Kernel Methods for Calmodulin Binding and Binding Site Prediction

Michael Hamilton
Computer Science Department
Colorado State University
Fort Collins, CO 80523-1873
hamiltom@cs.colostate.edu

A.S.N. Reddy
Department of Biology
Colorado State University
Fort Collins, CO 80523
reddy@lamar.colostate.edu

Asa Ben-Hur
Computer Science Department
Colorado State University
Fort Collins, CO 80523-1873
asa@cs.colostate.edu

ABSTRACT

Calmodulin (CaM) is a calcium-binding protein that is involved in a variety of cellular processes, interacting with many proteins. Since many CaM interactions are calcium-dependent, they are difficult to detect using high-throughput methods like yeast-two-hybrid. Furthermore, detection of CaM binding sites requires a significant experimental effort. Using a collection of CaM binding sites extracted from the Calmodulin Target Database we trained SVM-based classifiers to detect CaM binding sites using a variety of sequence features; our best classifier achieved an area under the ROC curve of 0.89 for detecting binding site locations at the amino acid level. We apply our classifiers to the problem of detecting CaM binding proteins in *Arabidopsis*; at a false-positive level of 0.05 we detected 638 novel putative CaM binding proteins. These proteins share overrepresented Gene Ontology terms associated with the functions of known CaM binders.

1. INTRODUCTION

Calmodulin (CaM) is a ubiquitous, highly-conserved calcium-binding protein found in all eukaryotes [5]. Calcium signals are important for many cellular processes, including ion transport, enzyme activation, phosphorylation, and dephosphorylation of proteins [16]. Calmodulin acts as a major calcium sensor and interacts with diverse proteins and regulates their function [17]. Finding the interaction partners of CaM is therefore important for the understanding of the processing of calcium signaling in the cell.

It is difficult to find interactions of CaM using high-throughput methods like yeast-two-hybrid [17]. A recent experiment using a protein array that targeted a thousand *Arabidopsis* proteins has revealed CaM interactions for around 200 of the proteins spotted on the array [14]. Although the choice of proteins was biased towards proteins that are likely to interact with CaM (kinases and transcription factors), these results indicate that CaM interacts with a large number of proteins in the *Arabidopsis* proteome.

Despite the increasing availability of protein-protein interactions, binding site location remains hard to obtain, and this type of data is still relatively scarce. The need to address this gap has led to the development of a variety of methods for identifying interaction sites (see [22] for a review). The majority of binding site prediction methods use structural information on surface residues to predict the likelihood that they belong to a protein-protein interface. Yet, due to the lack of available structural models for most proteins, the applicability of these methods is limited. Therefore it is of interest to develop prediction methods that use features inferred from sequence alone, and the work of Ofran et al. [12] suggests that it is indeed possible. Ofran et al. focused on predicting binding sites in general; in contrast, our work focuses on binding sites of a specific protein, namely Calmodulin.

Previous work has shown the feasibility of predicting CaM binding sites from sequence [15]. Their method uses a complex neural network architecture that is trained using features that characterize sequence characteristics of a window of amino acids. We obtain slightly improved performance using a conceptually simpler approach that uses support vector machines (SVM); we explore the effect of different representations on classifier performance, and obtain better performance than the previous work of Radivojac et al. [15].

2. APPROACH AND RELATED WORK

We address two problems: identification of CaM binding sites within known CaM binders and the identification of proteins that interact with CaM. Prediction of CaM binding sites is the problem of identifying which amino acids in a protein are likely involved in the interaction, whereas prediction of CaM binding proteins is a problem at the protein level and classification addresses the binary response of whether a protein has the potential to interact with CaM.

The problem of predicting CaM binding sites is an instance of the label-sequence learning problem [7]. Given a protein sequence $\mathbf{x}$ of length n, we want to associate with it a sequence of labels $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ where y_i, indicates whether position i is part of a binding site. A standard approach for solving this problem is the sliding window method [7] which reduces the problem to a two-class classification problem: Each position in the sequence is predicted to either belong to a binding site or not, independently of its neighbors. At this level one can apply standard classifiers that use features computed on the basis of a fixed-length

window of amino acids. This is the approach taken in [15], and is the strategy of our binary SVM method.

