

Manuscript version: Author's Accepted Manuscript

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:

<http://wrap.warwick.ac.uk/129168>

How to cite:

Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.

CaMELS: *In silico* Prediction of Calmodulin Binding Proteins and their Binding Sites

Wajid Arshad Abbasi¹, Amina Asif Shah¹, Saiqa Andleeb² and Fayyaz ul Amir Afsar Minhas^{1,*}

¹Biomedical Informatics Research Laboratory, Department of Computer and Information Sciences (DCIS), Pakistan Institute of Engineering and Applied Sciences (PIEAS), Nilore, Islamabad, Pakistan.

²Biotechnology Laboratory, Department of Zoology, University of AJ&K, Muzaffarabad, AK, Pakistan.

*Corresponding author email: fayyazafsar@gmail.com or afsar@pieas.edu.pk

Short Title: Calmodulin Interaction Learning System

Keywords: Calmodulin interaction prediction; Calmodulin binding site prediction; Multiple Instance Learning; CaMELS; protein-protein interaction.

ABSTRACT

Due to Ca²⁺ dependent binding and the sequence diversity of Calmodulin (CaM) binding proteins, identifying CaM interactions and binding sites in the wet-lab is tedious and costly. Therefore, computational methods for this purpose are crucial to the design of such wet-lab experiments. We present an algorithm suite called CaMELS (CalModulin intEraction Learning System) for predicting proteins that interact with CaM as well as their binding sites using sequence information alone. For CaM interaction prediction, CaMELS uses protein features coupled with a large-margin classifier and gives significantly improved prediction accuracy in comparison to existing techniques. CaMELS can not only identify whether a protein binds CaM or not, it can also predict CaM-binding residues in those proteins. It models the binding site prediction problem using multiple instance machine learning with a novel optimization algorithm. In comparison to conventional classification techniques, our proposed stochastic sub-gradient solver for multiple instance learning allows more effective training with a data set containing imprecisely annotated CaM-binding sites. We benchmarked the performance of CaMELS using a non-redundant set of binding proteins and binding sites in the CaM target database as well as the *A. thaliana* proteome. As a case study, we have used CaMELS for predicting the binding sites of Adenylyl cyclase domain from *B. pertussis* and have found our sequence-only prediction to be in close agreement with the known structure of the protein complex. Our interaction prediction results for the *A. Thaliana* proteome also show a high degree of overlap of gene ontology enrichment with known CaM targets. Python code for training and evaluating CaMELS together with a webserver implementation is available at the URL:

<http://faculty.pieas.edu.pk/fayyaz/software.html#camels>.

INTRODUCTION

Calmodulin (CaM) is a 149 amino acid long multifunctional calcium (Ca^{2+}) binding protein that is highly conserved across all eukaryotes.¹ CaM mediates many vital processes like immune response, muscle contraction, metabolism, nerve growth, and intracellular movement.² CaM is able to do all this by binding various targets in the cell including a large number of enzymes, ion channels and other proteins.^{3,4} Many CaM binding proteins are mostly unable to bind Ca^{2+} directly and therefore use CaM as a signal transducer and calcium sensor.^{5,6} Due to the involvement of CaM in different important biological processes, identification of proteins that bind CaM and the location of CaM binding sites within a protein can help biologists in elucidating underlying biological processes at the molecular level. Due to Ca^{2+} -dependent binding and the large sequence diversity of its targets, identifying CaM interactions and binding sites in the wet lab is very costly and time consuming.⁷ Therefore, there is an utmost need for computational techniques to support wet-lab experiments by predicting CaM binding proteins and their binding sites. This work presents a highly accurate *in-silico* CaM binding site and interaction prediction method that relies only on protein sequences.

A number of algorithms have been proposed for CaM interaction and binding site prediction in the literature.⁸⁻¹³ DeGrado *et al.* suggested an algorithm that finds amphiphilic α helix in a peptide sequence for CaM binding site prediction.¹³ Mruk *et al.* proposed a method called calmodulation meta-analysis for CaM binding site prediction by scoring the existence of canonical motifs in a given protein sequence.¹² Both the methods proposed by DeGrado *et al.* and Mruk *et al.* were designed using a limited dataset and cannot predict whether a protein will interact with CaM or not. Radivojac *et al.* and Hamilton *et al.* used a sliding-window classification approach for CaM binding site prediction based on the contiguous nature of CaM binding sites.⁸ These methods use

a conventional Support Vector Machine (SVM) classifier and do not explicitly handle imprecisions in binding site annotations in the training data. Annotations of CaM binding sites in proteins available in the literature typically span more residues than the minimal set of contiguous residues responsible for the interaction.¹⁰ Such imprecisions result from limitations of experimental procedures and time or cost considerations in identifying individual binding residues. Furthermore, all annotated binding site residues may not contribute equally to the binding energy. To counter such uncertainties, Minhas and Ben-Hur formulated this problem as a Multiple Instance learning (MIL) problem.¹⁰ Their approach, called MI-1, was designed primarily for CaM binding site prediction and offers very good accuracy for this task. However, the accuracy of MI-1 for CaM interaction prediction is very low. This is because MI-1 simply uses the predicted score of the most likely binding window in a protein as its CaM interaction propensity. However, a putative CaM-binding sequence in a protein will result in an interaction only if the three dimensional structure of the protein allows for it.¹³ Furthermore, MI-1 uses a heuristics approach to solve the MIL problem which may not converge to its optimal solution.

In this paper, we present CaMELS (CaModulin intEraction Learning System) for machine learning based CaM interaction and binding site prediction. CaMELS models interaction and binding site prediction as two different classification problems. For CaM interaction prediction, we use protein level features instead of window level features used in previous studies.^{9, 10} This led to a large improvement in the accuracy of interaction prediction. The biological significance of these results was verified through a Gene Ontology (GO) enrichment analysis of the *Arabidopsis thaliana* proteome.¹⁴ For CaM binding site prediction, we developed a stochastic sub-gradient optimization method for solving the MIL problem. This led to a substantial improvement in

binding site prediction accuracy. An analysis of the trained machine learning model for CaMELS revealed a significant correspondence between the model and known CaM binding site motifs.¹⁵

METHODS

Dataset and Preprocessing

Binding Site Dataset

For CaM binding site prediction, our dataset and its pre-processing follows our previous work.¹⁰ A set of 157 CaM binding proteins was taken from the CaM target database.¹⁵ Each of these proteins has one or more annotated binding sites and a total of 191 binding sites were identified in these proteins. These proteins were selected in such a way that no two proteins have more than 40% sequence identity in overall or in regions annotated as binding sites.

Interaction Data Set

For CaM interaction prediction, we used a set of 241 known CaM binding proteins from *Arabidopsis thaliana* as the positive set.¹⁶ We used CD-HIT¹⁷ to obtain a non-redundant set of 12, 217 proteins from the *Arabidopsis thaliana* proteome which is used as the negative set. Keeping the sequence diversity of known CaM binding proteins into account, the proteins in the negative set share less than 30% sequence similarity with the proteins in the positive set and less than 40% among themselves.

Classifiers

Binding Site Prediction

In CaM binding site prediction, the objective is to find the region of a protein that is involved in its binding with CaM. For this purpose, we adopt a sliding window approach in which each protein sequence is divided into overlapping windows of length 21. We represent the sequence of a

window starting at residue i in the protein by x_i and denote its associated label by $y_i \in \{+1, -1\}$ indicating whether x_i belongs to an annotated binding site (+1) or not (-1). This problem can be posed as a classification problem through a discriminant function $f(x) = \mathbf{w}^T \boldsymbol{\phi}(x)$, where $\boldsymbol{\phi}(x)$ represents the feature vector of window x and \mathbf{w} is the weight vector that needs to be learned. Residues involved in the binding of a protein with CaM can be identified based on the values of the discriminant function for the window centered at these residues.

