Journal of Cell and Molecular Biology 8(2): 41-49, 2010
Haliç University, Printed in Turkey.
http://jcmb.halic.edu.tr

Research Article

Prediction of EF-hand calcium-binding proteins and identification of calcium-binding regions using machine learning techniques

Kunal JAISWAL*, Chandan KUMAR and Pradeep Kumar NAIK

**Department of Bioinformatics and Biotechnology, Jaypee University of Information Technology, Waknaghat, Solan, Himachal Pradesh, India*

*(*author for correspondence; kunal.jaiswal@yahoo.com)*

Received: 02 September 2007; Accepted: 17 March 2009

Abstract

Predicting calcium binding proteins and identifying calcium-binding sites is an important problem in the field of proteomics. Most of the currently used methods employ structural protein data to predict calcium-binding sites. Here we present a method developed to predict calcium-binding proteins and identify calcium-binding sites from protein sequence data using machine learning techniques such as neural networks and support vector machines. We have developed the novel application CalPred, having nine implemented algorithms divided into two filters. The first filter predicts proteins as a whole i.e. whether they are calcium-binding proteins or not, and the second filter predicts the specific calcium-binding sites in the proteins which have passed through the first filter. The tool was able to pick sequences that were calcium-binding in nature but were not picked up by pattern sited as calcium-binding domain in PROSITE database. We also scanned four whole proteomes for potential calcium binding proteins. Addition supplementary information is available at http://www.juit.ac.in/calpred/user_docs.html.

Availability: The CalPred tool is available for free use to non-commercial users and can be downloaded to be used in-house as a stand alone server from http://www.juit.ac.in/calpred/index.html

Keywords: EF-hand Calcium binding proteins, artificial neural networks, support vector machines, machine learning, CalPred.

Makina öğrenimi tekniklerini kullanarak EF-el kalsiyum bağlanma proteinlerinin tahmini ve kalsiyum bağlanma bölgelerinin tayini

Özet

Kalsiyum bağlanma proteinlerini tahmin etmek ve kalsiyum bağlanma bölgelerini saptamak proteomik alanında önemli bir problemdir. Günümüzde çoğu metod kalsiyum bağlanma bölgelerini tahmin etmek için yapısal protein verilerini kullanır. Bu makalede nöral ağ ve destek vektör makinaları gibi makina öğrenimi tekniklerini kullanarak, protein dizi verilerinden kalsiyum bağlanma proteinlerini tahmin etmek ve kalsiyum bağlanma bölgelerini saptamak için geliştirilen bir metod sunulmuştur. İki filtreye bölünmüş dokuz algoritma ile CalPred orijinal uygulamasını geliştirdik. İlk filtre proteinleri bir bütün olarak tahmin eder, yani kalsiyum bağlanma proteini olup olmadıklarını belirler; ikinci filtre ise birinci filtreden geçen proteinlerin spesifik kalsiyum bağlanma bölgelerini tahmin eder. Bu araç, doğada kalsiyum bağlanma özelliği olan ama PROSITE veri tabanınca kalsiyum bağlanma bölgesi motifi olarak seçilmeyen dizileri de seçer. Aynı zamanda dört proteomu kalsiyum bağlanma proteini potansiyeli için taradık. İlave destek bilgiye http://www.juit.ac.in/calpred/user_docs.html. sitesinden ulaşılabilir.

Kullanılabilirlik: CalPred aracı ticari olmayan kullanıcılar için ücretsizdir ve evde kullanım için http://www.juit.ac.in/calpred/index.html adresinden indirilebilir.

Anahtar sözcükler: EF-el kalsiyum bağlanma proteinleri, yapay nöral ağ, destek vektör makinaları, makina öğrenme, CalPred

