Improved prediction of Multi-domains in protein chains using a Support Vector Machine

Piyali Chatterjee1, Subhadip Basu2, Mahantapas Kundu2, Mita Nasipuri2*, Dipak Kumar Basu2

1 Department of Computer Science & Engg., Netaji Subhash Engineering College, Garia, India
Email: chatterjee_piyali@yahoo.com
2 Department of Computer Science & Engg., Jadavpur University, Kolkata, India
Email: {subhadip8@yahoo.com, nasipuri@vsnl.com}

Abstract— A two pronged strategy, one involving the Support Vector Machine (SVM) as the classifier and the other including physicochemical properties as additional features, is proposed and implemented here for improved prediction of multi-domains in protein chains. It is experimentally observed to have achieved an accuracy of 76.46 after 25 fold cross validation of results on curated data, derived from CATH database.

Index Terms— Protein domain boundary, physicochemical properties, conformational flexibility, amino acid linker index, linker region.

I. INTRODUCTION

A domain refers to a segment of a polypeptide chain that can fold into a three dimensional structure irrespective of the presence of other segments of the chain. 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, a domain is a structural and functional unit of protein.

A. The Past Work

Out of the recent protein domain prediction methods, DOMpro [2] is an important one. A combination of evolutionary information, in the form of profiles, predicted secondary structures, predicted solvent accessibility of the protein chains are utilized to achieve accuracy of 69%.

Armadillo [3], another domain predictor uses Domain Linker propensity Index (DLI) to predict domains and domain boundaries. The work is finally reported to have achieved 37% sensitivity for multi-domain proteins.

The PSSM computed by PSI-BLAST of the proteins has also been used for domain boundary prediction by PPRODO [4] to achieve overall accuracy of 67%.

CHOPnet [5] uses evolutionary information, predicted one-dimensional structure (secondary structure, solvent accessibility), amino acid flexibility and amino acid composition for predicting domains in protein chains to have accuracy of 69% on all proteins.

GALZITSKAYA et al [6] [11] 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 [11].

In the work of Sikder and Zomaya [7], the performance of DomainDiscovery of protein domain boundary assignment is improved significantly by including the inter domain linker index value. The method is reported to have achieved 70% accuracy for multi-domain proteins.

A method DOMCUT [8] has been developed to predict linker regions among functional domains. The sensitivity and the selectivity, as achieved by this method, are 53.5% and 50.1% respectively.

A hybrid domain prediction web service DOMAC, integrating template-based and ab-initio methods, has been developed by Cheng [10]. As a result, the overall domain number prediction accuracies of the template-based and ab-initio methods are 75% and 46% respectively.

A new machine learning based domain predictor, viz., DomNet [12] is trained using a novel compact domain profile, secondary structure, solvent accessibility information and inter-domain linker index to have achieved 71% accuracy on the benchmark_2 dataset.

This is all about some of the important protein domain prediction techniques published in the recent past. In the light of the above discussion, it appears that there is still some scope for improvement in protein domain prediction.

II. MOTIVATION

An idea for a two-pronged strategy, one for making the feature set more powerful and the other for making the domain predicting classifier more accurate has motivated the present work. The Support Vector Machine (SVM) in place of the other kind of neural network classifiers appears to have potentials for implementation of this idea. The rationale behind the use of physico-chemical

© 2009 ACEEE
DOI: 01.IJRTET.02.03.373
properties of amino acids in prediction of domain boundaries, under the present work, may be explained as follows. Studies say that folding is driven by the burial of hydrophobic side chains into the interior of the molecule so to avoid contact with aqueous environment. So how the polar and non-polar side chains are distributed is the most important factor governing the folding of a protein into 3D structure. For protein domains, the domain boundary is conditioned by amino acid residues with a small value of side chain entropy, which correlates with the side chain size.

III. PRESENT WORK

Being motivated by the two pronged strategy, mentioned above, an attempt has been made under the present work to employ Support Vector Machines (SVMs) with different kernel functions, viz., linear, polynomial of degree 3 and radial basis function, for protein domain prediction on the basis of powerful feature set, consisting of six different features, viz., predicted secondary structure, predicted solvent accessibility, predicted conformational flexibility profile, amino acid composition, position specific scoring matrix (PSSM) and physico-chemical properties of amino acids.

In doing so, a 13 residue window is slid over the protein chain every time by one residue position until the end of chain is covered by the window. Once a window is fixed, all the above mentioned features are extracted from each residue covered by the window. The feature values are then fed into the input of the SVM, which decides about whether the central residue of the window is part of a domain or a domain linker (i.e., a non-domain) by making its output 1 or -1. The process is repeated until the window covers the last residue of the protein chain under consideration and in this process, domain and linker regions in the protein chain are identified.

B. The Features Set

It has been already mentioned that six types of features, viz., PSSM, predicted secondary structure, predicted solvent accessibility, predicted conformational flexibility profile, amino acid composition, and physico-chemical properties, are used for this work. All these features are extracted from each residue of the 13-residue window. Here is a brief description of features.

The position Specific Scoring Matrix

To predict domain region in protein, it is useful to exploit evolutionary information in the form of position specific scoring matrix (PSSM) or profile. So, the PSSM of each protein of the dataset has been computed using PSI-BLAST [1].

The Predicted Secondary Structure

It is useful to have the knowledge of the secondary structure of a protein when determining the domain boundaries across the chain. In this work, predicted secondary structures using SSPro(4.0) is used as features.

The Predicted Solvent Accessibility

The Predicted relative solvent accessibility is extracted from each residue using ACCpro 4.0 (c: exposed, b: buried at 25% threshold). It is encoded into 1 or 0 as it is exposed or buried.

The Predicted Conformational Flexibility Profile

The structural flexibility enables this motion of protein molecule and is thereby associated with various biological processes such as molecular recognition and catalytic activity. The prediction of flexibility may help to unravel protein function. For the present work, a conformational flexibility profile is computed using the CFP server tool.

The Amino Acid Linker Index

To represent the preference for amino acid residues in linker or regions, a parameter called the linker index is defined by Sumaya and Ohara [8]. This we have used here as a feature.

Physicochemical properties of Amino Acid

It has been already been explained in Sec.II how the physicochemical properties of amino acids, viz., hydrophobicity, polarity and molecular weight/size affect formation of the domain boundaries in protein chains. All these three are considered here to boost the power of the feature set, consisting of traditional features.

C. The SVM as a classifier

On the basis of the features, discussed above, 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. Depending on which side of the hyper plane or surface any point or sample of an unknown class lies, the class membership of the point is determined by the classifier. The subject of Machine learning comes to play its role in helping the pattern classifiers to learn one such separating plane or hyper surface, also known as the decision surface.

Out of various learning machines, 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 non domain residues of target proteins. This learning machine was developed by Vapnik [13]. Out of the two class data, those representing the class of interest are called positive data and the others negative data. In the present case, all the sample residues representing domain regions are called positive examples and the others representing linker regions called negative examples.
In an SVM, the input data are nonlinearly mapped into a high dimensional feature space, hidden from both the input and output. In the said feature space a separating hyperplane is determined that maximizes the margin between the two classes of data. The margin is defined as the distance of the closest point, in each class, to the separating hyperplane. Obviously, one such hyperplane can best separate the two classes compared to all other separating hyperplanes. This is the key to the superb generalization abilities of the SVM on two class data. Conceptually, the support vectors are those data points that lie closest to the decision surface and are therefore most difficult to classify. As such they have a direct bearing on the optimal location of the decision surface.

For a given dataset \( \{ x_i, y_i \}_{i=1}^p \) of \( p \) examples \( x_i \) with labels \( y_i \in \{-1, +1\} \), the separating hyperplane can be represented as a linear combination of the training examples and classifying an unknown test pattern \( x \) is done using the following expression which is shown in equation (1).

\[
f(x) = \sum_{i=1}^p \alpha_i y_i k(x_i, x) + b
\]  

(1)

Where \( k(x_i, y_i) \) is the kernel function, \( b \) the bias.

Constructing the optimal hyperplane is equivalent to finding the nonzero \( \alpha_i \). Any data point \( x_i \) corresponding to \( \alpha_i \) is termed as support vector. Finally, the sign of \( f(x) \) determines the class membership of \( x \). Appropriately chosen kernel functions help SVMs to handle nonlinearly separable pattern classes. Popular kernel functions are polynomials of arbitrary degrees, Gaussian RBFs etc. In practice, an SVM is implemented as a 2-layer feed forward neural network.

D. Experimental Results and Discussion

The protein domain prediction technique presented here is applied on a curated data set, derived from the CATH database, version 2.5.1. In CATH database, pairwise identity of proteins is less than 25%. The SVMs designed here with kernel functions polynomials of degrees 1 and 3 and a radial basis function are referred to as SVM-I, SVM-II and SVM-III respectively. The performances of each SVM classifier designed here are evaluated after 25 fold cross validation of results on the test samples. To show these results in a compact form, average values of the three performance measures, viz., Accuracy, Precision and Recall, for the three SVM classifiers, with polynomials of degree 1 and 3 and a radial basis function as the kernel functions, are shown in Table I after 25 fold cross validation of results.

The curves showing variations of performance measures, viz., Accuracy, Precision and Recall, for all of the training folds for each SVM classifier with a kernel function as 1st degree or 3rd degree polynomial, or radial basis function are shown in Figure(1-3) respectively. Going through the results shown in Table I, it can be concluded that the SVM classifier with a degree 3 polynomial kernel has shown the best performances in terms of accuracy and precision. These two measures are computed as 76.46% and 86.82% respectively. In Figure 4, the variations of obtained accuracies using SVM-I, SVM-II and SVM-III are shown for all of the folds.

<table>
<thead>
<tr>
<th>TABLE I. PERFORMANCE MEASURES OF THREE SVMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of SVM</td>
</tr>
<tr>
<td>--------------</td>
</tr>
<tr>
<td>SVM-I (linear)</td>
</tr>
<tr>
<td>SVM-II (polynomial)</td>
</tr>
<tr>
<td>SVM-III (radial basis)</td>
</tr>
</tbody>
</table>

Figure1. Precision –Recall rate (%) using SVM-I
Figure2. Precision –Recall rate (%) using SVM-II
The CATH database is used for testing the performances of the methods, viz., DOMpro, Armadillo and CHOPNet. Though all the methods, discussed in Sec.I(A) of this paper, are not tested on the same database, accuracies obtained by them on standard databases can at least provide some idea about their performances. Out of the three performance measures, Compared to these methods, the accuracy of the present method, experimentally observed as 76.46 is the best. It is noteworthy that the accuracy of the present method is even better than that of the Domain Discovery, which employs an SVM classifier like the present one and uses SCOP database, different from the one used here.

CONCLUSIONS

Finally, it can be concluded that the two-pronged strategy, involving the SVM as a powerful classifier and including physicochemical properties as additional features, as initiated at the beginning of this work, has become fruitful with improved domain boundary prediction results on the standard database.

ACKNOWLEDGMENT

Authors are thankful to the “Center for Microprocessor Application for Training Education and Research”, of Computer Science & Engineering Department, Jadavpur University, for providing infrastructure facilities during progress of the work. One of the authors, Mrs. P. Chatterjee, is thankful to Netaji Subhash Engineering College, Garia for permitting her to carry out research work.

REFERENCES