Available Online through
www.ijpbs.com (or) www.ijpbsonline.com

IJPBS |Volume 5| Issue 2|APR-JUN|2015|245-266

Research Article
Biological Sciences

DOM_SVM: PROTEIN DOMAIN BOUNDARY PREDICTION USING SUPPORT VECTOR MACHINE CLASSIFIER

Priya Chowdhury[1], Nilesh KR Jha[1]and Piyali Chatterjee[1*]

[1]Department of Computer Science & Engineering,
Netaji Subhash Engineering College, Kolkata-700152, India.

*Corresponding Author Email: priyachowdhury09@gmail.com

ABSTRACT

Domain boundary prediction plays a very important part in all the tasks involved in Proteomics, like homology based predictions, protein structure predictions etc. Domains are distinct structural units of a protein that can evolve and function independently. The accurate prediction of protein domain linkers and boundaries is often regarded as the initial step of protein tertiary structure and function predictions. Here, we are presenting a tool, DOM_SVM that robustly identifies the linker regions of multi-domain proteins. A two-pronged strategy is used to distinguish the linker and domain regions of a protein chain, i.e., use of a strong feature-set consisting of physicochemical properties of Amino acids taken from the AAIndex as well as Protein Secondary Structure and Relative Solvent Accessibility, and using the Support Vector Machine as a best two class classifier. The Support Vector Machine(SVM) is explored for accurate prediction of domain and linker regions by training on the curated dataset obtained from the CATH database, with the consideration of getting maximum recall in case of domain boundary prediction. The software is then tested on 90 target proteins of the CASP-11 dataset in order to evaluate its prediction accuracy using three-fold cross-validation experiments. We have observed significant performance of this classifier in prediction of domain regions of CASP-11 targets. DOM_SVM achieves a high accuracy, recall and precision for most of the target proteins of the CASP-11 dataset. Hence, it can be concluded that in most cases, DOM_SVM achieves better performance compared to the existing state-of-the-arts.

KEY WORDS

Domain boundary, Multi-domain protein, Physicochemical properties, Protein Secondary Structures, Support Vector Machine

INTRODUCTION

Protein and genes are two important macromolecules present in cells of any living organism. Proteins in different forms are involved in most of the cell functions, apart from providing structural support to cell bodies. For synthesizing proteins, cells require information about the polypeptide chains or amino acid sequences that constitute different proteins. The information is encoded inside the genes through Deoxyribonucleic Acid (DNA) sequences. To know about cell functions very closely, it is necessary to map and sequence the genomes of different organisms. Success of Human Genome Project, aided by rapid DNA sequencing techniques, has become a landmark event in this regard. Under this project it has been possible to prepare the complete sequence of the human genome, estimated to contain about 3 billion base pairs of nucleotides in double helices of DNA molecular structures. Encouraged by this success, various research efforts were launched,

in the early period of 2000s, to map and sequence the genomes of a variety of organisms. A *domain* is a segment of a protein chain that can fold into a three dimensional structure independently. The overall 3D structure of the polypeptide chain is referred to as the protein's tertiary structure, whereas the domain is the fundamental building block of tertiary structure. It contains an individual hydrophobic core built from secondary structural units separated by loop regions. Due to the evolution, multi-domain proteins are likely to have emerged to create new functions. As a result, various proteins have been diverged from common ancestors by different combinations and associations of domains. To predict the tertiary structure of a protein, it is useful to segment the protein by identifying domain boundaries in it. The knowledge of domains is used to classify proteins and understand their structures, functions and evolution. So, it can be said that a *domain* is a structural and functional unit of protein. A number of methods so far have been developed to identify protein domains starting from their primary sequences which are mainly developed for prediction of multi-domains in protein chains.

Galzitskaya et al. [1] have developed a method based on finding the minima in a latent entropy profile. This method correctly predicts the domain boundaries for about 60% proteins. A method DOMCUT[2], based on the difference in amino acid compositions between domain and linker regions, has been developed to predict linker regions among domains. The sensitivity and the selectivity, as achieved by this method, are 53.5% and 50.1% respectively. CHOPnet[3] uses evolutionary information, predicted secondary structure, solvent accessibility, amino acid flexibility and amino acid composition for predicting domains in protein chains. It achieves prediction accuracy of 69% on all proteins. Armadillo[4] is the another domain predictor which uses any amino acid index to convert a protein sequence to a smoothed numeric profile. The work is finally reported to have achieved 37% sensitivity for multi-domain proteins. PPRODO[5] uses evolutionary information in the form of the position specific scoring matrix of the target protein, which can be obtained through PSI-BLAST. This information has also been used for domain boundary prediction. Artificial neural network has been used there as a classifier. The overall accuracy of domain boundary prediction as achieved by PPRODO is 67%. Machine Learning based ab-initio domain predictor DOMpro[6] uses recursive neural networks (1D- RNNs) to predict domains in a protein chain. To test the prediction accuracy of DOMpro, a curated dataset, derived from the CATH database is used. DOMpro is found to be correctly predicted the domains from the combined dataset of single and multi-domain proteins in 69% of the cases. In the work of *Sikder and Zomaya*[7] , the inter-domain index value uplifts the performance of *Domain Discovery* of protein domain boundary assignment. The method has achieved 70% accuracy for multi-domain proteins. Cheng proposed a hybrid domain prediction web service, called DOMAC[8], by integrating *template-based* and *ab-initio* methods. The template-based method is used in DOMAC to predict domains for proteins having homologous template structures in protein Data Bank. As a result, the overall domain number prediction accuracies of the *template-based* and *ab-initio* methods are 75% and 46% respectively. To achieve a more accurate and stable predictive performance, a new machine learning based domain predictor, viz., DomNet [14] is trained using effective feature sets like a novel compact domain profile, predicted secondary structure, solvent accessibility information and inter-

domain linker index. It is observed to have achieved 71% accuracy. FIEFDom[9] is other type of multi-domain prediction where prediction is done using fuzzy mean operator. This fuzzy operator assigns a membership value for each residue as belonging to a domain boundary thus finding contiguous boundary regions. Eickholt *et al.* [10]propose a new method DoBo where machine learning approach with evolutionary signals is used. It achieves 60% recall and 60% precision. Another SVM predictor *DROP*[11] with 25 optimal features distinguish domains from linkers very effectively. They use random forest algorithm to evaluate features. Based on creating a hinge region strategy, a new approach *DomHR* predicts domain boundary by computing profiles of domain Hinge-boundary (DHB) features. In the work of *Sadowski*[12], prediction of domain boundaries is done from inverse covariance's using kernel smoothing based method and alpha carbon models.

