BRNN-SVM: Increasing the Strength of Domain Signal to Improve Protein Domain Prediction Accuracy

Kalsum U. Hassan¹, Razib M. Othman², Rohayanti Hassan², Hishammuddin Asmuni³, Jumail Taliba² and Shahreen Kasim⁴

¹Department of Information Technology, Kolej Poly-Tech MARA Batu Pahat, Tingkat 3, Bangunan Tabung Haji, Batu Pahat,
²Laboratory of Computational Intelligence and Biotechnology, Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia, UTM Skudai,
³Department of Software Engineering
Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia, UTM Skudai,
⁴Department of Web Technology,
Faculty of Computer Science and Information Technology
Universiti Tun Hussien Onn, UTHM Parit Raja,
Malaysia

1. Introduction

A protein domain is the basic unit of protein structure that can develop itself by using its own shapes and functions, and exists independently from the rest of the protein sequence. Protein domains can be seen as distinct functional or structural units of a protein. Protein domains provide one of the most valuable information for the prediction of protein structure, function, evolution, and design. Protein domain is detected from protein structure that is predicted from protein sequence of amino acid. The protein sequence may be contained of single-domain, two-domain, or multiple-domain with different or matching copies of protein domain. A protein domain comprises of protein domain boundary that relates to a part in amino acid residue where each residue in the protein chain is defined as domain position. Each shape of protein domain is a compacted and folded structure that is independently stable. It exists independently since the protein domain is a part of the protein sequence. The independent modular nature of protein domain means that it can often be found in proteins with the same domain content, but in different orders or in different proteins. The knowledge of protein domain boundaries is useful in analysing the different functions of protein sequences.

Several methods have been developed to detect the protein domain, which can be categorized as follows: (1) Methods based on similarity and used multiple sequence alignments to represent domain boundaries, e.g. KemaDom (Lusheng et al., 2006) and Biozon (Nagaranjan
and Yona, 2004); (2) Methods that depend on known protein structure to identify the protein domain, e.g. AutoSCOP (Gewehr et al., 2007) and DOMpro (Cheng et al., 2006); (3) Methods that used dimensional structure to assume protein domain boundaries, e.g. GlobPlot (Linding et al., 2003), Mateo (Lexa and Valle, 2003), and Dompred-DPS (Marsden et al., 2002); (4) Methods that used comparative model such as Hidden Markov Models (HMM) to identify other member of protein domain family, e.g. HMMPfam (Bateman et al., 2004) and HMMSMART (Ponting et al., 1999); and (5) Methods that are solely based on protein sequence information, e.g. Armadillo (Dumontier et al., 2005) and SBASE (Kristian et al., 2005). However, these methods only produce good results in the case of single-domain proteins.

There is no sign to indicate when a protein domain starts and ends. Protein sequence with closely related homologues can reveal conserved regions which are functionally important (Elhefnawi et al., 2010). Nowadays, it is not only important to detect a protein domain accurately from large numbers of protein sequences with unknown structure, but it is also essential to detect protein domain boundaries of the protein sequence (Chen et al., 2010). Protein domain boundaries are important to understand and analyse the different functions of protein (Paul et al., 2008) as shown in Fig. 1. The difficulty in protein domain prediction lies in the detection of the protein domain boundaries in the protein sequences, since the protein sequences alone contain the structural information but it is only available in small portion along the protein space. The secondary structure provides the sequence information used in protein domain prediction such as the similarity of protein chain, the potential of protein domain region and boundaries. Methods that used secondary structure information in protein domain prediction, such as DOMpro and KemaDom has shown improvement in predicting the protein domain compared to other protein domain predictors.