We have solved this classification problem using a conventional support vector machine (SVM)¹⁸ as well as a multiple instance learning (MIL) framework.¹⁹ We use the conventional SVM as a baseline for our results by taking the annotated binding site windows in a protein as positive class examples and the remaining ones as negatives.^{9, 10}

- **Multiple Instance learning (MIL)**

As discussed in the introduction section, the annotated CaM binding sites in the binding dataset are imprecise due to limitations in experimental procedures and include residues that may not be involved in binding. A classical supervised classification approach such as an SVM cannot be used effectively with such ambiguously labeled training examples.¹⁹ To cope with these challenges we formulated the binding site prediction problem as a MIL problem.¹⁰

MIL is a generalization of supervised learning where labels are available for bags or sets of examples and not for individual examples.^{19, 20} A bag is labeled -1 if it is known that all instances in it are negative. A bag carries a label of $+1$ if it contains at least one positive instance. The remaining instances in a positive bag can be negative. The objective of MIL is to learn a discriminant function that discriminates between positive and negative examples using bag level labels only. The problem of binding site prediction with our data can be mapped to MIL by defining

a positive and a negative bag for every protein. The positive bag contains all annotated binding site windows in that protein whereas the negative bag contains all the remaining widows. In case of a protein with multiple annotated binding sites, all the windows belonging to those binding sites constitute a single positive bag.

A number of methods for solving the MIL problem exist in the literature.^{10, 20, 21} Of particular interest in this domain are large margin MIL solutions because they can handle high dimensional feature spaces and non-linear classification boundaries.^{10, 20} The formulations use a discriminant function of the form $f(x) = \mathbf{w}^T \phi(x)$ and are inspired from SVM style large margin classification. Intuitively, the differences between these approaches lie in the way they enforce classification and multiple instance learning constraints. In the mi-SVM technique proposed by Andrews *et al.*, the discriminant function is such that at least one example from every positive bag receives a score larger than +1 whereas all negative examples have scores less than -1. In contrast, in MI-1 by Minhas and Ben-Hur, an example in a positive bag must rank higher than all negative examples from the same protein with some margin.^{10, 20} In this work, we have improved the MI-1 formulation further using a stochastic sub-gradient solver. Henceforth, we give the mathematical formulation for MI-1 SVM. For this purpose, we denote the positive bag of all windows in the annotated binding sites in a protein p by B_p . The set of non-binding site windows in the protein is represented by N_p . The large margin MIL formulation for the binding site prediction can be written as:

$$\min_{\mathbf{w}, \xi \geq 0} \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} + \sum_p \xi_p \quad (1)$$

such that for all proteins p in the training set:

$$\max_{i \in B_p} f(x_i) \geq \max_{j \in N_p} f(x_j) + 1 - \xi_p.$$

Here, $\lambda > 0$ is the regularization parameter which controls the trade-off between constraint violation and margin maximization and ξ_p is the extent of margin violation.

The MI-1 formulation given above is a more concise and direct model of the binding site prediction problem in comparison to a conventional classifier such as an SVM or other existing MIL solutions. MI-1 also leads to faster training in comparison to previous models. This is because the number of slack variables (ξ_p) in MI-1 SVM is equal to the number of proteins and not the number of training examples as in previous formulations.

- **Stochastic sub-gradient optimization (SSGO) for MIL**

Unlike a conventional SVM, the MI-1 formulation is combinatorial in nature due to the \max function in its constraints. As a consequence, a specialized optimization method is needed for its solution. Similar to other large margin MIL classifiers such as miSVM, Minhas and Ben-Hur solved the optimization problem in Eq. 1 using a heuristic approach based on iterative retraining of a conventional SVM.¹⁰ Due to its heuristic nature, this approach may not lead to an optimal solution. To overcome these issues, we have developed a stochastic sub-gradient algorithm for Multiple Instance Learning inspired from the Pegasos solver for conventional binary SVMs by Shalev-Shwartz *et al.*²² Henceforth, we describe the proposed stochastic sub-gradient optimization algorithm for MI-1.

Based on the principle of structured risk minimization, we represent the constrained optimization problem in Eq. 1 as an unconstrained one as follows:^{22, 23}

$$\min_{\mathbf{w}} \frac{\lambda}{2} \mathbf{w}^T \mathbf{w} + \sum_p l(\mathbf{w}; p) \quad (2)$$

here, $l(\mathbf{w}, p)$ is the hinge loss function and can be written as:

$$l(\mathbf{w}; p) = \max \left\{ 0, 1 - \left(\max_{i \in B_p} f(x_i) - \max_{j \in N_p} f(x_j) \right) \right\}. \quad (3)$$

The stochastic sub-gradient solver for this problem operates iteratively by choosing a protein p randomly in each iteration t and estimates the sub-gradient of the objective function given in Eq. 2 based only on the chosen protein. This sub-gradient can be written as:

$$\Delta_t = \lambda \mathbf{w}^T - \mathbb{I}(f(x_{i^*}) - f(x_{j^*}) < 1) (\boldsymbol{\phi}(x_{i^*}) - \boldsymbol{\phi}(x_{j^*})) \quad (4)$$

Here,

$$\begin{aligned} i^* &= \operatorname{argmax}_{i \in B_p} f(x_i) \\ j^* &= \operatorname{argmax}_{j \in N_p} f(x_j) \end{aligned}$$

and $\mathbb{I}(\cdot)$ is the indicator function such that $\mathbb{I}(\cdot) = 1$ if its argument is true and 0 otherwise. The weight vector is updated in a direction opposite to the direction of the sub-gradient by $\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \mu_t \Delta_t$ using a step size of $\mu_t = \frac{1}{\lambda t}$. The complete optimization algorithm is given in Table 1.

Interaction Prediction

In CaM interaction prediction, the objective is to predict whether a given protein interacts with CaM or not. For a protein p in the interaction dataset, we indicate its associated feature representation by $\psi(p)$. All proteins in the interaction dataset have binary labels indicating whether they interact with CaM (+1) or not (-1). This problem can be posed as a classification

problem through a discriminant function $z(p)$. For CaM interaction prediction, we used the following two strategies:

- **Discriminant function scoring (DFS)**

The trained classifier for CaM binding site prediction can be used for CaM interaction prediction. This can be done by using the score of the most likely binding site in the protein as the interaction propensity of that protein. Mathematically, the CaM interaction score for a protein p is given by $z(p) = \max_{x \in p} \mathbf{w}^T \boldsymbol{\phi}(x)$. This approach was used in previous studies to predict CaM interactions of proteins in the *A. thaliana* proteome and is used as a baseline.^{9, 10} The fundamental assumption behind this approach is that the presence of a binding site within a protein is predictive of its interaction with CaM.

- **SVM with protein level features**

In this work, we hypothesize that the whole sequence of the protein carries information that can be useful for binding prediction. As a consequence, we extracted protein level features, $\psi(p)$, of a protein and used a standard binary SVM for classification.¹⁸ We used SVM with the Gaussian kernel, $k(p', p) = \exp(-\gamma \|\psi(p') - \psi(p)\|^2)$, here, the parameter γ controls the spread of the Gaussian Kernel.²⁴ We used the Scikit-learn package for implementation of the SVM.²⁵

Feature Extraction

We have a number of protein and window level features representation denoted by $\psi(\cdot)$ and $\boldsymbol{\phi}(\cdot)$, respectively. All the features are normalized to unit norm.

Window Level Features

For window level features of protein sequences, we sweep a window of length 21 across the entire length of the protein and extract features from individual windows.

- Amino Acid Composition Features (AAC)

The amino acid composition of a given sequence is a 20 dimensional vector containing the counts of occurrences of individual amino acids in it. AAC is used both at the window and protein levels for binding site and interaction prediction, respectively.

- Position Dependent 1-Spectrum Features (PD-1)

This feature representation captures the position and type information of different amino acids in a window. It is a 420 dimensional vector $\phi(x)$ such that its component $\phi^{a,k}(x)$ is set to one if amino acid a occurs at position k in the window and zero otherwise.

- Position Dependent BLOSUM-62 Features (PD-Blosum)

In order to model the substitutions of physio-chemically similar amino acids in proteins sequence, we expressed each protein sequence using the BLOSUM-62 substitution matrix.²⁶ The 420-dimensional PD-Blosum feature vector for a sequence window of 21 residues is obtained by stacking the columns of the BLOSUM-62 matrix corresponding to each residue in the window.

