We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

7

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

Kalsum U. Hassan[1], Razib M. Othman[2], Rohayanti Hassan[2], Hishammuddin Asmuni[3], Jumail Taliba[2] and Shahreen Kasim[4]
[1]*Department of Information Technology, Kolej Poly-Tech MARA Batu Pahat, Tingkat 3, Bangunan Tabung Haji, Batu Pahat,*
[2]*Laboratory of Computational Intelligence and Biotechnology, Faculty of Computer Science and Information Systems Universiti Teknologi Malaysia, UTM Skudai,*
[3]*Department of Software Engineering Faculty of Computer Science and Information Systems Universiti Teknologi Malaysia, UTM Skudai,*
[4]*Department 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.

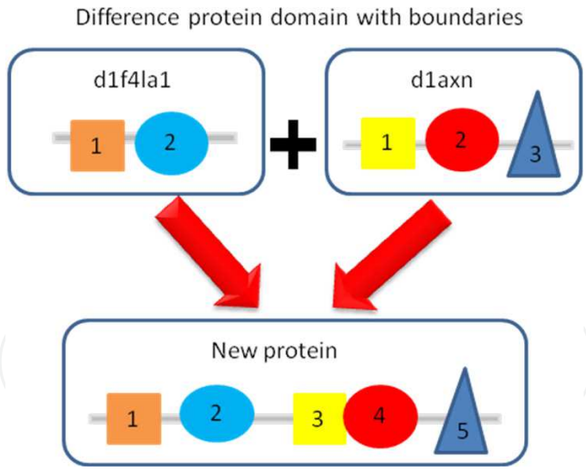


Fig. 1. An example of constructing a new protein from different protein domain boundaries.

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 (HMMPfam 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:

$$O_i = N_o(X_i, F_i, B_i) . \tag{1}$$

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 $B_i \in \mathbb{Z}^m$ where $m = n$. To obtain the composite the F_i and B_i, the BRNN equation is applied as follows:

$$F_i = N_f(X_i, F_{i-1}), \tag{2}$$

$$B_i = N_b(X_i, B_{i+1}), \tag{3}$$

where $N_f(X_i, F_{i-1})$ and $N_b(X_i, B_{i+1})$ are learnable non-linear state transition function. The boundary condition for F_i and B_i can be set to 0, for example $F_i = F_{n+1} = 0$ where n is 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-sheet, 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:

$$I_k = \sum_i W_{ik} Y_i + b_k, \tag{4}$$

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:

$$O_k = \frac{1}{1 + e^{-X_k}}, \tag{5}$$

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: Renaund 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.

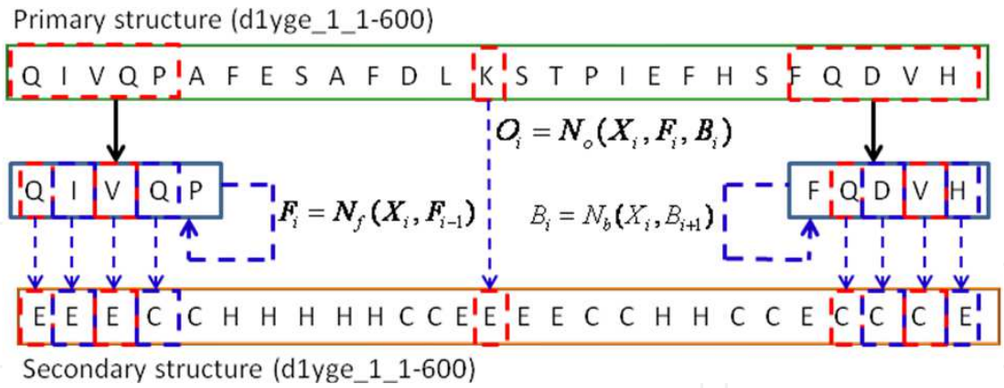


Fig. 2. BRNN architecture with left (forward) and right (backward) context associated with two recurrent networks ("tool"). The left and right contexts are produced by two similar recurrent networks which intuitively can be thought in term of two "tools" that are shifted along the protein chain.