Introduction

Calcium plays an important role in many biological processes including cell signaling (Carafoli, 2002), apoptosis (Orrenius et al., 2003) and cell differentiation (Hennings et al., 1980). It performs its various functions by binding with Ca+2 receptors called the calcium binding proteins or the CaBPs. Thus it is important to predict the CaBPs and identify the regions in these proteins where Ca+2 ions bind. Attempts have been made to solve the problem of identifying protein calcium-binding sites or in general metal-binding sites in proteins previously, but most of them used protein structure information. (Deng et al., 2005; Wei et al., 1999; Sodhi et al., 2004; Liang et al., 2003) Here we present a method to differentiate between CaBPs / non-CaBPs and find calcium-binding sites using protein sequence data. The CaBPs often share a common motif known as the EF-hand motif. The EF-hand motif is a helix-turn-helix structural motif that is twelve to thirteen residues long and is cited in PROSITE database, (Hulo et al., 2006) under entry number PS00018. The simplest way to predict the EF-hand calcium binding proteins would be using pattern matching, but in many organisms EF-hand-like calcium-binding proteins with different structural elements around the Ca+2 ions binding loop regions have been identified (Rigden et al., 2003; Rigden et al., 2003); and many of them have flexible lengths of Ca+2-binding loops that are different from loops present in EF-hand motifs. Thus, there is a need to apply techniques on CaBPs problem; that are not entirely dependent on pattern matching techniques. These could be using statistical / machine learning techniques that try to capture global information of the protein sequences.

Various machine learning techniques have been applied to biological problems related to proteins including protein structure prediction (Rost, 1996), protein fold recognition (Ding and Dubchak, 2001), protein sub-cellular localization prediction (Hua and Sun, 2001) and prediction of proteasome cleavage motifs (Kesmir et al., 2002). Here we applied two machine learning techniques namely, artificial neural networks (ANN) and support vector machines (SVM) to the problem of calcium-binding protein prediction and calcium-binding region identification.

Materials and methods

Encoding methods

In this study, three types of encoding methods are used, these are "pepstats", "binary" and "pssm encoding" methods.

Using the "pepstats encoding" method, the protein sequence is encoded into its physicochemical properties using pepstats application of EMBOSS suite of programs. Out of all the properties, 51 properties are then used to create an input vector for training and testing of the machine learning modules. These properties are normalized, prior to creation of the input vectors. The properties not used for encoding methods are the number of residues which depend on protein length, charge on protein which is dependent on residue type, A280 Molar Extinction Coefficient (i.e similar to A280 Extinction Coefficient 1mg/ml) and improbability of expression in inclusion bodies that depends upon many other factors, such as E. coli strain, incubation temperature, type of expression vector, strength of promoter and medium. The list of properties used for training and testing the modules and their associated normalization factors that are used in this study; are given in the supplementary material.

The "binary encoding" method takes a window of thirteen amino acid residues of a protein sequence at a time, and encodes every amino acid by a group of 20 units, each for a possible amino acid type that can be present at a particular position; thus creating thirteen binary vectors (1, 0, 0, . . .) and for a particular window totaling to 260 (13*20) units. The window is then shifted by per residue position so as to create N window frames where N is the number of residues in a protein sequence. The binary models are then trained on N training vectors for each protein sequence.

The "pssm encoding" method uses position specific scoring matrices created by PSI-BLAST using three iterations and with default e-value threshold of 0.001. The matrix has 20 * M elements, where M is the length of query protein; and each

element represents the frequency of occurrence of each of the 20 amino acids at one position in the alignment (Altschul et al, 1997). This encoding method then uses overlapping windows of 13*20 elements to create input vectors for machine learning modules. Similar to above, here too the window shift method is used to create M training vectors for each protein sequence.

In both the binary and pssm encoding methods, the protein sequence is tagged with a stretch of 7 "Xs", such that information on the first and last seven residues is not lost. In these methods, the X's are encoded as vectors of twenty zero's.

Modules and filter layers

In this study along with the two machine learning techniques, we used three types of protein sequence encoding methods i.e. "pepstats", "binary" and "pssm encoding" methods. Using a combination of these three methods, and the two techniques nine modules were created namely; $ANN_{pepstats}$, ANN_{binary}, ANN_{pssm}, $SVM_{pepstats\_linear}$, $SVM_{pepstats\_polynomial}$, $SVM_{pepstats\_radial\_basis}$, $SVM_{pepstats\_sigmoidal\_tanh}$, SVM_{binary} and SVM_{pssm}. The nomenclature of these modules is according to the rule that first word in the name indicates the machine learning technique say ANN or SVM, and the second subscripted word signifies the encoding method; "pepstats", "binary" or "pssm encoding". In case of SVM, a third subscripted word was used indicating the kernel type. If it is absent as in case of SVMbinary and SVMpssm the kernel type, defaults to linear kernel. The modules associated with pepstats encoding method constitute the first level filter of CalPred tool (Figure 1). The "pepstats" (Rice et al., 2000) is a program from EMBOSS suite that calculates physicochemical properties of a protein sequence as a whole and thus it is used to predict the nature of a protein, whether it's a CaBP or non-CaBP. The other encoding methods work on a single amino acid of the sequence at a time, using the information from that particular residue and the seven neighbour residues on each side. So the modules associated with "binary" and "pssm" encoding methods are incorporated in second-level filter that is used to predict calcium binding regions.

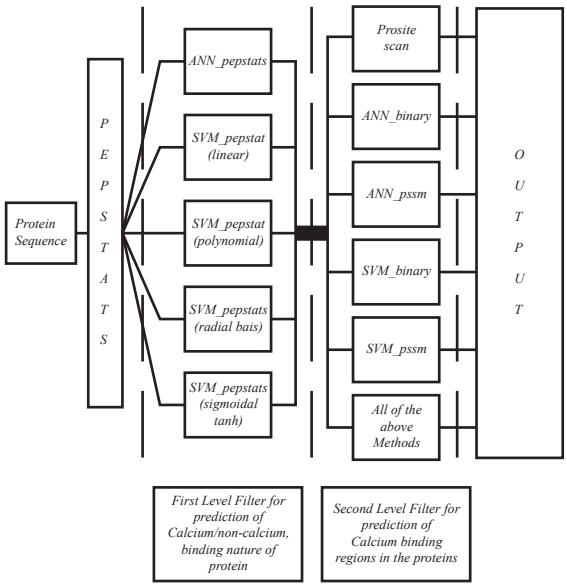

Figure 1. Overall framework of the CalPred tool.

Workflow

Each protein sequence that is being queried with the CalPred tool goes through the first-level filter of the tool that is used for prediction of calcium/non-calcium binding nature of protein. Here the user can use any of the pepstats related classifier or a combination of all. If the protein passes the first-level filter, it then goes through second-level filter where it is processed using PSSM or binary-encoding related classifiers, the output of which is shown to end-user along with the regions predicted to be calcium-binding regions. In the second phase of prediction, user is again given the choice to select the classifier or to use the combination of all the classifiers.

Datasets

The initial dataset consisted of 306 CaBPs and 358 non-CaBP sequences. The CaBPs were obtained from EF-hand calcium-binding proteins data library available at http://structbio.vanderbilt.edu/cabp_database/ and the non-CaBPs were obtained from Entrez protein database. The protein datasets used in the training-testing cycles of the nine modules were hecked for sequence similarity to remove

redundancy. The initial two datasets were filtered so that no protein sequence in the final datasets had more than 90% sequence similarity with sequence in that dataset. This was done using CD-HIT program (Li and Godzik, 2006) that clustered the sequences of a dataset in different clusters such that each cluster had sequences with more than 90% sequence similarity. Representative sequence from each cluster was taken to form the final datasets. After removing redundancy, the final datasets consisted of 188 CaBPs and 214 non-CaBP sequences. The sequence identifiers for these datasets are available in the supplementary material.

Five-fold cross validation

A newly developed statistical procedure must be checked for its validity. We have used five-fold cross validation technique to check the validity of all the nine modules that have been developed. For this purpose a dataset partitioning method was used to create the five sub-datasets. This partitioning method is similar to the previously used methods (Bendtsen et al., 2004). Here five sub datasets of sequences were created by randomly assigning a sequence to a sub-dataset such that each sub-dataset had approximately equal number of CaBPs and non-CaBPs and all five sub-datasets had approximately equal number of sequences.

Each of the nine methods is trained and tested five times where, in each instance of training - testing cycle; four sub-datasets are used for training and the remaining one for testing purpose. The performance measures given have been averaged over the five testing sub-datasets.

Artificial Neural Networks

All the three neural network modules were implemented using the Stuttgart neural network simulator (Zell and Mamier, 1997). A feed-forward neural network with standard back-propagation algorithm is utilized in all cases but the architecture differs as their function differs.

• In $ANN_{pepstats}$ module the neural network used had an architecture as 51-4-1 i.e. it had 51 nodes in input layer representing the values of physicochemical properties from the pepstats encoding method, 4 nodes in hidden layer and 1 node in output layer showing whether a given protein is CaBP or non-CaBP. The cut-off value used for prediction in this module is 0.9, i.e. a query protein is regarded as belonging to CaBP family if its score is greater than or equal to 0.9.

• The ANN_{binary} and ANN_{pssm} modules were incorporated in the second level filter of the CalPred tool; that functions to predict calcium binding regions in the given protein. The calcium-binding domain entry given in PROSITE database had a calcium-binding motif of thirteen residues length. Using this information; the window size in ANN_{binary} and ANN_{pssm} modules was fixed at thirteen amino acids. In these modules, thirteen residues of a protein sequence are taken at a time; and the prediction is done for the middle residue i.e. seventh residue. Both the binary as well as pssm encoding methods, encodes the thirteen residues into 260 numeric values and thus input layer of these modules consisted of 260 nodes. The hidden layer in both modules had 20 nodes. The output layer of these modules consisted of a single node predicting whether the Ca+2 ions will bind to the seventh residue or not. Thus, the ANN_{binary} and ANN_{pssm} modules share the same architecture of 260-20-1. The cut-off value used for prediction in both of these modules is 0.5, i.e. the seventh residue of a window is regarded as a Ca+2 ion binding residue if its score is greater than or equal to 0.5.

Support Vector machines

The support vector machines used in the SVM related modules first tried to map the input vector into high dimensional feature space, either linearly or by methods depending on kernel type chosen; such that error is minimized over the training dataset. Then an optimized division is sought between the positive and negative classes say “a CaBP and non-CaBP protein” or “a Ca+2 ion binding residue and non-Ca+2 ion binding residue”. This was done by constructing a hyperplane that separated these two classes by the largest margin (Vapnik, 1998). Here we have used SVMlight software (Joachims, 1998) for implementing the support vector machine related modules. The SVMlight software allows users to choose from a

number of available modes and kernel functions. In all the modules, a default classification mode is used while kernel functions are varied. The kernel functions available are linear, polynomial, radial basis and sigmoidal. In all the support vector machines the cut-off value used for prediction is 0, i.e. a query vector is regarded as member of positive dataset if its score is greater than 0 and is regarded as member of negative dataset if its score is less than 0. The ones having scored equal to zero are regarded as undefined.

Performance Measures

The performance measures used to evaluate the nine modules are listed below. These measures have been calculated for each module using five test sub-datasets; and the final measures (Table 1) given have been averaged over these five sub datasets.

• *Accuracy:* The accuracy of the modules have been calculated as:

Where P and N refer to true positives and true negatives say correctly predicted calcium-binding

$$Q_{ACC} = \frac{P+N}{P+N+O+U}$$

and non calcium-binding proteins respectively; and O and U refer to false positives and false negatives i.e. incorrectly predicted calcium-binding and non calcium-binding proteins.

• *Specificity (Q_{spec})* and sensitivity *(Q_{sens})* of the modules are defined as:

• The Matthews correlation coefficient (MCC) is defined as:

$$Q_{spec} = \frac{N}{N+O} \qquad Q_{sens} = \frac{P}{P+U}$$

• Q_{Pred} *(Probability of correct prediction)* is defined as:

$$MCC = \frac{(P \times N) - (O \times U)}{\sqrt{(P+U) \times (P+O) \times (N+U) \times (N+O)}}$$

Results and Discussion

Performance over test dataset

$$Q_{pred} = \frac{P}{P+O} \times 100$$

The performance of the modules over the test dataset is given in Table 1. The accuracy of first-level filter modules on the test dataset varies from 57% to 94%. Support vector machine's polynomial kernel type performed the best on the test dataset amongst all the classifiers / modules in first-level filter. The other measures of SVMpepstats_polynomial module are also better than the modules in the first-level filter, with sensitivity of 0.917 and specificity of 1.00.

The performance measures for modules in second-level filter, used for predicting calcium-binding regions are also calculated. There is no perfect method to calculate the specific calcium-binding regions in a protein, as one can observe from the pattern matching technique; using PROSITE database's calcium-binding domain entry which also gives proteins that are calcium-binding but are not picked up by the pattern. Therefore the performance measures of test dataset, for the second-level filter are calculated assuming; each residue belonging to a calcium-binding protein which is predicted as a calcium-binding site by the second-level filter is a true positive sample and a residue belonging to a non calcium-binding protein which is predicted as a non calcium-binding site by the second-level filter is a true negative sample. Similarly, false positives and false negatives were calculated. Though this assumption does not make second-level filter an efficient filter, but due to lack of a standard method of finding calcium-binding regions in the protein; we made this assumption for the purpose of calculating the performance measures in the same manner as they were calculated in first-level filter. Due to this assumption, SVM related modules of the second-level filter show high performance measures on the test datasets (Table 1).

Table 1. Summary of the prediction results of the nine modules used in first and second layers on the test dataset.

Name of the module	Accuracy	Specificity	Sensitivity	Probability of correct prediction (QPred)	Matthews correlation coefficient (MCC)
$ANN_{pepstats}$	0.8815	0.8536	0.9176	79.64	0.7579
ANN_{binary}	0.7131	0.7527	0.5907	43.64	0.3141
ANN_{pssm}	0.8364	0.8790	0.7538	75.59	0.6327
$SVM_{pepstats\_linear}$	0.9406	0.9109	1.00	87.58	0.8927
$SVM_{pepstats\_polynomial}$	0.9453	0.9170	1.00	88.51	0.9006
$SVM_{pepstats\_radial\_basis}$	0.7075	0.5535	0.7451	89.91	0.4487
$SVM_{pepstats\_sigmoidal\_tanh}$	0.5761	0.6508	0.9116	22.64	0.1763
SVM_{binary}	0.9101	0.9239	0.8797	83.89	0.7935
SVM_{pssm}	0.9994	0.9993	0.9994	99.87	0.9986

Performance over human proteome

The human proteome was analyzed to predict the calcium-binding proteins in them. The proteome was taken from Entrez protein database and analyzed using first-level filter modules. All modules having accuracy greater than 65% were used in this analysis and the detailed results of proteome; as data files and tabulated statistics are given in the supplementary material. A protein is regarded as a calcium-binding protein if any of its score from first-level modules is above the cut-off score of that module. Performance measures are not calculated based on the data from whole human proteome, due to the large difference in number of sequences present in positive and negative dataset.

The Human proteome: A case study

In the human proteome we analyzed 34,122 non-calcium binding proteins and 58 calcium-binding proteins. The plots for distribution of output scores of $ANN_{pepstats}$, $SVM_{pepstats\_linear}$, $SVM_{pepstats\_polynomial}$ and $SVM_{pepstats\_radial\_basis}$ module for human proteome are given as Figures 2-5. It is clear from the plot (Figure 2) that $ANN_{pepstats}$ module separates the two datasets i.e. calcium-binding proteins and non calcium-binding proteins to a large extend. Mostly the calcium-binding proteins score above 0.9 value and the non calcium-binding

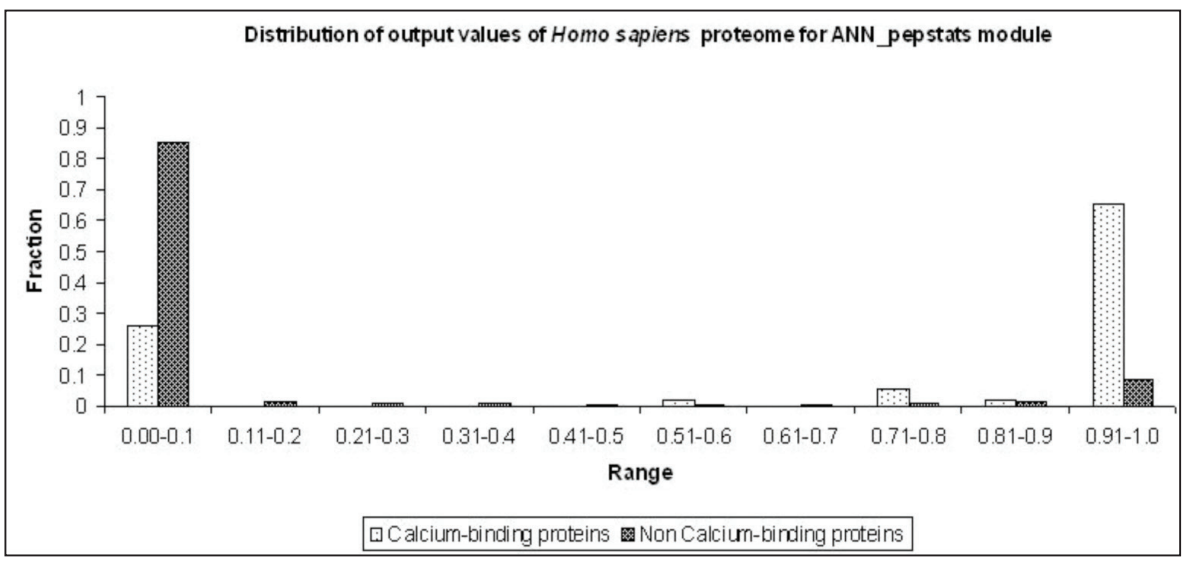

Figure 2. Distribution of output values of *Homo Sapiens* proteome for $ANN_{pepstats}$ module.

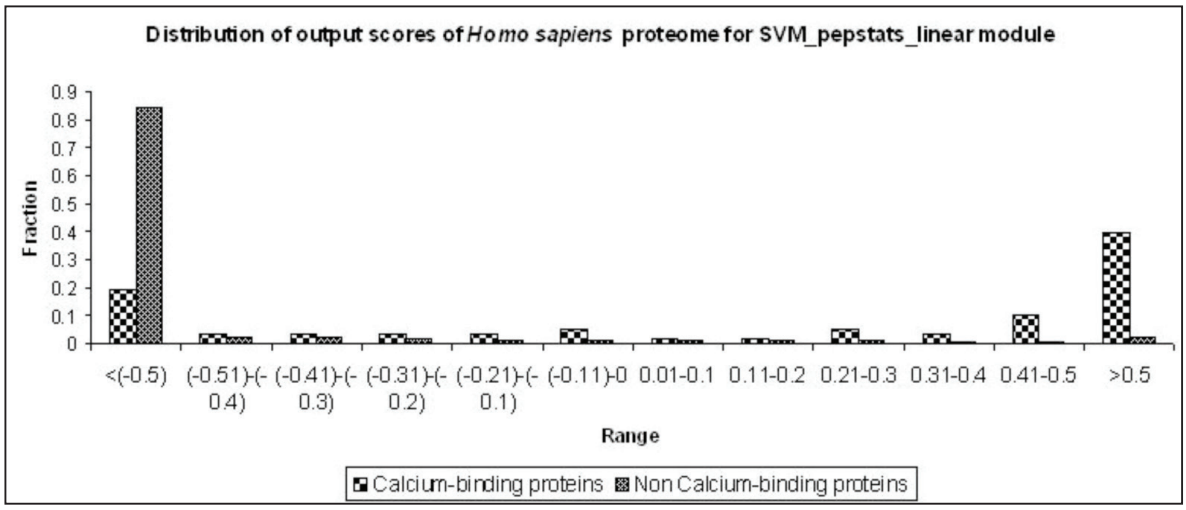

Figure 3. Distribution of output values of Homo Sapiens proteome for $SVM_{pepstats\_linear}$ module.

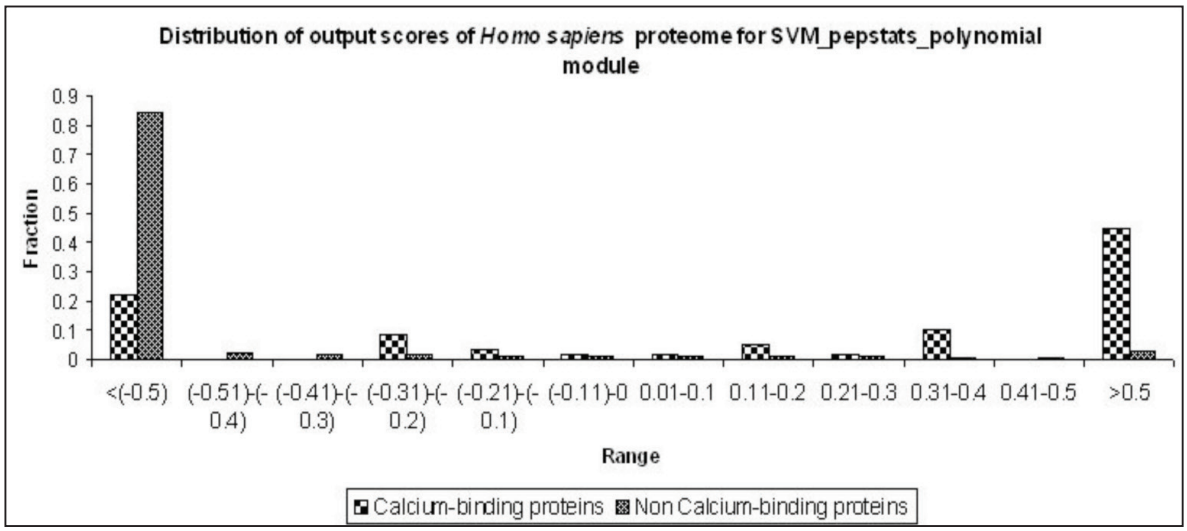

Figure 4. Distribution of output values of Homo Sapiens proteome for $SVM_{pepstats\_polynomial}$ module.

proteins score below 0.1.

The $SVM_{pepstats\_linear}$ and $SVM_{pepstats\_polynomial}$ modules having zero threshold values, for classifying the proteins also identified the non calcium-binding proteins effectively (Figure 3 and 4) but were found to be lesser efficient in screening calcium-binding proteins as the scores for the later were scattered over the whole range of -1 to 1, though majority of the scores were present above the threshold cut-off value. The $SVM_{pepstats\_radial\_basis}$ module was not able to distinguish the proteins efficiently in the proteome (Figure 5); this was expected as the module showed lesser accuracy on test dataset too.

Performance over PROSITE dataset

The calcium-binding proteins that were not picked up by the pattern cited as calcium-binding domain entry in PROSITE database with ID PS00018 were analyzed with the first- and second-level filters. There were 85 such proteins and these are listed in the entry itself. The first-level filter was able to identify 45 of the calcium-binding proteins. The scores of these entries for first-level filter are given as data file in supplementary material. Here too we have used, only those first-level filter modules that showed greater than 65% accuracy on test dataset. The plots for distribution of scores of these entries

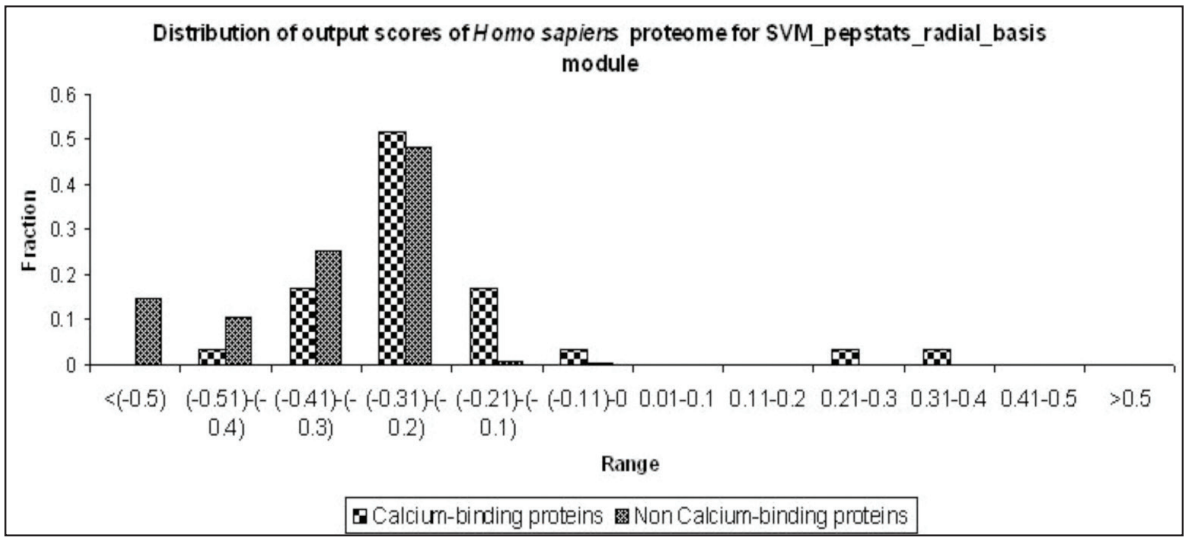

Figure 5. Distribution of output values of Homo Sapiens proteome for $SVM_{pepstats\_radial\_basis}$ module.

with first-level filter for different modules are given in supplementary material. The detailed results for second-level filter are given in the supplementary material.

Conclusions

Analyzing calcium-binding proteins from sequence data is important in calcium-mediated biological studies. Existing methods use protein structural data (Deng et al., 2005; Wei et al., 1999; Sodhi et al., 2004; Liang et al., 2003) to predict protein binding sites. The approach is vital in cases where the structural data for proteins is unavailable. Intelligent systems like the ones used in this study, utilize global information of protein sequence data instead of using simple pattern matching techniques. These can significantly increase the accuracy of calcium-binding proteins related studies. Applying this to human proteome gives us an insight that how well such systems are suitable for proteome-wide studies where proteins are not categorized but are sequenced. As these classifiers, are clearly able to separate the positive and the negative datasets therefore; such systems can also be used for high throughput experiments in novel proteomes. This study gives an insight into development of similar tools and techniques for other metal-binding proteins or for all metal-binding proteins in general.

Future Work

We have presented a simple framework based on sequence data for classification of calcium binding protein and region identification. Future work on the topic should involve usage of optimized support vector machine models of the framework to get the real accuracy values of related classifiers. The combination of structural data within the proposed framework i.e. by adding a third filter layer could also lead to better results.

Acknowledgements

We are thankful to Nelson et al, authors of EF-hand calcium-binding protein data library for making the data library publicly available and bioinformatics.org for providing us free web space for creation of the mirror for static pages of CalPred tool.

References

Altschul S.F., Madden T.L., Alejandro A.S., Zhang J., Zheng Z., Miller W., and Lipman D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.,* 25: 3389-3402, 1997.

Bendtsen J.D., Jensen L.J., Blom N., Von H.G. and Brunak S. Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng. Des Sel,* 17: 349-356, 2004.

Carafoli E. Calcium signaling: A tale for all seasons. *Proc Natl Acad Sci,* 99(3): 1115-1122, 2002.

Ding C.H.Q. and Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics,* 17(4): 349-358, 2001.

Deng H., Liu H. and Zhang Y. Mining Calcium binding Sites from Protein Structure Graphs. ICNN&B '05 *International Conference on Neural Networks and Brain,* 1980-1985, 2005.

Hennings H., Michael D., Cheng C., Steinert P., Holbrook K. and Yuspa S.H. Calcium regulation of growth and differentiation of mouse epidermal cells in culture. *Cell,* 19: 245-254, 1980.

Hulo N., Bairoch A., Bulliard V., Cerutti L., De Castro,E., Langendijk-Genevaux P.S., Pagni M., Sigrist C.J.A. The PROSITE database. *Nucleic Acids Res.,* 34: D227-D230, 2006.

Hua S. and Sun Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics,* 17(8): 721-728, 2001.

Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proceedings of the European Conference on Machine Learning,* 1998.

Kesmir C., Nussbaum A.K., Schild H., Detours V., and Brunak S. Prediction of proteasome cleavage motifs by neural networks. *Protein Engineering.* 15(4): 287-296, 2002.

Li W. and Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* 22: 1658-1659, 2006.

Liang M.P., Banatao D.R., Klein T.E., Brutlag D.L. and Altman R.B. WebFEATURE: an interactive web tool for identifying and visualizing functional sites on macromolecular structures. *Nucleic Acids Res.,* 34: 3324-3327, 2003.

Orrenius S., Zhivotovsky B. and Nicotera P. Regulation of cell death: The calcium-apoptosis link. *Nat Rev Mol Cell Biol.,* 4(7): 552-565, 2003.

Rice P., Longden I. and Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.,* 16(6):276-277, 2000.

Rigden D.J., Jedrzejas M.J. and Galperin M.Y. An extracellular calcium-binding domain in bacteria with a distant relationship to EF-hands. *FEMS Microbiol Lett.,* 221(1): 103-110, 2003.

Rigden D.J., Jedrzejas M.J. and Moroz O.V. Structural diversity of calcium-binding proteins in bacteria: single-handed EF-hands? *Trends Microbiol.,* 11(7): 295-297, 2003.

Rost B. PHD: predicting 1D protein structure by profile based neural networks. *Meth. in Enzym.,* 266: 525-539, 1996.

Sodhi J.S., Bryson K., McGuffin L.J., Ward J.J., Wernisch L. and Jones D.T. Predicting Metal-binding Site Residues in Low-resolution Structural Models. *J. Mol. Biol.,*342: 307-20, 2004.

Vapnik V.N. *Statistical learning theory.* Wiley-Interscience, New York, 1998.

Wei L., Huang E.S. and Altm R.B. Are predicted structures good enough to preserve functional sites? *Structure,* 7: 643-650, 1999.

Zell A. and Mamier G. *Stuttgart neural network simulator Version 4.2.* University of Stuttgart, Stuttgart, Germany, 1997.