In the light of the above discussion, it appears that there is still scope for improvement for domain boundary prediction. Physicochemical properties of Amino Acids, predicted protein secondary structure, solvent accessibility and the use of SVM classifier appear to have the potential for the implementation of these ideas.

MATERIALS AND METHODS

The present work, DOM_SVM uses a strategy, i.e. designing of a strong feature set, and using SVM as a best classifier for a two class classification. The current experiment is conducted in two stages. In the first stage 354 protein chains of the CATH database (version 2.5.1) are used to perform a three-fold cross validation experiments on the three kernels of the SVM, namely, Linear, Polynomial and Radial basis function. Based on the evaluation metrics, the most suited and accurate SVM is chosen. This SVM is then used to test 90 CASP 11 target proteins in the second stage of the experiment.

the datasets:

CATH Dataset:- This experiment is trained on 354 protein chains of the CATH database(version 2.5.1). The average chain length of these amino acids lie between 300-500 residues; and each protein chain consists of, at an average, two domains connected by a linker.Here, we consider domain boundary region as within ± 20 residues from the true boundary assignment.

CASP Dataset : DOM_SVM is tested on the target proteins of CASP 11 dataset available on www.predictioncenter.org. The CASP11 dataset consists of 90 target proteins with an average chain length of 250-350 residues.

the feature set:

The features taken for this work are a combination of structural information (like Protein Secondary Structure) and sequence information (like Average Flexibility indices, Hydrophobicity, Linker Propensity Index, Hydrophobicity scales, Polarity, Linker Propensity from Helical, etc) taken from the AAINDEX[13]. For a brief description of the features in the feature set, please refer Table 1. Description and relevance of choosing the features in the feature set are as given below.

1. The Normalized Flexibility Parameters (B-Values) (F8)

   The one of the reasons for flexibility and mobility is the presence of multiple domains in proteins. Structures of proteins in changing environment affect domain motions. Proteins are dynamic molecules that are in constant motion. As a consequence, the structural flexibility is the responsible factor for protein's mobility which is thereby associated with various biological processes such as molecular recognition and catalytic activity. Very often, it is found that linkers are composed

of flexible residues such that its adjacent domains can move independently. The Debye-Waller factor (B-value), which measures local residue flexibility, is widely used to measure residue flexibility. The prediction of flexibility may help to locate the position of linker and domain. In this work, Normalized average flexibility parameters (B-values) from the AAINDEX dataset have been taken as features as presence of multiple domains increases protein flexibility.

2. The Polarity (F7)
   The combination of polar and non-polar side chains constitutes the protein chain which governs the folding of a protein into 3D structure. As domain is a unit of 3-D structure so, polarity has been considered as a feature from the AAINDEX dataset.
3. The Amino Acid Linker Index (F3)
   To represent the preference for amino acid residues in linker or regions, a parameter called the linker index is defined by Sumaya and Ohara. The linker index S_i for amino acid residue is defined as follows:

$$S_i = -ln \frac{f_i^{linker}}{f_i^{domain}}$$

   where, f_i^{linker} or f_i^{domain} , is the frequency of amino acid residue *i* in the linker or domain region. The negative value of S_i indicates that the amino acid residue *i* preferably belongs to a linker region. From the AAINDEX dataset, linker index has been used as a feature.
4. Hydrophobicity (F2)
   The arrangement of hydrophobic side chains into the interior of the molecule to avoid contact with aqueous environment determines the nature of folding. The average hydrophobicity for linkers is observed as 0.65± 0.09. Small linkers show an average hydrophobicity of 0.69 ± 0.11, while large linkers are more hydrophobic with 0.62± 0.08. The more exposed the linker, the more likely it is to contain hydrophilic residues. Greater hydrophobicity is found in more linker connections between two domains.
5. Hydrophobicity Scale (F5)
   Hydrophobicity scales are values that define relative hydrophobicity of amino acid residues. The more positive the value, the more hydrophobic are the amino acids located in that region of the protein. These scales are commonly used to predict the Trans membrane alpha-helices of membrane proteins. When consecutively measuring amino acids of a protein, changes in value indicate attraction of specific protein regions towards the hydrophobic region inside lipid bi-layer.
6. Linker Propensity From Helical And Non-Helical (F9 & F10)
   Two main types of linker are identified; helical and non-helical. Helical linkers are thought to act as rigid spacers separating two domains. Non-helical linkers are rich in Prolines, which also leads to structural rigidity and isolation of the linker from the attached domains. This means that both linker types are likely to act as a scaffold to prevent unfavourable interactions between folding domains.
7. Protein Secondary Structure (F12)
   The secondary structure of the protein sequence is also taken as one of the features. Three different secondary structure information (helix, sheet and coil) of every amino acid in the protein sequence has been considered here as one of the features.

support vector machine as classifier:

The Support Vector Machine, which is known for its superb generalization abilities with two class

data, is selected here to act as a classifier of domain and linker residues of target proteins. This learning machine was developed by *Vapnik* [14]. The central residue of each window of amino acids, constituting the target protein, can be represented as a point vector in the input feature space. Considering such points corresponding to all the residues of the protein, two clusters, one representing the domain region and the other non-domain linker region, are ideally formed in the input feature space. Traditionally, a pattern classifier finds a hyper-plane or hyper-surface in the input feature space separating the two clusters. Out of the two class data, those representing the class of interest are called positive data and the other - negative data. In addition to performing linear classification, SVMs can efficiently perform a non linear classification by implicitly mapping their inputs into high-dimensional feature spaces. SVM represents an extension to nonlinear models of the generalized portrait algorithm developed by *Vladimir*[14]. Some of the popularly used Kernel functions of the SVM are given below:

1. Radial basis function Kernel

   In machine learning, the (Gaussian) radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernel-based learning algorithms. In particular, it is commonly used in support vector machine classification.