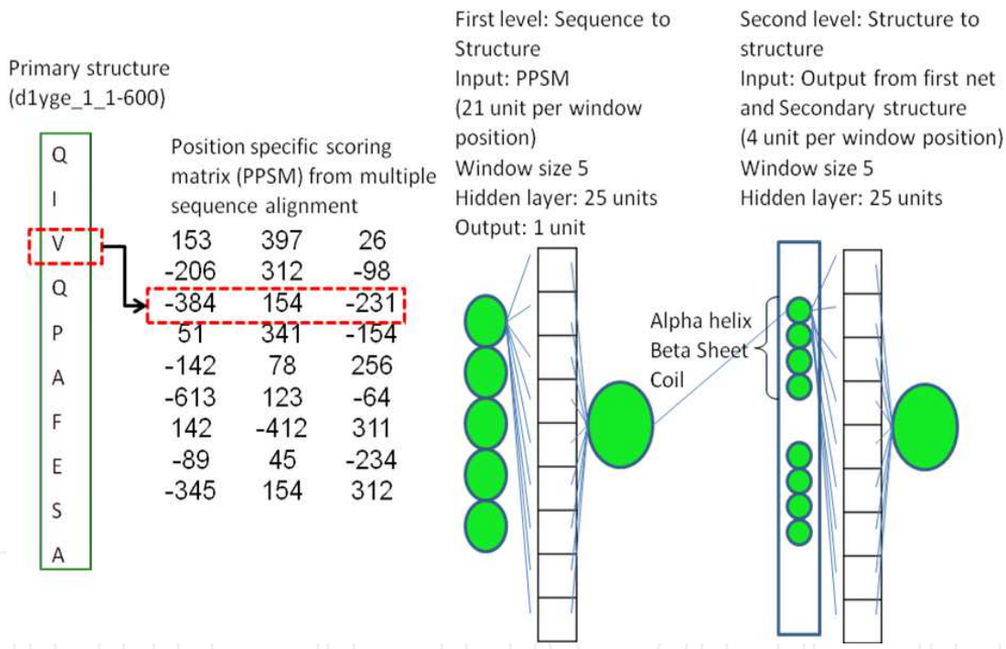


Fig. 3. An example of secondary structure prediction using BRNN.

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.

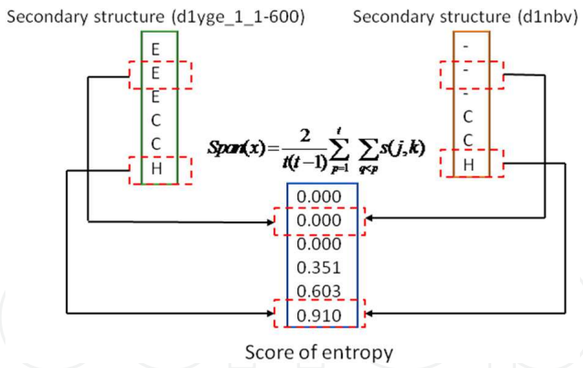


Fig. 4. An example of entropy calculation.

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

$$Span(x) = \frac{2}{t(t-1)} \sum_{p=1}^{t} \sum_{q<p} s(j,k) , \tag{6}$$

where $s(j,k)$ is $s(\alpha_{px}, \alpha_{qx})$. $Span()$ 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. α_{px} and α_{qx} 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).

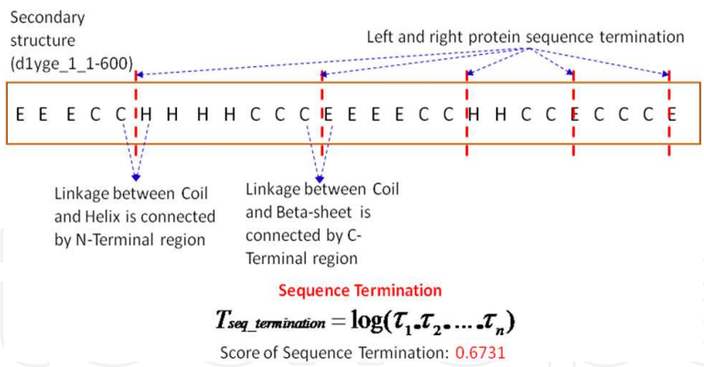


Fig. 5. An example of protein sequence termination calculation.

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:

$$T_{seq\_termination} = \log(\tau_1 \bullet \tau_2 \bullet \ldots \bullet \tau_n) , \tag{7}$$

where τ_n is the e-value of the n protein sequence.

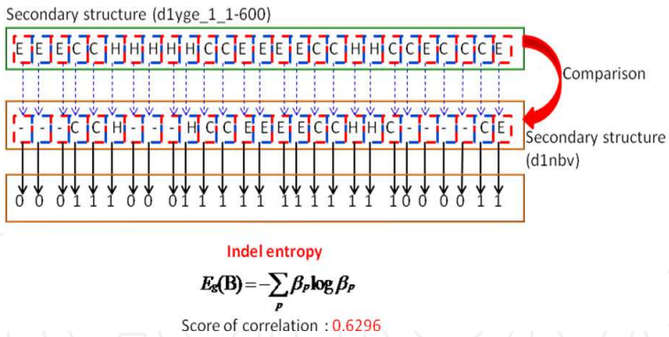


Fig. 6. An example of correlation calculation.

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:

$$E_g(\mathrm{B}) = -\sum_p \beta_p \log \beta_p \,, \tag{8}$$

where β_p is the various indel lengths seen at a position.

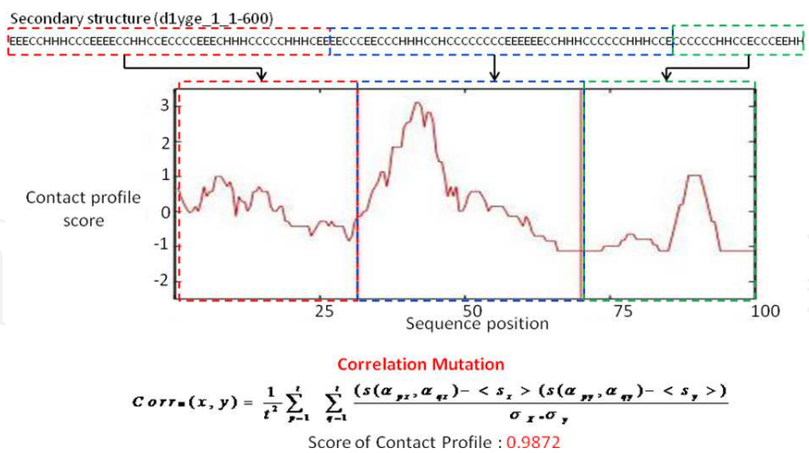


Fig. 7. An example of contact profile calculation.