It is also worth mentioning methods that predict binding at the level of conserved domains or motifs [20]: methods which aim to explain an observed network of interactions in terms of interactions between these features. Such methods, while using global information on the interaction network, ignore properties of the individual proteins, and only provide a coarse-grained prediction of binding site location. The proposed methods in this work are much more specific. By training a classifier on examples of interaction partners for a specific protein (CaM), it is possible to capture properties of the binding sites.

3. METHODS

3.1 Binary Classification with Support Vector Machines

The support vector machine (SVM) is a state-of-the-art classifier that constructs a maximum-margin boundary separating two classes of data [6]. It is being widely used in bioinformatics because of its accuracy and ability to deal with a variety of data with the help of kernels [4].

The data for a binary classification problem is composed of N examples $\mathbf{x}_i$, each belonging to some input space $\mathcal{X}$, with labels $y_i \in \{1, -1\}$. Finding the maximum margin linear classifier whose discriminant function is $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$ is formulated as a maximum-margin problem using the following optimization problem [6, 4]:

$$\begin{aligned} \underset{\mathbf{w},b}{\text{minimize}} \quad & \frac{1}{2}\,||\mathbf{w}||^2 + C\sum_{i=1}^{N}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ \ i = 1, \ldots, N, \end{aligned} \tag{1}$$

where ξ_i are slack variables that allow examples to fall within the margin (margin violations) or be misclassified, and C is a positive constant that specifies the penalty associated with margin violations and errors.

3.2 Sliding-Window Classifiers

To apply a binary SVM to the problem of CaM binding site prediction we take advantage of the fact that CaM binding sites are contiguous in sequence [13], allowing us to use the sliding window approach. In what follows we denote the i^{th} amino acid in protein $\mathbf{x}$ by $\mathbf{x}[i]$. The event of binding at position i of a protein $\mathbf{x}$ is represented by a window of length $2k + 1$ that incorporates features computed on neighboring amino acids $\mathbf{x}[i-k], \ldots, \mathbf{x}[i+k]$. For amino acids that lie at the beginning and end of the sequence, the window is shorter in length. For instance, the window associated with the first amino acid will only consist of the amino acids up to position k in the sequence.

3.3 p-spectrum Kernel

Since known CaM binding sites are of varying lengths, we use sequence kernels that represent a sequence in a fixed dimensional feature space, independent of its length. In order to leverage the amino acid content of a protein, we use the p-spectrum kernel [11] for $p = 1$ and $p = 2$. The p-spectrum of a string over an alphabet Σ is a vector whose components count occurrences of all length-p substrings. For protein sequences, Σ is the 20-letter amino acid alphabet. In our experiments, the p-spectrum kernel is normalized using the cosine kernel,

$$k_{\text{cosine}}(\mathbf{x}, \mathbf{x}') = \frac{k(\mathbf{x}, \mathbf{x}')}{\sqrt{k(\mathbf{x}, \mathbf{x})k(\mathbf{x}', \mathbf{x}')}}. \tag{2}$$

3.4 Gappy Pair Kernel

We propose a simple extension of the 2-spectrum kernel that considers all pairs of amino acids that are up to a given distance apart, which will allow us to capture short motifs that are present in CaM binding proteins. Let $\phi^d_{uv}(\mathbf{x})$ be the number of times the amino acid u is followed by the amino acid v at a distance of at most d. The feature map $\Phi^d_{\text{gappypair}}(\mathbf{x})$ is the vector with components $\phi^d_{uv}(\mathbf{x})$. The Gappy Pair Kernel is then defined as

$$k_{\text{gappypair}}(\mathbf{x}, \mathbf{x}') = \Phi^d_{\text{gappypair}}(\mathbf{x})^T\Phi^d_{\text{gappypair}}(\mathbf{x}'). \tag{3}$$

Normalization is performed using the cosine kernel, Equation (2). While the dimensionality of the Gappy Pair feature space is $d\,|\Sigma|^2$, a sequence $\mathbf{x}$ has only $O(d|\mathbf{x}|)$ nonzero features. Therefore it is feasible to explicitly represent the feature vector $\Phi^d_{\text{gappypair}}$, leading to kernel computation which is $O(d * (|\mathbf{x}| + |\mathbf{x}'|))$ for sequences $\mathbf{x}$ and $\mathbf{x}'$. In this work we used a sum of gappy pair kernels with d in the range $0, \ldots, 9$.