   The RBF kernel on two samples x and x', represented as feature vectors in some *input space*, is defined as:

   $$k(x, x') = exp\left(\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

   $\|x - x'\|^2$ may be recognized as the squared Euclidean distance between the two feature vectors. σ is a free parameter.

2. Polynomial Kernel

   The polynomial kernel looks not only at the given features of input samples to determine their similarity, but also combinations of these. For degree-*d* polynomials, the polynomial kernel is defined as

   $$k(x, y) = (x^T + c)^d$$

   where *x* and *y* are vectors in the *input space*, i.e. vectors of features computed from training or test samples and $c \geq 0$ is a free parameter trading off the influence of higher-order versus lower-order terms in the polynomial. When $c = 0$, the kernel is called homogeneous.

3. Linear Kernel

   Given some training data D, a set of n points of the form

   $$D = \left((x_i, y_i) \vee x_i \in R^p, y_i \in (-1,1)\right)\frac{n}{i=1}$$

   where the y_i is either 1 or −1, indicating the class to which the point x_i belongs. Each x_i is a p dimensional real vector. We want to find the maximum margin hyper plane that divides the points having $y_i = 1$ from those having $y_i = -1$.

RESULTS AND DISCUSSION

The experimentation for this study has two parts. In the first part, Experiment-I, the Support Vector machine is trained with an unbalanced dataset with positive and negative samples in the ratio of 1:5. In the second part, a balanced dataset, i.e., ratio of positive and negative samples is 1:1, formed as a result of under sampling of the negative data is used to train the Support Vector Machine. The experimentation in both the parts is done in two stages. The first stage consists of three-fold cross-validation experiments with three kernels of the Support Vector machine, namely- Linear, Polynomial and Radial Basis Function, followed by the selection of the most accurate model. The second stage of

the experimentation encompasses the testing of 90 CASP-11 target proteins on the selected model.

evaluation metrics:

The evaluation metrics used here are as follows:

Accuracy (A) = $\frac{TP+TN}{TP+TN+FP+FN}$

Precision (P) = $\frac{TP}{TP+FP}$

Recall (R) = $\frac{TP}{TP+FN}$

Sensitivity^{+}(Se^{+}) = $\frac{TP}{TP+FN}$

Sensitivity $^{-}$(Se^{-}) = $\frac{TN}{TN+FP}$

Specificity $^{+}$(Sp^{+}) = $\frac{TP}{TP+FP}$

Specificity $^{-}$(Sp^{-}) = $\frac{TN}{TN+FN}$

where TP, TN, FP, FN stand for True Positive, True Negative, False Positive and False Negative respectively.

experimentation

experiment-I

354 protein chains of the CATH database (version 2.5.1) have been considered for this experimentation, out of which, some proteins have been excluded as they contained unidentified regions. Over 1 lakh samples were derived from these proteins, among which 90,000 samples were negative and 20,000 samples were positive. To avoid any loss of data, 3-fold cross validation experiments were performed with all the available samples, in which each data set had positive and negative samples in the ratio 1:5. The detailed view is presented in the Table 2.

After the preparation of dataset, in the first stage of our experiment, 3-fold cross validation experiment were performed on three kernels of the Support Vector Machine i.e. Linear, Polynomial (Degree=2), RBF (Gamma=0.0901). The results of 3-fold cross validation are given in Table 3.

From Table 3, it is clear that the radial basis function kernel yields the best results. Here the results were found to be most accurate for SVM-I. Hence this case was further tested with different values of Gamma so as to properly tune the Support Vector Machine and to find the value of Gamma for which the best results can be obtained. The results can be found on Table 4. From Table 4, it has been found that the optimum value of Gamma is 0.013 where the best overall performance among all the values of Gamma has been achieved. This optimal Gamma value is used in the next stage of the experiment where this SVM is applied on 90 targets of CASP 11 dataset to evaluate its prediction accuracy. Table 5 lists performance measures for the 90 target proteins of the CASP 11 dataset.

From Table 5, it can be stated that DOM_SVM achieves an average accuracy of 94.5% on 90 targets of CASP 11 dataset. DOM_SVM shows promising results for 9 targets of the CASP 11 dataset, and identifies linkers in the following 9 targets - T0760, T0770, T0777, T0805, T0826, T0830, T0836, T0848 and T0851.

experiment-II

In the previous part of the experimentation, the SVM was trained with an unbalanced dataset (ratio positive: negative is 1:5). Training a classifier with an unbalanced dataset is generally a challenging issue. SVMs are no exception to this, as in this case, the SVM becomes *biased* towards the majority class [15]. So, to eliminate this problem, the second part of our experimentation is done with a balanced dataset, i.e. ratio of positive and negative sample is 1:1. The dataset is derived from the same CATH dataset as before. The detailed view is presented in the Table 6.

After the preparation of dataset, in the first stage of the experimentation, 3-fold cross validation experiment were performed on three kernels of the Support Vector Machine i.e. Linear, Polynomial (Degree=2), RBF

(Gamma=0.3). The results of 3-fold cross validation are given in Table 7.

From the Table 7, it is clear that the radial basis function kernel yields the best results. Here the results were found to be most accurate for SVM-III. Hence this case was further tested with different values of Gamma so as to properly tune the Support Vector Machine and to find the value of Gamma for which we can get the best results. The results are given in Table 8. From Table 8, it has been found that the optimum value of Gamma is 0.25, where the best overall performance among all the values of Gamma has been achieved. This optimal Gamma value is used in the next stage of the experiment where this SVM is applied on 90 targets of CASP 11 dataset to evaluate its prediction accuracy. Performance measures for the 90 target proteins of the CASP 11 dataset is given in Table 9.

From Table 9, it can be stated that DOM_SVM achieves an accuracy of 94.2%, precision of 0.9468 and recall of 0.9954 on an average on 90 targets of CASP 11 dataset. DOM_SVM shows promising results for 12 targets of the CASP 11 dataset, and identifies linkers in the following 12 targets – T0760, T0770, T0780, T0801, T0805, T0813, T0819, T0826, T0830, T0836, T0843 and T0851.