![Fig. 1. An example of constructing a new protein from different protein domain boundaries.](image)

Previously, Neural Network (NN) is used as a classifier to detect protein domain such as in the work of Armadillo, Biozon, Dompred-DPS, and DOMpro. Of late, Support Vector Machines (SVM) is perceived as a strong contender to NN in protein domain classification. Unlike NN, SVM is much less affected by the dimension of the input space and employs structural risk minimization rather than empirical risk minimization. SBASE (Kristian et al., 2005) and KemaDom are examples that apply SVM in protein domain prediction. The results from these methods are more accurate compared to NN.
2. BRNN-SVM algorithm

The BRNN-SVM begins with seeking the seed protein sequences using BLAST (Altschul et al., 1997) in order to generate a dataset. The dataset is split into training and testing sets. Multiple-alignment is performed using ClustalW (Larkin et al., 2007), where the alignments are represented as a protein sequence of alignment column that is associated to one position in the seed protein sequence. Bidirectional Recurrent Neural Network (BRNN) is used to generate secondary structure from alignment of protein sequence in order to highlight the signal of protein domain boundaries. The protein secondary structure is predicted into three types: alpha-helices, beta-sheet, and coil. The information of secondary structure are extracted using six measures (which are entropy, protein sequence termination, correlation, contact profile, physio-chemical properties, intron-exon information, and score of secondary structure) to increase the domain signal. This extracted information will be used for SVM input for the protein domain prediction. SVM processes the information and classify the protein domain into single-domain, two-domain, and multiple-domain. The BRNN-SVM is evaluated by comparing it with other existing methods either based on similarity and multiple sequence alignment (Biozon and KemaDOM), known protein structure (AutoSCOP and DOMpro), dimensional structure (GlobPlot, Mateo, and Dompred-DPS), comparative model (HMMFam and HMMSMART), and sequence alone (Armadillo and SBASE). An analysis of the results has demonstrated that the BRNN-SVM shows outstanding performance on single-domain, two-domain, and multiple-domain. The steps involved in BRNN-SVM can be simplified as follows: (1) Generate training and testing sets using BLAST; (2) Perform multiple sequence alignment using ClustalW; (3) Predict secondary structure by BRNN; (4) Extract information from protein secondary structure; (5) Classify the protein domain by SVM; and (6) Evaluate the performance using sensitivity and specificity, and accuracy.

3. Secondary structure prediction by BRNN

For each protein sequence, the secondary structure information is predicted based on an ensemble of BRNNs. The input for predicting secondary structure is a single protein sequence from a multiple sequence alignment. Then, BRNN derives protein sequence information from PSI-BLAST (Altschul et al., 1997) to include homology structure that is used in the protein secondary structure information prediction. Subsequently, the protein secondary structure information is divided into three classes: alpha-helices, beta-sheets, and coils.

The BRNN is described in Fig. 2–3. This BRNN involves a set of $i$ protein sequences as input $X_i$ variable, a forward $F_i$, and backward $B_i$, a chain of hidden variables, and a set of $O_i$ as an output variable. The relationship between these variables is implemented using feed-forward NN. Three NNs $N_o$, $N_f$, and $N_b$ are used to implement BRNN. The output $O_i$ (Chen and Chaudhari, 2007) is as follows:

$$O_i = N_o(X_i, F_i, B_i).$$

The output $O_i$ depends on input $X_i$ at the position $i$, the forward $F_i$ (Chen and Chaudhari, 2007) is the hidden context in the vector $F_i \in \mathbb{Z}^n$ and the backward $B_i$ (Chen and Chaudhari, 2007) is the hidden context in the vector $B_i \in \mathbb{Z}^n$ where $m = n$. To obtain the composite the $F_i$ and $B_i$, the BRNN equation is applied as follows:
Recurrent Neural Networks and Soft Computing

\[ F_i = N_f(X_i, F_{i-1}), \]  
\[ B_i = N_b(X_i, B_{i+1}), \]

where \( N_f(X_i, F_{i-1}) \) and \( N_b(X_i, B_{i+1}) \) are learnable non-linear state transition functions. The boundary condition for \( F_i \) and \( B_i \) can be set to 0, for example \( F_i = F_{i+1} = 0 \) where \( n \) is the length of the protein sequence being processed.

The \( N_f \) and \( N_b \) are assigned to be a “tool” that can be shifted along the protein sequence. For the prediction class at the position \( i \), the “tool” is shifted in the opposite direction starting from the \( N \) and \( C \) terminus, up to position \( i \). Then, the “tool” output at position \( i \) is combined with the input \( X_i \) to compute the output \( O_i \) using \( N_o \). From the output \( O_i \), the membership probability of the residue at the position \( i \) is computed to predict the domain boundary.