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:

$$Corr_m(x,y) = \frac{1}{t^2}\sum_{p=1}^{t}\sum_{q=1}^{t}\frac{(s(\alpha_{px},\alpha_{qx}) - < s_x >)(s(\alpha_{py},\alpha_{qy}) - < s_y >)}{\sigma_X \bullet \sigma_y}, \quad (9)$$

where α_{px} and α_{qx} represent the amino acids in position x and the α_{py} and α_{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 α_{px}, α_{qx}, α_{py}, and α_{qy} refer to the scoring matrix BLOSUM50. The $< s_x >$ and $< s_y >$ are the average similarity of amino acids in position x and y. The σ_X and σ_y are standard deviations and t is the number of protein sequences that are indicated in the columns.

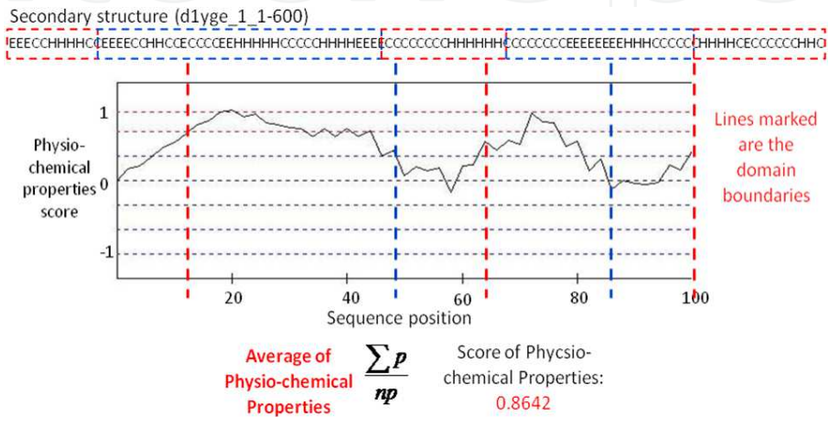


Fig. 8. An example of physio-chemical properties calculation.

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 hydropobicity 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.

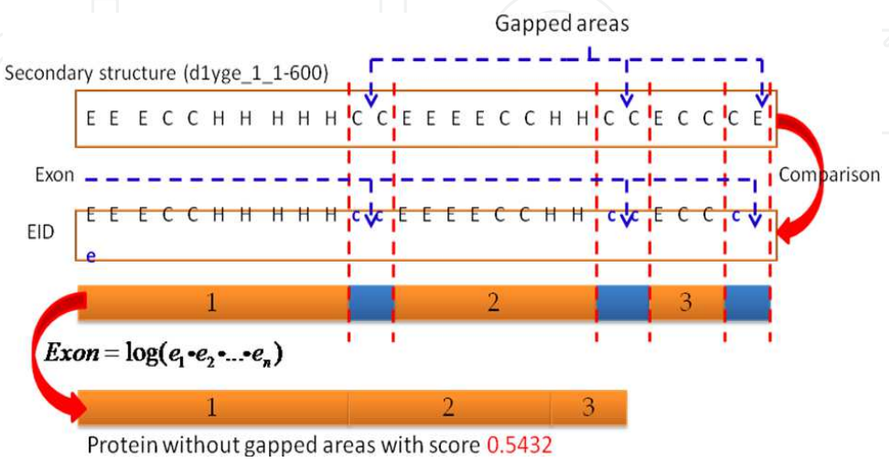


Fig. 9. An example of intron-exon calculation.

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:

$$E_{\text{exon\_termination}} = \log(\mathcal{E}_1 \bullet \mathcal{E}_2 \bullet \ldots \bullet \mathcal{E}_n), \tag{10}$$

where $\mathcal{E}_n$ 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:

$$\sum \frac{(Score\_of\_measures)}{n}, \tag{11}$$

where $Score\_of\_measures$ is obtain 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.