The difference between the predicted structure of these proteins, and the ones available in the CASP dataset is shown graphically in the following figures- Figure 1 and Figure 2 (T0760) and Figure 3 and Figure 4 (T0848).

comparison of DOM_SVM with the existing state-of-the-arts:

DOM_SVM is designed with the motive of demarcating domains and linkers in a given protein sequence. So far, it has been tested on 90 CASP 11 targets, and it has been able to detect linkers in some of them as mentioned before. Those targets in which linkers were found by DOM_SVM, have been tested with other domain-linker predictor tools available, like DROP[11], DOMPred, DoBo[10] etc. The results of this comparison have been tabulated in Table 10 (for the results of experiment I) and Table 11 (for the results of experiment II).

Table 1. ***Amino Acids features selected from the AAIndex database***

Feature no.	AAIndex Accession_no	Brief Feature Description
F1	BHAR880101	Average flexibility indices
F2	JOND750101	Hydrophobicity
F3	SUYM030101	Linker propensity index
F4	GEOR030103	Linker propensity from 2-linker dataset
F5	PONP930101	Hydrophobicity scales
F6	GEOR030106	Linker propensity from medium dataset
F7	ZIMJ680103	Polarity
F8	VINM940101	Normalized flexibility parameters (B-values), average
F9	GEOR030108	Linker Propensity from Helical
F10	GEOR030109	Linker Propensity from Non-Helical
F11	BAEK050101	Linker Index
F12	-	Protein Secondary Structure
F13	-	Relative solvent accessibility

Table 2. ***Data Set showing Positive and Negative Samples***

Data Set	Positive Sample(Linker)	Negative Sample (Domain)
Data Set 1	6200	31000
Data Set 2	6200	31000
Data Set 3	6500	32500

Table 3. ***Performance measures of SVM-I, SVM-II, SVM-III***

SVM	Kernel	Accuracy(%)	Precision	Recall
SVM-I	**Linear**	49.55	0.494	0.548
	Polynomial	48.78	0.491	0.643
	Radial	48.10	0.490	0.934
SVM-II	**Linear**	45.23	0.161	0.5416
	Polynomial	47.51	0.160	0.518
	Radial	80.13	0.192	0.598
SVM-III	**Linear**	47.04	0.1709	0.5652
	Polynomial	45.13	0.1651	0.5652
	Radial	78.46	0.1571	0.669

Table 4.

Performance measures of SVM I with varying values of γ

γ	Accuracy (%)	Specificity$^+$	Specificity$^-$	Sensitivity$^+$	Sensitivity$^-$
0.01	58.74	0.1835	0.844	0.4278	0.6192
0.013	59.97	0.1839	0.8434	0.4075	0.6382
0.015	60.74	0.1843	0.8432	0.3959	0.6497
0.018	62.05	0.1852	0.843	0.3757	0.6695
0.02	63.09	0.1874	0.8433	0.3642	0.6843
0.023	64.52	0.1887	0.843	0.3422	0.706
0.025	65.67	0.1914	0.8432	0.3286	0.7222
0.027	66.45	0.1902	0.8422	0.311	0.7353
0.03	67.76	0.1914	0.8417	0.2895	0.7554
0.04	71.85	0.1968	0.8404	0.2237	0.8175
0.05	74.91	0.2028	0.8393	0.1725	0.8644
.0901	81.04	0.2138	0.8353	0.0514	0.9622

Table 5. *Performance measures of 90 CASP 11 targets*

S.No.	Target	PDB Id	Accuracy (%)	Sp$^+$	Sp$^-$	Se$^+$	Se$^-$
1	T0759	4928	96.77	-	0.972477	0	1
2	T0760	4pqx	90.27	1	0.792453	0.120	1
3	T0761	4pw1	79.93	-	0.810526	0	1
4	T0762	4qst	94.32	-	0.946429	0	1
5	T0763	4q0y	85.03	-	0.865031	0	1
6	T0764	4q34	96.31	-	0.964809	0	1
7	T0765	4pwu	67.86	-	0.593750	0	1
8	T0766	4q53	87.72	-	0.892308	0	1
9	T0767	4qpv	84.11	-	0.849057	0	1
10	T0768	4oju	90.26	-	0.911765	0	1
11	T0769	2mq8	92.71	-	0.794643	0	1
12	T0770	4q69	92.80	0	0.949791	0	0.978448
13	T0771	4qeo	80.32	-	0.740196	0	1
14	T0772	4quz	82.73	-	0.837736	0	1
15	T0773	-	96.72	-	0.766234	0	1
16	T0774	4qb7	93.11	-	0.891821	0	1
17	T0775	gp34	100	-	1	-	1
18	T0776	4qga	90.73	-	0.886719	0	1
19	T0777	-	96.86	0	0.975275	0	0.994398

20	T0780	4qdy	78.60	-	0.737452	0	1
21	T0781	4qaw	92.08	-	0.923810	0	1
22	T0782	4qrl	85.71	-	0.874074	0	1
23	T0783	4cvw	100	-	1	-	1
24	T0784	4qcy	84.17	-	0.858065	0	1
25	T0785	4d0v	100	-	1	-	1
26	T0786	4qvu	87.50	-	0.821970	0	1
27	T0789	4w4i	96.42	-	0.911864	0	1
28	T0790	4l4w	92.78	-	0.877133	0	1
29	T0791	4kxr	100	-	1	-	1
30	T0792	-	100	-	1	-	1
31	T0793	-	94.65	-	0.921008	0	1
32	T0794	4cyf	99.56	-	0.995745	0	1
33	T0795	-	100	-	1	-	1
34	T0796	-	99.66	-	0.938511	0	1
35	T0797	4ojk	100	-	1	-	1
36	T0798	4ojk	90.11	-	0.828283	0	1
37	T0799	-	100	-	1	-	1
S.No.	**Target**	**PDB Id**	**Accuracy (%)**	**Sp^+**	**Sp^-**	**Se^+**	**Se^-**
38	T0800	4qrk	88.31	-	0.890688	0	1
39	T0801	4piw	100	-	1	-	1
40	T0802	-	100	-	1	-	1
41	T0803	4oqw	100	-	1	-	1
42	T0804	-	100	-	1	-	1
43	T0805	-	99.49	0	1	-	0.995327
44	T0806	-	100	-	1	-	1
45	T0807	4wqw	100	-	1	-	1
46	T0808	4quw	97.51	-	0.976077	0	1
47	T0810	-	95.91	-	0.960836	0	1
48	T0811	-	100	-	1	-	1
49	T0812	-	95.21	-	0.877451	0	1
50	T0813	4wji	100	-	1	-	1
51	T0814	4r7f	96.57	-	0.966981	0	1
52	T0815	4u13	100	-	1	-	1
53	T0816	-	100	-	1	-	1
54	T0817	4wed	94.70	-	0.948571	0	1
55	T0818	4r1k	86	-	0.873494	0	1
56	T0819	4wkt	100	-	1	-	1
57	T0820	-	100	-	1	-	1
58	T0821	4r7s	95.75	-	0.960000	0	1
59	T0822	-	100	-	1	-	1