3.5 Physico-Chemical Kernel

The Amino Acid Index Database provides access to several physico-chemical properties of amino acids [10]. The physico-chemical mapping used in this work represents a sequence of amino acids as a 60-dimensional feature vector where each component is the average of an individual property across the sequence. To account for the different scales of different properties, each feature vector is standardized, where the means and standard deviations are computed by sampling windows over all training data. This type of representation has been used by [15] for prediction of CaM binding sites.

3.6 `PSI-BLAST` Kernel

Given that CaM is highly conserved across a variety of species, and as it has been shown that interacting proteins tend to co-evolve [8], it is likely that CaM interaction sites share similar evolutionary histories. To represent the evolutionary history of an amino acid sequence, position-specific scoring matrices (PSSMs) are built for each protein using `PSI-BLAST` [1]. For a protein sequence $\mathbf{x}$ of length n we obtain

$$\text{PSSM}(\mathbf{x}) = \begin{bmatrix} p_1(x[1]) & p_1(x[2]) & \ldots & p_1(x[n]) \\ p_2(x[1]) & p_2(x[2]) & \ldots & p_2(x[n]) \\ \vdots & \vdots & \ddots & \vdots \\ p_{20}(x[1]) & p_{20}(x[2]) & \ldots & p_{20}(x[n]) \end{bmatrix}, \tag{4}$$

where $p_m(\mathbf{x}[i])$ is the level of conservation of amino acid m for the i^{th} position of $\mathbf{x}$.

For this work, PSSMs for all training and test sequences were generated running `PSI-BLAST` for five iterations using default parameters against the non-redundant UniRef50 database [19].

We define a feature mapping that represents a PSSM as a length-400 feature vector where each feature represents the average level of conservation of an amino acid with respect

to another:

$$\phi_{m,s}(\mathbf{x}) = \frac{1}{|\mathbf{x}|}\sum_{i=1}^{|\mathbf{x}|} I_m(\mathbf{x}[i])p_s(\mathbf{x}[i]), \quad (5)$$

where $I_m(\mathbf{x}[i])$ is the indicator function that equals 1 if $\mathbf{x}[i] = m$, and 0 otherwise. Each feature $\phi_{m,s}(\mathbf{x})$ corresponds to the average level of conservation of the amino acid s, over columns of the PSSM where an amino acid m is encountered in $\mathbf{x}$. Let $\Phi_{\text{psi}}(\mathbf{x})$ be the vector with components $\phi_{m,s}(\mathbf{x})$. The kernel is then defined as

$$k_{\text{psi}}(\mathbf{x}, \mathbf{x}') = \Phi_{\text{psi}}(\mathbf{x})^T \Phi_{\text{psi}}(\mathbf{x}'). \quad (6)$$

Normalization is then performed using Equation (2).

3.7 CaM binding and binding site prediction

In what follows we assume a single CaM binding site per sequence. A binding site is predicted at the location for which the discriminant function achieves its highest value along the length of the protein.

We use the classifiers trained to predict CaM binding sites to rank potential CaM interaction partners. A protein is assigned a score which is the maximum score of all windows in the protein. These scores are used to rank potential CaM binders.

3.8 Experimental Setup

3.8.1 Datasets

For the prediction of CaM binding sites we use the Calmodulin Target Database, which provides a central data repository for CaM binding site data, containing nearly 200 proteins from a variety of species [21]. From this database, a non-redundant dataset of 153 proteins containing 185 binding sites was extracted by [15]. This dataset was constructed such that no two proteins share more than 40% sequence identity, and no two binding sites more than 50% identity. The known binding site or sites in each protein serve as positive examples for the binary SVM; we generated negative examples by sliding a length-21 window at 10 amino acid increments.

Our second experiment is aimed at prediction of CaM binding proteins in *Arabidopsis*. We train a CaM binding site predictor on the data from the Calmodulin Target Database and use it to rank *Arabidopsis* proteins for the likelihood of an interaction with CaM. All *Arabidopsis* proteins that are not in the training set were used to test this classifier. The 241 proteins found in [14] to interact with at least one of the *Arabidopsis* CaMs were used as positive examples, and all other *Arabidopsis* proteins were used as negative examples.

3.8.2 Leave-One-Protein-Out Cross Validation

Following [15], we use a leave-one-protein-out cross validation (LOPOCV) procedure. For each protein, a classifier was trained on the remaining proteins, and the prediction on the test protein was noted. This procedure was repeated until every protein was involved in testing.