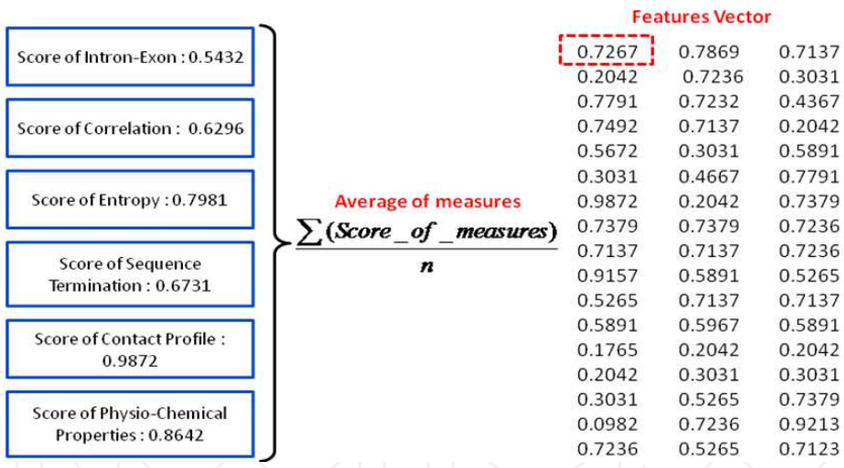


Fig. 10. An example of features vector calculation.

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 basic 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:

$$l_s \in \{+1, -1\}\,, \tag{12}$$

where I_s is the input space with corresponding predefined labels (Dong et al., 2003):

$$y_i \in I_s (i = 1, ..., n)\,, \tag{13}$$

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 σ^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:

$$K(\vec{y}_i, \vec{y}_j) = \exp(\frac{-r \mid \mid \vec{y}_i - \vec{y}_j \mid \mid^2}{2\sigma^2})\,, \tag{14}$$

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

The best pair of parameter of *C* and σ 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 σ 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 is divided 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.

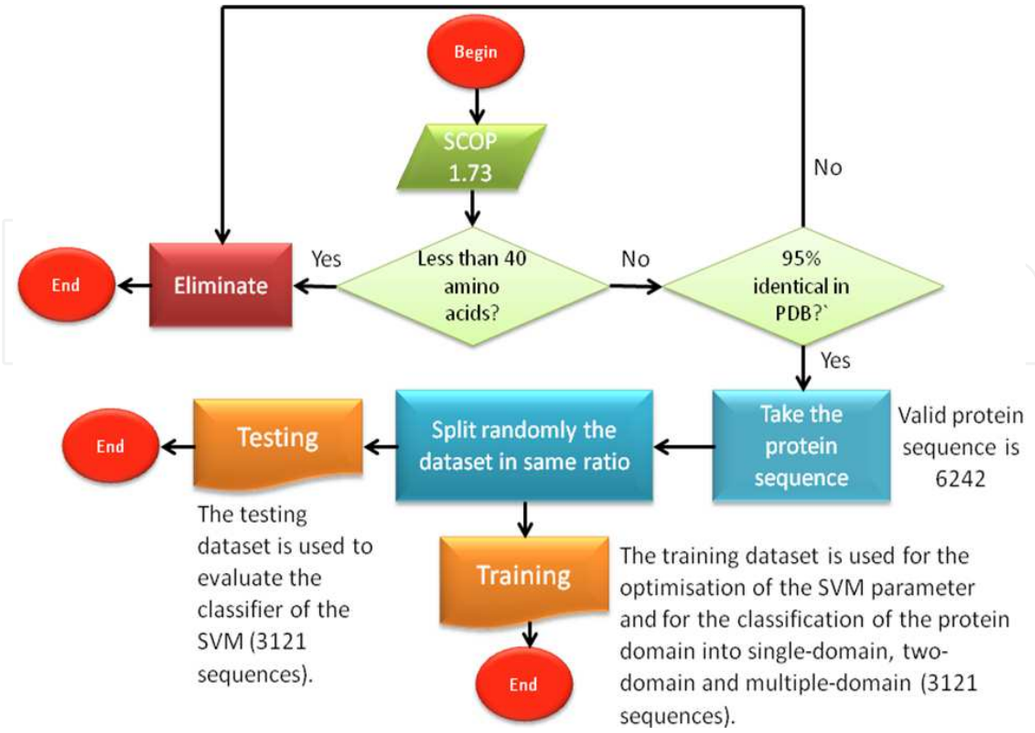


Fig. 11. Dataset generation process.

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:

$$SN = \frac{TP}{TP + FN} \times 100 \,, \tag{15}$$

$$SP = \frac{TP}{TP + FP} \times 100 \,. \tag{16}$$

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:

$$AC = \frac{TP - TN}{TP + FN + TN + FP} \times 100 \,. \tag{17}$$

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, physio-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 Armadilo. 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).

	Single-Domain		Two-Domain		Multiple-Domain		
Method	SN	SP	SN	SP	SN	SP	AC
BRNN-SVM	0.87	0.79	0.73	0.76	0.81	0.79	0.83
Similarity and multiple sequence alignment							
Biozon	0.27	0.93	0.33	0.23	0.21	0.35	0.38
KemaDom	0.82	0.76	0.70	0.73	0.78	0.76	0.79
Known protein structure							
AutoSCOP	0.80	0.65	0.62	0.57	0.73	0.72	0.69
DOMpro	0.85	0.80	0.43	0.55	0.79	0.73	0.71
Dimensional structure							
GlobPlot	0.78	0.74	0.32	0.58	0.59	0.67	0.69
Mateo	0.57	0.74	0.21	0.25	0.47	0.53	0.45
Dompred-DPS	0.55	0.73	0.52	0.43	0.67	0.66	0.62
Comparative model							
HMMPfam	0.65	0.60	0.53	0.59	0.35	0.33	0.62
HMMSmart	0.77	0.69	0.66	0.63	0.23	0.20	0.71
Sequence alone							
SBASE	0.86	0.77	0.69	0.74	0.76	0.76	0.80
Armadillo	0.31	0.89	0.29	0.21	0.17	0.35	0.27

Table 1. Performance comparison between BRNN-SVM and other protein domain prediction methods.

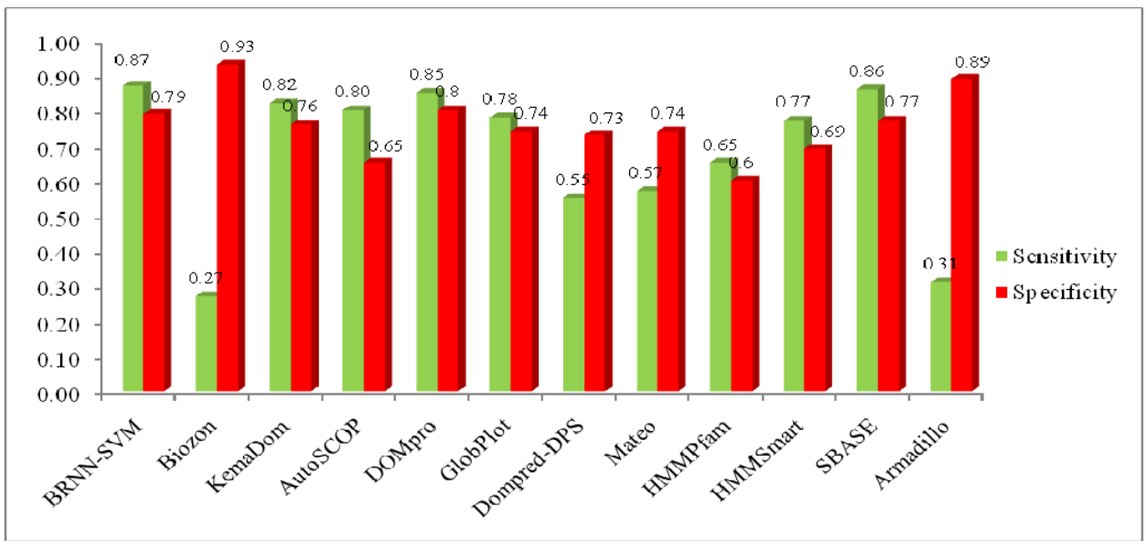


Fig. 12. Performance comparison between BRNN-SVM and other protein domain prediction methods on single-domain. The best sensitivity is BRNN-SVM with 87% and the best specificity is Armadillo with 89% since the BRNN-SVM classifies the protein sequence with no predicted protein domain boundaries as a single-domain.

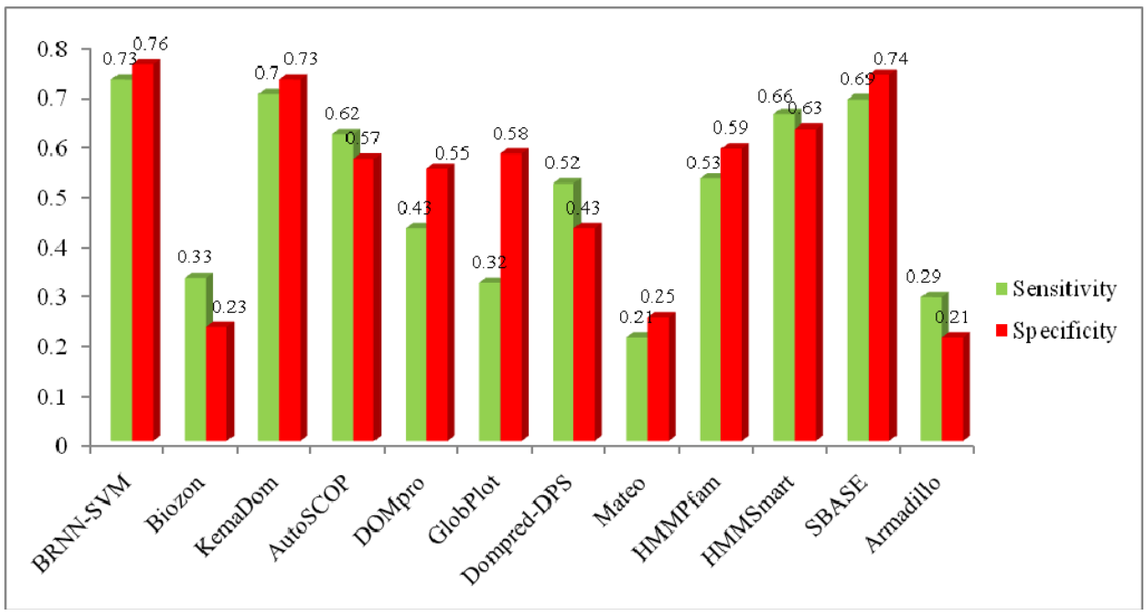


Fig. 13. Performance comparison between BRNN-SVM and other protein domain prediction methods on two-domain. The best performance for two-domain prediction is BRNN-SVM with 73% for sensitivity and 76% for specificity since the secondary structure information has given a strong signal to assign protein boundaries because the protein secondary structure predicted is based on interaction between long-range interactions of the amino acid.