- Propy Features (propy)

To capture biophysical properties of amino acids and sequence-derived structural features, we used a feature extraction package called propy.²⁷ Propy gives a 1,537 dimensional feature representation of a window. This representation includes different features such as pseudo-amino acid compositions (PseAAC), autocorrelation descriptors, sequence-order-coupling number, quasi-sequence-order descriptors, amino acid composition, transition and the distribution of

various structural and physicochemical properties.^{28, 29} This feature representation is also used at the protein level for CaM interaction prediction. Each of these features is standardized to have zero mean and unit standard deviation across all examples.

Protein Level Features

Here, we describe the features extracted using whole protein sequences.

- Local Alignment Features (SWIPE)

To capture sequence similarities of a given protein to known CaM binders, we performed local sequence alignment of a protein with the 157 proteins in our binding site dataset. This leads to a 157 dimensional feature vector of alignment scores. Proteins in the binding site dataset share less than 40% sequence identity with CaM binders in the interaction dataset. We used SWIPE³⁰ with the BLOSUM-62 substitution matrix to perform fast smith-waterman local alignments.³¹ We used gap insertion and extension penalties as 10 and 0.5, respectively.

- Averaged BLOSUM-62 Features (Blosum)

This 20 dimensional feature representation is built by averaging the columns of the BLOSUM-62 matrix²⁶ corresponding to all amino acids occurring in the given protein sequence.

Performance Evaluation

Interaction Prediction

For CaM interaction prediction, we have used 10-fold stratified cross validation.³² The fold-wise average of the following metrics has been used to quantify the performance of different models.

- Area Under ROC Curve (AUC-ROC)

The ROC curve is obtained by plotting true positive rate (TPR) against false positive rate (FPR) at different thresholds on the output of the classifier.³³ The area under the ROC curve (AUC-ROC), expressed as a percentage, is reported.

- Area Under 10% ROC Curve (AUC-ROC_{0.1})

AUC-ROC_{0.1} is obtained by plotting true positive rate (TPR) in ROC curves up to the first 10% false positives. This measure gives us a sense of how many true positives are produced at low false positive rates.

- Area Under Precision-recall Curve (AUC-PR)

The precision-recall (PR) curve is obtained by plotting precision against recall at different thresholds for the discriminant function values. The area under the PR curve is a useful metric in classification problems involving imbalanced data such as ours.³³

- Tversky Index for GO Enrichment Analysis (GOTI)

We used the Gorilla tool for Gene Ontology (GO) term enrichment analysis of known CaM binders and top scoring *A. Thaliana* proteins from different interaction prediction methods.³⁴ To quantify the degree of correspondence of GO terms between predicted and known CaM binding proteins, we used the Tversky Index.³⁵ Tversky Index for GO Enrichment Analysis (GOTI) is computed as follows:

$$GOTI = \frac{|M \cap N|}{|M \cap N| + |M - N| + |N - M|}$$

Here, M and N are the sets of GO enrichment terms for known CaM binders in the interaction dataset and top 240 predictions, respectively. The proteome of *A. thaliana* was used as the background set except for known CaM binders and their close homologs.

Binding Site Prediction

For CaM binding site prediction, we have used Leave One Protein Out (LOPO) cross-validation. In this protocol, the classifier is tested on all residue-level windows of a protein after training it on the data from all other proteins. This process is repeated for all the proteins in the binding site dataset. In addition to AUC-ROC, AUC-ROC_{0.1}, and AUC-PR, we report the following biologist-centered performance metrics as well.³⁶

- True Hit Rate (THR)

The true hit rate is the percentage of proteins in the binding set in which the top scoring residue predicted by a classifier lies in an annotated binding site. An ideal classifier would always have THR=100%.

- False Hit Rate (FHR)

This metric represents the percentage of non-binding site residues that score higher than the highest scoring residue in the annotated binding site of a protein. An ideal classifier would always have FHR=0%.

- Median Rank of the First Positive Prediction (MRFPP)

To get an intuition for the distribution of false negatives in comparison to the top scoring true positive, we used MRFPP. This metric is the median rank of the first true positive prediction across all proteins. An ideal predictor should have MRFPP =1, i.e., for at least 50% proteins, the top

scoring prediction by the predictor is a true positive.^{36, 37} In comparison to AUC-PR, this measure is more intuitive to biologists as it reveals directly how often the top scoring predictions can be expected to be a binding site.

Model Selection

We used grid search over training data to find the optimal values of hyper-parameters of different classifiers. For SSGO based MIL, the values of λ was selected from the set {0.1,0.01,0.001} with 1000 training epochs using AUC-PR as the metric for selection in LOPO cross-validation. For the SVM in interaction prediction, the values of $C \in \{0.01,0.1,1,10,100,1000,10000\}$ and $\gamma \in \{0.1,0.25,0.5,2,4,8,16\}$ are used in the grid search. AUC-PR is employed as the performance metric in 10 fold cross-validation.

RESULTS AND DISCUSSION

In this section, we present and discuss the results and major outcomes of our study.

Interaction Prediction

For CaM interaction prediction, we adopted two different approaches: discriminant function scoring based on MIL and SVM with protein level features. The results of these two approaches across different features are shown in Table 2 and Fig. 1a. DFS gives a maximum AUC-PR of 6.0% with an AUC-ROC of 74.0%. These results do not show any improvement in predictive performance in comparison to the MI-1 method.¹⁰ On the other hand, the protein level SVM based technique proposed in this work provides a big improvement in prediction performance. The maximum AUC-PR and AUC-ROC obtained with this classification scheme are 55% and 86.7%, respectively (Table 2; Fig. 1a). A marked increase is also noted in AUC_{0.1} from 26.0% to 65.1%.

The violin plot in Fig. 2 shows the densities of the scores obtained from the DFS and SVM based methods for examples from both classes. It can be easily noticed that the degree of overlap between distributions of scores of positive and negative class examples is significantly larger for DFS in comparison to CaMELs. These results show that the features of the whole protein improve the performance of CaM interaction prediction in comparison to features at the window level. This can potentially be explained by the fact that the mere presence of a CaM binding site in a protein is not predictive of its interaction with CaM.^{13, 38, 39}

- **Analysis of features**

In CaMELs, we have used different protein level feature representations for CaM interaction prediction such as propy, SWIPE, AAC and Blosum. We obtained the best performance using propy features in comparison to other feature representations (Table 2; Fig. 1a). We expect this to be a consequence of incorporating k-mer features and different correlation factors in a protein chain in the propy feature representation.²⁷ We tested this hypothesis by taking different combinations of propy features. With 20-dimensional amino acid composition and 400-dimensional dipeptide frequency features we obtained an AUC-PR of 47.0% whereas the 720-dimensional sequence correlation features alone produced an AUC-PR of 52.3%. Please note that using all the 1,537 propy features results in an AUC-PR of 55.0%. This shows that the auto-correlation features of physicochemical properties are responsible for the improvement in prediction accuracy.

- **Performance comparison**

We have also compared the interaction prediction performance of CaMELs with the previous state of the art method MI-1¹⁰ and a general purpose protein interaction predictor called iLoops.⁴⁰ We compared the performance of these techniques using a reduced data set of 5000 randomly sampled proteins not in the positive set. This reduction was done due to the limitation of the iLoops server.

The PR curves for this comparison are shown in Fig. 1b. CaMELS gives an AUC-PR of 58.3% whereas, the AUC-PR obtained through MI-1 and iLoops are 14.8% and 8.3%, respectively.

- **Biological significance**

We verified the biological relevance of our results by performing gene ontology enrichment analysis of the top 240 predictions from CaMELS from the proteome of *A. Thaliana*. Table 3 shows the results of this analysis. We observed significant overlap between the GO terms for molecular function and biological process ontologies between known and predicted CaM binding proteins. The Tversky Indices (GOTI) for these analyses are 68% and 34% for the biological process and molecular function ontologies, respectively. GO term enrichment analysis of the top predictions from DFS reveals no overlap between the enriched GO terms of the predictions with known CaM binders (Table 3). The enriched terms include phosphorylation, signal transduction, signaling, kinase activity, etc. and correlate with the known functions of CaM binders. We provide the ranked list of the top predictions from CaMELS used in this analysis in the online supplementary material.

Binding Site Prediction

Table 4 and Fig. 3 show the results of CaMELS for binding site prediction. Table 4 also shows the best results of SVM, mi-SVM and MI-1 from our previous study using the same evaluation protocol.¹⁰ CaMELS gives an AUC-PR of 87.0% with AUC-ROC of 99.2% (Table 4; Fig. 3a). The results show a large improvement in the performance of CaMELS in comparison to MI-1 with AUC-ROC of 96.9%. (Table 4). A notable increase is also seen in AUC_{0.1} from 59.0% to 79.0% (Table 4; Fig. 3b). True and False hit rates of 77% and 1.0% also show the high prediction accuracy of CaMELS. The median rank of the first positive prediction (MRFPP) is 1.0 for all features used

with CaMELS. The difference in the prediction accuracy between the conventional SVM and multiple instance learning based methods (mi-SVM, MI-1 and CaMELS) shows the effectiveness of modeling CaM binding site prediction problem through multiple instance learning. Furthermore, the improved inaccuracy of CaMELS with respect to other MIL based techniques (MI-1 and mi-SVM) is a consequence of solving the MIL optimization problem through the proposed stochastic sub-gradient optimization (SSGO).

- **Analysis of features**

CaMELS used different window level feature representations for CaM binding site prediction such as PD-Blosum, PD-1, AAC, a combination of PD-1 with AAC (AAC+PD-1) and propy. The PD-Blosum feature representation gives the best results in comparison to other features (Table 4, Fig. 3a & b). These improvements are due to the use of feature representation which models the position-specific substitution behavior of different amino acids within the protein.

Fig. 4 shows the weight vectors obtained from training CaMELS with the AAC and PD-1 feature representations. These weight vectors show the importance of individual amino acids in determining CaM binding sites within a protein sequence. The weight vector of AAC feature representation, shown in Fig. 4a, depicts large positive weights for positively charged amino acids Arginine (R), Lysine (K) and the hydrophobic amino acid Tryptophan (W). Amino acids such as Aspartic acid (D), Glutamic acid (E), Proline (P) and Tyrosine (Y) have large negative weights. These amino acid propensities in CaM binding sites are in close agreement with previous studies and also with known CaM binding motifs.^{9, 10, 15, 41}

The weight vector of position dependent feature representation, shown in Fig. 4b, illustrates the role of different amino acids in binding site prediction with respect to their positions in a given window. For example, Lysine (K) shows large positive weights at the end of the window but small

in the middle; Arginine (R) shows large positive weights at positions 8 and 18; Tryptophan (W) has large positive weights at middle and negative weights at the corner of the window; Aspartic acid (D), Glutamic acid (E), Proline (P) show their negative role in CaM binding with large negative weights in the middle. This position dependent learning behavior of the classifier is in close agreement with known CaM binding motifs and can be used to extract more biologically relevant motifs.¹⁵

- **Biological significance**

We have also verified the accuracy of binding site prediction of CaMELS by predicting the binding site of the Adenylyl cyclase domain from *Bordetella pertussis*. The crystal structure of this toxin in complex with CaM is available in the Protein Data Bank as 1YRT⁴² and is not part of our training set. Fig. 5 shows this structure along with the predicted CaM binding sites from CaMELS. The predicted binding site overlaps significantly with the residues of the Adenylyl cyclase that occur within 5Å of CaM in the complex structure. The highest scoring region coincides with a Tryptophan residue at position 242 in the protein which is known to stabilize this complex.⁴²

Webserver for CaMELS

We have developed and deployed a webserver of CaMELS. This webserver takes a query protein sequence in plain or fasta format and performs CaM interaction and binding site prediction for it. The user interface of the CaMELS webserver is shown in Fig. 6. After the successful submission of a protein sequence, the users will be redirected to a page showing CaMELS predicted scores for CaM interaction and binding site predictions. For CaM interaction prediction, the predicted score shows the interaction propensity of the submitted protein with CaM. Similarly, for CaM binding site prediction, residue level scores of all windows are shown. A plot of residue level scores of all

windows for binding site prediction with the location of predicted binding site is also shown on this page. The webserver is available at the following URL.

<http://faculty.pieas.edu.pk/fayyaz/software.html#camels>.

CONCLUSIONS

We have presented a set of models for CaM interaction and binding site prediction called CaMELS. CaMELS uses protein sequence information only and offers state of the art accuracy both for interaction and binding site prediction. For interaction prediction, CaMELS achieved significant improvement in performance using protein level features in comparison to earlier methods that used information derived only from the most likely CaM binding site in a protein. This shows that sequence information of the whole protein is predictive of its interaction with CaM. We have also presented a multiple instance learning model for solving the binding site prediction problem. Our results show near perfect classification accuracy for this problem with the use of a stochastic gradient solver. The proposed suite of algorithms is expected to be very helpful to biologists working on analyzing the functions and interaction behavior of CaM and its target proteins.

ACKNOWLEDGEMENTS

The authors are thankful to Mr. Naveed Akhtar, High Performance Computing (HPC) lab, PIEAS, Pakistan for technical support.

FUNDING

Wajid A. Abbasi is supported by a grant under indigenous 5000 Ph.D. fellowship scheme from the Higher Education Commission (HEC) of Pakistan.

Conflict of Interest: none declared.

REFERENCES

1. Bouché N, Yellin A, Snedden WA, Fromm H. Plant-specific calmodulin-binding proteins. *Annu Rev Plant Biol* 2005;56:435–466.
2. Chin D, Means AR. Calmodulin: a prototypical calcium sensor. *Trends in Cell Biology* 2000;10(8):322–328.
3. Yamniuk AP, Vogel HJ. Calmodulin's flexibility allows for promiscuity in its interactions with target proteins and peptides. *Mol Biotechnol* 2004;27(1):33–57.
4. Reichow SL, Clemens DM, Freites JA, Németh-Cahalan KL, Heyden M, Tobias DJ, Hall JE, Gonen T. Allosteric mechanism of water-channel gating by Ca²⁺-calmodulin. *Nat Struct Mol Biol* 2013;20(9):1085–1092.
5. Vogel HJ. Calmodulin: a versatile calcium mediator protein. *Biochem Cell Biol* 1994;72(9-10):357–376.
6. Möller W, Brown DM, Kreyling WG, Stone V. Ultrafine particles cause cytoskeletal dysfunctions in macrophages: role of intracellular calcium. Part Fibre Toxicol 2005;2:7.
7. Reddy ASN, Ben-Hur A, Day IS. Experimental and computational approaches for the study of calmodulin interactions. *Phytochemistry* 2011;72(10):1007–1019.
8. Radivojac P, Vucetic S, O'Connor TR, Uversky VN, Obradovic Z, Dunker AK. Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition. *Proteins* 2006;63(2):398–410.
9. Hamilton M, Reddy ASN, Ben-Hur A. Kernel Methods for Calmodulin Binding and Binding Site Prediction. In: Proceedings of the 2Nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine; 2011; New York, NY, USA. ACM. p 381–386.

10. Minhas F ul AA, Ben-Hur A. Multiple instance learning of Calmodulin binding sites. *Bioinformatics* 2012;28(18):i416–i422.
11. Wang L, Liu Z-P, Zhang X-S, Chen L. Prediction of hot spots in protein interfaces using a random forest model with hybrid features. *Protein Eng Des Sel* 2012;25(3):119–126.
12. Mruk K, Farley BM, Ritacco AW, Kobertz WR. Calmodulation meta-analysis: Predicting calmodulin binding via canonical motif clustering. *J Gen Physiol* 2014;144(1):105–114.
13. Degrado WF, Erickson-Viitanen S, Wolfe HR, O’Neil KT. Predicted calmodulin-binding sequence in the γ subunit of phosphorylase b kinase. *Proteins* 1987;2(1):20–33.
14. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene Ontology: tool for the unification of biology. *Nat Genet* 2000;25(1):25–29.
15. Yap KL, Kim J, Truong K, Sherman M, Yuan T, Ikura M. Calmodulin Target Database. *J Struct Func Genom* 2000;1(1):8–14.
16. Popescu SC, Popescu GV, Bachan S, Zhang Z, Seay M, Gerstein M, Snyder M, Dinesh-Kumar SP. Differential binding of calmodulin-related proteins to their targets revealed through high-density *Arabidopsis* protein microarrays. *Proc Natl Acad Sci U S A* 2007;104(11):4730–4735.
17. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinform* 2010;26(5):680–682.
18. Cortes C, Vapnik V. Support-Vector Networks. *Machine Learning* 1995;20(3):273–297.
19. Dietterich T, Lathrop RH. Solving the Multiple-Instance Problem with Axis-Parallel Rectangles. *Artif Int* 1997;89:31–71.

20. Andrews S, Tschantaridis I, Hofmann T. Support vector machines for multiple-instance learning. In: ; 2003; . MIT Press. p 561–568.
21. Leistner C, Saffari A, Bischof H. MIForests: Multiple-Instance Learning with Randomized Trees. In: Daniilidis K, Maragos P, Paragios N, editors. Computer Vision – ECCV 2010. , Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2010. p 29–42.
22. Shalev-Shwartz S, Singer Y, Srebro N, Cotter A. Pegasos: primal estimated sub-gradient solver for SVM. *Math Program* 2011;127(1):3–30.
23. Vapnik VN. An overview of statistical learning theory. *IEEE Trans Neural Netw* 1999;10(5):988–999.
24. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. Support Vector Machines and Kernels for Computational Biology. *PLoS Comput Biol* 2008;4(10):e1000173.
25. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 2011;12:2825–2830.
26. Eddy SR. Where did the BLOSUM62 alignment score matrix come from? *Nat Biotech* 2004;22(8):1035–1036.
27. Cao D-S, Xu Q-S, Liang Y-Z. propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics* 2013;29(7):960–962.
28. Li ZR, Lin HH, Han LY, Jiang L, Chen X, Chen YZ. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucl Acids Res* 2006;34(suppl 2):W32–W37.

29. Limongelli I, Marini S, Bellazzi R. PaPI: pseudo amino acid composition to score human protein-coding variants. *BMC Bioinform* 2015;16:123.
30. Rognes T. Faster Smith-Waterman database searches with inter-sequence SIMD parallelisation. *BMC Bioinform* 2011;12(1):221.
31. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;147(1):195–197.
32. Alpaydin E. Design and analysis of machine learning experiments, in *Introduction to Machine Learning*. Cambridge, Massachusetts: MIT Press; 2010.
33. Davis J, Goadrich M. The Relationship Between Precision-Recall and ROC Curves. In: *Proceedings of the 23rd International Conference on Machine Learning*; 2006; New York, NY, USA. ACM. p 233–240.
34. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinform* 2009;10:48.
35. Tversky A. Features of similarity. *Psychol Rev* 1977;84(4):327–352.
36. Abbasi WA, Minhas FUAA. Issues in performance evaluation for host–pathogen protein interaction prediction. *J Bioinform Comput Biol* 2016;14(03):1650011.
37. Minhas F, Geiss BJ, Ben-Hur A. PAIRpred: Partner-specific prediction of interacting residues from sequence and structure. *Protein Struct Funct Bioinform* 2014;82(7):1142–1155.
38. Kessel A, Ben-Tal N. *Introduction to Proteins: Structure, Function, and Motion*. Boca Raton, FL, USA: CRC Press; 2010.
39. Petsko GA, Ringe D. *Protein Structure and Function*. Sinauer Associates, USA: New Science Press; 2004.

40. Planas-Iglesias J, Marin-Lopez MA, Bonet J, Garcia-Garcia J, Oliva B. iLoops: a protein–protein interaction prediction server based on structural features. *Bioinformatics* 2013;29(18):2360–2362.
41. Rhoads AR, Friedberg F. Sequence motifs for calmodulin recognition. *FASEB J* 1997;11(5):331–340.
42. Guo Q, Shen Y, Lee Y-S, Gibbs CS, Mrksich M, Tang W-J. Structural basis for the interaction of *Bordetella pertussis* adenylyl cyclase toxin with calmodulin. *EMBO J* 2005;24(18):3190–3201.

Figure Legends

Figure 1. (a) Precision-recall curves for interaction prediction across all models; (b) Precision-recall curves for the comparison of CaMELS with MI-1 and iLoops. The area under the curve is shown in parenthesis as mean across different folds.

Figure 2. Violin plot showing the predictive performance of DFS and CaMELS. Density distributions of CaM Interacting (+1) and non-interacting (-1) proteins with respect to DFS and CaMELS scores are shown. Dotted lines show the first, second and third quartiles of these densities.

Figure 3. (a) Precision-recall curves for binding site prediction across all models; (b) ROC_{0.1} curves for binding site prediction across all models. The area under the curve is shown in parenthesis as mean across

Figure 4. Weight vectors for AAC and PD-1. (a) Weights of different amino acids in the (position-independent) AAC feature representation; (b) Heat map of the weights of different amino acids against their position from position-dependent 1-spectrum (PD-1) feature representation.

Figure 5. The 3D Structure (PDB ID: 1YRT) of Adenylyl cyclase (colored in brown) in complex with CaM (colored in cyan). The predicted binding site from CaMELS is shown in purple. Residues of Adenylyl cyclase within 5Å of CaM are shown in stick form.

Figure 6. The user interface of the webserver for CaMELS. (a) The user can submit fasta file or plain sequence of a protein of interest for CaM interaction and binding site prediction; (b) CaMELS prediction scores for CaM interaction and binding site shown on a redirected page.

Table 1. MIL algorithm with SSGO training for CaM binding site prediction

Inputs: λ, T

Initialize: set $w_0 = 0$

For $t = 1, 2, \dots, T$

Select a protein p uniformly at random

$$i^* = \operatorname{argmax}_{i \in B_p} w_t^T \phi(x_i)$$

$$j^* = \operatorname{argmax}_{j \in N_p} w_t^T \phi(x_j)$$

$$\text{Set } \mu_t = \frac{1}{\lambda t}$$

$$\text{If } w_t^T \phi(x_{i^*}) - w_t^T \phi(x_{j^*}) < 1:$$

$$\text{Set } w_{t+1} \leftarrow \left(1 - \frac{1}{t}\right) w_t + \mu_t (\phi(x_{i^*}) - \phi(x_{j^*}))$$

else:

$$\text{Set } w_{t+1} \leftarrow \left(1 - \frac{1}{t}\right) w_t$$

Output: $w = w_{T+1}$

Table 2. Interaction prediction results for all models.

<i>Method</i>	<i>Features</i>	<i>AUC-ROC</i>	<i>AUC-ROC_{0.1}</i>	<i>AUC-PR</i>
CamELs	propy	86.7	65.1	55.0
	SWIPE	78.3	51.9	40.2
	AAC	74.7	40.4	26.8
	Blosum	78.4	32.6	11.4
	PD-Blosum	68.0	18.0	6.0
DFS	Blosum	74.0	26.0	4.0
	AAC	72.0	24.0	4.0
	PD-1	69.0	16.6	3.0
	AAC+PD-1	71.0	17.0	2.6

Table 3. GO term enrichment analysis results

Method	GOTI		
	Process	Function	Component
CaMELS	0.34	0.68	0.40
DFS	0.01	0.0	0.0

Table 4. Binding site prediction results for all models. AUC-PR and MRFPP were not available for MI-1, mi-SVM and SVM.