Classifier performance was measured using the area under the Receiver Operating Characteristic (ROC) curve. In our LOPOCV procedure an ROC curve is generated by averaging the ROC curves produced for each left-out protein. The area under the curve (AUC) is then computed to produce an overall classifier score. We also used AUC_{50} scores which are computed by considering the ROC curve until the first 50 false positives.

Kernel	AUC	AUC_{50}
1-spec	0.87	0.62
2-spec	0.87	0.57
pchem	0.87	0.58
psi	0.88	0.61
gappy-pair	0.89	0.61

Table 1: Classifier performance at the amino acid level for the 1-spectrum and 2-spectrum (1-spec, 2-spec), Physico-Chemical (pchem), PSI-BLAST (psi), and the Gappy Pair (gappy-pair) kernels.

3.8.3 Model selection

The size of the sliding window was chosen as 21, as this is the average length of the CaM binding sites in our data, and the maximum distance for the gappy pair kernel was chosen as 10. When running LOPOCV the SVM slack penalty parameter C was chosen for each left out protein using 3-fold cross-validation using the values $\{10^{-4}, 10^{-3}, \ldots, 10^{3}\}$. By using the AUC as the performance criterion, the C with the highest score is used for overall classifier assessment. SVM experiments were conducted using `PyML` which provides a machine learning framework for kernel-based learning [2].

4. RESULTS

4.1 Binding Site Prediction

ROC curves computed using our LOPOCV procedure are shown in Figure 1, and the associated AUC and AUC_{50} scores for each kernel are provided in Table 1. We observe that the PSI-BLAST and gappy-pair kernels perform slightly better than the kernels defined using the 1-spectrum, 2-spectrum, and physico-chemical properties. This highlights the added benefit of incorporating evolutionary information, and the flexibility of the gappy-pair kernel, which allows the classifier to capture simple motifs associated with CaM binding (see details below). The differences are evident in the lefthand panel in Figure 1, which zooms on the highest ranking predictions. The fact that the 1-spectrum kernel worked almost as well as our more sophisticated kernels suggests that binding site selection is driven by amino acid composition. This is not surprising, as it is well known that CaM binding regions are often disordered, a property that is largely determined by amino acid composition [15].

Adding kernels often leads to improved performance [3], but has not been the case here. This is the result of the similarity in the information captured by the different kernels, as illustrated in Figure 2, which shows binary SVM discriminant values plotted along the length of a known CaM binding protein.

4.2 Classifier Weight Analysis

The weights learned by a classifier provide useful information on the relative importance of the input features for the different kernels we studied.

1-spectrum kernel.