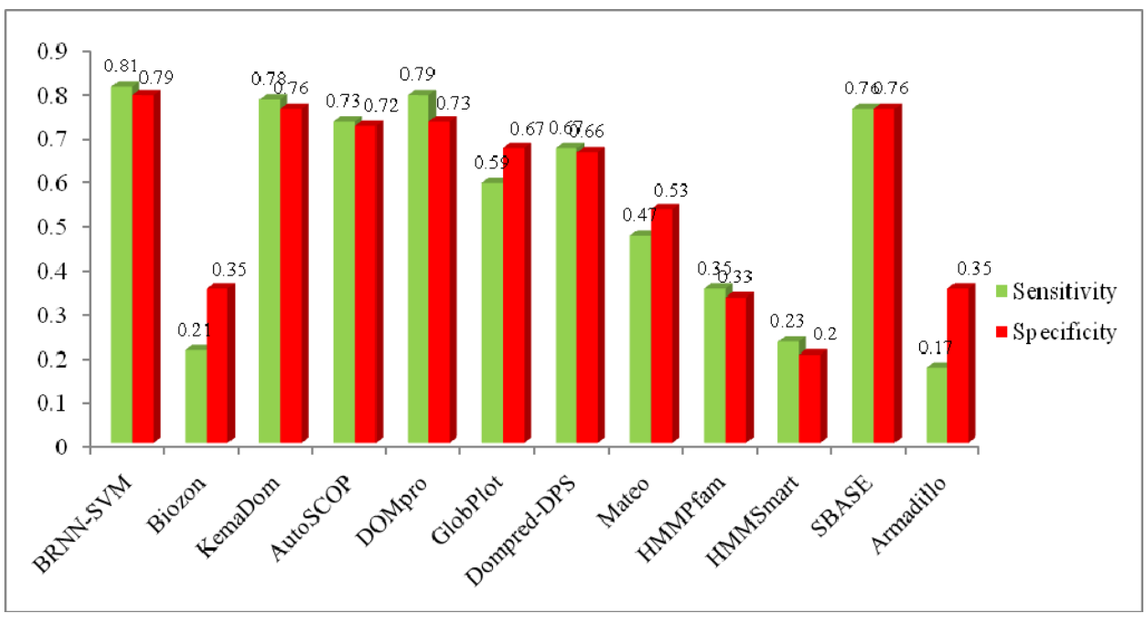


Fig. 14. Performance comparison between BRNN-SVM and other protein domain prediction methods on multiple-domain. The best performance of multiple-domain prediction is BRNN-SVM with 81% sensitivity and 79% specificity since the BRNN is a transaction between an input and an output sequence of variable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

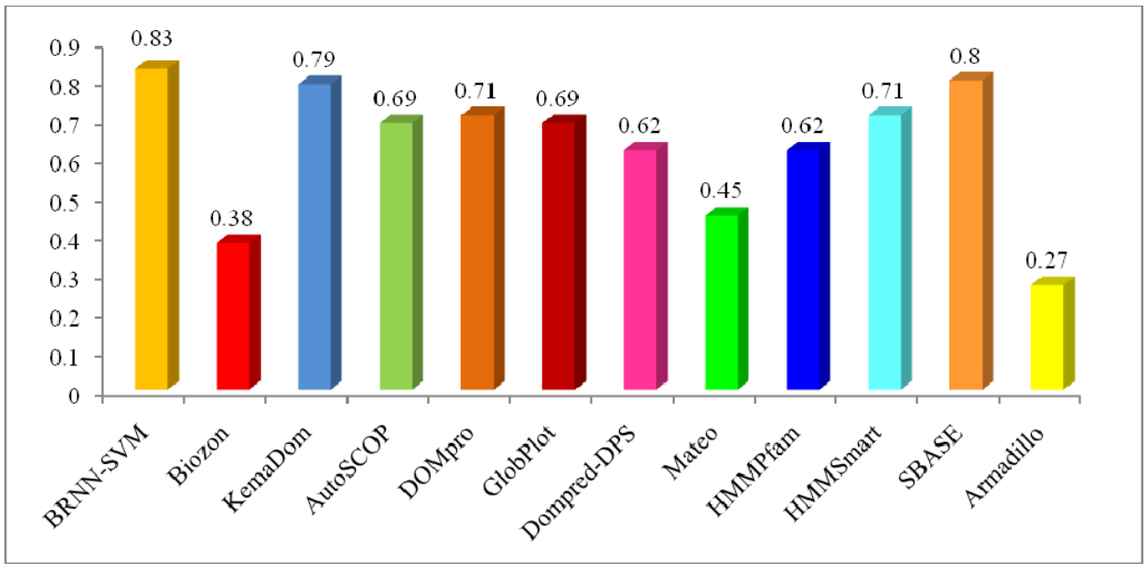


Fig. 15. Performance comparison between BRNN-SVM and other protein domain prediction methods on accuracy. The best accuracy of protein domain prediction is BRNN-SVM with 83% since the protein secondary structure is predicted using BRNN and the information of secondary structure is extracted from features extraction which increases the protein domain signal.