<i>Method</i>	<i>Features</i>	<i>AUC-ROC</i>	<i>AUC-ROC_{0.1}</i>	<i>AUC-PR</i>	<i>THR</i>	<i>FHR</i>
CaMELS	PD-Blosum	99.2	79.0	87.0	77	1.0
	AAC+PD-1	98.9	77.6	85.6	75	1.0
	PD-1	98.4	76.2	84.1	72	2.0
	propy	98.0	74.7	81.2	68	2.0
	AAC	97.9	72.3	80.7	68	2.0
MI-1	AAC+PD-1	96.9	59.0	--	75	1.2
mi-SVM	AAC+PD-1	96.2	55.6	--	68	1.9
SVM	AAC+PD-1	95.9	55.1	--	65	2.1

Figure 1. (a)

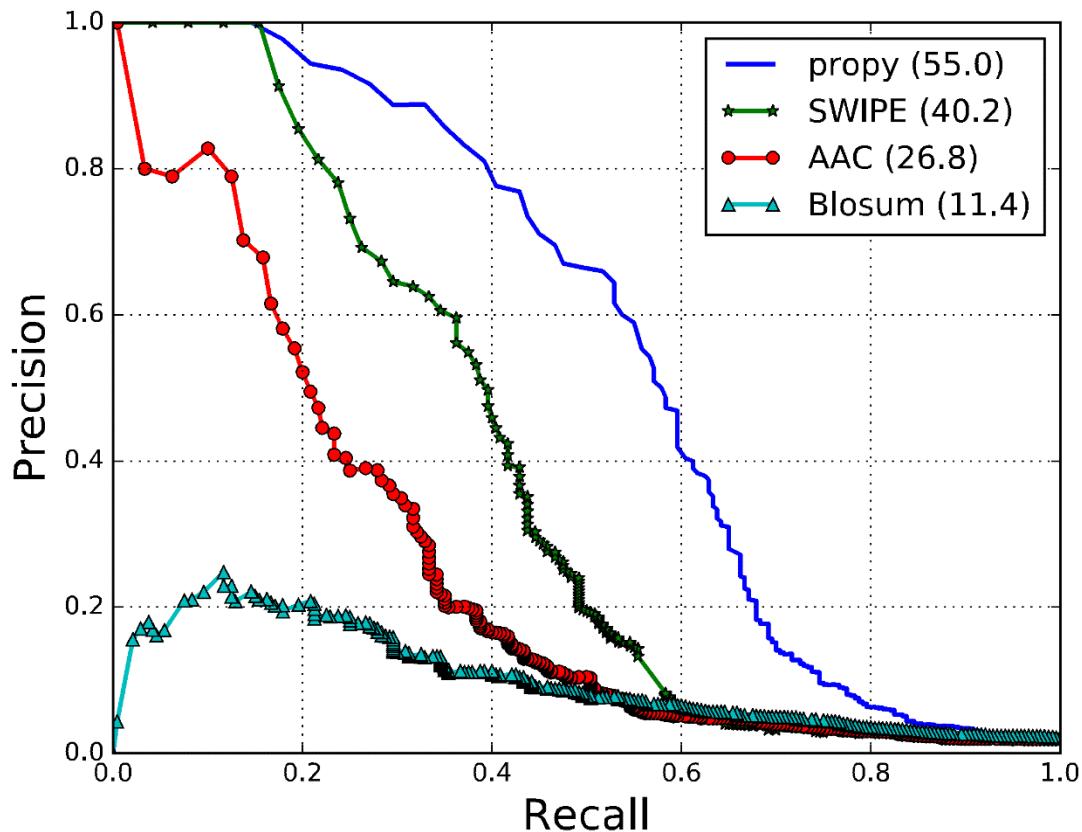


Figure 1. (b)

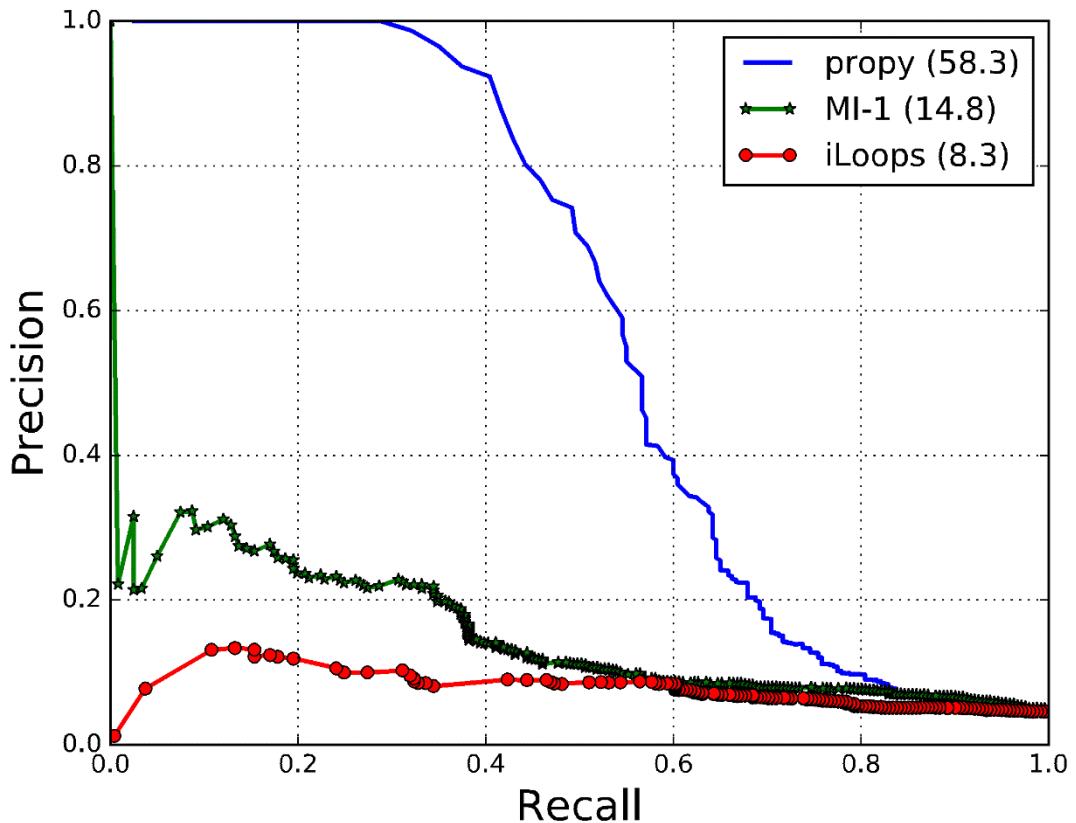


Figure 2

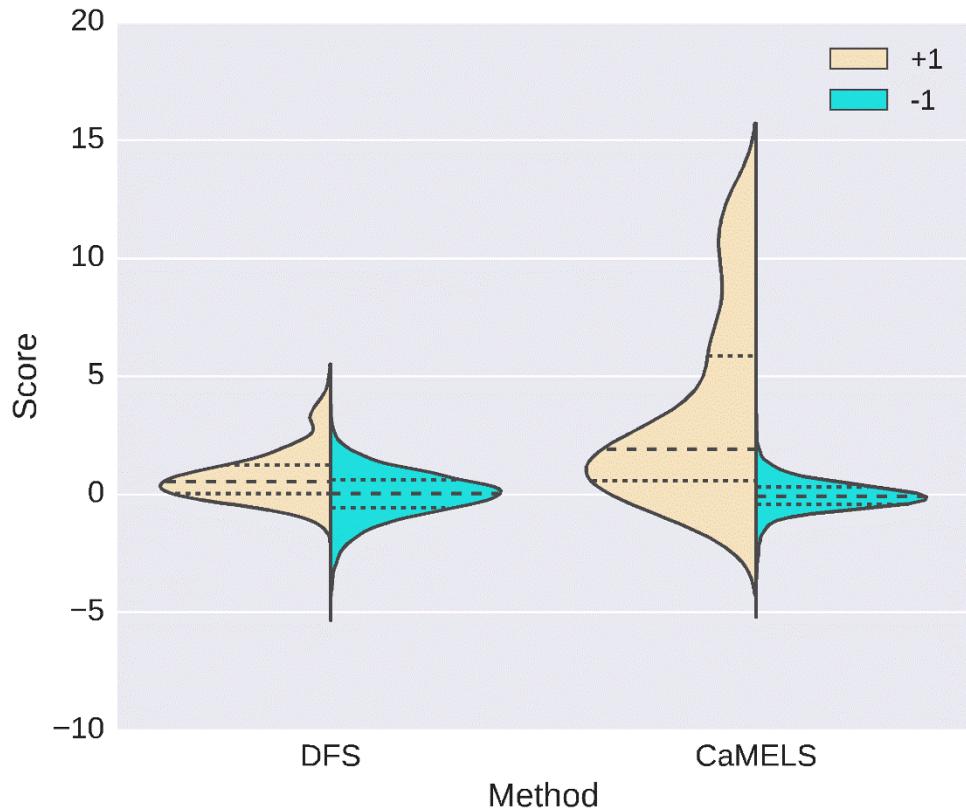


Figure 3. (a)

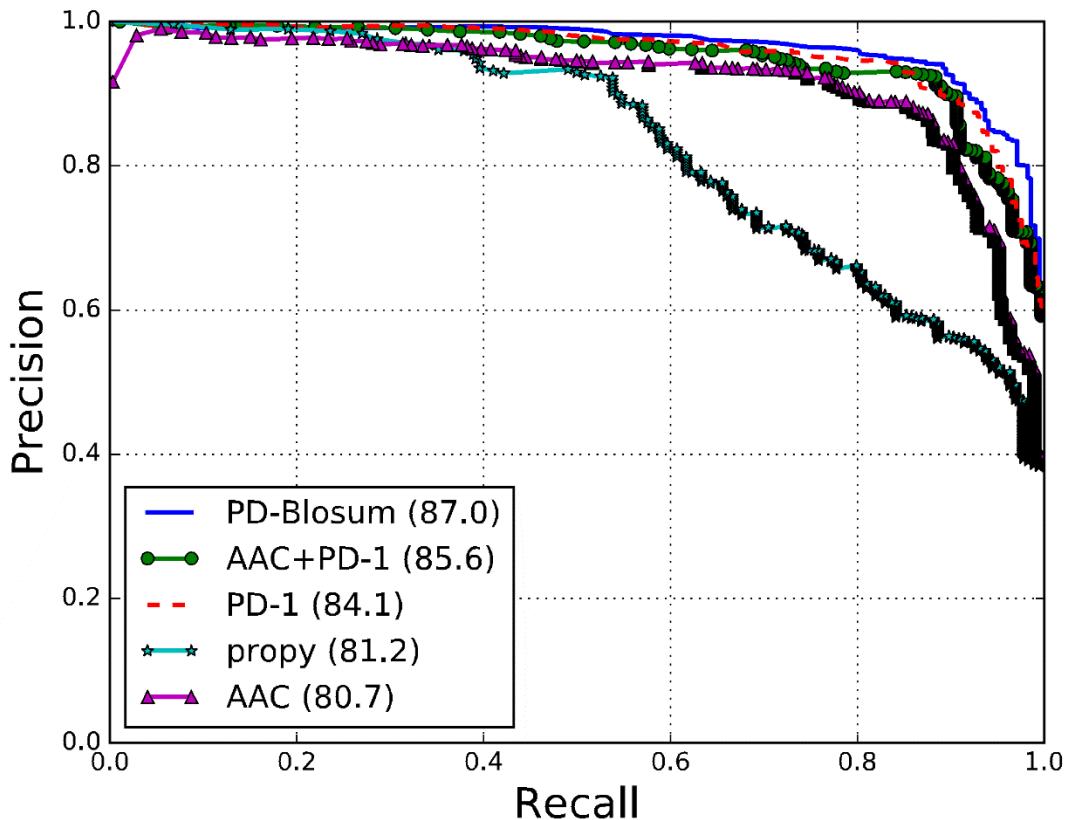


Figure 3. (b)

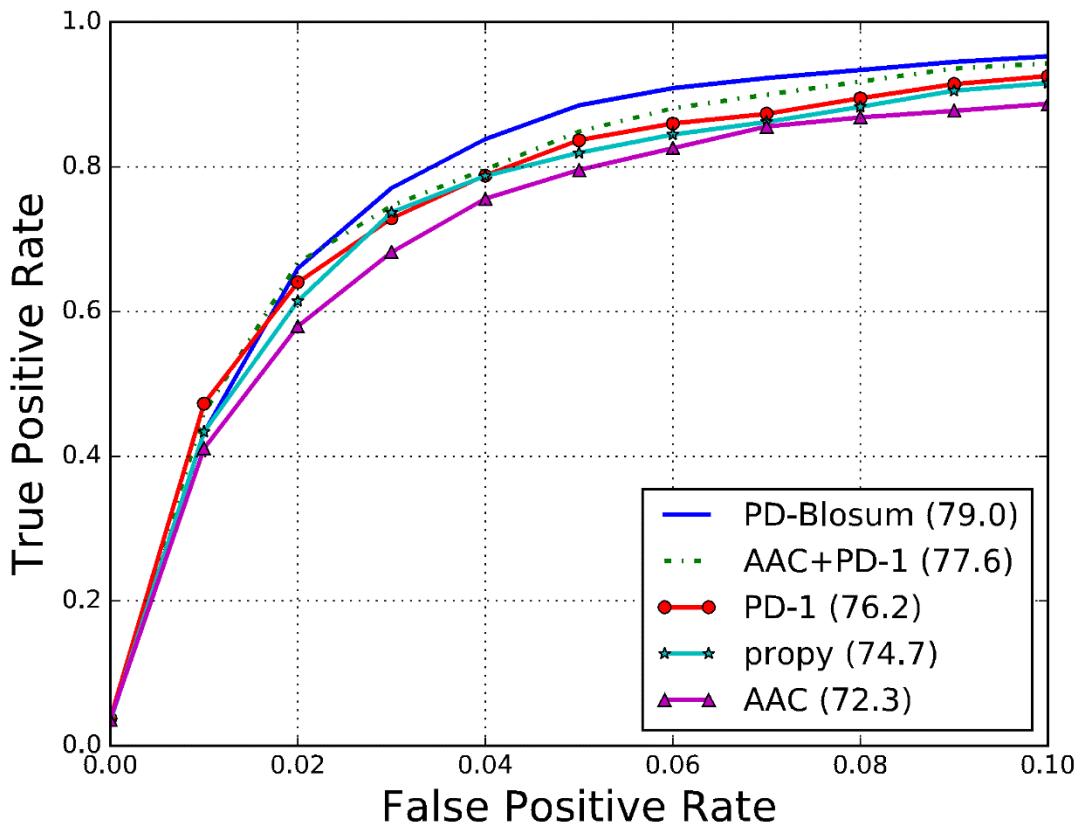


Figure 4. (a)

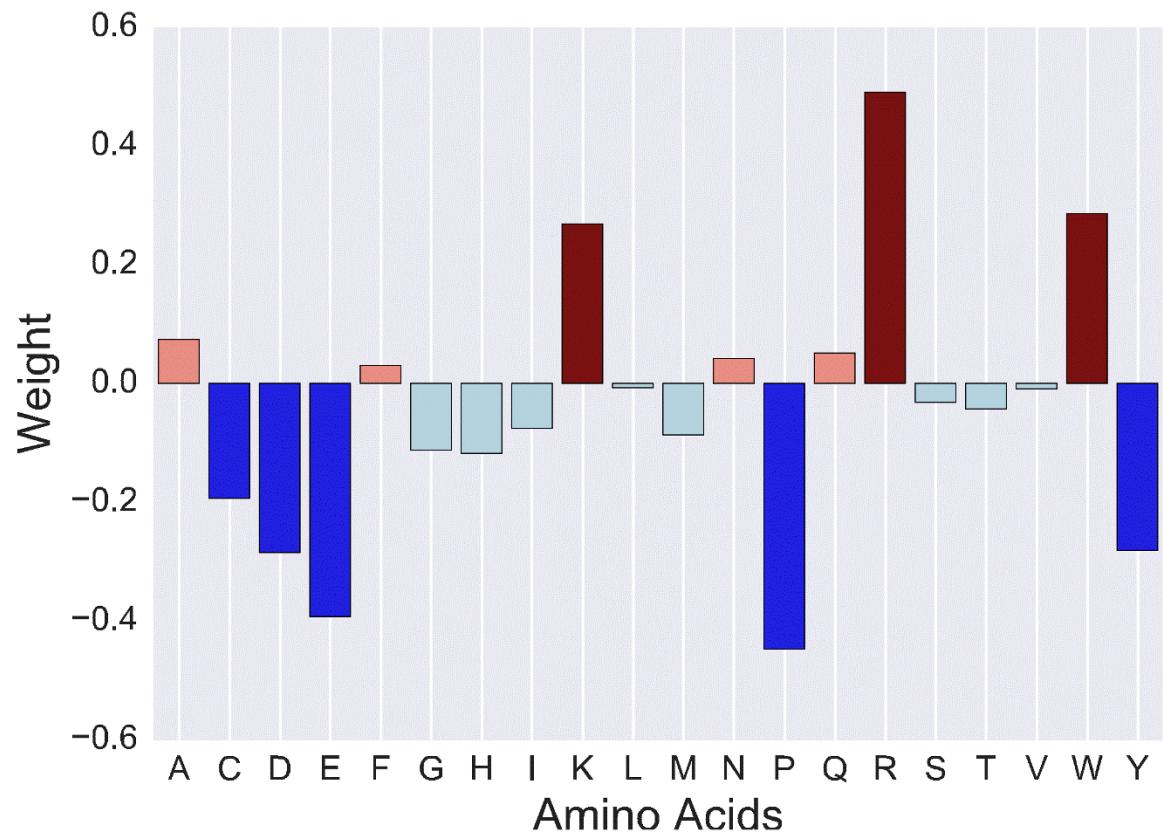


Figure 4. (b)

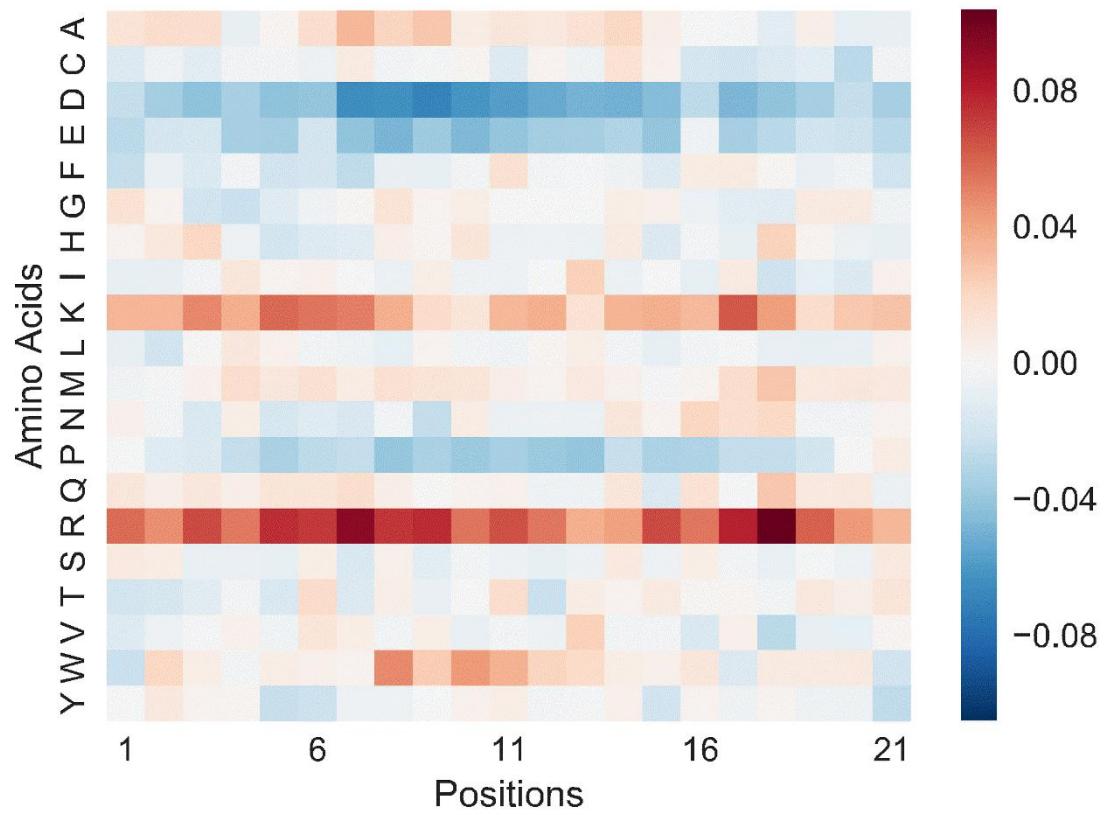


Figure 5

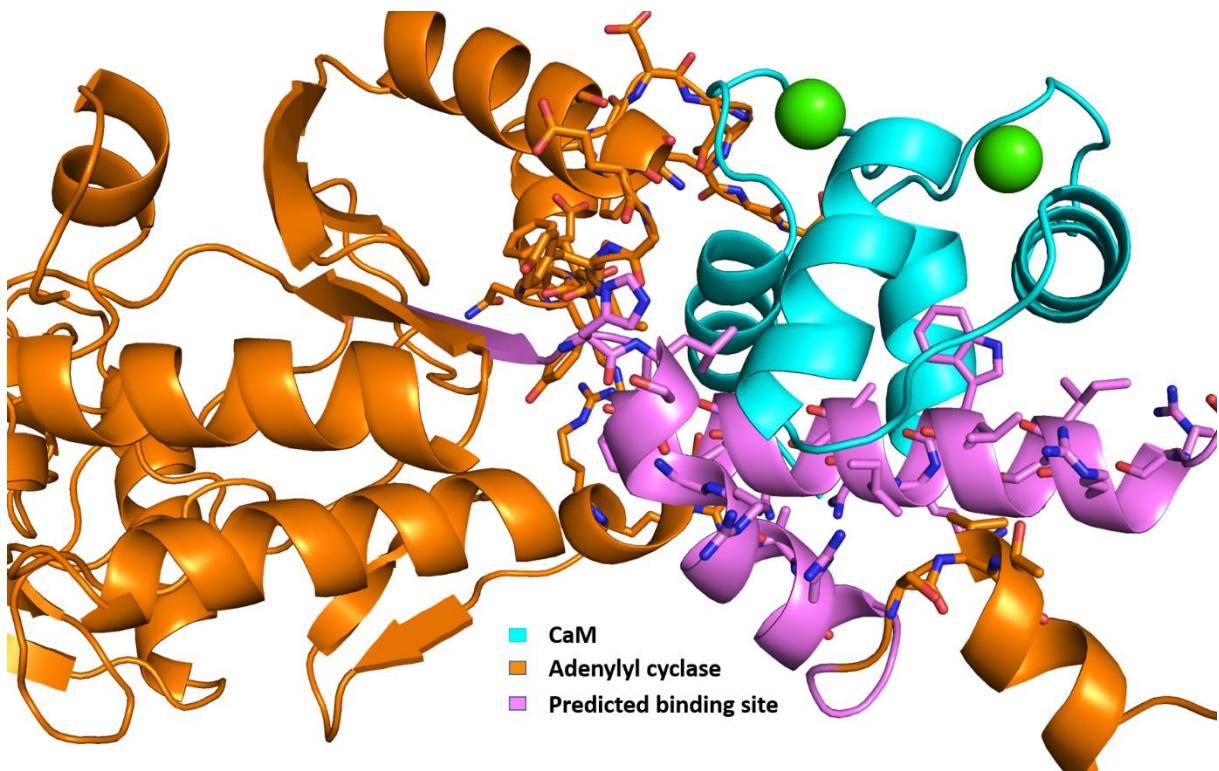
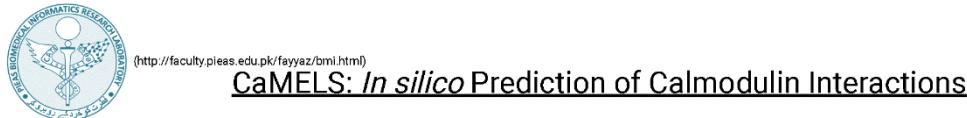


Figure 6. (a)



What is CaMELS?

We present a set of novel algorithms, called CaMELS (CalModulin intEraction Learning System) for predicting both the binding sites and the possibility of interactions of proteins with Calmodulin (CaM) using sequence information alone. The proposed algorithms give state of the art classification results and are available as a cloud based webserver. You can use it to make predictions for proteins of your interest.

Citation details:

If you use CaMELS please cite Wajid Arshad Abbasi (<http://faculty.pieas.edu.pk/fayyaz/wajid/index.html>) and Fayyaz-ul-Amir Afzar Minhas (<http://faculty.pieas.edu.pk/fayyaz/index.html>) (2016), "CaMELS: *In silico* Prediction of Calmodulin Binding Proteins and their Binding Sites", DCIS, PIEAS, (preprint release).

Select fasta file or paste the required sequence

(Sequence length>21 required)

Prediction type:

Input file in fasta format (containing only one protein): No file chosen

(or) Paste your protein sequence in plain text:

SUBMIT

Figure 6. (b)

The results of your submission generated through CAMELS are:

Interaction Prediction Results:

2.76

Binding Site Prediction Results:

