BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net **Prediction model**

Volume 8(19)

Structure based function prediction of proteins using fragment library frequency vectors

Akshay Yadav[1]* & Valadi Krishnamoorthy Jayaraman[2]*

[1]38/Adwait, Pooja park, Paud road, Kothrud, Pune 411038; [2]Centre for Development of Advanced Computing(C-DAC), Pune; Akshay Yadav – Email: akshayy.yadav@gmail.com; Valadi Krishnamoorthy Jayaraman Email: jayaramanv@cdac.in
*Corresponding authors

Received Sepetember 14, 2012; Accepted September 19, 2012; Published October 01, 2012

Abstract:
The function of the protein is primarily dictated by its structure. Therefore it is far more logical to find the functional clues of the protein in its overall 3-dimensional fold or its global structure. In this paper, we have developed a novel Support Vector Machines (SVM) based prediction model for functional classification and prediction of proteins using features extracted from its global structure based on fragment libraries. Fragment libraries have been previously used for abintio modelling of proteins and protein structure comparisons. The query protein structure is broken down into a collection of short contiguous backbone fragments and this collection is discretized using a library of fragments. The input feature vector is frequency vector that counts the number of each library fragment in the collection of fragments by all-to-all fragment comparisons. SVM models were trained and optimised for obtaining the best 10-fold Cross validation accuracy for classification. As an example, this method was applied for prediction and classification of Cell Adhesion molecules (CAMs). Thirty-four different fragment libraries with sizes ranging from 4 to 400 and fragment lengths ranging from 4 to 12 were used for obtaining the best prediction model. The best 10-fold CV accuracy of 95.25% was obtained for library of 400 fragments of length 10. An accuracy of 87.5% was obtained on an unseen test dataset consisting of 20 CAMs and 20 NonCAMs. This shows that protein structure can be accurately and uniquely described using 400 representative fragments of length 10.

Keywords: Support vector machines, Fragment libraries, Protein fragments, Cell Adhesion Molecules, Function prediction

Background:

The structure function relationship of proteins is generally more reliable than the sequence-function relationship, for function annotation. Structure based identification of proteins are often superior to sequence based approaches because the folding pattern is retained even if the sequence similarity drops **[1].** There are several methods for structure based protein function prediction ranging from the analysis of the overall fold to identification of highly specific three-dimensional clusters of functional residues **[2].** Proteins with similar functions have similar folds and hence finding the structural neighbours is the first step for structure based function prediction. There are several existing methods for structural alignment with best known like DALI **[3],** and others that include SSM **[4],** GRATH **[5],** VAST **[6]** and CE **[7],** each having different algorithms for structural alignments. A new faster approach has been recently developed called FAST **[8]** uses a directionality-based scoring scheme to align structures at the residue-residue level rather than by secondary structure.

Fragment library is collection of representative fragments of a particular length. These libraries are constructed by clustering the fragments of CA traces (of particular length) of 200 accurately determined structures and taking representative fragment from each cluster **[9].** Fragment libraries have been shown to model protein structures accurately by representing the polypeptide chain by a sequence of rigid fragments, concatenated without any degrees of freedom **[9].** The main concept behind fragment library based modelling is to discretize the protein conformational space so that any chain has a finite number of spatial arrangements. This discretization is characterised by the accuracy with which it models the native protein conformations as well as number of allowed states per residue. Inbal Budowski-Tal *et al.* also used the same fragment libraries for retrieving the structural neighbours of the protein using an algorithm called FragBag **[10].** It has been also shown that a protein structure can be described by a sequence of 12,903 clusters or conformational types of overlapping peptide fragments prepared from 1.2 million amino acid residues in 4849 PDB structures. These conformation types can be considered equivalent to building blocks that is used by nature to build protein structures **[11]**