BRNN is used to predict protein secondary structure into alpha-helices, beta-sheets, or coils. The BRNN consists of an input layer, hidden layer, and output layer. The protein sequences are fed into the input layer. The protein secondary structure is encoded into the output layer as follows:

- (1, 0) = Alpha-Helices
- (0, 1) = Beta-Sheets
- (0, 0) = Coil

The input layer (John et al., 2006) is defined as follows:

\[ I_k = \sum_i W_{ik} Y_i + b_k, \]

where \( W_{ik} \) is the sum of all the input to the unit, \( Y_i \) is the connection strength, \( b_k \) is the bias from the protein sequence, \( i \) is the number of protein sequence, and \( k \) is the number of output from the protein sequence. The output layer (John et al., 2006) is defined as follows:

\[ O_k = \frac{1}{1 + e^{-X_k}}, \]

where \( X \) is a real number between -8 and 8. This has been experimentally determined as the best range. \( k \) represents the number of outputs from the protein sequence.

The alpha-helices measure is divided into two types: amphipathic helices and hydrophobic helices. To predict an amphipathic helices region, Helical Wheel Representation (HWR: Renaud and McConkey, 2005) is applied. The HWR predicts the residues from the solvent and side chains interaction of protein sequence with amphipathic helices. Then, the score of amphipathic helices and hydrophobic helices are merged to predict the alpha-helices region for the protein sequence. The beta-sheets are assigned using Kabsch and Sander’s program (Kabsch and Sander, 1983). The extension of beta-sheets is situated and connected to form theatre-backbone H-bonds according to the Pauling pairing rules (Pauling and Corey, 1951). When two H-bond is formed or surrounded by two H-bond in the sheet, this formation is defined as beta-sheet (E). If only one amino acid fulfils the criteria, the sheet will be called beta-bridge (B). The residues that are neither alpha-helices nor beta-sheets are classified as coils.
BRNN-SVM: Increasing the Strength of Domain Signal to Improve Protein Domain Prediction Accuracy

Fig. 2. BRNN architecture with left (forward) and right (backward) context associated with two recurrent networks (“tool”). The left and right contexts are produced by two similar recurrent networks which intuitively can be thought in term of two “tools” that are shifted along the protein chain.

Fig. 3. An example of secondary structure prediction using BRNN.

4. Features extraction

Features extraction in BRNN-SVM is important to obtain the protein domain information from the predicted secondary structure. The secondary structure information is used to compute the change of the protein sequence position that constitutes a part of the protein domain boundary. This information is believed to reflect the protein structural properties that have informative protein domain structure and is used to detect the protein domain boundaries. The information as shown in Fig. 4–9 is entropy, protein sequence termination, correlation, contact profile, physio-chemical properties and intron-exon information.
Fig. 4. An example of entropy calculation.

The effective entropy measure takes into account the similarity of amino acids. An evolutionary pressure is used to calculate the evolutionary span (Nagarajan and Yona, 2004) defined as:

$$\text{Span}(x) = \frac{2}{t(t-1)} \sum_{p=1}^{t} \sum_{q\neq p} s(j,k),$$

(6)

where $s(j,k)$ is $s(\alpha_{px}, \alpha_{qk})$. Span() is used to compare the sum of pairwise similarity of amino acids. The $x$ is an alignment from the multiple sequence alignment and $t$ is the number of protein sequences that has participated in $x$. $\alpha_{px}$ and $\alpha_{qk}$ represent the amino acids in position $x$. $s(j,k)$ is the similarity score of amino acids where $j$ and $k$ refer to the scoring matrix BLOSUM50 (Henikoff and Henikoff, 1992).

Fig. 5. An example of protein sequence termination calculation.

