pair is \(w*20+10\).

We first consider the problem where the bonding status of the given Cystine is known. As the bonding status is known, the problem reduces to 2-class classification. We first apply Support Vector Machines with \(C = 10\) and a polynomial kernel with \(D = 2\). We then use Fuzzy Support Vector Machines with the same parameters. The SVM implementation we used was SVMHeavy \(^2\) [25], [26] based on Incremental training of support vector machines. The same software was extended to incorporate fuzziness in feature space. The second problem involves prediction of bonding patterns, without the knowledge of bonding states. As it can be seen clearly, as the number of Cystines increase, the number of possible bonds increase exponentially. We follow the same procedure as in our earlier experiments with the same set of parameters. Five fold cross validation was performed for each experiment reported.

We measure the performance of the proposed technique with the earlier work with the help of specificity, sensitivity and accuracies \(Q_c\) and \(Q_p\). Specificity is the ability to reject false positive matches given by \(TN/(TP+TN)\); Sensitivity is the ability to detect true positive matches given by \(TP/(TP+FP)\) (TP = True Positive; FP = False Positive; TN = True Negative). \(Q_c\) is given by \((TP = TN)/(TP+TN+FP+FN)\) and \(Q_p\) is the protein level accuracy.

VI. Results and Discussion

As can be seen clearly from table I, the proposed method outperforms earlier methods. The normal SVM gives 4% higher sensitivity and 8% higher specificity than the earlier method of Cheng et al. [4]. Although overall our results were higher, during our experiments we found that there was a lot of variation in results for different folds (for a 5 fold experiment). By plotting a smaller set of features we found that this variation was due to the presence of outliers. Hence we extended our work to Fuzzy Support Vector Machines where we used the fuzzy membership function based on centroid of different classes for classification. As we expected, the results were higher for FSVM than for normal SVM. The overall sensitivity of the FSVM was 9% higher than the earlier methods, and the specificity 12% higher, clearly demonstrating the superiority of the proposed method.

The same was true for the prediction problem. We measured the performance using the overall accuracy at bridge level as well as at protein level. In bridge pattern prediction, the SVM results were not particularly impressive compared to earlier results (in fact, they were almost the same). However, FSVM outperformed the earlier methods by 7%. The results are summarized in table II. The higher accuracy was due to better prediction of the entire pattern in proteins with fewer cystines. There was no appreciable change in prediction for proteins with larger number of cystines.

To demonstrate the improvement in training in FSVM compared to normal SVM we compared the number of support vectors obtained by both the methods (a support vector is a training vector \(x_i\) for which \(\alpha_i \neq 0\)). This result is summarised in table III. As can be seen from the table, FSVM yields fewer support vectors than standard SVM, demonstrating the usefulness of fuzzification in feature space.

\[\text{TABLE III}
\]

<table>
<thead>
<tr>
<th></th>
<th>Fuzzy SVM</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>0.38</td>
<td>0.47</td>
</tr>
<tr>
<td>Prediction</td>
<td>0.43</td>
<td>0.55</td>
</tr>
</tbody>
</table>

\(^1\)http://au.expasy.org/tools/protscale.html

\(^2\)Can be downloaded from http://www2.ee.mu.oz.au/pgrad/apsh/svm/
TABLE I
RESULTS FOR DISULPHIDE BRIDGE CLASSIFICATION AND COMPARISON WITH RECENT LITERATURE ON SPX DATASET

<table>
<thead>
<tr>
<th>Number of Bridges</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.48</td>
<td>0.71</td>
<td>0.61</td>
<td>0.65</td>
<td>0.65</td>
<td>0.70</td>
</tr>
<tr>
<td>2</td>
<td>0.63</td>
<td>0.63</td>
<td>0.63</td>
<td>0.61</td>
<td>0.67</td>
<td>0.64</td>
</tr>
<tr>
<td>3</td>
<td>0.67</td>
<td>0.62</td>
<td>0.66</td>
<td>0.60</td>
<td>0.66</td>
<td>0.70</td>
</tr>
<tr>
<td>4</td>
<td>0.55</td>
<td>0.50</td>
<td>0.61</td>
<td>0.51</td>
<td>0.66</td>
<td>0.56</td>
</tr>
<tr>
<td>5</td>
<td>0.41</td>
<td>0.37</td>
<td>0.56</td>
<td>0.38</td>
<td>0.63</td>
<td>0.42</td>
</tr>
<tr>
<td>6</td>
<td>0.33</td>
<td>0.29</td>
<td>0.59</td>
<td>0.37</td>
<td>0.66</td>
<td>0.42</td>
</tr>
<tr>
<td>7</td>
<td>0.36</td>
<td>0.31</td>
<td>0.47</td>
<td>0.36</td>
<td>0.60</td>
<td>0.35</td>
</tr>
<tr>
<td>8</td>
<td>0.32</td>
<td>0.30</td>
<td>0.44</td>
<td>0.32</td>
<td>0.54</td>
<td>0.38</td>
</tr>
<tr>
<td>9</td>
<td>0.71</td>
<td>0.61</td>
<td>0.55</td>
<td>0.35</td>
<td>0.67</td>
<td>0.39</td>
</tr>
<tr>
<td>10</td>
<td>0.40</td>
<td>0.37</td>
<td>0.59</td>
<td>0.45</td>
<td>0.72</td>
<td>0.62</td>
</tr>
<tr>
<td>12</td>
<td>0.55</td>
<td>0.50</td>
<td>0.60</td>
<td>0.50</td>
<td>0.75</td>
<td>0.55</td>
</tr>
<tr>
<td>14</td>
<td>0.62</td>
<td>0.57</td>
<td>0.65</td>
<td>0.58</td>
<td>0.70</td>
<td>0.60</td>
</tr>
<tr>
<td>16</td>
<td>0.23</td>
<td>0.22</td>
<td>0.43</td>
<td>0.25</td>
<td>0.55</td>
<td>0.43</td>
</tr>
<tr>
<td>17</td>
<td>0.40</td>
<td>0.35</td>
<td>0.51</td>
<td>0.31</td>
<td>0.67</td>
<td>0.40</td>
</tr>
<tr>
<td>25</td>
<td>0.40</td>
<td>0.24</td>
<td>0.63</td>
<td>0.30</td>
<td>0.66</td>
<td>0.28</td>
</tr>
<tr>
<td>26</td>
<td>0.73</td>
<td>0.42</td>
<td>0.69</td>
<td>0.30</td>
<td>0.71</td>
<td>0.27</td>
</tr>
<tr>
<td>Overall</td>
<td>0.54</td>
<td>0.55</td>
<td>0.62</td>
<td>0.59</td>
<td>0.66</td>
<td>0.64</td>
</tr>
</tbody>
</table>

The first set of sensitivity and specificity are the results from Chang et al. [4].
The second set is the result for Normal SVM
The third set is the result of FSVM

TABLE II
RESULTS FOR DISULPHIDE BRIDGE PREDICTION AND COMPARISON WITH RECENT LITERATURE ON SPX DATASET

<table>
<thead>
<tr>
<th>Number of Bridges</th>
<th>Accuracy at Bridge Level</th>
<th>Accuracy at Protein Level</th>
<th>Accuracy at Bridge Level</th>
<th>Accuracy at Protein Level</th>
<th>Accuracy at Bridge Level</th>
<th>Accuracy at Protein Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>-</td>
<td>0.59</td>
<td>0.65</td>
<td>0.53</td>
<td>0.73</td>
<td>0.62</td>
</tr>
<tr>
<td>2.00</td>
<td>-</td>
<td>0.59</td>
<td>0.59</td>
<td>0.50</td>
<td>0.67</td>
<td>0.58</td>
</tr>
<tr>
<td>3.00</td>
<td>-</td>
<td>0.54</td>
<td>0.61</td>
<td>0.56</td>
<td>0.63</td>
<td>0.57</td>
</tr>
<tr>
<td>4.00</td>
<td>-</td>
<td>0.34</td>
<td>0.63</td>
<td>0.46</td>
<td>0.64</td>
<td>0.48</td>
</tr>
<tr>
<td>Overall</td>
<td>-</td>
<td>0.51</td>
<td>0.63</td>
<td>0.52</td>
<td>0.68</td>
<td>0.58</td>
</tr>
</tbody>
</table>

The first set are the results from Chang et al. [4].
The second set is the result for Normal SVM
The third set is the result of FSVM

VII. CONCLUSION AND FUTURE WORK
Disulphide bridge pattern prediction is a hard problem on which relatively little work has been undertaken. In this paper, we have proposed a novel method to predict disulfide bridges based on Fuzzy Support Vector Machines. The two important problems of bridge classification and bridge pattern prediction have been discussed. The use of profiles along with statistical and physico-chemical parameters as features has been investigated. The effect of outliers on normal support vector machines has demonstrated and a fuzzy membership function incorporated into the feature space to reduce this effect. The method results in appreciable improvement in sensitivity, specificity and recognition accuracy. In future, we plan to further extend our approach to classify sequences containing disulfide bonds and also those not containing them. We also plan to consider the problem of predicting the number of bridges present in a given sequence.

ACKNOWLEDGMENT
The authors would like to thank Prof. David Jones for kindly providing the pffilt program. They are also grateful to NCBI for the PSI-BLAST program and Chang et. al. for making disulphide bond data SPX available on the web.

REFERENCES


Author/s:
Jayavardhana, Rama G. L.; SHILTON, ALISTAIR; PARKER, MICHAEL; PALANISWAMI, MARIMUTHU

Title:
Disulphide Bridge Prediction using Fuzzy Support Vector Machines

Date:
2005

Citation:

Publication Status:
Published

Persistent Link:
http://hdl.handle.net/11343/34074