60	T0823	-	100	-	1	-	1
61	T0824	-	100	-	1	-	1
62	T0826	-	99.43	0	0.996317	0	0.998155
63	T0827	-	97.44	-	0.975430	0	1
64	T0828	-	58.58	-	0.587097	0	1
65	T0829	4rgi	100	-	1	-	1
66	T0830	-	95.71	0	0.968366	0	0.989228
67	T0831	4qul	100	-	1	-	1
68	T0832	4rds	86.72	-	0.813230	0	1
69	T0833	4r03	83.33	-	0.852941	0	1
70	T0834	4r7q	90.15	-	0.835616	0	1
71	T0835	-	97.06	-	0.971698	0	1
72	T0836	-	96.28	0	1	-	0.965686
73	T0837	-	100	-	1	-	1
74	T0838	-	86.23	-	0.876623	0	1
75	T0839	-	100	-	1	-	1
76	T0840	-	97.70	-	0.977578	0	1
77	T0841	-	100	-	1	-	1
78	T0843	4xau	100	-	1	-	1
S.No.	**Target**	**PDB Id**	**Accuracy (%)**	**Sp^+**	**Sp^-**	**Se^+**	**Se^-**
79	T0845	4rs0	96.76	-	0.962060	0	1
80	T0847	4urj	100	-	1	-	1
81	T0848	4r4q	93.20	1	0.934659	0	1
82	T0849	4w66	100	-	1	-	1
83	T0851	4w01	99.32	0	1	-	0.993421
84	T0852	4wqr	90.45	-	0.869565	0	1
85	T0853	2mqb	100	-	1	-	1
86	T0854	4m3	100	-	1	-	1
87	T0855	2mqd	100	-	1	-	1
88	T0856	4qt6	100	-	-	-	-
89	T0857	2mqc	100	-	-	-	-
90	T0858	-	94.35	-	-	-	-
	Avg		**94.5**		**0.90**		**0.9**

Table 6. ***Data Set showing Positive and Negative Samples***

Data Set	Positive Sample(linker)	Negative Sample (Domain)
Data Set 1	6200	12400
Data Set 2	6200	12400
Data Set 3	6500	13000

Table 7. ***Performance measures of SVM-I, SVM-II, SVM-III***

SVM	Kernel	Accuracy(%)	Precision	Recall
SVM-I	Linear	49.55	0.494	0.548
	Polynomial	48.78	0.491	0.643
	Radial	48.10	0.490	0.934
SVM-II	Linear	52.04	0.519	0.560
	Polynomial	51.19	0.517	0.571
	Radial	49.35	0.488	0.267
SVM-III	Linear	54.24	0.541	0.557
	Polynomial	54.56	0.558	0.441
	Radial	78.46	0.157	0.669

Table 8. ***Performance measures of SVM-III with varying values of*** **γ**

S.No.	γ	Accuracy (%)	Precision	Recall
1	0.01	51.33	0.5132	0.5168
2	0.09	51.45	0.5316	0.2448
3	0.2	51.34	0.5673	0.1128
4	0.23	51.88	0.5797	0.1368
5	0.245	51.33	0.5413	0.5571
6	0.248	51.32	0.5307	0.228
7	0.25	50.60	0.5044	0.6818
8	0.251	50.62	0.5036	0.3602
9	0.252	50.23	0.5013	0.9012
10	0.255	50.25	0.5014	0.9203
11	0.253	50.22	0.5012	0.9197
12	0.26	50.12	0.5006	0.9649
13	0.27	49.99	0.5	0.9832
14	0.4	50	0.5	1
15	0.5	50	0.5	1
16	1.5	50	0.5	1

Table 9. ***Performance measures of 90 CASP 11 targets***

S.No.	Target	PDB Id	Accuracy (%)	Precision	Recall
1	T0759	4928	96.77	0.9677	1

2	T0760	4pqx	89.82	0.8973	1
3	T0761	4pw1	79.93	0.7993	1
4	T0762	4qst	94.32	0.9432	1
5	T0763	4q0y	85.03	0.8503	1
6	T0764	4q34	96.31	0.9631	1
7	T0765	4pwu	67.86	0.6786	1
8	T0766	4q53	87.72	0.8772	1
9	T0767	4qpv	84.11	0.8411	1
10	T0768	4oju	90.26	0.9026	1
11	T0769	2mq8	92.71	0.9271	1
12	T0770	4q69	93.22	0.9483	0.9821
13	T0771	4qeo	80.32	0.8032	1
14	T0772	4quz	82.73	0.8273	1
15	T0773	-	96.72	0.9672	1
16	T0774	4qb7	93.11	0.9311	1
17	T0775	gp34	100	1	1
18	T0776	4qga	74.60	0.925	0.794
19	T0777	-	97.43	0.9743	1
20	T0780	4qdy	79.01	0.7893	1
21	T0781	4qaw	92.08	0.9208	1
22	T0782	4qrl	85.71	0.8571	1
23	T0783	4cvw	100	1	1
24	T0784	4qcy	84.17	0.8417	1
25	T0785	4d0v	100	1	1
26	T0786	4qvu	87.50	0.875	1
27	T0789	4w4i	96.42	0.9642	1
28	T0790	4l4w	92.78	0.9278	1
29	T0791	4kxr	100	1	1
30	T0792	-	100	1	1
31	T0793	-	94.65	0.9465	1
32	T0794	4cyf	99.56	0.9956	1
33	T0795	-	100	1	1
34	T0796	-	99.66	0.9966	1
35	T0797	4ojk	100	1	1
36	T0798	4ojk	90.11	0.9011	1
37	T0799	-	100	1	1
38	T0800	4qrk	88.31	0.8831	1
39	T0801	4piw	99.72	1	0.9972
40	T0802	-	100	1	1
41	T0803	4oqw	100	1	1
42	T0804	-	100	1	1
43	T0805	-	98.48	1	0.9848
44	T0806	-	100	1	1
S.No.	**Target**	**PDB Id**	**Accuracy (%)**	**Precision**	**Recall**
45	T0807	4wqw	100	1	1
46	T0808	4quw	97.51	0.9751	1
47	T0810	-	95.91	0.9591	1
48	T0811	-	100	1	1