In a multiple sequence alignment, the protein sequence termination is not necessarily displayed. The left and right protein sequence termination score is calculated for each protein sequence with an e-value that is larger than 0. The scores of protein sequence termination are then used to identify the strong signal of the protein domain boundary. Left and right protein sequence terminations score (Menachem and Chen, 2008) are defined as:

$$T_{\text{seq.termination}} = \log(\tau_1, \tau_2, \ldots, \tau_n),$$

(7)

where $\tau_n$ is the e-value of the $n$ protein sequence.
Correlation is two random protein sequences that are positively correlated if high values of one are likely to be associated with high values of the other. Possible correlations range is 1 or 0. A zero correlation indicates that there is no relationship between the sequences. A correlation of 1 indicates a perfect positive correlation, meaning that both sequence move in the same direction together. The correlation of amino acids with protein secondary structure information is used to predict the protein structure. It is also important to understand the force that causes the flexibility of a protein structure. Every protein sequence in a multiple sequence alignment contains information of structural flexibility. To find a position that is more flexible in a protein sequence, indel entropy (Zou et al., 2008) based on the distribution of protein sequence lengths is used:

$$E_{\beta}(B) = -\sum_{p} \beta_{p} \log \beta_{p} \ ,$$  \hspace{1cm} (8)

where $\beta_{p}$ is the various indel lengths seen at a position.

The predicted contact profile of a protein sequence is obtained by getting the structural flexibility information. Then, the number of pairwise contact profile is counted for each protein sequence. The contact profile between residues in a protein sequence is predicted based on correlated mutations. Correlated mutations (Pazos et al., 1997) between two columns $x$ and $y$ are defined as:

$$Corr_{x,y}(x, y) = \frac{1}{N} \sum_{i=1}^{N} \frac{(s(\alpha_{x,i}, \alpha_{y,i}) - <s_{x}>)(s(\alpha_{x,i}, \alpha_{y,i}) - <s_{y}>)}{\sigma_{x}\sigma_{y}}$$

Score of Contact Profile : 0.9872
Recurrent Neural Networks and Soft Computing

\[ \text{Corr}_{m}(x, y) = \frac{1}{t^2} \sum_{p=1}^{t} \sum_{q=1}^{t} \frac{(s(\alpha_{px}, \alpha_{qx}) - \langle s_x \rangle)(s(\alpha_{py}, \alpha_{qy}) - \langle s_y \rangle)}{\sigma_x \sigma_y}, \]  

(9)

where \( \alpha_{px} \) and \( \alpha_{qx} \) represent the amino acids in position \( x \) and the \( \alpha_{py} \) and \( \alpha_{qy} \) represent the amino acids in position \( y \). The \( s(\alpha_{px}, \alpha_{qx}) \) and \( s(\alpha_{py}, \alpha_{qy}) \) are the similarity score of amino acids and \( \alpha_{px}, \alpha_{qx}, \alpha_{py}, \) and \( \alpha_{qy} \) refer to the scoring matrix BLOSUM50. The \( \langle s_x \rangle \) and \( \langle s_y \rangle \) are the average similarity of amino acids in position \( x \) and \( y \). The \( \sigma_x \) and \( \sigma_y \) are standard deviations and \( t \) is the number of protein sequences that are indicated in the columns.

Fig. 8. An example of physio-chemical properties calculation.

Physio-chemical properties are information that is used to predict protein domain boundaries. Hydrophobicity is used to display the distribution of protein sequence residue that in turn, used for the detection of physio-chemical properties. In BRNN-SVM, the score of hydrophobicity and molecular weight (Black and Mould, 1991) is used to predict physio-chemical properties in protein sequence. The average hydrophobicity and molecular weight for each measure of protein sequence is calculated to determine the physio-chemical properties that are affecting the protein domain boundary detection.

Fig. 9. An example of intron-exon calculation.
The intron-exon data contains intron-exon structure at Deoxyribonucleic Acid (DNA) level that is related to protein domain boundaries in which folded protein domain boundaries exist independently. Each protein domain defines the intron-exon position. The intron-exon data is taken from the EID database (Saxonov et al., 2000). Then, each protein sequence is compared with the database and the gapless matching protein sequence is kept. The similarity of the protein sequence is calculated in order to define the exon boundary using an equation defined as the sequence termination. Finally, the exon termination score (Saxonov et al., 2000) is calculated as follows:

$$E_{\text{exon\_termination}} = \log(E_1 \cdot E_2 \cdot \ldots \cdot E_n),$$

(10)

where $E_n$ is the e-value of the $n$ protein sequence. After that, the average of measures score from features extraction's phase is calculated in order to generate the features vector and used as input to SVM as follows:

$$\frac{\sum (\text{Score\_of\_measures})}{n},$$

(11)

where $\text{Score\_of\_measures}$ is obtained from features extraction (entropy, protein sequence termination, correlation, contact profile, physio-chemical properties, and intron-exon information) score and $n$ refer to quantity of features extraction measurements where it could be seven. Fig. 10 has shown the example of features vector calculation.

5. Domain detection by SVM

SVM is a machine learning technique based on statistical learning theory that trains multiple functions such as polynomial functions, radial basis functions, and spines to form a single classifier. The SVM is applied to identify the protein domain boundaries position. The SVM works by: (1) Mapping the input vector into a feature space which is relevant to the kernel function; and (2) Seeking an optimized linear division from multiple $n$-separated hyperplane, where $n$ is classes of protein sequence in the dataset. The input (Dong et al., 2003) vector is defined as follows:
where \( I_s \) is the input space with corresponding predefined labels (Dong et al., 2003):

\[
y_i \in I_s(i = 1, \ldots, n),
\]

where \(+1\) and \(-1\) are used to stand, respectively, for the two classes. The SVM is trained with Radial Basic Function (RBF) kernel, a function that is often used in pattern recognition. The parameters of SVM training are \( \sigma^2 \), the RBF kernel smoothing parameter and \( C \), the learning variable to trade-off between under- and over-generalization. The RBF (Zou et al., 2008) is defined as follows:

\[
K(\tilde{y}_i, \tilde{y}_j) = \exp\left( \frac{-r \mid \mid \tilde{y}_i - \tilde{y}_j \mid \mid^2}{2\sigma^2} \right),
\]

where \( \tilde{y}_i \) is labels and \( \tilde{y}_j \) is input vector. The input vector will be the centre of the RBF and \( \sigma \) will determine the area of influence this input vector has over the feature space. A larger value of \( \sigma \) will give a smoother decision surface and a more regular decision boundary since the RBF with large \( \sigma \) will allow an input vector to have a strong influence over a larger area.

The best pair of parameter of \( C \) and \( \sigma \) is search via k-fold cross-validation scheme to safeguard unbiased tweaking. In this study, \( k = 10 \) is applied where the protein sequence is split into \( k \) subsets of approximately equal size portions. The best combinations of \( C \) and \( \sigma \) obtained from the optimization process were used for training the final SVM classifier using the entire training set. The SVM classifier is subsequently used to predict the testing datasets. The SVM training detects the protein domain boundaries based on scores that corresponds to the domain information or different domain information. The SVM classified the protein domain into single-domain, two-domain, and multiple-domain. Various quantitative metrics were obtained to measure the effectiveness of the BRNN-SVM: true positives (TP) for the number of correctly classified protein domain; false positives (FP) for the number of incorrectly classified protein domain; true negatives (TN) for the number of correctly classified non protein domain; and false negatives (FN) for the number of incorrectly classified non protein domain.

6. Dataset and evaluation measure

To test the BRNN-SVM, seed protein sequences obtained from the PDB database (Berman et al., 2000) are selected with their corresponding domain structure that exists in SCOP database (Andreeva et al., 2008) version 1.73. The SCOP 1.73 with 40% less identity in PDB contains 9,536 protein sequences. The protein sequences are reconstructed from which short protein sequences that are less than 40 amino acids are removed. Then, the protein sequences are searched from the NR database (Henikoff et al., 1999) using BLAST and protein sequences that have more than 20 hits are kept. Hence, the number of protein data retained is 6,242. The dataset are randomly split into training and testing sets. Training set is used for optimizing the SVM parameters and for training the SVM classifier to predict unseen protein domain boundaries. Testing set is used for evaluating the performance of the SVM. The dataset are randomly split into training and testing sets in the same ratio which is 3,121 protein sequences respectively. The process of generating the dataset is shown in Fig. 10.
Based on the classification output of SVM, a series of statistical metrics were computed to measure the effectiveness of the BRNN-SVM. Sensitivity (SN: Zaki et al., 2006) and specificity (SP: Zaki et al., 2006), which indicates the ability of the prediction system to correctly classify the protein domain and not protein domain respectively; the SN and SP are defined as follows:

\[
SN = \frac{TP}{TP + FN} \times 100 ,
\]

\[
SP = \frac{TP}{TP + FP} \times 100 .
\]

To provide an indication of the overall performance of the system, we computed accuracy (AC: Zaki et al., 2006), for the percentage of the correctly predicted protein domain. The AC is defined as follows:

\[
AC = \frac{TP - TN}{TP + FN + TN + FP} \times 100 .
\]

7. Computational results

The BRNN-SVM is tested and compared its performance with other methods such as based on similarity and multiple sequence alignment (Biozon and KemaDOM), known protein structure (AutoSCOP and DOMpro), dimensional structure (GlobPlot, Mateo, and
Dompred-DPS), comparative model (HMMPfam and HMMSMART), and sequence alone (Armadillo and SBASE). The properties of protein sequence are derived from a protein secondary structure using several measures such as entropy, correlation, protein sequence termination, contact profile, physico-chemical properties and intron-exon boundaries measures. The protein secondary structure generates a strong signal of protein domain boundaries and is used to locate the protein domain regions using the following procedures. Firstly, the BRNN-SVM starts by searching large protein sequences and comparing them with the NR database to generate multiple sequence alignments. Secondly, the secondary structure is predicted for each protein sequence using BRNN. Thirdly, some of the scores from several measures are calculated as input in the SVM training. Finally, the results generated by SVM are evaluated. This evaluation provides a clear understanding of strengths and weaknesses of an algorithm that has been designed.

The datasets obtained from SCOP 1.73 that have been defined in the previous section are used to test and evaluate the BRNN-SVM and other protein domain prediction methods. The results of the prediction accuracy compared with other protein domain prediction methods including sensitivity and specificity for single-domain, two-domain and multiple-domain are presented in Table 1 and Fig. 11-14. It is easy to see that predicting two-domain or multiple-domain is more difficult than predicting single-domain. The results depict the higher sensitivity and specificity represent better achievement and the priority is given to sensitivity in order to determine the achievement of protein domain prediction since sensitivity measures the proportion of actual positives which are correctly identified for protein domain prediction. The BRNN-SVM achieved a higher sensitivity of 87% for single-domain, 73% for the two-domain and 81% for the multiple-domain compared to other methods. The BRNN-SVM achieved a higher specificity of 76% for the two-domain and 79% for the multiple-domain compared to other methods. The BRNN-SVM increases of 83% for accuracy as compared to KemaDom method with 79% and SBase method with 80%.

The properties of protein sequence have given a strong signal to assign protein boundaries because the protein secondary structure predicted is based on interaction between long-range interactions of the amino acid. The use of protein secondary structure prediction based on BRNN involves informative communion between an input and an output sequence of variable length. The BRNN is based on the forward, backward and hidden Markov chains that transmit information in both directions along the sequence between the input and output. This shows that interaction exists in protein folding and plays an important role in the formation of protein secondary structure. The information does have an effect on the protein domain boundaries prediction. The BRNN-SVM relies on scores of measures to detect the protein domain region in order to classify a domain for the protein sequence.

However, the prediction of specificity for a single-domain prediction is 79% which is 14% lower compared to the Biozon and 10% lower compared to Armadillo. The reason is that the BRNN-SVM classifies the protein sequence with no predicted protein domain boundaries as a single-domain. Therefore, the number of protein domain for the protein sequence is from the start until the end. The situation is aggravated when the protein sequence is too long. To solve this problem, the protein sequence can be split into protein sub-sequences before predicting the protein domain (Kalsum et al., 2009).
Table 1. Performance comparison between BRNN-SVM and other protein domain prediction methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Single-Domain</th>
<th>Two-Domain</th>
<th>Multiple-Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SN</td>
<td>SP</td>
<td>SN</td>
</tr>
<tr>
<td>BRNN-SVM</td>
<td>0.87</td>
<td>0.79</td>
<td>0.73</td>
</tr>
<tr>
<td><strong>Similarity and multiple sequence alignment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Biozon</td>
<td>0.27</td>
<td>0.93</td>
<td>0.33</td>
</tr>
<tr>
<td>KemaDom</td>
<td>0.82</td>
<td>0.76</td>
<td>0.70</td>
</tr>
<tr>
<td><strong>Known protein structure</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AutoSCOP</td>
<td>0.80</td>
<td>0.65</td>
<td>0.62</td>
</tr>
<tr>
<td>DOMpro</td>
<td>0.85</td>
<td>0.80</td>
<td>0.43</td>
</tr>
<tr>
<td><strong>Dimensional structure</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GlobPlot</td>
<td>0.78</td>
<td>0.74</td>
<td>0.32</td>
</tr>
<tr>
<td>Mateo</td>
<td>0.57</td>
<td>0.74</td>
<td>0.21</td>
</tr>
<tr>
<td>Dompred-DPS</td>
<td>0.55</td>
<td>0.73</td>
<td>0.52</td>
</tr>
<tr>
<td><strong>Comparative model</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMMPFam</td>
<td>0.65</td>
<td>0.60</td>
<td>0.53</td>
</tr>
<tr>
<td>HMMMSmart</td>
<td>0.77</td>
<td>0.69</td>
<td>0.66</td>
</tr>
<tr>
<td><strong>Sequence alone</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBASE</td>
<td>0.86</td>
<td>0.77</td>
<td>0.69</td>
</tr>
<tr>
<td>Armadillo</td>
<td>0.31</td>
<td>0.89</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Fig. 12. Performance comparison between BRNN-SVM and other protein domain prediction methods. The best sensitivity is BRNN-SVM with 87% and the best specificity is Armadillo with 89% since the BRNN-SVM classifies the protein sequence with no predicted protein domain boundaries as a single-domain.
Fig. 13. Performance comparison between BRNN-SVM and other protein domain prediction methods on two-domain. The best performance for two-domain prediction is BRNN-SVM with 73% for sensitivity and 76% for specificity since the secondary structure information has given a strong signal to assign protein boundaries because the protein secondary structure predicted is based on interaction between long-range interactions of the amino acid.

Fig. 14. Performance comparison between BRNN-SVM and other protein domain prediction methods on multiple-domain. The best performance of multiple-domain prediction is BRNN-SVM with 81% sensitivity and 79% specificity since the BRNN is a transaction between an input and an output sequence of variable length. This shows that interaction exists in protein folding and plays an important role in the formation of protein secondary structure.
8. Conclusion

An algorithm named BRNN-SVM has been developed in order to solve the problem of weak domain signal. The algorithm begins with searching the seed protein sequences as dataset from SCOP 1.73. The dataset is split into training and testing sets. Then, multiple sequence alignment is performed prior to the prediction of protein secondary structure using BRNN. Several measures such as entropy, protein sequence termination, correlation, contact profile, physio-chemical properties and intron-exon data are used to increase the strength of domain signal from protein secondary structure. SVM classified the prediction into single-domain, two-domain and multiple-domain. Lastly, the results from SVM are evaluated in term of sensitivity and specificity. BRNN is based on forward, backward and hidden Markov chains that transmit information in both directions along the sequence between the input and output. Therefore, it increases accuracy of protein secondary prediction and well as providing strong domain signal from this protein secondary structure based on the generated measures. This is believed to be the reason why BRNN-SVM is a good method for protein domain predictors especially in two-domain and multiple-domain.

9. Acknowledgements

We would like to express our appreciation to the reviewers of this paper for their valuable suggestions. This work is supported by the Malaysian Ministry of Science, Technology and Innovation (MOSTI) under Grant No. 02-01-06-SF0230.
10. References


www.intechopen.com