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

Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, Vol.25, No.17, (July 1997), pp.3389-3402, ISSN 0305-1048

Andreeva, A.; Howorth, D.; Chandonia, J.M.; Brenner, S.E.; Hubbard, T.J.P.; Chothia, C. & Murzin, A.G. (2008). Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Research*, Vol.36, No.Database Issue, (November 2007), pp.D419-D425, ISSN 0305-1048

Bateman, A.; Birney, E.; Cerruti, L.; Durbin, R.; Etwiller, *L.*; Eddy, S.R.; Griffiths-Jones, S.; Howe, K.L.; Marshall, M.; Sonnhammer, E.L.; David, J.; Studholme, C.Y. & Sean, R.E. (2004). The Pfam protein families database. *Nucleic Acids Research*, Vol.32, No.Database Issue, (January 2004), pp.D138-D141, ISSN 0305-1048

Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliand, G.; Bhat, T.N.; Weissing, H.; Ilya, N.S. & Bourne, P.E. (2000). The protein data bank. *Nucleic Acids Research*, Vol.28, No.1, (January 2000), pp.235-242, ISSN 0305-1048

Black, S.D. & Mould, D.R. (1991). Development of hydrophobicity parameters to analyze proteins which bear post or contranslation modification. *Analytical Biochemistry*, Vol.193, No.1, (February 1991), pp.72-82, ISSN 0003-2697

Chen, J. & Chaudhari, N. (2007). Cascaded bidirectional recurrent neural networks for protein secondary structure prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol.4, No.4, (October 2007), pp.572-582, ISSN 1545-5963

Chen, P.; Liu, C.; Burge, L.; Li, J.; Mohammad, M.; Southerland, W.; Gloster, C. & Wang, B. (2010). DomSVR: domain boundary prediction with support vector regression from sequence information alone. *Amino Acids*, Vol.39, No.3, (February 2010), pp.713-726, ISSN 0939-4451

Cheng, J.; Sweredoski, M.J. & Baldi, P. (2006). DOMpro: protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks. *Data Mining and Knowledge Discovery*, Vol.13, No.1, (July 2006) pp.1-10, ISSN 1384-5810

Dong, L.; Yuan, Y. & Cai, T. (2006). Using bagging classifier to predict protein domain structural class. *Journal of Biomolecular Structure and Dynamics*, Vol.24, No.3, (December 2006), pp.239-242, ISSN 0739-110

Dumontier, M.; Yao, R.; Feldman, H.J. & Hogue, C.W. (2005). Armadillo: domain boundary prediction by amino acid composition. *Journal of Molecular Biology*, Vol.350, No.5, (July 2005), pp.1061-1073, ISSN 0305-1048

Elhefnawi, M.M; Youssif, A.A; Ghalwash, A.Z & El Behaidy, W.H. (1 January 2010). *An Integrated Methodology for Mining Promiscuous Proteins: A Case Study of an Integrative Bioinformatics Approach for Hepatitis C Virus Non-structural 5a Protein*, Springer, Retrieved from http://www.springerlink.com/content/ l067380601040028/

Gewehr, J. E.; Hintermair, V. & Zimmer, R. (2007). AutoSCOP: automated prediction of SCOP classification using unique pattern-class mapping. *Bioinformatics*, Vol.23, No.10, (March 2007), pp.1203-1210, ISSN 1367-4803

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks, *PNAS*, Vol.89, No.22, (November 1992), pp.10915-10919, ISSN 0027-8424

Henikoff, S.; Henikoff, J.G. & Pietrokovski. S. (1999). Block+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, Vol.15, No.6, (June 1999), pp.471-479, ISSN 1471-2105

John, B.; Ryan, M. & Fernando, P. (2006). Domain adaptation with structural correspondence learning, *Proceedings of the Empirical Methods in Natural Language*, pp. 120-128, ISBN 1-932432-73-6, Sydney, Australia, (July 22-23), 2006

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, Vol.22, No.12, (December 1983), pp.2577-2637, ISSN 0305-1048

Kalsum, H.U.; Shah, Z.A.; Othman, R.M.; Hassan, R.; Rahim S.M.; Asmuni, H.; Taliba, J. & Zakaria, Z. (2009). SPlitSSI-SVM: an algorithm to reduce the misleading and increase the strength of domain signal. *Computer in Biology and Medicine*, Vol.39, No.11, (November 2009), pp.1013-1019, ISSN 0010-4825

Kristian, V.; Laszlo, K.; Vilmos, A. & Sandor, P. (2005). The SBASE domain sequence resource, release 12: prediction of protein domain-architecture using support vector machines. *Nucleic Acids Research*, Vol.33, No.Database Issue, (January 2005), pp.D223-D225, ISSN 0305-1048