49	T0812	-	95.21	0.9521	1
50	T0813	4wji	99.66	1	0.9966
51	T0814	4r7f	96.57	0.9657	1
52	T0815	4u13	100	1	1
53	T0816	-	100	1	1
54	T0817	4wed	94.70	0.947	1
55	T0818	4r1k	86	0.86	1
56	T0819	4wkt	99.72	1	0.9972
57	T0820	-	100	1	1
58	T0821	4r7s	95.75	0.9575	1
59	T0822	-	100	1	1
60	T0823	-	100	1	1
61	T0824	-	100	1	1
62	T0826	-	99.05	0.9962	0.9943
63	T0827	-	97.44	0.9744	1
64	T0828	-	58.58	0.5858	1
65	T0829	4rgi	100	1	1
66	T0830	-	92.67	0.9664	0.9575
67	T0831	4qul	100	1	1
68	T0832	4rds	86.72	0.8672	1
69	T0833	4r03	83.33	0.8333	1
70	T0834	4r7q	90.15	0.9015	1
71	T0835	-	97.06	0.9706	1
72	T0836	-	89.36	1	0.8936
73	T0837	-	100	1	1
74	T0838	-	86.23	0.8623	1
75	T0839	-	100	1	1
76	T0840	-	97.70	0.977	1
77	T0841	-	100	1	1
78	T0843	4xau	99.72	1	0.9972
79	T0845	4rs0	96.76	0.9676	1
80	T0847	4urj	100	1	1
81	T0848	4r4q	92.6	0.926	1
82	T0849	4w66	100	1	1
83	T0851	4w01	99.09	1	0.9909
84	T0852	4wqr	90.45	0.9045	1
85	T0853	2mqb	100	1	1
86	T0854	4m3	100	1	1
87	T0855	2mqd	100	1	1
S.No.	**Target**	**PDB Id**	**Accuracy (%)**	**Precision**	**Recall**
88	T0856	4qt6	100	1	1
89	T0857	2mqc	100	1	1
90	T0858	-	94.35	0.9435	1
	Average		**94.2**	**0.9468**	**0.9954**

Table 10. ***Comparison of PDP_SVM with existing state-of-the-arts for Experiment-I***

S.No.	Target	Sequence Length	Linkers Observed in Domain Definition	Linker Predicted By				
				PDP_SVM	DROP[11]	DoBo [10]	DomPred	DomSSEA
1	T0760	242	1-32	22-25	101-127	47	127	-
2	T0770	488	1-32	133-138, 420-422, 430	435-456	114,357, 421-422,433, 435	222,344	183,195, 366,263, 404,420
3	T0777	366	1-17, 363-366	136-137	160-172	52-53,100,102-103,300, 301-302,307-308,310-311,314, 316,317	-	236,291
4	T0805	214	1	142	75-96	66-67,69-74,97,99, 111-112, 152,155-157, 159,161-162, 167,170	75	50,58,89
5	T0826	544	1-10,370-544	84	208-232	47-48,50-52,54-57,63,66-68,70-65,79,227,238,483	228	96,147,176,183,194,257,269,355
6	T0830	575	1-26,574-575	204, 210-212, 244, 352	529-555	48,51,356,363-364,367-368,391, 394,396, 398,401, 510	88,257,312	156,222, 364,393

7	T0836	204	-	101-102, 106-110	63-89	51,75,79 - 80,105,111- 112,118, 120,121, 125,134 - 138,141	-	-
8	T0848	354	1-33	14-15	305-322	51,111,142- 145,149, 155-158,160, 163,169, 173	146	168,219, 244,247, 257,312
9	T0851	456	851	274-276	373-390	373-374,380, 383,408, 411	207,312	16,140,245,325

a) no linker found by the predictor.

Table 11. ***Comparison of PDP_SVM with existing state-of-the-arts for Experiment-II***

S. No.	Target	PDB Id	Sequence Length	Linkers Observed in Domain Definition	Linker Predicted By				
					PDP_SVM	DROP[11]	DoBo[10]	DomPred	DomSSEA
1	T0760	pqx	242	1-32	15-16	101-127	47	127	-
2	T0770	4q69	488	1-32	133-138, 349, 430	435-456	114,357,421-422,433,435	222,344	183,195,366,263,404,420
3	T0780	4qdy	259	1-39,231-259	14	138-256	NA	136	117,154
4	T0801	4piw	376	1-2, 376	82	268 - 288	NA	264	72,93,108,172, 190
5	T0805	-	214	1	139-141	75-96	66-67,69-74,97,9	75	50,58,89

							9, 111-112, 152,155-157, 159,161-162, 167,170		
6	T0813	4wji	307	303-307	192	147-156	48,57,61,66,69-70,245, 247,251,255,258,259,262-263	172	180,190,206
7	T0819	4wkt	373	372-373	232	200-215		158,284	131,241,248,258,271
8	T0826	-	544	1-10,370-544	23-24, 84	208-232	47-48,50-52,54-57,63,66-68,70-65,79,227,238, 483	228	96,147, 176,183,194,257,269,355
9	T0830	-	575	1-26,574-575	93, 195-209, 213-214, 244, 289, 292, 349, 353	529-555	48,51,356,363-364,367-368,391,394,396,398,401,510	88,257, 312	16,140, 245,325
10	T0836	-	204	-	48-55, 97-99, 101, 103-110	63-89	51,75,79-80,105, 111-112,118	-	-

							,120,121,125,134-138,141		
11	T0843	4xau	369	-	83	271-289	251-255,265,275,279,282,286,297-298,300,305,312,314,316,319,320	270	109,122,135
12	T0851	4w01	456	851	26-27, 274-275	373-390	373-374,380,383,408,411	207,312	16,140, 245,325

a) NA: Result not available in the predictor
b) no linker found by the predictor.

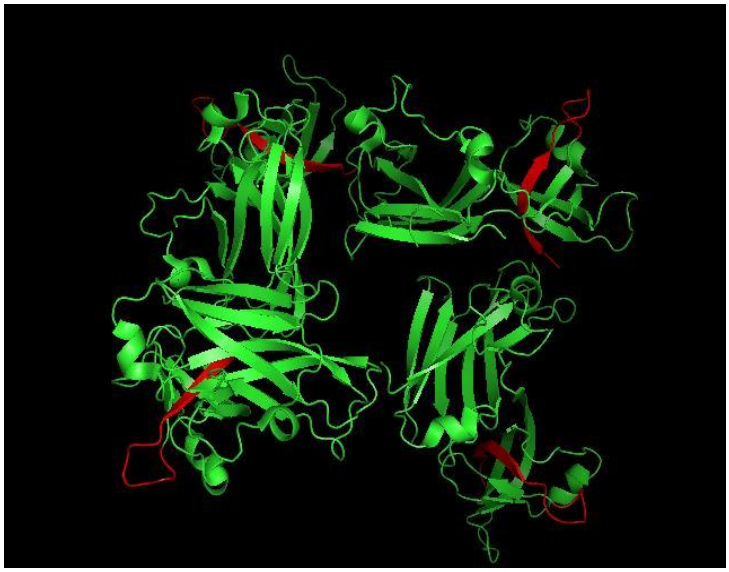

Figure 1. ***Structure of T0760. The green part is the domain region. The red part is the non-domain (linker) region.***

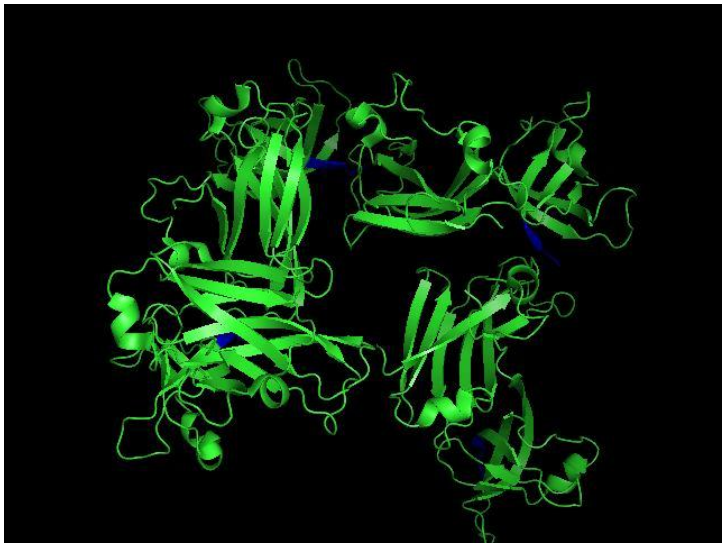

Figure 2. ***Structure of T0760 as predicted by DOM_SVM in experiment I. The green part is the domain region. The blue part is the linker region.***

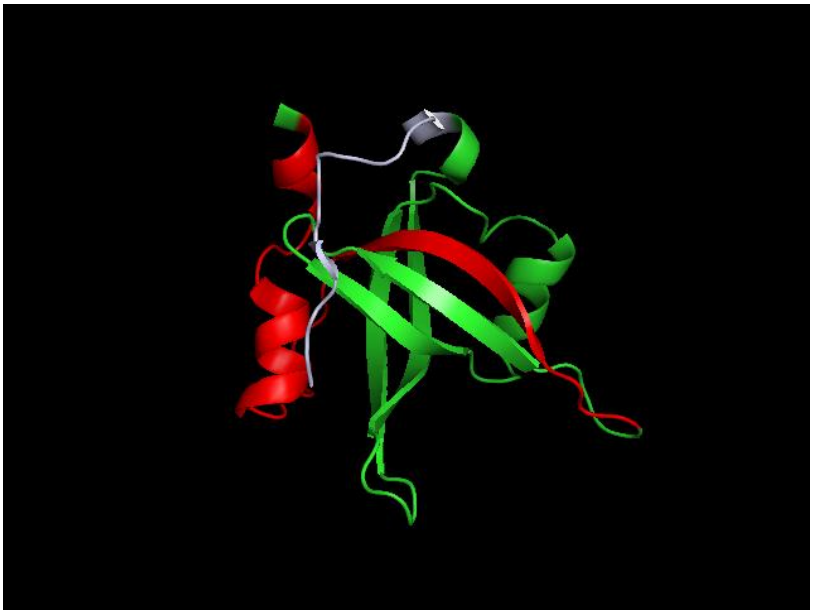

Figure 3. ***Structure of T0848. The green part and blue-white part are the two domain regions distinctly. The red part is the non-domain (linker) region.***

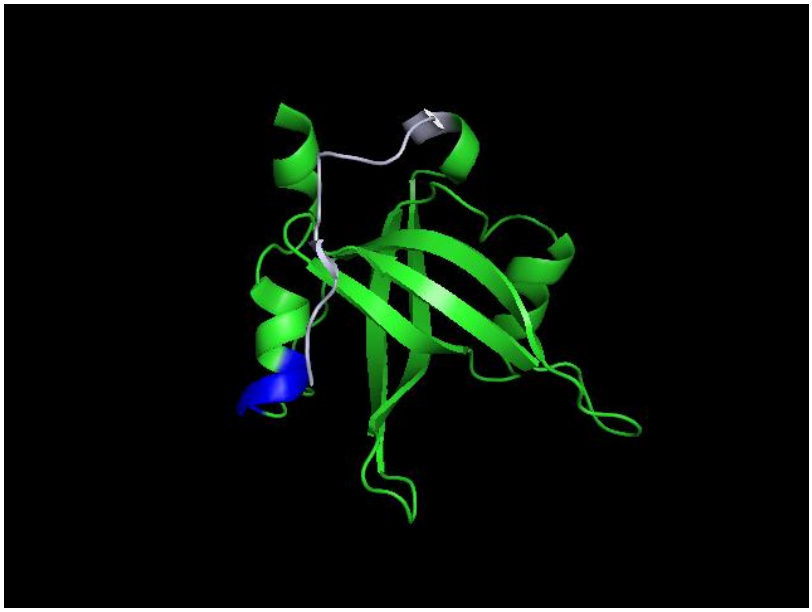

Figure 4. ***Structure of T0848 as predicted by DOM_SVM in experiment I. The green part and blue-white part are the two domain regions distinctly. The blue part is the linker region.***

CONCLUSIONS

The work presented in this study, DOM_SVM, addresses a very important issue of Bioinformatics, namely, prediction of Protein Domain Boundaries. This information will further facilitate the prediction of protein function. Machine learning techniques involving classifier like Support Vector Machine (SVM) has been developed for this purpose. In doing so, it has been observed that the choice of appropriate feature set and classifier is very important for the success of this technique. Keeping this in mind, DOM_SVM uses a two pronged strategy, i.e..designing a strong feature set (relevant physiochemical properties of Amino acids from the AAIndex database, predicted secondary structure, and predicted solvent accessibility) and using Support Vector Machine as best classifier for a 2-class classifier problem. DOM_SVM uses a SVM as a classifier, trained on different folds of training data. Curated data derived from CATH database is considered here for training the Support Vector Machine classifier. The model is then tested on 90 CASP 11 target proteins, and the results are evaluated on the basis of various metrics.

A number of works have already been done in this field, but there is still scope for improvement, that DOM_SVM tries to achieve. It uses a different strategy that combines a very strong feature set, with the genius of the best two- class classifier-The Support Vector Machine. DOM_SVM has been experimented with various ratios of training data. In case of an unbalanced dataset, the classifier is likely to become biased. Hence, DOM_SVM has been trained with balanced dataset as well. We have thus achieved a high rate of accuracy, precision and recall with DOM_SVM on 90 proteins of CASP 11 database.

In order to test the accuracy and relevance of the results of DOM_SVM, its performance on CASP 11 targets have been compared with that of other existing state-of-the-arts, like DROP[11], DOMPred, DoBo[10] etc on the same.

However, DOM_SVM does not take into account the ranking of features. Random Forest Classifier or any other T-test based strategy can

be used to rank these features that may further enhance its performance. The same can also be done by the use of evolutionary approach along with a set of strong feature set. In this work, CATH (2.5.1) has been taken as training dataset containing only 354 proteins. The latest version of CATH dataset (version 4.1) is available that can be used as training dataset as future work.

ACKNOWLEDGEMENTS

The authors are really grateful to Netaji Subhash Engineering College, Kolkata-700152, India, for pursuing this research work in the department of Computer Science and Engineering.

REFERENCES

[1] S. O. Garbuzynskiy, "Prediction of domain boundaries and unfolded regions in protein chain," *Proteins*, no. June, 2006.

[2] M. Suyama and O. Ohara, "DomCut: Prediction of inter-domain linker regions in amino acid sequences," *Bioinformatics*, vol. 19, no. 5, pp. 673–674, Mar. 2003.

[3] J. Liu and B. Rost, "Sequence-based prediction of protein domains," *Nucleic Acids Res.*, vol. 32, no. 12, pp. 3522–3530, 2004.

[4] M. Dumontier, R. Yao, H. J. Feldman, and C. W. V Hogue, "Armadillo: domain boundary prediction by amino acid composition.," *J. Mol. Biol.*, vol. 350, no. 5, pp. 1061–73, Jul. 2005.

[5] J. Sim, S.-Y. Kim, and J. Lee, "PPRODO: prediction of protein domain boundaries using neural networks.," *Proteins*, vol. 59, no. 3, pp. 627–32, May 2005.

[6] J. Cheng, M. J. Sweredoski, and P. Baldi, "DOMpro: Protein domain prediction using profiles, secondary structure, relative solvent accessibility, and recursive neural networks," *Data Min. Knowl. Discov.*, vol. 13, no. 1, pp. 1–10, 2006.

[7] A. R. Sikder and A. Y. Zomaya, "Improving the performance of DomainDiscovery of protein domain boundary assignment using inter-domain linker index.," *BMC bioinformatics*, vol. 7 Suppl 5. p. S6, 2006.

[8] J. Cheng, "DOMAC: An accurate, hybrid protein domain prediction server," *Nucleic Acids Res.*, vol. 35, no. SUPPL.2, pp. 354–356, 2007.

[9] R. Bondugula, M. S. Lee, and A. Wallqvist, "FIEFDom: A transparent domain boundary recognition system using a fuzzy mean operator," *Nucleic Acids Research*, vol. 37, no. 2. pp. 452–462, 2009.

[10] J. Eickholt, X. Deng, and J. Cheng, "DoBo: Protein domain boundary prediction by integrating evolutionary signals and machine learning.," *BMC Bioinformatics*, vol. 12, no. 1, p. 43, 2011.

[11] T. Ebina, H. Toh, and Y. Kuroda, "DROP: An SVM domain linker predictor trained with optimal features selected by random forest," *Bioinformatics*, vol. 27, no. 4, pp. 487–494, 2011.

[12] M. I. Sadowski, "Prediction of protein domain boundaries from inverse covariances.," *Proteins*, vol. 81, no. 2, pp. 253–60, Feb. 2013.

[13] S. Kawashima, "AAindex: Amino Acid index database," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 374–374, Jan. 2000.

[14] V. N. Vapnik, "The nature of statistical learning theory," Jun. 1995.

[15] A. Ben-hur and J. Weston, "A User ' s Guide to Support Vector Machines Preliminaries : Linear Classifiers," pp. 1–18.

*Corresponding Author:
priyachowdhury09@gmail.com