SVM (Vapnik, V. N., 2000) has gained popularity in comparison to other machine learning techniques for pattern recognition and prediction in biological data **[12-15]** because of their ability to very effectively handle noise and large datasets. In this present study, a novel Support Vector Machine (SVM) based algorithm has been developed for structure based function prediction using fragment libraries. As an example, this algorithm was applied for prediction of Cell Adhesion Molecules (CAMs). CAMs are proteins through which direct cell-to-cell contacts are made between cell surfaces and between cells and extracellular matrix. These contacts are required for the differentiation of cell structures during development, tissue formation and various interactions with cells like immune responses etc. Most of the CAMs belong to group of proteins called immunoglobulin superfamily (IgSF) having the characteristic immunoglobulin like domain **[16].** These domains mainly have a similar core structure with two β-sheets packed face-to-face. Multiple libraries of different sizes and fragment lengths were investigated for obtaining the best prediction accuracy for identification of CAMs.

Methodology:

Dataset

A training set containing well curated high resolution structures 100 CAMs and 132 NonCAMs were compiled from the Protein Data Bank (PDB) **[17].** All the proteins were single subunit proteins having only one chain. In case of multi-subunit CAMs, the subunit annotated with "Cell Adhesion" as function was extracted. An independent test set was prepared containing 20 CAMs and 20 NonCAMs which were previously not included in the training set.

Fragment Libraries

Fragment libraries are collection of representative protein fragments of particular length. All the fragment libraries were taken from previous study **[9].** There the fragment libraries are prepared by following algorithm. Extract all overlapping CA traces from 200 accurately determined protein structures of a particular length; Cluster the resulting fragments using k-means simulated annealing technique; Prepare a library by selecting a representative from each cluster.

In total of 34 libraries of fragment lengths ranging from 4 to 12 and sizes ranging from 4 to 400 were investigated for searching best CAM prediction model. The number of input features in each model were equal to the number clusters corresponding to that library.

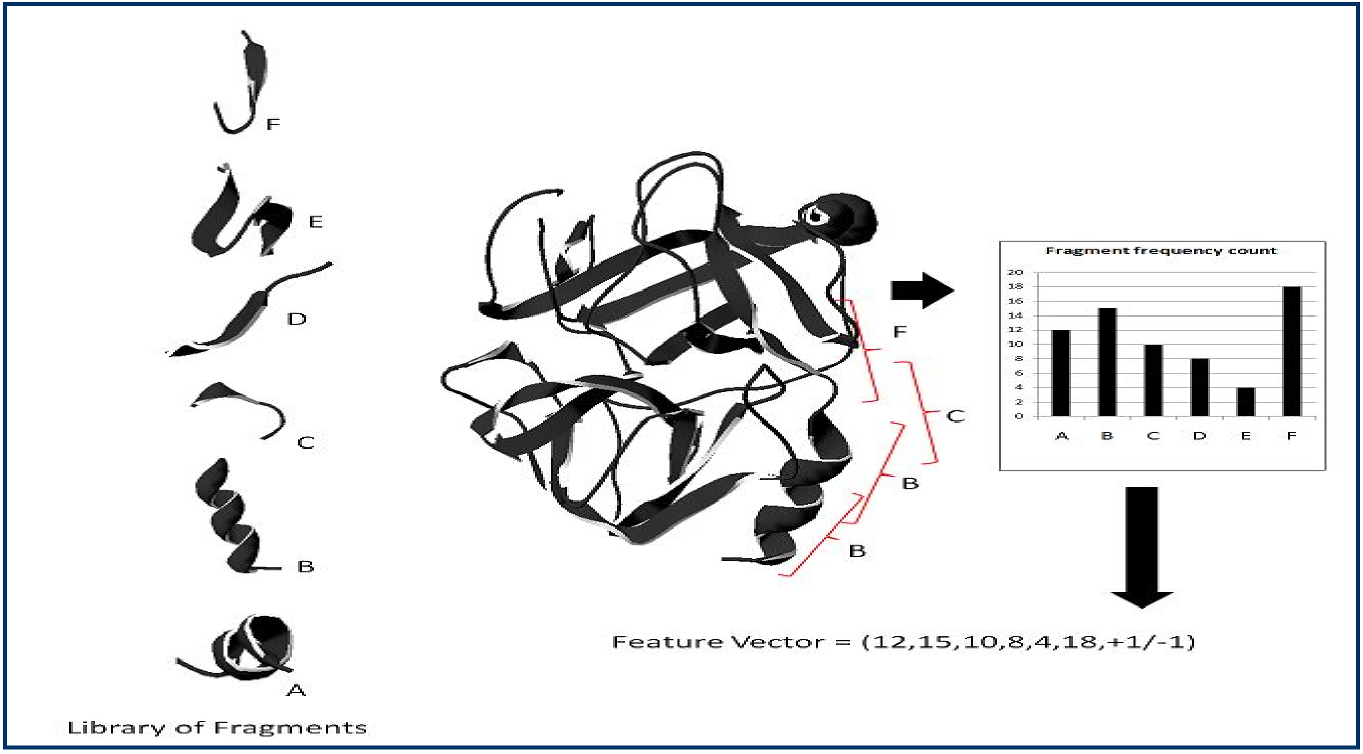

Figure 1: For illustration, we consider a library of 6 fragements. Each (over lapping) contiguous CA segment in the backbone is associated with its most similar library fragment appeared in the bag.In this example, feature vector = (12, 15, 10, 8, 4, 18), corresponding to the fragments (A, B, C, D, E, F). The last coordinate in the feature vector is +1 for CAMs and -1 for NonCAMs.

Fragment library frequency vectors generation

Each protein in the training dataset is broken down into collection of overlapping CA fragments. Then, for a particular fragment library, each fragment from the collection is compared to each library fragment using RMSD as a measure of similarity. This procedure is used to define a vector that counts the number of occurrences each library fragment in a given protein **(Figure 1).** Hence, the given protein can be described by this frequency vector which can be used as input features training and classification. The maxcluster [http://www.sbg.bio.ic.ac.uk/maxcluster/index.html#toc] program was used to calculate RMSD between the fragments and the whole procedure was automated using C code developed in house.

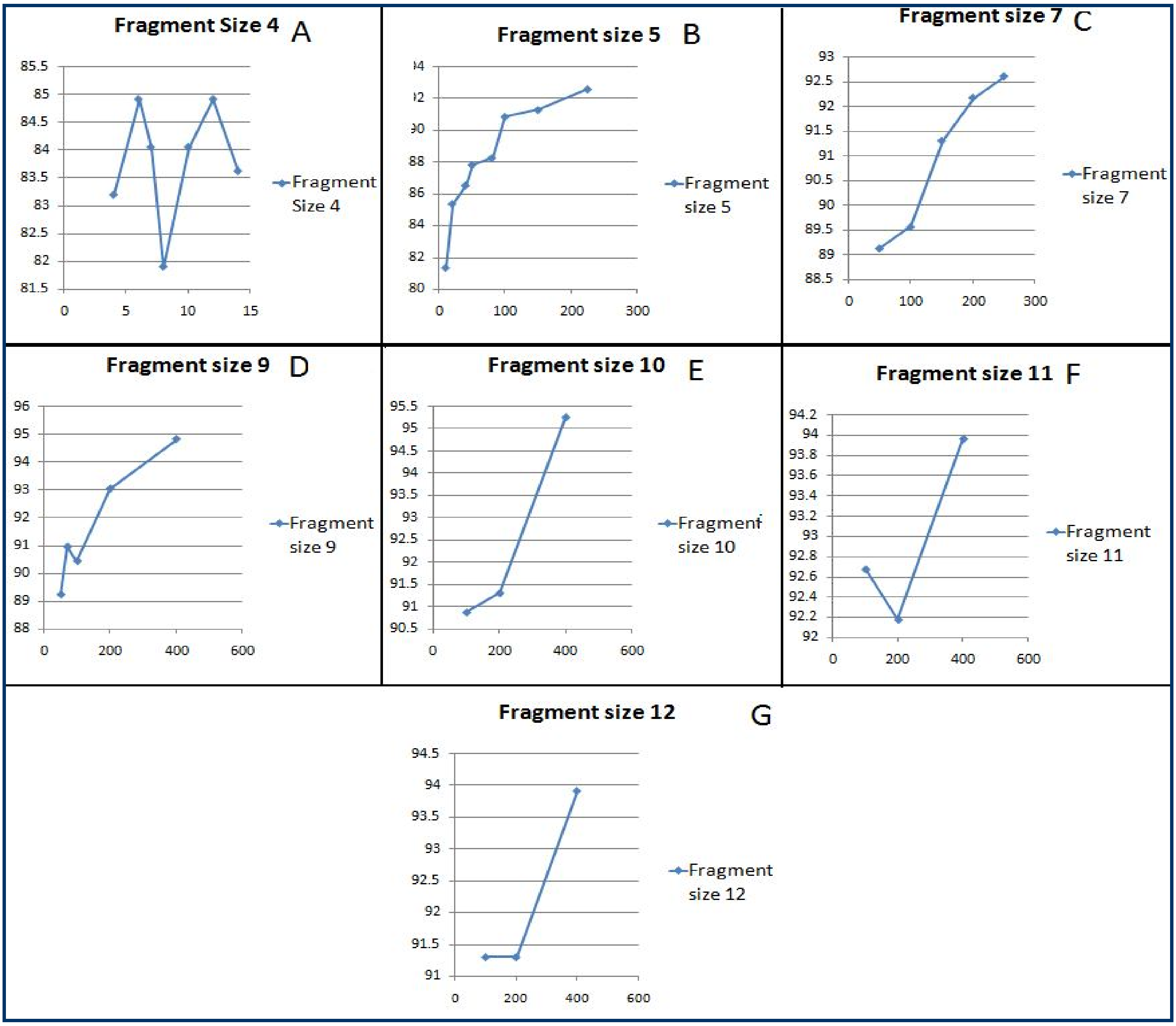

Figure 2: The variation of 10-fold Cross Validation % accuracy as a function of library size for fragment length 4 **(A),** fragment length 5; **(B),** fragment length **7; (C),** fragment length 9, **(D),** fragment length 10; **(E),** fragment length 11 **(F)** and fragment length 12 **(G).** The X axis represents the library size and the Y-axis represents the 10-fold Cross Validation % accuracy.

SVM training and classification

SVMs are a class of machine learning algorithms based on the theory of statistical learning and the principle of structural risk minimization **[11, 18]** that are used for pattern recognition and regression. SVM attempts to find an optimal hyperplane that maximally separates the training datasets by maximizing the margin between them. It non-linearly transforms the original input space into a higher-dimensional feature space by means of kernel functions to make the data linearly separable in higher dimensional feature space.

The training dataset is of the form $\{(x_i, y_i)\}$ i=1, 2. . . N. Here x_i is the vector representing the features (count of each fragment in the given library) for the *i*-th protein in the training dataset, y_i is the corresponding class of the protein and N is the total number of proteins in the training dataset. For CAMs y_i=+1 and for NonCAMs y_i=-1.

The SVM-based classification is dependent on the sign of f(x), which is calculated as

$$f(\mathbf{x}) = \sum_{i=1}^{m} y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b$$

Where m is the number of input data having non-zero values of Lagrange multipliers (α_i) (usually less than N) obtained by solving a quadratic optimization problem, K (x_i, x) is the kernel matrix and b is the bias term. Kernel matrix calculations were performed with kernel functions. The Gaussian Radial Basis Function (RBF) kernel was used model building and optimization. LIBSVM software **[19]** was used for training and testing the SVM classifiers. The user defined parameters for RBF kernel viz. gamma and regularisation parameter C were optimized for obtaining the best 10-fold Cross Validation (CV) accuracy using the grid search method with gamma ranging from (0.0001-1) and C ranging from (1-10). The performance of the model with maximum CV accuracy was tested on unseen test dataset after training with optimal parameters.

Discussion:

The SVM classifier models were built for structure-based prediction and classification of CAMs using each type of fragment library. The RBF kernel parameters were optimised using the grid search method for obtaining the maximum 10-Fold CV accuracy. The results were grouped according to fragment sizes and are shown in table 1-7. Any specific trend was not observed in %CV accuracy with increasing library size for fragment length of 4 **(Figure 2 A).** The %CV accuracy increased by increasing the library size, for fragment size 5 **(Figure 2 B).** This trend was also observed for fragment sizes 7, 9,10,11,12 **(Figure 2 C-G).** This result is consistent with the fact that increasing the library size increases the complexity (Number of allowed states per residue) which is able to model protein structure more accurately **[9].** Libraries of low complexity tend to have a lower accuracy than libraries of high complexity **[20]**. Highest CV accuracy of 95.2586% was obtained for the library of size 400 and fragment length 10 with γ = 0.0025 and Regularisation parameter C = 2. This classifier was used for testing the model on unseen test data consisting of 20 CAMs and 20 NonCAMs. The test accuracy was 87.5 % with 5 misclassifications out of 40. The sensitivity and specificity were 0.80 and 0.95 respectively.

Conclusion:

Fragment libraries can be used describe the protein structures accurately, by discretization of protein conformational space. The structural features derived using fragment library of size 400 and fragment length 10, can be effectively used for structure based classification and function annotation of proteins as the function is correlated to the structure. Fragments smaller than or equal to 4 cannot represent the structural information accurately, even if the library size is increased. This approach can be reliably used for structure based classification for other classes of proteins. The classification accuracy can be further improved by selecting the top ranking features for training.

Acknowledgement:

Dr. VKJ gratefully acknowledges Department of Science and Technology, New Delhi for financial support.

Reference:

[1] Whisstock JC & Lesk AM, *Q Rev Biophys*. 2003 **36**: 307 [PMID: 15029827]

[2] Watson JD *et al. Curr Opin Struct Biol*. 2005 **15**: 275 [PMID: 15963890]

[3] Holm L & Sander C, *J Mol Biol.* 1993 **233**: 123 [PMID: 8377180]

[4] Krissinel E & Henrick K, *Acta Crystallogr D Biol Crystallogr*. 2004 **60**: 2256 [PMID: 15572779]

[5] Harrison A *et al. Bioinformatics.* 2003 **19**: 1748 [PMID: 14512345]

[6] Madej T *et al. Proteins.* 1995 **23**: 356 [PMID: 8710828]

[7] Shindyalov IN & Bourne PE, *Protein Eng.* 1998 **11**: 739 [PMID: 9796821]

[8] Zhu J & Weng Z, *Proteins*. 2005 **58**: 618 [PMID: 15609341]

[9] Kolodny R *et al. J Mol Biol.* 2002 **323**: 297 [PMID: 12381322]

[10] Budowski-Tal I *et al. Proc Natl Acad Sci U S A.* 2010 **107**: 3481 [PMID: 20133727]

[11] Tendulkar AV *et al. J Mol Biol.* 2004 **338**: 611 [PMID: 15081817]

[12] Bhasin M & Raghava GP, *Nucleic Acids Res*. 2004 **32**: W414 [PMID: 15215421]

[13] Brown MP *et al. Proc Natl Acad Sci U S A***.** 2000 **97**: 262 [PMID: 10618406]

[14] Byvatov, E & Schneider G, *Appl Bioinformatics*. 2003 **2**: 67 [PMID: 15130823]

[15] Ding CH & Dubchak I, *Bioinformatics*. 2001 **17**: 349 [PMID: 11301304]

[16] Chothia C & Jones EY, *Annu Rev Biochem*. 1997 **66**: 823 [PMID: 9242926]

[17] http://www.pdb.org/pdb/home/home.do.

[18] Muller KR *et al. IEEE Trans Neural Netw.* 2001 **12**: 181 [PMID: 18244377]

[19] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[20] Park BH & Levitt M, *J Mol Biol*. 1995 **249**: 493 [PMID: 7783205]

Edited by P Kangueane

Citation: Yadav & Jayaraman, Bioinformation 8(19): 953-956 (2012)