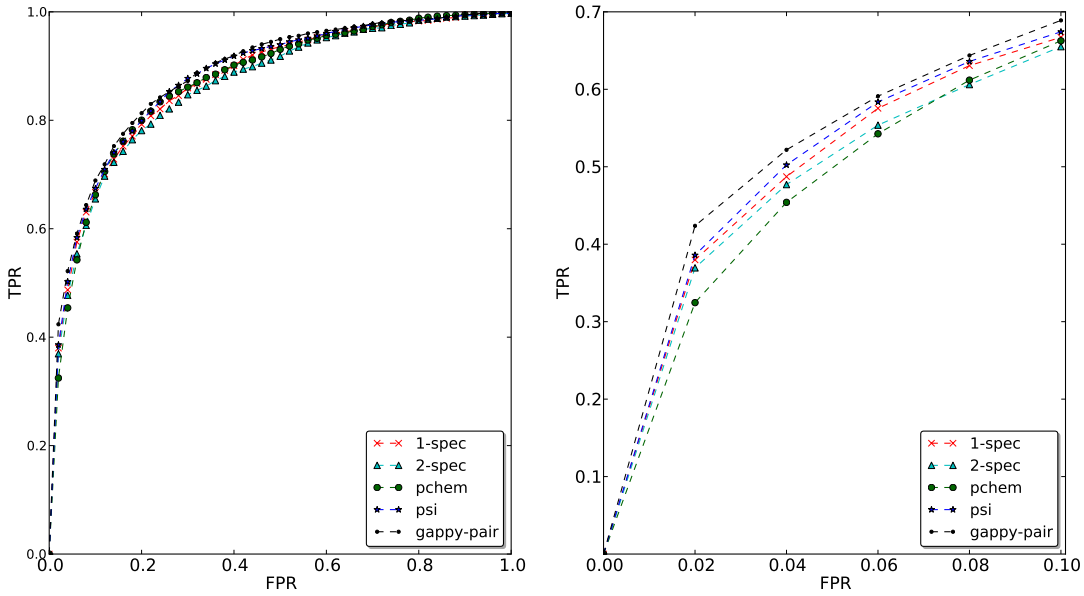

Figure 1: ROC curves for binding site prediction. On the left is the full ROC curve, and on the right we zoom in on the left-most section of the curve. The kernels used are: 1-spectrum (1-spec), 2-spectrum (2-spec), physico-chemical (pchem), PSI-BLAST (psi), and Gappy-pair (gappy-pair). The corresponding AUC scores are found in Table 1

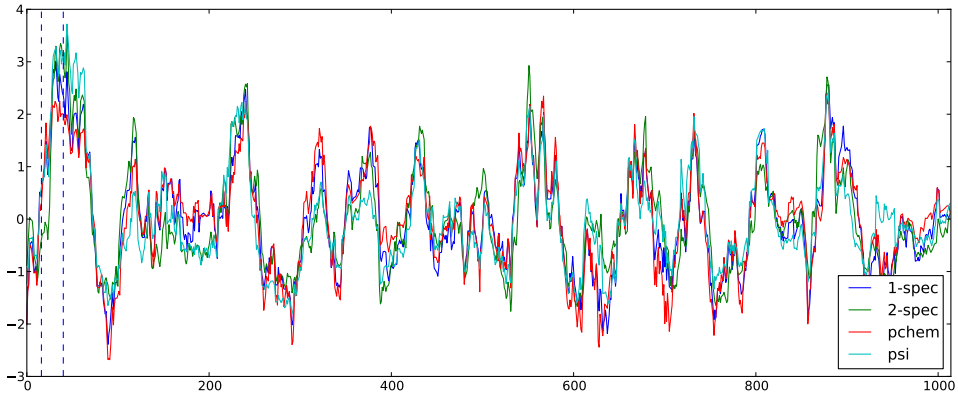

Figure 2: Discriminant values for each amino acid for the *Arabidopsis* protein AT4G37640 computed using the binary SVM with several kernels. The dashed lines indicate the location of the known binding site. For each kernel, the maximum value across the protein occurs within the known binding site.

Certain amino acids have a higher tendency to occur in CaM binding sites, and the binary SVM trained using the 1-spectrum kernel assigned these features weights that are in agreement with these tendencies. We found high positive weights for arginine (R), lysine (K), and tryptophan (W) and negative weights of aspartic acid (D), glutamic acid (E), and proline (P); this is in agreement with the propensity of CaM binding to occur at amphiphilic alpha helices.

Physico-Chemical kernel.

The top feature chosen by the SVM for physico-chemical properties kernel was the flexibility parameter, suggesting that CaM binding sites are more flexible than non-binding regions. A high degree of flexibility often coincides with a region that is disordered, a property shared by several CaM binding sites [15]. Other top features correspond to propensities of forming amphiphilic alpha-helices, which as mentioned previously, agrees with the literature on CaM binding sites.

Gappy Pair kernel.

The gappy pair kernel allows the classifier to represent short motifs. Alignment of CaM binding sites has revealed a few weak motifs [18], and calcium-independent interactions were characterized by the so-called IQ motifs. The top gappy-pair feature was the motif `Q...R`, where '.' denotes a gap. This motif forms the beginning of most of the motifs that represent calcium-independent interactions. The rest of the top features were dominated by the amino acids arginine (R) and lysine (K), which are enriched in CaM binding proteins.

4.3 Interaction Prediction

Kernel	AUC
1-spec	0.71
2-spec	0.71
pchem	0.64
psi	0.74
gappy-pair	0.71

Table 2: AUC scores for the task of predicting proteins that interact with CaM in *Arabidopsis*.

We used our binding site classifiers to predict CaM binding by assigning a score to each protein as described in Section 3.7. We tested the classifiers by testing them on the *Arabidopsis* proteome (27,379 proteins), where the known CaM binders in *Arabidopsis* were used as positive test examples and the remaining proteins were used as negative test examples. AUC scores for each classifier were computed and the results are summarized in Table 2. Figure 4 features the ROC curves for the binary SVM. In this task the binary SVMs performed better for all kernels with the exception of the Physico-Chemical kernel. There was also a clear advantage to the PSI-BLAST kernel, which performed best. Although the AUC score of 0.74 for this kernel may not seem impressive, at a threshold where the false positive rate is expected to be around 2.5%, 747 potential CaM interactions are identified, of which 638 are novel (this threshold is marked by a star in Figure 4). As our knowledge of CaM binders is only partial, we believe that actual classifier accuracy is higher than indicated by these results. A demonstration of this phenomenon in the context of protein function prediction is provided in [9]. Protein-level annotation of our 750 top predictions with potential CaM binding sites are found on a companion website at `http://combi.cs.colostate.edu/supplements/cam/`.

4.4 Comparison to Previous Work

Previous work by Radivojac et al. [15] achieved an AUC of 0.89 for binding site prediction using a post-processing step that rejects regions that are underrepresented in the training data. We are able to achieve the same level of performance for all potential binding sites. We obtained the raw predictions made in [15], and as shown in Figure 3, our most accurate kernels perform slightly better across all possible thresholds, especially for highly confident predictions. The difference in performance is much more pronounced in the *Arabidopsis* CaM binding task. Radivojac et al. have kindly provided us with their method's predictions on the *Arabidopsis* proteome from which we computed protein-level predictions using the same strategy we applied to our amino-acid level classifiers. Improved performance of our methods is readily observable in the ROC curves shown in Figure 4. At the chosen false positive level our method achieves a true positive rate that is several times better than Radivojac et al.'s.

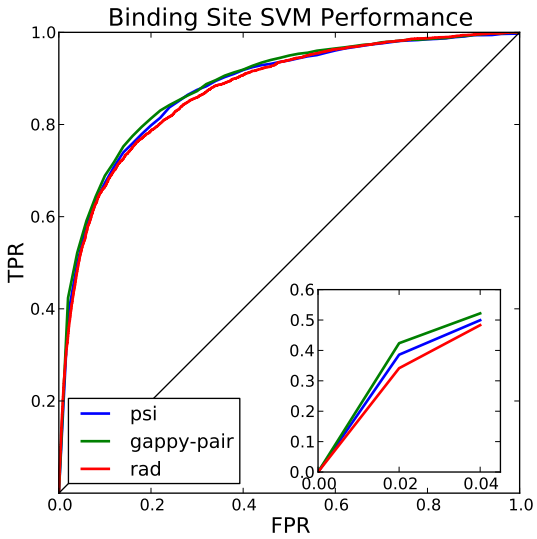

Figure 3: ROC curves for the top performing kernels and previous results (rad) by Radivojac et al. [15]. The inset represents highly confident predictions.

5. CONCLUSION

In this work we presented an SVM-based method for predicting CaM binding proteins and identifying CaM binding sites. Performance using amino acid composition alone was nearly as accurate at predicting CaM binding sites as more sophisticated kernels, suggesting that the problem is driven by amino acid composition. Predicting CaM binding proteins in *Arabidopsis* proved to be a more challenging problem, and in this case the kernel that uses sequence conservation computed using PSI-blast showed much better

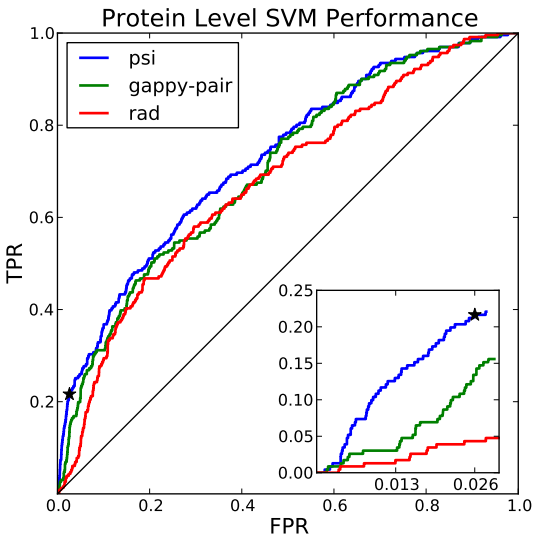

Figure 4: ROC curves for the binary SVM and previous results (rad) by Radivojac et al. [15] tested at the protein level to predict CaM binding in *Arabidopsis*. The inset represents highly confident predictions and the star marks the threshold used for the GO analysis.

performance. In addition, our methods are shown to perform better than the work of Radivojac et al. [15] especially in the protein interaction prediction task.

Our results on predicting CaM binding activity directly from sequence show the promise of our approach. At a very conservative threshold, our classifiers identified 638 novel CaM binding proteins in *Arabidopsis*, adding to around 200 known CaM binders. This large number of interactions highlights the key role of CaM in the plant cell. The interested reader can find binding site predictions for these predicted CaM binders in a supplemental website at `http://combi.cs.colostate.edu/supplements/cam/`.

6. ACKNOWLEDGMENTS

The authors wish to thank Keith Dunker and Predrag Radivojac for providing the processed CaM binding datasets, *Arabidopsis* predictions, and their cross-validation results.

7. REFERENCES

[1] S. F. Altschul et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, September 1997.

[2] A. Ben-Hur. *PyML - machine learning in Python*, 2009. Software available at `http://pyml.sourceforge.net/`http://pyml.sourceforge.net/.

[3] A. Ben-Hur and W. Noble. Kernel methods for predicting protein–protein interactions. *Bioinformatics*, 21(suppl 1):i38, 2005.

[4] A. Ben-Hur, C. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch. Support vector machines and kernels for computational biology. *PLoS Computational Biology*, 4(10), 2008.

[5] N. Bouche, A. Yellin, W. Snedden, and H. Fromm. Plant-specific calmodulin-binding proteins. *Annual Review of Plant Biology*, 56(1):435–466, 2005.

[6] C. Cortes and V. Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[7] T. G. Dietterich. Machine learning for sequential data: A review. In *Proceedings of the Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition*, pages 15–30, London, UK, 2002. Springer-Verlag.

[8] C.-S. Goh et al. Co-evolution of proteins with their interaction partners. *Journal of Molecular Biology*, 299(2):283–293, 2000.

[9] M. A. Hibbs, C. L. Myers, C. Huttenhower, D. C. Hess, K. Li, A. A. Caudy, and O. G. Troyanskaya. Directing experimental biology: A case study in mitochondrial biogenesis. *PLoS Comput Biol*, 5(3):e1000322, 03 2009.

[10] S. Kawashima, H. Ogata, and M. Kanehisa. AAindex: Amino acid index database. *Nucleic Acids Research*, 27(1):368–369, 1999.

[11] C. S. Leslie et al. The spectrum kernel: a string kernel for SVM protein classification. In *Pacific Symposium on Biocomputing*, pages 566–575, 2002.

[12] Y. Ofran and B. Rost. Isis: interaction sites identified from sequence. *Bioinformatics*, 23(2):e13–e16, 2007.

[13] K. T. O'Neil and W. F. DeGrado. How calmodulin binds its targets: sequence independent recognition of amphiphilic alpha-helices. *Trends in Biochemical Sciences*, 15(2):59–64, 1990.

[14] S. C. Popescu et al. Differential binding of calmodulin-related proteins to their targets revealed through high-density *Arabidopsis* protein microarrays. *PNAS*, 104(11):4730–4735, March 2007.

[15] P. Radivojac et al. Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition. *Proteins: Structure, Function, and Bioinformatics*, 63(2):398–410, 2006.

[16] A. Reddy. Calcium: silver bullet in signaling. *Plant Sci*, 160(3):381–404, 2001.

[17] A. Reddy, A. Ben-Hur, and I. S. Day. Experimental and computational approaches for the study of calmodulin interactions. *Phytochemistry*, 2011.

[18] A. R. Rhoads and F. Friedberg. Sequence motifs for calmodulin recognition. *The FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology*, 11(5):331–340, April 1997.

[19] B. E. Suzek et al. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007.

[20] H. Wang et al. Insite: a computational method for identifying protein-protein interaction binding sites on a proteome-wide scale. *Genome Biology*, 8(9):R192, 2007.

[21] K. L. Yap et al. Calmodulin target database. *Journal of Structural and Functional Genomics*, 1(1):8–14, March 2000.

[22] H. Zhou and S. Qin. Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics*, 23(17):2203–2209, June 2007.