Larkin, M.A.; Blackshields, G.; Brown, N.P.; Chenna, R.; McGettigan, P.A.; McWilliam, H.; Valentin, F.; Wallace, I.M.; Wilm, A.; Lopez, R.; Thompson, J.D.; Gibson, T.J. & Higgins, D.G. (2007). ClustalW and ClustalX version 2.0. *Bioinformatics*, Vol.23, No.21, (November 2007), pp.2947-2948, ISSN 1367-4803

Lexa, M. & Valle, G. (2003). Pimex: rapid identification of oligonucleotide matches in whole genomes. *Bioinformatics*, Vol.19, No.18, (May 2003), pp.2486-2488, ISSN 1367-4803

Linding, R.; Russell, R.B.; Neduva, V. & Gibson, T.J. (2003). GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Research*, Vol.31, No.13, (July 2003), pp.3701-3708, ISSN 0305-1048

Lusheng, C.; Wei, W.; Shaoping, L.; Caiyan, J. & Fei, W. (2006). KemaDom: a web server for domain prediction using kernel machine with local context. *Nucleic Acids Research*, Vol.34, No.Web Server Issue, (July 2006), pp.W158-W163, ISSN 1362-4962

Marsden, R.L.; McGuffin, L.J. & Jones, D.T. (2002). Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Science*, Vol.11, No.12, (December 2002), pp.2814-2824, ISSN 0961-8368

Menachem, F. & Chen, Y. (2008). A computational framework to empower probabilistic protein design. *Bioinformatics*, Vol.24, No.13, (July 2008), pp.I214-222, ISSN 1367-4803

Nagaranjan, N. & Yona, G. (2004). Automatic prediction of protein domain from sequence information using a hybrid learning system. *Bioinformatics*, Vol.20, No.1, (February 2004), pp.1335-1360, ISSN 1367-4803

Paul, D.Y.; Abdur R.S.; Bing B.Z. & Albert Y.Z. (2008). Improving general regression network for protein domain boundary prediction. *BMC Bioinformatic*, Vol.9, No.1, (February 2008), pp.S12, ISSN 14712105

Pauling, L. & Corey, R. B. (1951). Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets, *PNAS*, Vol.37, No.11, (November 1951), pp.729-740, ISSN 0027-8424

Pazos, F.; Helmer-Citterich, M.; Ausiello, G. & Valencia, A. (1997). Correlated mutation contain information about protein-protein interaction. *Journal of Molecular Biology*, Vol.271, No.4, (June 1997), pp.511-523, ISSN 00222836

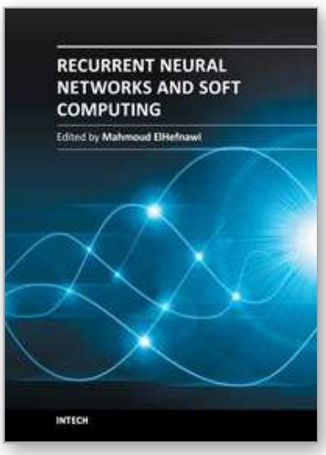
Ponting, P.; Schultz, J.; Milpetz, F. & Bork, P. (1999). SMART: identification and annotation of domains from signaling and extracellular protein sequences. *Nucleic Acids Research*, Vol.27, No.1, (January 1999), pp.229-232, ISSN 0305-1048

Renaund, G. & McConkey, B.J. (2005). Ab initio secondary structure prediction using inter-residue contacts, *Proceedings of the Research in Computational Molecular Biology*, pp.1-2, ISBN 3-540-25866-3, Cambridge, USA, (May 14-18), 2005

Saxonov, S.; Daizadeh, I.; Fedorov, A. & Gilbert, W. (2000). EID: the Exon-Intron Database - an exhaustive database of protein-coding intron-containing genes. *Nucleic Acids Research*, Vol.28, No.1, (January 2000), pp.185-190, ISSN 0305-1048

Zaki, N.; Deris, S. & Alashwal, H. (2006). Protein-protein interaction detection dased on substring sensitivity measure. *International Journal of Biomedical Science*, Vo.1, No. 1, (January 2006), pp.1216-1306, ISSN 2010-3832

Zou, S.; Huang, Y.; Wang, Y. & Zhou, C. (2008). A novel method for predicition of protein domain using distance-based maximal entropy. *Journal of Bionic Engineering*, Vol.5, No.3, (April 2008), pp. 215-223, ISSN 1672-6529



Recurrent Neural Networks and Soft Computing
Edited by Dr. Mahmoud ElHefnawi

ISBN 978-953-51-0409-4
Hard cover, 290 pages
Publisher InTech
Published online 30, March, 2012
Published in print edition March, 2012

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Kalsum U. Hassan, Razib M. Othman, Rohayanti Hassan, Hishammuddin Asmuni, Jumail Taliba and Shahreen Kasim (2012). BRNN-SVM: Increasing the Strength of Domain Signal to Improve Protein Domain Prediction Accuracy, Recurrent Neural Networks and Soft Computing, Dr. Mahmoud ElHefnawi (Ed.), ISBN: 978-953-51-0409-4, InTech, Available from: http://www.intechopen.com/books/recurrent-neural-networks-and-soft-computing/brnn-svm-increasing-the-strength-of-domain-signal-to-improve-protein-domain-prediction-accuracy

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen