Calmodulin Signaling: Analysis and Prediction of a Disorder-Dependent Molecular Recognition

Predrag Radivojac,[1] Slobodan Vucetic,[2] Timothy R. O'Connor,[3] Vladimir N. Uversky,[4] Zoran Obradovic,[2] and A. Keith Dunker[4]*

[1] *School of Informatics, Indiana University, Bloomington, Indiana*
[2] *Center for Information Science and Technology, Temple University, Philadelphia, Pennsylvania*
[3] *Department of Electrical Engineering and Computer Science, Washington State University, Pullman, Washington*
[4] *Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, Indiana*

ABSTRACT **Calmodulin (CaM) signaling involves important, wide spread eukaryotic protein–protein interactions. The solved structures of CaM associated with several of its binding targets, the distinctive binding mechanism of CaM, and the significant trypsin sensitivity of the binding targets combine to indicate that the process of association likely involves coupled binding and folding for both CaM and its binding targets. Here, we use bioinformatics approaches to test the hypothesis that CaM-binding targets are intrinsically disordered. We developed a predictor of CaM-binding regions and estimated its performance. Per residue accuracy of this predictor reached 81%, which, in combination with a high recall/precision balance at the binding region level, suggests high predictability of CaM-binding partners. An analysis of putative CaM-binding proteins in yeast and human strongly indicates that their molecular functions are related to those of intrinsically disordered proteins. These findings add to the growing list of examples in which intrinsically disordered protein regions are indicated to provide the basis for cell signaling and regulation. Proteins 2006;63:398–410.**

Key words: protein–protein interactions; protein function; unfolded; unstructured

INTRODUCTION

Calmodulin (CaM) is a 148-residue long, 16.7-kDa, intracellular protein involved ubiquitously in numerous eukaryotic regulatory processes and highly conserved throughout the eukaryotic kingdom.[1–3] CaM is characterized by a unique dumbbell-like structure, with a flexible linker connecting globular lobes at the two termini.[4,5] The N- and C-terminal globules are homologous and independent,[6] each consisting of a pair of helix-loop-helix (EF-hand) structural motifs. The central flexible linker becomes an α-helix in the crystalline state.[7,8]

CaM is a major transducer of calcium signals via its interactions with many partners and so is abundantly expressed in all eukaryotic cells so far studied.[1] The modes of CaM interaction with binding partners are very diverse: CaM binds short peptides and proteins both in the presence or the absence of Ca^{2+}, reversibly and irreversibly, as an inhibitor or an activator.[9] Perhaps even more remarkably, CaM regulates the activity of kinases[10] and phosphatases,[11] protein classes with opposite functions. CaM also exhibits the capacity to bind small drug-like molecules,[12] although it is unclear whether this capacity relates to biological function in any direct way.

The interactions between CaM and its binding targets (CaMBTs) involve disorder-to-order transitions for the CaM molecule. The flexible linker,[4,5] which becomes structured upon complex formation, allows the globular domains to wrap around the CaMBT as shown in Figure 1. The helix–helix interactions within the two globular domains are not completely rigid, so the helix–helix packing interfaces in these domains vary in a manner that depends on the detailed interactions with the different CaMBTs.[3,13] Finally, the CaM target-binding surface is rich in methionines, which adopt different configurations when CaM associates with CaMBTs having different sequences. The end result of this structural plasticity is that CaM binds to a very large number of different sequences with high affinity.

Most of the CaM's binding targets are regions of about 20 residues in length, typically in an α-helical conformation. The Calmodulin Target Database[14] classifies CaM-binding targets into five distinct motifs: 1–10, 1–14, 1–16, IQ, and Other. Motifs 1–10, 1–14, and 1–16 are named according to the positions of large hydrophobic residues. Binding to these motifs is predominantly Ca^{2+} dependent. On the other hand, binding to IQ motifs is typically independent of Ca^{2+} concentration. Nevertheless, peptides that bind to CaM regardless of the $[Ca^{2+}]$ have also been encountered. Interestingly, some CaM-binding tar-

This article contains Supplementary Materials, which can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: National Institutes of Health; Grant number: 1R01 LM07688 (A.K.D. and Z.O). Grant sponsor: the Indiana Genomics Initiative, funded in part by the Lilly Endowment (A.K.D. and P.R.)

*Correspondence to: A. Keith Dunker, Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 714 N. Senate Avenue EF 250, Indianapolis, IN 46202. E-mail: kedunker@iupui.edu

Received 20 June 2005; Revised 19 September 2005; Accepted 21 September 2005

Published online 21 February 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20873

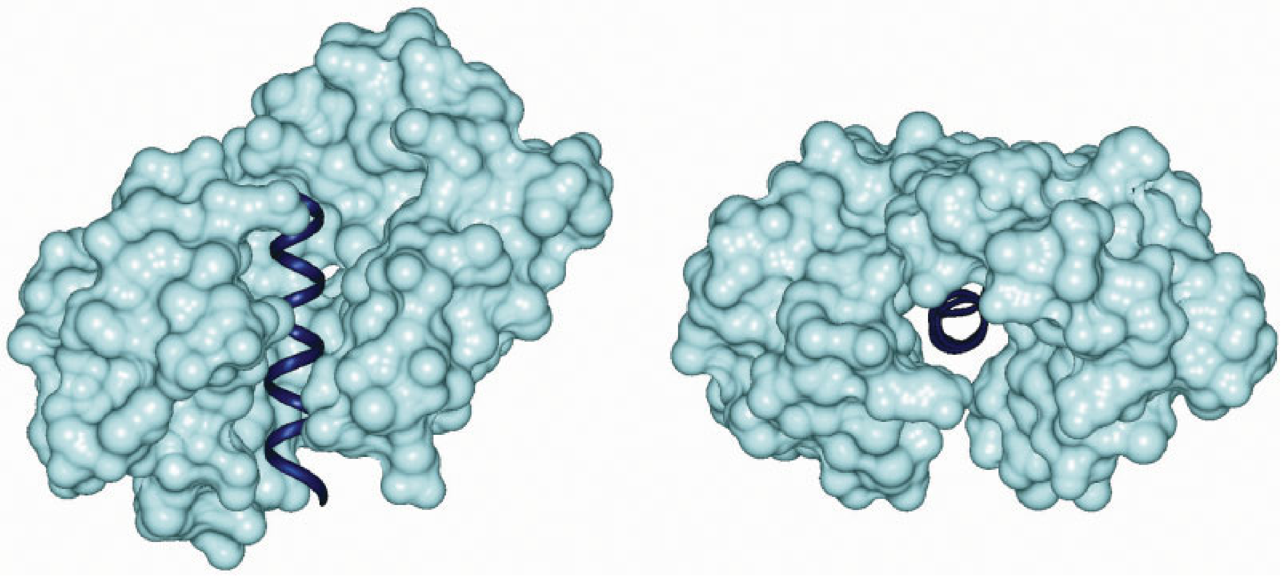

Fig. 1. The CaMBT peptide bound to CaM. This figure illustrates (especially the axial view of the helix) the spatial separation from a parent globular domain of the CaMBT peptide. Because of this behavior, we hypothesize intrinsic disorder enables the CaMBT peptide to be spatially separated from the rest of the protein in the unbound state that permits CaM to subsequently bind in such a manner.

gets are not helical in their entire lengths. For example, in the CaM-CaMKK complex, CaMKK's hairpin loop residues are essential for the interaction.[15] Another rule-breaking motif is a β-sheet region of iNOS, which accommodates both α- and β-like conformations in different modes of binding to CaM.[16]

Three lines of evidence suggest that, before their associations with CaM, CaMBTs exhibit conformational flexibility or even intrinsic disorder, which is a protein feature that is receiving increasing attention.[17–21] First, in a collection of more than 20 examples reported since the late 1970s, CaM-stimulated enzymes were also stimulated by limited trypsin digestion. The digested proteins were no longer stimulated by CaM, and if tested, the digested proteins no longer bound CaM; here we provide a few of the pioneering references.[22–26] Because structured proteins are orders of magnitude more resistant to trypsin digestion than are unfolded proteins,[27] such digestion studies can be interpreted as indicators that the CMBTs are structurally disordered before association with CaM. Second, the manner in which CaM wraps around the CaMBTs suggests that there has to be enough physical space to accommodate CaM. A piece of protein, even a helix, in such an open environment would very likely sample an ensemble of structures and conformations over time, as experimentally verified for some CaMBTs.[28,29] Third, in a limited number of examples, the CaMBT regions of proteins are missing from the electron density maps of the corresponding protein crystal structures, suggesting that these regions lack specific 3-D structures and are instead intrinsically disordered. We have found four such examples in the Protein Data Bank (PDB) that matched CaM-binding regions from the Calmodulin Target Database: calcineurin (1aui|a), N-Nos (1tll), Irs-1 (1qqg), and Rala (2bov|a).

Although experimental data indicates that several CaMBTs are disordered before association with CaM, the generality of these findings has not been tested. In the present study we used bioinformatics approaches to carry out a systematic analysis over a large number of examples in order to further test the hypothesis that CaMBTs are intrinsically disordered or are flanked by disordered regions before their associations with CaM. To our knowledge, no large-scale systematic studies have been undertaken so far on the structural properties of CaMBTs, on their abundance, nor on the functional properties of their associated proteins. Here, we first performed an analysis of sequence and physicochemical properties of CaMBTs. Then, we constructed a predictor of CaM-binding targets and carried out a detailed evaluation on all types of known binding motifs: our results strongly indicate that CaMBTs are predictable from an amino acid sequence with a useful accuracy and that disordered, compared with ordered, regions have a significantly higher propensity for containing CaMBTs. Overall, these results support the hypothesis that CaMBTs are disordered (i.e., they sample multiple conformations before their associations with CaM).

The pioneering studies by Depaoli-Roach et al.[22] and others[23–26] was one of several early harbingers in the late 1970s, indicating the possibility of functional roles for highly flexible, perhaps even unfolded regions of proteins. Years later, Rose and coworkers helped to further stimulate the growing interest in unfolded proteins with their computational analysis of the steric interactions of unfolded protein.[30,31] Although this and other work[32] has focused on unfolded proteins as the initial state in the protein folding reaction, the recent findings that many proteins use unfolded regions for biological function[17,21,33] means that the seminal work of Rose, his coworkers, and

TABLE I. Characteristics of the Dataset[†]

Motif type[a]	No. of regions	No. of proteins	No. of binding residues
1–10	15	12	239
1–12	7	6	135
1–14	39	38	774
1–16	1	1	23
IQ	49	35	977
Other	87	74	1955

[†]Note that, because several proteins contain multiple binding regions and because some binding regions overlap, the total number of proteins exceeds 157 and the total number of binding residues exceeds 4088.

[a]Motif type 1–10 includes 1–10, 1–5–10, and basic 1–5–10 from Calmodulin Target Database (CTD).[14] Motif type 1–14 includes 1– 14, 1–5– 8 –14, 1– 8 –14, and basic 1–14. Motif type IQ includes IQ and IQ-like motifs. Motif type Other includes basic and others from CTD. Motif type 1–12 was separated form its original CTD classification Other.

others on unfolded proteins will have a wider applicability than protein folding alone.

MATERIALS AND METHODS

Datasets

A set of proteins containing CaM-binding targets was assembled from the Calmodulin Target Database[14] located at http://calcium.uhnres.utoronto.ca/ctdb/. This database contains the accession numbers of the calmodulin-binding proteins (CaMBPs) and the sequences of the isolated binding targets (CaMBTs). Several of these motifs have structural representatives in PDB that indicate a common binding mechanism among them (e.g., 1cdl, 1ckk, 1g4y, 1iq5, 1mxe, 2bbm).

The set of 210 CaMBPs containing 287 CaMBTs was then filtered for similarity to prevent over-representation of any particular sequence and binding site during predictor construction. A nonredundant set was selected such that no two CaM-binding proteins or peptides were significantly similar. To achieve this, we used a 40% sequence identity threshold at a protein level and 50% for the binding targets. In several cases, based on visual inspection and the fact that CaMBTs were located in the divergent regions, we allowed the global sequence identity to exceed 40%. At a binding target level, only 55 of 19,503 pairs (0.3%) had sequence identity $>40\%$ (and $\leq 50\%$). Because these thresholds lay well below those providing accurate functional inference "by homology transfer,"[34] we considered our dataset to be nonredundant. The resulting dataset contained 157 proteins in which 198 regions were labeled as targets. The total residue count in the filtered dataset was 132,709, of which 4088 were involved in CaM binding (Table I).

Several other datasets were also used at various stages of model building. A set of 154 intrinsically disordered proteins was taken from DisProt,[35] whereas a set of nonredundant globular proteins with high-resolution 3-D structures was selected by Smith et al.[36] To estimate functional characteristics of proteins, we have extracted all yeast and human proteins from Swiss-Prot.[37] Finally, to estimate the reliability of CaMBT prediction, we constructed a SwissRep dataset that consisted of 54,846 sequences obtained as a random subset of UniRef50, which is listed in the release 44.0 of Swiss-Prot. All datasets are available on request.

Predictor Architecture and Data Representation

The predictor consists of two stacked layers, one working at an amino acid level with the other at the predicted region level. Outputs from the first layer were grouped into runs of consecutive positive predictions and, using an appropriate data representation, fed into the second layer, which then estimates a probability that the whole region is CaM-binding. In the first layer, 92 sequence features were collected for each residue of the 157 selected proteins using symmetric sliding windows of length $w_{in} \in \{1, 7, 11, 21\}$. To account for the residues near protein ends, the window was allowed to expand or collapse near the N- or C-terminus, respectively. The first 20 features were the amino acid compositions only within $w_{in} \in \{11, 21\}$. Another set of features consisted of sequence complexity[38] and physicochemical properties such as net charge, total charge, and aromatic content within the same windows. The final set of features was constructed using outputs of several sequence-based predictors: hydrophobic moment,[39] secondary structure,[40] and intrinsic disorder.[41–43] The second stage inputs also contain length of the predicted region, percentage of predicted disorder in the flanking regions within $w_{flank} \in \{10, 20, 30, 40\}$ at both sides of the predicted region, overall predicted globularity of the protein, and charge/hydropathy ratio.[19] Also, unrealistically short predicted binding regions (targets) were filtered out. The predictor model is shown schematically in Figure 2.

To construct a dataset appropriate for model training, examples (vectors) corresponding to the CaM-binding residues (predicted regions in the second layer) were labeled as positives, whereas all remaining residues (false positive regions in the second layer) from the same set of proteins were labeled as negatives. In this setting, the task of a predictor is to separate the CaM-binding residues (regions) from the sequence "background."

Model Selection and Training

Because of the small size of the positive dataset, we chose to use an ensemble of 20 logistic regression models. Note that the positive dataset was effectively smaller than 4088 because of the data representation: a shift of w_{in} by one residue induces only a slight change in feature space, eventually causing a significant number of examples to be mutually dependent. Each model was trained on a balanced set using all the positive examples and a random selection of negatives, with the final prediction being an arithmetic average of all 20 models. Unpromising features were filtered out using the t-test, whereas the dimensionality was further reduced to 95% of variance using principal component analysis. Balanced training sets are known to be a good choice in cases with training data that exhibit asymmetric noise.[44] Overprediction on the majority (here negative) class can be easily addressed by changing deci-

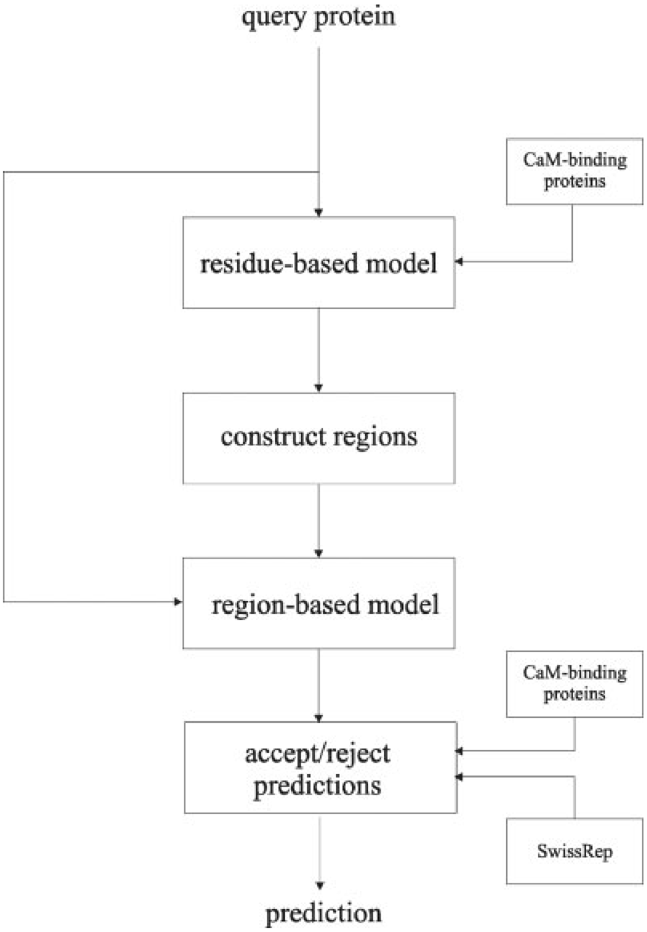

Fig. 2. CaMBT prediction mechanism. CaM-binding residues were first predicted on a query protein. After the regions are constructed, they are accepted or rejected based on the second layer. The model can also decide to refuse prediction based on the reliability model which is incorporated in the second layer.

sion thresholds or adjusting the outputs of the predictor[45] if the goal is to minimize the total number of misclassified examples.

Model evaluation was performed using a per protein leave-one-out strategy. In particular, one protein at a time was selected as a test set, whereas the remaining 156 sequences were used for the training. After the predictor was constructed, its performance was evaluated on the test protein. Each protein was used only once as a test set and the performance results were averaged over all test proteins.

Performance Measures

We measured accuracy on both per-residue and per-region levels. Per-residue performance was measured using standard definitions of sensitivity (*sn*), specificity (*sp*), and precision (*pr*). Sensitivity represents the percentage of true positives predicted to be positive (CaM-binding residues), specificity represents the percentage of true negatives predicted to be negative, whereas the precision represents the percentage of positively predicted residues that are in fact positives. Each statistic (*sn*, *sp*, *pr*) was measured for each test protein and finally averaged over all 157 sequences. In addition to *sn*, *sp*, and *pr*, we also report accuracy (*acc*) on a balanced sample, where *acc* is defined as the average of *sn* and *sp*, and visualize tradeoffs between the true (*sn*) and false ($1 - sp$) positives by plotting the receiver operating characteristic (ROC) curve. The area under the ROC curve (*AUC*) provides another useful measure of accuracy. Both *acc* and *AUC* are essentially unaffected by the disparity in class sizes.

At a per-region level, we measured sensitivity (or recall) and precision because a negative CaM-binding region cannot be defined. A situation where a predicted and a true region overlapped was considered a hit, whereas noncovered true regions were considered missed regardless of the distance of the nearest positively predicted residue. Care was taken to prevent one residue/region counted twice as a true positive. Sensitivity and precision at a protein level were estimated based on 198 available binding regions over all proteins.

Assessment of Prediction Reliability

Given a relatively small set of 157 nonredundant proteins used for the training of the CaMBT predictor, it is highly likely that the sequences used to develop CaMBT predictor are not a good representative of the protein feature space and so the resulting predictor is likely to be biased. To reduce the adverse effects of statistical inference on the residues with features underrepresented by the CaMBT training set, we developed a model to assess the reliability of each prediction outputted by the CaMBT predictor. In case this reliability predictor signals that the input is coming from an underrepresented part of the sequence space, the predictions made by the CaMBT predictor were simply disregarded. Although this resulted in the reduced coverage by the CaMBT predictor, it also allowed for an improved quality of statistical inference on the remaining residues.

To develop a reliability estimator, we constructed a SwissRep dataset and used it as an unbiased representative of the protein sequence space. The model consisted of an ensemble of 20 feed-forward neural networks with 10 hidden nodes and one output neuron, all using a sigmoidal transfer function. Each classifier was trained on a balanced set with 10,000 randomly selected data points from SwissRep (a positive set) and 10,000 randomly selected points from the 157 nonredundant CaBMT sequences (a negative set). A residue was represented by three sets of features within sliding windows of lengths $w_{in} \in \{11, 21, 41\}$. Each feature set consisted of 20 amino acid frequencies, sequence complexity,[38] and average hydropathy,[46] flexibility,[47] and coordinate values.[48] After transforming each feature to its *z*-score, the resulting set of 72 normalized features was transformed to a set of principal components retaining about 95% of variance. The reliability predictor is incorporated into the second stage of the overall predictor (Fig. 2).

It is easy to show[49] that a prediction near 0.5 by the reliability estimator indicates that the given residue is represented equally well by SwissRep and CaMBT se-

TABLE II. Analysis of the CaMBTs Whose Coordinates are Present in PDB†

Swiss-Prot id	PDB id (no. chains)	CaMBT position	Sequence identity (%)	Comments
Q14012	1ao6|a (2)	300–319	97	A300–L302, S304–A306, V310–E311, K313–D314 are in crystal contacts.
P48736	1e8z|a (1)	678–693	100	CaMBT overlaps with two α-helices at R678–N687 and R689–S705.
O88935	1px2|b (2)	123–138	99	K133 is in crystal contacts.
O97754	1tki|a (2)	318–337	96	R309, L311, H318, Q343–V348 are in crystal contacts.
P19065	1sfc|i (12)	73–92	100	Largely unfolded in solution.[51] Residues G72–L92 are in contacts with chains A, B, J, L, and K.
Q9QYF3	1w7j|a (2)	770–789	93	Residues A769–Y786 are in contact with multiple residues of myosin light chain 1. Residues 755–795 are likely disordered.[101] Last three CaMBTs were not expressed.
		793–812		
		818–837		
		866–885		
P11017	1omw|b (3)	26–45	90	K43 is in contact with chain A. I29, M31, R34–R35 are in contact with chain G
Q8IXV9	1gg3|a (3)	76–92	100	Residues M1–N209 were not expressed. S469–R471 are involved in interchain contacts.
		390–406		
		473–489		
P11023	1zbd|a (2)	51–84	96	V68–I69, D71–F72, V74, K85, Q87, and W89 are in contact with chain B.
P27322	3hsc (1)	263–283	81	R264, S282, T284 are in crystal contacts.
Q99LL8	1omw|a (3)	16–38 610–630	98	No crystal or interchain contacts at CaMBT residues.
P30301	1ymg|a (1)	223–235	92	L218–Y219, L222–F224, R226, K238 are in crystal contacts.
Q9TU34	1xzz|a (1)	49–81 106–128	99	Q102, K109–V114, Q116, G118–V120, L124, L126–K127, N129 are in crystal contacts. Region 49–81 was not expressed.
P11531	1dxx|d (4)	18–42 104–125	95	E12–D15, K18–K19, S30, G33–H36, E38–N39, F41–S42, Q45, D101–L106 are in crystal contacts.

†All coordinates are presented with respect to original Swiss-Prot or TrEMBL sequences. The first column indicates the original sequence given in the Calmodulin Target Database. The second column represents the PDB chain closest to the original sequence, and the corresponding sequence identity is in column four. Crystal contacts were calculated using CryCo program[100] with default threshold of 10 Å.

quences; lower or higher values indicate that the residue is over- or underrepresented, respectively, in CaMBT sequences. In our approach, the CaBMT predictor was not applied on the residues with reliability prediction above a given threshold.

RESULTS

CaM-binding Targets Present in PDB

In order to investigate structural properties of CaM-binding regions, we searched for the structures of known CaM-binding proteins[14] in PDB. We were specifically interested in those that were in the monomeric form and, if in a complex, not associated with CaM. We also required $\geq$70% sequence identity as a reasonable threshold for the functional inference by similarity.[34]

As mentioned in the Introduction, we found four proteins with missing atom coordinates in the place of CaM-binding regions (1aui|a, 1qqg, 1tll, 2bov|a). Such regions can be considered disordered with high confidence, as it has been argued previously that the proportion of wobbly ordered regions among all regions with missing electron density is small.[50] In another 24 sequences that matched CaM-binding proteins (1azs|a, 1byy|a, 1efn|b, 1gc1|g, 1iss, 1jkt, 1joc, 1kgd|a, 1khu, 1khx|a, 1n4k|a, 1n9d|a, 1pb7|a, 1qav|a, 1qom, 1rzj|g, 1taz|a, 1u4q, 1w7j|a, 1wgp|a, 1xi4|r, 1yfo, 2bf1|a, 3nos), the actual CaMBTs were removed from the expression construct before crystallization. This is frequently done in cases of the hard-to-crystallize fragments such as disordered regions or multidomain proteins. Finally, 14 CaM-binding proteins had CaMBTs present in the coordinate lists and were analyzed in more detail (Table II). Five of these proteins were in complexes in which CaMBTs are in direct contact with a binding partner and as such cannot interact with CaM without releasing the current partner and possibly undergoing structural change. In fact, one of these cases, synaptobrevin (1sfc|i), is an experimentally verified disordered protein in its monomeric form.[51] In seven cases, CaMBT residues were directly involved in crystal contacts and so their observed structure could not be trusted as native.

In total, out of 210 CaM-binding proteins, 42 had representatives in PDB with sufficient sequence identity, and of those, only in three cases have we found an ordered form for a CaM-binding target (1e8z|a, 1gg3, 1omw|a) that did not have CaMBT residues directly involved in crystal contacts or in interchain interactions.

Dataset Inconsistencies

Because the large majority of crystallized CaM-binding targets are known to be helical, with about 22 residues in length on the average, we have analyzed nonredundant CaMBTs available in our dataset with respect to their

TABLE III. Comparison of Sequence Properties

	CaMBTs	Non-CaMBTs	Globular
Net charge	0.1779 ± 0.0080	−0.0148 ± 0.0014	−0.0156 ± 0.0016
Total charge	0.3094 ± 0.0073	0.2517 ± 0.0013	0.2226 ± 0.0016
Hydropathy	−0.6322 ± 0.0509	−0.3886 ± 0.0086	−0.2871 ± 0.0116
Disorder score	0.5363 ± 0.0102	0.4596 ± 0.0022	0.3181 ± 0.0018
Aromatic content	0.0819 ± 0.0040	0.0774 ± 0.0008	0.0927 ± 0.0011
Helix score	5.460 ± 0.173	3.969 ± 0.031	3.088 ± 0.041
Sheet score	1.024 ± 0.099	1.320 ± 0.019	2.088 ± 0.033
Coil score	2.213 ± 0.131	3.319 ± 0.026	3.419 ± 0.035

length; 154 CaMBTs were shorter than 22 residues, whereas 44 were longer. The shortest CaMBT in our dataset was 8, and the longest one was 36 residues long. Overall, if all CaMBTs with lengths shorter than 22 residues were actually mislabeled as shorter than in truth, then such a mislabeling would produce 504 false negative residues (12.3% of the positive dataset). On the other hand, 251 residues (6.1%) of the CaM-binding residues are those that exceed 22 residues and could be false positives. Finally, because many proteins in our dataset contain multiple binding regions, it is possible that there exist many other yet nonannotated CaMBTs scattered across the set of currently available proteins. Thus, our dataset is likely to contain a certain amount of class-label noise, which poses a limitation for the CaMBT analysis and predictor performance.

Statistical Properties of CaM-binding Regions

Several sequence and physicochemical properties of the CaM-binding proteins were analyzed. Here, CaM-binding proteins were separated into CaM-binding residues and remaining (background) residues and then compared to the set of globular proteins.[36] The results of this analysis are shown in Table III.

Table III indicates that CaM-binding residues are distinct from other residues in several respects: they have significantly higher net charge and helical propensity, as observed before,[52] but also higher total charge and propensity of being intrinsically disordered. On the other hand, despite this tendency for conformational flexibility these regions also have increased aromatic content. These properties, seemingly conflicting, could be explained by the longer window sizes that were used to obtain disorder scores (41 residues), whereas the aromatic content was calculated only within CaM-binding residues. Thus, in addition to the presence of charged residues, the overall disorder propensity of CaMBTs is also influenced by the flexible flanking regions. Interestingly, the disorder score of the non-CaMBT residues was on average higher that that of globular proteins. Even though this score can partly be attributed to the CaMBT flanking regions and yet-to-be-annotated additional CaMBTs, we believe that biological functions of CaMBPs may simply require more conformational flexibility than that present in compact globular proteins.

TABLE IV. Estimated Prediction Accuracy (per residue and per region) and Standard Error of Our Predictor of CaM-binding Targets

Accuracy per residue (%)				Accuracy per region (%)	
sn	*sp*	*acc*	*pr*	*sn*	*pr*
73.2 ± 2.7	88.5 ± 0.8	80.8	29.1 ± 2.1	80.0	22.8

Predictor Performance

The overall performance of the predictor was estimated both on a per residue and per region basis (see Performance Measures). Results of this estimation, shown in Table IV, indicate that the balanced accuracy of the predictor is 80.8%. Shown in Figure 3 is an ROC curve for the prediction data. The *AUC* was calculated to be 89.1%.

To gain insight into the predictor's ability to detect an unknown motif type, we performed the following experiment: all proteins containing a particular motif type were excluded from the training set, whereas the prediction accuracy was estimated using sequences only from the left-out motifs and the remaining negative residues. Thus, residues belonging to a nonexcluded motif type were not used in the evaluation. Prediction accuracies on the unseen motifs are shown in Table V. These results indicate that our model is general, but these results also emphasize the similarity of all motif types in our feature space.

To our knowledge, computational prediction of CaM-binding targets was first proposed by DeGrado et al.,[52] who searched for the motifs of appropriate length, hydrophobicity, hydrophobic moment, and charge. The features used to develop this original model are a subset of those used herein. Another predictor of CaMBTs can be found at the Calmodulin Target Database web site.[14] It consists of hidden Markov models designed to recognize the observed classes of motifs. Hidden Markov models belong to so-called generative predictors. On the other hand, our discriminative model was trained on all types of motifs simultaneously and as such may have increased generalizability, traded off for a somewhat decreased specificity. Therefore, generative and discriminative approaches complement each other. A direct comparison between these models is not possible because the former was trained on all available sequences.

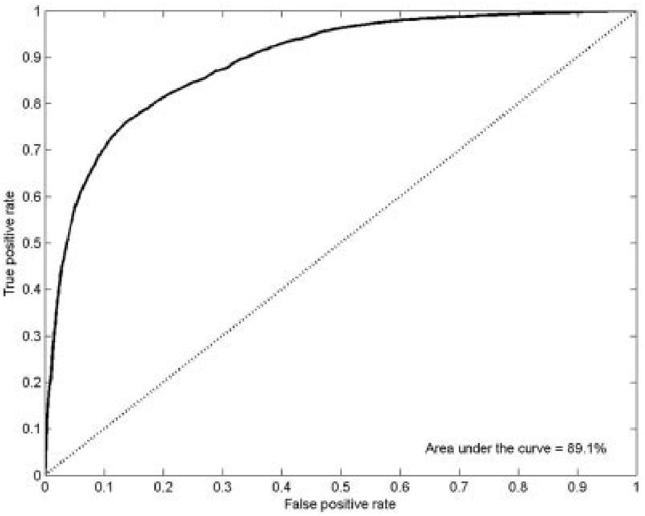

Fig. 3. The receiver operating characteristic (ROC curve) of the CaMBT predictor (solid line) and the ROC curve of a random predictor (dotted line). The area under the curve (*AUC*) was calculated to be 89.1%. The *x*-axis represents the false positive rate (1 − *sp*), whereas the *y*-axis represents the true positive rate (*sn*).

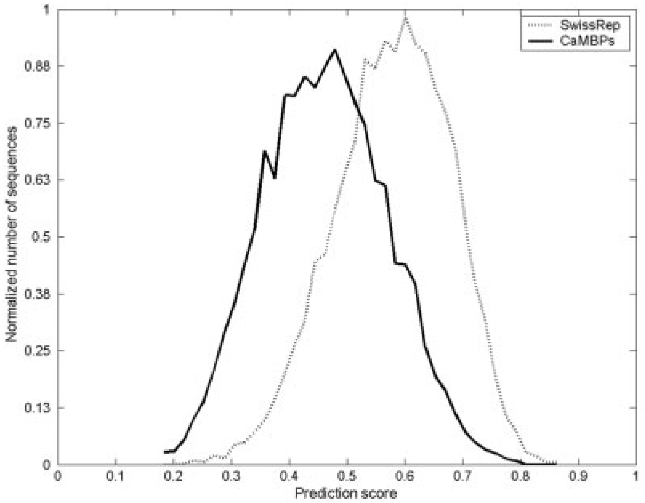

Fig. 4. Estimated probability densities of scores for CaMBPs (solid line) and SwissRep dataset (dotted line) of the reliability model. Reliability scores are assigned on a per-residue basis, thus a CaMBT prediction is accepted only for reliability scores below 0.598.

TABLE V. Prediction Accuracy of CaM-binding Site Predictor when All Proteins Containing a Specific Motif Type Were Excluded from Training†

Motif	Prediction accuracy (%)[a]				No. of
type	*sn*	*sp*	*acc*	*pr*	proteins
1–10	75.3 ± 9.0	88.2 ± 15.4	81.8	24.9 ± 5.8	12
1–12	85.0 ± 9.2	71.5 ± 15.4	78.3	58.5 ± 13.8	6
1–14	80.6 ± 4.9	87.2 ± 1.6	83.9	25.1 ± 4.5	38
1–16	95.7	96.1	95.9	53.7	1
IQ	72.9 ± 6.6	86.0 ± 1.7	79.4	17.8 ± 3.7	35
Other	59.6 ± 4.0	92.2 ± 0.6	75.9	30.4 ± 3.0	74

†Only the residues of this motif type and negative residues were used for testing. Note that standard error is greatly influenced by the number of proteins containing particular motifs.

[a]Values are means ± SE.

Analysis of Reliability

The developed reliability assessment model was applied on all residues of SwissRep and CaM-binding proteins. The estimated distributions of reliability scores are shown in Figure 4. By selecting the false negative rate (number of CaMBP residues rejected by the reliability assessment model) of 10, 5, and 1%, with rejection thresholds of 0.598, 0.636, and 0.700, a total of 41.8, 28.4, and 9.4% residues from SwissRep were selected as underrepresented by CaMBP training sequences. The significant difference in the distribution of the reliability scores among SwissRep and CaMBPs confirmed the anticipated bias of CaMBPs compared to the overall distribution of the protein feature space and validated the use of the reliability estimator in conjunction with our CaMBT predictor.

Application of the Predictor

In this section we present two illustrative examples that support our main hypothesis and demonstrate the usefulness of the CaMBT predictor. First, we discuss in some detail the interaction of CaM with one of its binding partners, caldesmon (CaD), and then we focus on how our predictor can be utilized as a companion to proteomics experiments.

We selected CaD as an illustrative example to show that even proteins that are disordered over their entire lengths can be associated with important functions and that these functions can be modulated by CaM binding. CaD interacts with actin, tropomyosin,[53] and CaM at multiple sites along its amino acid sequence.[54–57] For example, chicken gizzard CaD contains at least four CaM-binding sites. One of them was assigned to a long N-terminal region 26–199, which also binds myosin,[58] whereas C-terminal domain was shown to contain three additional CaM-binding sites, centers A (close to Trp674), B (close to Trp707), and B′ (close to Trp737).[54] Only the first of these three sites was originally present in our positive dataset, whereas others were included as negatives (as they were not present in the Calmodulin Target Database).

CaD has been shown experimentally to be mostly if not entirely disordered.[59] In agreement with this experimental data, CaD is predicted to be disordered over its entire length by the VL3 order/disorder predictor [Fig. 5(a), dotted line]. The CaMBT predictor identifies several potential CaMBTs in CaD [Fig. 5(a), solid line]. Importantly, all three experimentally verified CaMBTs at the C-terminal part of CaD are recognized by the CaMBT predictor as CaM-binding sites. The N-terminal region of CaD (residues 26–199) is too large to be considered as a single CaMBT. However, three CaM-binding regions are predicted within this long stretch, suggesting that this long previously identified binding region possibly contains three specific target regions.

Because all three C-terminal CaMBTs of CaD are correctly predicted and the exact locations of the N-terminal CaMBTs are unknown, it seems reasonable to consider the

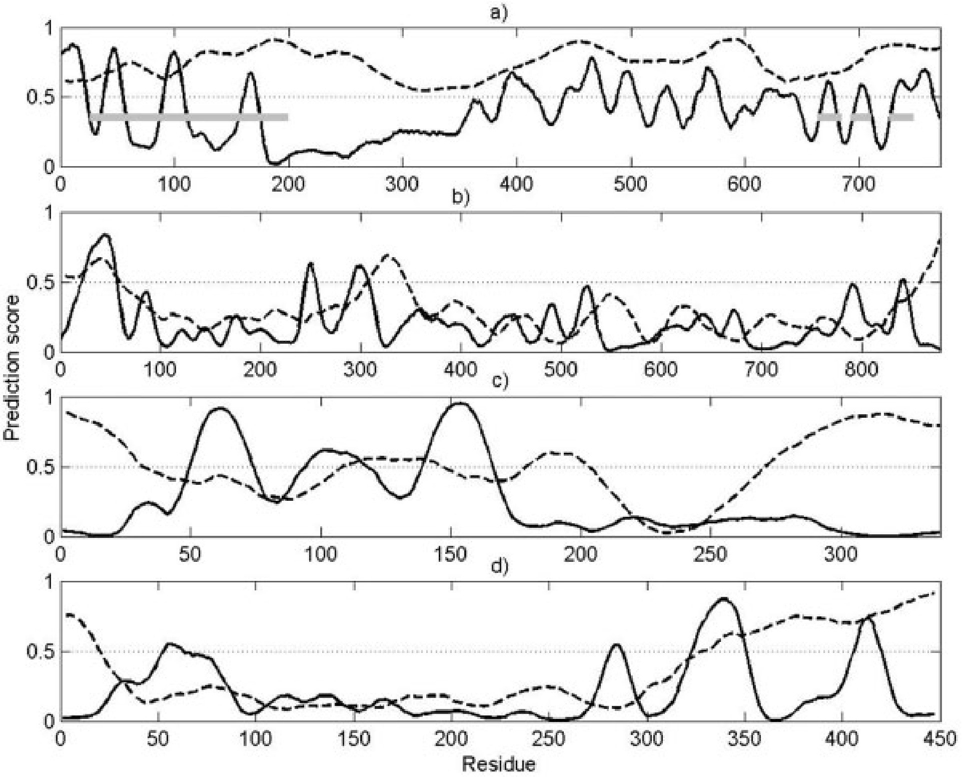

Fig. 5. CaMBT prediction (solid line) and disorder prediction (dotted line) for four selected CaMBPs. (**a**) Chicken caldesmon (sp:P12957); thick grey lines indicate experimentally verified regions; the first and third C-terminal CaMBTs were visualized ±10 residues around the known aromatic residue; (**b**) yeast msh4, IQ motif predicted in Ref. 60 is at 380–388; (**c**) yeast reg2, IQ motif predicted in Ref. 60 is at 186–194; (**d**) yeast cmk2, IQ motif predicted in Ref. 60 is at 385–393. Proteins in (**b**–**d**) were selected from the study by Zhu et al.[60] The high predictions of CaM-binding sites are frequently correlated with high predictions of intrinsic disorder.

per residue sensitivity of prediction on CaD to be very high. Per residue specificity, on the other hand, is about 67%, assuming that there are no additional CaMBTs present in this protein. On a binding region level, sensitivity remains high, whereas precision is at least 23% in case that none of the three N-terminal predictions is correct (3 of 13). Thus, both per residue and per region accuracies agree well with our estimates from the Predictor Performance section.

Our second application is related to a proteome-wide study by Zhu et al.,[60] in which the authors detected 37 putative CaMBPs using protein chips. Only six of these proteins had previously been experimentally verified. For 14 of the detected proteins, the authors computationally found possible IQ-motifs using sequence alignments, whereas other hypothesized regions of interactions were not provided. Out of all 37 proteins, only two shared >30% global sequence identity with any of the proteins or peptides in our training data.

Of 14 proteins hypothesized by Zhu et al.[60] to contain IQ motifs, 13 are predicted as positives by our model. We predict only four motifs to be at the positions found by Zhu et al. (sps19, ipp1, rpl26b, cmk1), whereas in several other proteins (e.g., reg2, cmk2, rpb3, myo4) the predicted regions (with high scores) were shifted from the sites suggested by the authors [Fig. 5(b–d)] and, we believe, may not be IQ motifs. On the other hand, a motif search algorithm available at the Calmodulin Target Database web site identifies only two target regions (in msh4 and cmk1) that overlap with those found by Zhu et al.

On the basis of these results we suggest here that a combined use of generative and discriminative computational approaches may be beneficial in detecting CaM-binding partners. Such predictions could be especially useful when combined with laboratory experiments.

Analysis of Function of Putative CaMBPs

We applied our predictor to complete yeast (data not shown) and human proteomes available in Swiss-Prot. Each region that had a prediction score over 0.75 and reliability index below 0.598 was considered positive. These thresholds provide a smaller number of false positives and significantly reduce the length dependence in the target set. A set of function-related keywords was collected for each protein in two nonoverlapping sets: a set of putative target proteins and a set of the remaining proteins. Then, a p-value was calculated such that the functional category was over- or under-represented in the putative CaM-binding proteins. In addition to the p-value, we also calculated the difference and relative difference in

TABLE VI. Overrepresented (15 highest ranked) and underrepresented (10 highest ranked) Swiss-Prot keywords for putative CaMBPs in *H. sapiens*[†]

Category	p-value	Difference* (%)
Overrepresented in CaMBPs		
Nuclear protein	$1.1 \cdot 10^{-45}$	15.3 (56.9)
DNA binding	$2.1 \cdot 10^{-42}$	9.9 (63.3)
Coiled coil	$6.0 \cdot 10^{-40}$	5.0 (66.8)
Transcription regulation	$1.7 \cdot 10^{-39}$	7.9 (56.9)
Homeobox	$5.7 \cdot 10^{-38}$	3.0 (99.4)
Alternative splicing	$1.1 \cdot 10^{-37}$	10.5 (37.4)
Chromosomal protein	$5.0 \cdot 10^{-34}$	1.0 (97.9)
Developmental protein	$5.1 \cdot 10^{-34}$	2.6 (66.0)
Phosphorylation	$1.0 \cdot 10^{-33}$	5.5 (40.7)
Activator	$1.5 \cdot 10^{-33}$	2.7 (71.0)
ATP binding	$5.5 \cdot 10^{-33}$	5.1 (54.1)
Ribosomal protein	$2.1 \cdot 10^{-30}$	1.5 (81.8)
Repressor	$6.1 \cdot 10^{-28}$	1.6 (68.6)
Helicase	$4.6 \cdot 10^{-27}$	0.9 (82.4)
RNA binding	$1.9 \cdot 10^{-26}$	2.1 (58.9)
Underrepresented in CaMBPs		
Immunoglobulin V region	$6.8 \cdot 10^{-40}$	−2.7 (98.0)
Direct protein sequencing	$3.9 \cdot 10^{-35}$	−7.4 (38.6)
Lipoprotein	$1.0 \cdot 10^{-34}$	−3.1 (57.8)
Transmembrane	$1.6 \cdot 10^{-31}$	−7.5 (24.1)
Signal	$2.4 \cdot 10^{-30}$	−7.6 (29.7)
Glycoprotein	$4.6 \cdot 10^{-28}$	−6.5 (22.6)
GPI-anchor	$4.8 \cdot 10^{-26}$	−1.0 (74.6)
Keratin	$4.4 \cdot 10^{-25}$	−1.1 (76.3)
3D structure	$1.4 \cdot 10^{-24}$	−3.1 (22.9)
Pyrrolidone carboxylic acid	$1.5 \cdot 10^{-24}$	−1.1 (74.6)

[†]The table shows (i) the p-value that a particular keyword is equally likely for putative CaMBPs and remaining human proteins; (ii) the difference in percentages in the two groups, and (iii) an absolute relative difference in percentage. The relative difference, presented in parentheses, is normalized by the maximum between the two percentages.

the fractions of proteins containing a particular keyword between the two sets (Table VI).

DISCUSSION

Limitations to Model Accuracy and Future Improvements

Constructing a model and estimating its accuracy in the case of CaM-binding region predictor is difficult primarily because of the dataset noise. The major sources of noise are stemming from the difficulties in precisely determining CaM-binding regions and the possibility of multiple, non-annotated binding targets in the interacting protein. Thus, the ability to reduce mislabeled residues would quite likely contribute to an even higher prediction accuracy. Other limitations to the performance of our model are caused by the data representation and the exclusion of long-range interactions in predicting binding sites. In addition to denoising, further improvement in predicting CaM-binding targets, especially reducing the false positives, can be achieved by including other information sources such as protein network data and evolutionary and genomic information. Finally, integrating discriminative with generative prediction models would contribute to the reduction of false negatives.

Interaction of CaM with CaD

CaD has multiple CaM-binding sites so gaining further insight into the structure-function relationships of CaD and its interactions with CaM can serve as a model for the structures and functions of CaM target proteins.

CaD is a ubiquitous actin-binding protein involved in the regulation of smooth muscle contraction, nonmuscle motility, and cytoskeleton formation.[58,61,62] It was originally discovered in chicken gizzard smooth muscle as a protein that binds to CaM in a Ca^{2+}-dependent manner and to filamentous actin (F-actin) in a Ca^{2+}-independent manner.[63] Subsequently it was identified in different smooth muscle tissues and in a variety of nonmuscle cells and cell cultures (for reviews see Refs. 61, 62, 64, 65). CaD controls a thin-filament-linked regulation of smooth muscle contraction through its specific binding to F-actin and F-actin–tropomyosin with a concomitant inhibition of the actin-stimulated myosin ATPase.[58] The CaD action is reversed by interaction with several Ca^{2+}-dependent proteins, including CaM and caltropin. Furthermore, the functional activity of CaD is regulated by multiple phosphorylation,[62,66] and by calcium via Ca^{2+}-binding proteins, with CaD being alternatively bound either to F-actin or to CaM, depending upon the calcium concentration. This thin filament-based modulatory effect is assumed to provide additional "fine-tuning" to the well-established, myosin light chain phosphorylation-dependent, thick filament-based regulation of smooth muscle contraction.[67]

Structurally, a large hydrodynamic radius and a coil-like shape of the far-UV circular dichroism spectrum[59] suggest that CaD belongs to the family of natively disordered proteins. Functionally, CaD interacts specifically with numerous binding partners and could be divided into four independent domains: (i) the N-terminal domain that interacts with myosin and tropomyosin; (ii) the second domain that participates in the binding of tropomyosin; (iii) the third domain that is involved in the interaction with myosin, tropomyosin, and actin; and (iv) the C-terminal domain that plays the most important role in the functioning of CaD, interacting with actin, Ca^{2+}-binding proteins, myosin, tropomyosin, and phospholipids. Interestingly, the identified CaM-binding sites in the C-terminal region are correctly predicted and the long N-terminal region associated with CaM binding is predicted to be subdivided into three distinct CaM-binding regions [Fig. 5(a)]. These predictions could be useful for further experimental work.

Other Functions of CaM-binding Partners

Table VI indicates functional differences between putative CaM-binding proteins and the remaining proteins from *Homo sapiens*. Several important functional classes, such as transcription regulation, nucleic acid binding, kinases, phosphatases, and others are briefly discussed below.

Experimental research suggests that CaM is actively involved in transcription regulation by both directly binding transcription factors[68] or via various kinases and phosphatases.[69] These mechanisms are especially complex in plants where CaM is expressed in many isoforms. Furthermore, CaM can affect nuclear processes by entering the nucleus and interacting with DNA and/or nuclear proteins. In fact, recent reports suggest that, in addition to its familiar functions in the cytoplasm, CaM may participate in rapid signaling between cytoplasm and nucleus,[70] modulating nuclear processes indirectly via the active transport of many Ca^{2+} signals to the nucleus.[71] These facts can be used to explain the high abundance in the CaMBPs of such annotations as nuclear protein, regulation of transcription, chromosomal protein, DNA-binding, homeobox, and phosphorylation (Table VI). On the other hand, CaM has also been shown to be involved in regulation of cell growth and proliferation. A related observation is that the abundance of CaM increases significantly in cells undergoing division and differentiation.[9] This is consistent with the high representation of developmental proteins in the set of CaMBPs.

The high prediction of CaMBTs among the ribosomal proteins came as a surprise. On the one hand, because of the high positive charge of these proteins, it may easily be argued that the presence of ribosomal proteins is an artifact from the similar high positive charge of CaMBTs and so ribosomal proteins may simply belong to the set of false-positive predictions. On the other hand, it has been pointed out that the real biological role of ribosomal proteins is far from being completely understood. In fact, functions traditionally assigned to ribosomal proteins are the facilitation of the rRNA folding and the maintenance of an optimal ribosomal configuration, providing protein biosynthesis with the required speed and accuracy. However, there is evidence that a number of ribosomal proteins have moonlighting functions[72] apart from both the ribosome and protein synthesis.[73] These extraribosomal functions of these proteins include, but are not limited to, replication, transcription, RNA processing, DNA repair, autogenous regulation of translation, regulation of development, and malignant transformation. Furthermore, ribosomal proteins were reported to possess DNA-binding motifs such as the zinc finger,[74] the bZIP domain,[75] and the helix-turn-helix motif.[76] Recent findings by Mazumder et al.[77] confirmed the idea that much is yet to be understood about the function of ribosomal proteins. In particular, it has been shown that phosphorylated human L13a is able to inhibit ceruloplasmin at the RNA translation level. Thus, the abundance of CaMBT-like sites in ribosomal proteins found in our study may be related to the regulation of the extraribosomal activities of ribosomal proteins.

Many of the keywords that are overrepresented in putative CaMBPs correlate strongly with intrinsically disordered proteins. One such example is repeat regions. Such regions are known not only to carry out important protein functions,[33] but repeat region expansion may represent an important evolutionary mechanism for disordered proteins.[78] A second example is phosphorylation. Iakoucheva et al.[79] showed that phosphorylation sites are significantly more likely to occur in disordered than in ordered regions. The connection between the functions described by the keywords and CaM binding can be simple, with the function and CaM binding occurring in the same region of sequence, or the connection can be more complex, with different parts of the protein being responsible for the two functions.

CaM Signaling from a Structural Perspective

The available X-ray and NMR data suggest that several CaM-binding regions[80–83] and isolated peptides[28,29] are intrinsically disordered in their CaM-free state. In addition, other domains containing CaM-binding regions have been shown to undergo rapid protease digestion,[22–26] which is a hallmark of unfolded proteins.[27] Our analysis of sequence properties of a large number of nonredundant CaM-binding targets and the existence of only a small number of structured CaMBTs indicates that an unstructured form of these regions and/or its flanking residues is a rule, rather than an exception. This observation is further supported by the facts that the common binding mechanism between CaM and its partner requires steric accessibility to accommodate the interaction. Thus, CaM–CaMBT coupling is highly unlikely to occur within a static globular domain. Although it is possible that a globular domain could unfold before its association with CaM binding, the occurrence of such an event would be unlikely unless the globular domain were very unstable to begin with, such as a molten globular form. Finally, a scenario where an ordered α-helical segment moves away from a folded domain to interact with CaM is not very feasible unless, for example, the binding energy were very low. Such a mechanism would be facilitated if the binding helix were connected to the rest of the protein by a flexible tether.

The disorder-to-helix conformational change upon binding as exemplified by the CaMBTs likely represents a very common theme for signaling proteins in eukaryotic cells. For example, p53 has two such helix-forming binding regions: (i) a natively unfolded region near the amino terminus[84] becomes a helix upon binding to Mdm2,[85] and (ii) an unstructured region near the carboxyl terminus[86] likewise becomes helical upon binding to $S100B^{\beta\beta}$.[87] An analysis of several other experimentally verified disordered proteins provides 24 additional examples of disorder-to-helix conformational changes upon complex formation.[88]

The existence of helical molecular recognition elements has been observed in several cases to correspond to short, sharp dips of predicted order within longer regions of predicted disorder.[89] The VLXT predictor appears, at least so far, to be the most useful among the disorder predictors for identifying such binding sites. Indeed, two binding sites predicted by dips in the VLXT plots were subsequently confirmed experimentally.[90,91] Also, both of these regions were shown to adopt helical conformations upon binding to their partners[92] (B.F. Luisi, personal communication). We recently formalized the analysis of dips in VLXT plots as indicators of α-helix-forming molecular recognition elements and applied this algorithm to a set of signaling proteins: nearly half of the signaling proteins were indicated to contain segments that potentially correspond to regions that undergo disorder-to-helix changes upon complex formation.[93]

In light of these findings, the high predictions of helicity, high aromatic content, and yet high predictions of long disorder within CaMBTs coincide with the above-mentioned model of signaling in which CaM-binding targets are likely to undergo coupled binding and folding.[20,21,33,94–96] In addition to disorder-to-helix, disorder-to-sheet and disorder-to-irregular transitions have also been observed (reviewed in Ref. 97). This coupled binding and folding mechanism does not fit the simple lock-and-key idea[98] or the induced-fit hypothesis as originally proposed,[99] but requires some sort of extension of the latter such as "extreme induced fit." [94] By whichever terminology is used to describe such associations, this mechanism adds to the versatility and complexity of protein-protein interactions.

ACKNOWLEDGMENTS

We thank Giselle M. Knudsen and anonymous reviewers for their useful comments.

REFERENCES

1. Van Eldik LJ, Watterson DM, editors. Calmodulin and signal transduction; San Diego: Academic Press; 1998.
2. Stull JT. Ca2+-dependent cell signaling through calmodulin-activated protein phosphatase and protein kinases minireview series. J Biol Chem 2001;276(4):2311–2312.
3. Vetter SW, Leclerc E. Novel aspects of calmodulin target recognition and activation. Eur J Biochem 2003;270(3):404–414.
4. Barbato G, Ikura M, Kay LE, Pastor RW, Bax A. Backbone dynamics of calmodulin studied by 15N relaxation using inverse detected two-dimensional NMR spectroscopy: the central helix is flexible. Biochemistry 1992;31(23):5269–5278.
5. Wall ME, Clarage JB, Phillips GN. Motions of calmodulin characterized using both Bragg and diffuse X-ray scattering. Structure 1997;5(12):1599–1612.
6. Finn BE, Evenas J, Drakenberg T, Waltho JP, Thulin E, Forsen S. Calcium-induced structural changes and domain autonomy in calmodulin. Nat Struct Biol 1995;2(9):777–783.
7. Babu YS, Sack JS, Greenhough TJ, Bugg CE, Means AR, Cook WJ. Three-dimensional structure of calmodulin. Nature 1985; 315(6014):37–40.
8. Babu Y, Bugg CE, Cook WJ. Structure of calmodulin refined at 2.2 Å resolution. J Mol Biol 1988;203:191–204.
9. Chin D, Means AR. Calmodulin: a prototypical calcium sensor. Trends Cell Biol 2000;10(8):322–328.
10. Hook SS, Means AR. Ca(2+)/CaM-dependent kinases: from activation to function. Annu Rev Pharmacol Toxicol 2001;41:471–505.
11. Klee CB, Ren H, Wang X. Regulation of the calmodulin-stimulated protein phosphatase, calcineurin. J Biol Chem 1998; 273(22):13367–13370.
12. Osawa M, Swindells MB, Tanikawa J, Tanaka T, Mase T, Furuya T, Ikura M. Solution structure of calmodulin-W-7 complex: the basis of diversity in molecular recognition. J Mol Biol 1998;276(1): 165–176.
13. Albrecht K, Hart J, Shaw A, Dunker AK. Quaternion contact ribbons: a new tool for visualizing intra- and intermolecular interactions in proteins. Pac Symp Biocomput 1996;1:41–52.
14. Yap K, Kim J, Truong K, Sherman M, Yuan T, Ikura M. Calmodulin target database. J Struct Funct Genomics 2000;1:8–14.
15. Osawa M, Tokumitsu H, Swindells MB, Kurihara H, Orita M, Shibanuma T, Furuya T, Ikura M. A novel target recognition revealed by calmodulin in complex with Ca2+-calmodulin-dependent kinase kinase. Nat Struct Biol 1999;6(9):819–824.
16. Yuan T, Vogel HJ, Sutherland C, Walsh MP. Characterization of the Ca2+-dependent and -independent interactions between calmodulin and its binding domain of inducible nitric oxide synthase. FEBS Lett 1998;431(2):210–214.
17. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 1999;293:321–331.
18. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z. Intrinsically disordered protein. J Mol Graph Model 2001;19(1):26–59.
19. Uversky VN. What does it mean to be natively unfolded? Eur J Biochem 2002;269(1):2–12.
20. Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci 2002;27:527–533.
21. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6(3):197–208.
22. Depaoli-Roach AA, Gibbs JB, Roach PJ. Calcium and calmodulin activation of muscle phosphorylase kinase: effect of tryptic proteolysis. FEBS Lett 1979;105(2):321–324.
23. Toscano WA Jr, Westcott KR, LaPorte DC, Storm DR. Evidence for a dissociable protein subunit required for calmodulin stimulation of brain adenylate cyclase. Proc Natl Acad Sci USA 1979; 76(11):5582–5586.
24. Tucker MM, Robinson JB Jr, Stellwagen E. The effect of proteolysis on the calmodulin activation of cyclic nucleotide phosphodiesterase. J Biol Chem 1981;256(17):9051–9058.
25. Meijer L, Guerrier P. Activation of calmodulin-dependent NAD+ kinase by trypsin. Biochim Biophys Acta 1982;702(1):143–146.
26. Manalan AS, Klee CB. Activation of calcineurin by limited proteolysis. Proc Natl Acad Sci USA 1983;80:4291–4295.
27. Fontana A, de Laureto PP, de Filippis V, Scaramella L, Zambonin M. Limited proteolysis in the study of protein conformation. In: Sterchi EE, Stöcker W, editors. Proteolytic enzymes: tool and targets. Springer-Verlag; 1999. p 257–284.
28. O'Neil KT, Wolfe HR Jr, Erickson-Viitanen S, DeGrado WF. Fluorescence properties of calmodulin-binding peptides reflect alpha-helical periodicity. Science 1987;236(4807):1454–1456.
29. Krueger JK, Bishop NA, Blumenthal DK, Zhi G, Beckingham K, Stull JT, Trewhella J. Calmodulin binding to myosin light chain kinase begins at substoichiometric Ca2+ concentrations: a small-angle scattering study of binding and conformational transitions. Biochemistry 1998;37(51):17810–17817.
30. Pappu RV, Srinivasan R, Rose GD. The Flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. Proc Natl Acad Sci USA 2000;97(23):12565–12570.
31. Pappu RV, Rose GD. A simple model for polyproline II structure in unfolded states of alanine-based peptides. Protein Sci 2002; 11(10):2437–2455.
32. Rose GD, Richards FM, Eisenberg DS, Kuriyan J, editors. Unfolded proteins. Volume 62, Advances in protein chemistry. New York: Academic Press; 2002.
33. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. Biochemistry 2002; 41(21):6573–6582.
34. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y. Automatic prediction of protein function. Cell Mol Life Sci 2003;60(12):2637–2650.
35. Vucetic S, Obradovic Z, Vacic V, Radivojac P, Peng K, Iakoucheva LM, Cortese MS, Lawson JD, Brown CJ, Sikes JG, Newton CD, Dunker AK. DisProt: a database of protein disorder. Bioinformatics 2005;21(1):137–140.
36. Smith DK, Radivojac P, Obradovic Z, Dunker AK, Zhu G. Improved amino acid flexibility parameters. Protein Sci 2003;12: 1060–1072.
37. Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, Pilbout S, Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31:365–370.
38. Wootton JC, Federhen S. Analysis of compositionally biased regions in sequence databases. Methods Enzymol 1996;266:554–571.
39. Eisenberg D, Weiss RM, Terwilliger TC. The hydrophobic moment detects periodicity in protein hydrophobicity. Proc Natl Acad Sci USA 1984;81:140–144.

40. Rost B, Sander C, Schneider R. PHD—an automatic mail server for protein secondary structure prediction. Comput Appl Biosci (CABIOS) 1994;10(1):53–60.
41. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. Proteins 2001;42:38–48.
42. Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK. Predicting intrinsic disorder from amino acid sequence. Proteins 2003;53(S6):566–572.
43. Vucetic S, Brown CJ, Dunker AK, Obradovic Z. Flavors of protein disorder. Proteins 2003;52:573–584.
44. Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. Proceedings of the 11th Annual Conference on Computational Learning Theory; Madison, WI, 1998. p 92–100.
45. Vucetic S, Obradovic Z. Classification on data with biased class distribution. Proceedings of the 12th European Conference on Machine Learning, Freiburg, Germany. 2001. p 527–538.
46. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132.
47. Vihinen M, Torkkila E, Riikonen P. Accuracy of protein flexibility predictions. Proteins 1994;19:141–149.
48. Galaktionov SG, Marshall GR. Prediction of protein structure in terms of intraglobular contacts: 1D to 2D to 3D. In: States DJ, Agarwal P, Gaasterland T, Hunter L, Smith RF, editors. Proc Int Conf Intell Syst Mol Biol; June 12–15, 1996; Washington University Institute for Biomedical Computing, Center for Molecular Design, St. Louis, MO. AAAI Press. 1996. p 42.
49. Peng K, Obradovic Z, Vucetic S. Exploring bias in the Protein Data Bank using contrast classifiers. Pac Symp Biocomput 2004:435–446.
50. Garner E, Cannon P, Romero P, Obradovic Z, Dunker AK. Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization. Genome Inform Ser Workshop Genome Inform 1998;9:201–213.
51. Hazzard J, Sudhof TC, Rizo J. NMR analysis of the structure of synaptobrevin and of its interaction with syntaxin. J Biomol NMR 1999;14:203–207.
52. DeGrado WF, Erickson-Viitanen S, Wolfe HR Jr, O'Neil KT. Predicted calmodulin-binding sequence in the gamma subunit of phosphorylase b kinase. Proteins 1987;2(1):20–33.
53. Fraser ID, Copeland O, Bing W, Marston SB. The inhibitory complex of smooth muscle caldesmon with actin and tropomyosin involves three interacting segments of the C-terminal domain 4. Biochemistry 1997;36(18):5483–5492.
54. Gusev NB. Some properties of caldesmon and calponin and the participation of these proteins in regulation of smooth muscle contraction and cytoskeleton formation. Biochemistry (Mosc) 2001;66(10):1112–1121.
55. Huber PA, El-Mezgueldi M, Grabarek Z, Slatter DA, Levine BA, Marston SB. Multiple-sited interaction of caldesmon with Ca(2+)-calmodulin. Biochem J 1996;316(Pt 2):413–420.
56. Medvedeva MV, Kolobova EA, Huber PA, Fraser ID, Marston SB, Gusev NB. Mapping of contact sites in the caldesmon-calmodulin complex. Biochem J 1997;324(Pt 1):255–262.
57. Wang E, Zhuang S, Kordowska J, Grabarek Z, Wang CL. Calmodulin binds to caldesmon in an antiparallel manner. Biochemistry 1997;36(48):15026–15034.
58. Marston SB, Redwood CS. The molecular anatomy of caldesmon. Biochem J 1991;279(Pt 1):1–16.
59. Lynch WP, Riseman VM, Bretscher A. Smooth muscle caldesmon is an extended flexible monomeric protein in solution that can readily undergo reversible intra- and intermolecular sulfhydryl cross-linking. A mechanism for caldesmon's F-actin bundling activity. J Biol Chem 1987;262(15):7429–7437.
60. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T, Mitchell T, Miller P, Dean RA, Gerstein M, Snyder M. Global analysis of protein activities using proteome chips. Science 2001;293(5537):2101–2105.
61. Sobue K, Sellers JR. Caldesmon, a novel regulatory protein in smooth muscle and nonmuscle actomyosin systems. J Biol Chem 1991;266(19):12115–12118.
62. Matsumura F, Yamashiro S. Caldesmon. Curr Opin Cell Biol 1993;5(1):70–76.
63. Sobue K, Muramoto Y, Fujita M, Kakiuchi S. Purification of a calmodulin-binding protein from chicken gizzard that interacts with F-actin. Proc Natl Acad Sci USA 1981;78(9):5652–5655.
64. Huber PA. Caldesmon. Int J Biochem Cell Biol 1997;29(8-9):1047–1051.
65. Dabrowska R, Kulikova N, Gagola M. Nonmuscle caldesmon: its distribution and involvement in various cellular processes. Review article. Protoplasma 2004;224(1–2):1–13.
66. Morgan KG, Gangopadhyay SS. Invited review: cross-bridge regulation by thin filament-associated proteins. J Appl Physiol 2001;91(2):953–962.
67. Adelstein RS, Eisenberg E. Regulation and kinetics of the actin-myosin-ATP interaction. Annu Rev Biochem 1980;49:921–956.
68. Bouche N, Scharlat A, Snedden W, Bouchez D, Fromm H. A novel family of calmodulin-binding transcription activators in multicellular organisms. J Biol Chem 2002;277(24):21851–21861.
69. Luan S, Kudla J, Rodriguez-Concepcion M, Yalovsky S, Gruissem W. Calmodulins and calcineurin B-like proteins: calcium sensors for specific signal response coupling in plants. Plant Cell 2002;14 Suppl:S389–S400.
70. Liao B, Paschal BM, Luby-Phelps K. Mechanism of Ca2+-dependent nuclear accumulation of calmodulin. Proc Natl Acad Sci USA 1999;96(11):6217–6222.
71. Szymanski DB, Liao B, Zielinski RE. Calmodulin isoforms differentially enhance the binding of cauliflower nuclear proteins and recombinant TGA3 to a region derived from the Arabidopsis Cam-3 promoter. Plant Cell 1996;8(6):1069–1077.
72. Jeffery CJ. Molecular mechanisms for multitasking: recent crystal structures of moonlighting proteins. Curr Opin Struct Biol 2004;14(6):663–668.
73. Wool IG. Extraribosomal functions of ribosomal proteins. Trends Biochem Sci 1996;21(5):164–165.
74. Chan YL, Suzuki K, Olvera J, Wool IG. Zinc finger-like motifs in rat ribosomal proteins S27 and S29. Nucleic Acids Res 1993;21(3):649–655.
75. Chan YL, Olvera J, Gluck A, Wool IG. A leucine zipper-like motif and a basic region-leucine zipper-like element in rat ribosomal protein L13a. Identification of the tum-transplantation antigen P198. J Biol Chem 1994;269(8):5589–5594.
76. Rice PA, Steitz TA. Ribosomal protein L7/L12 has a helix-turn-helix motif similar to that found in DNA-binding regulatory proteins. Nucleic Acids Res 1989;17(10):3757–3762.
77. Mazumder B, Sampath P, Seshadri V, Maitra RK, DiCorleto PE, Fox PL. Regulated release of L13a from the 60S ribosomal subunit as a mechanism of transcript-specific translational control. Cell 2003;115(2):187–198.
78. Tompa P. Intrinsically unstructured proteins evolve by repeat expansion. Bioessays 2003;25(9):847–855.
79. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res 2004;32(3):1037–1049.
80. Kissinger CR, Parge HE, Knighton DR, Lewis CT, Pelletier LA, Tempczyk A, Kalish VJ, Tucker KD, Showalter RE, Moomaw EW, Gastinel LN, Habuka N, Chen X, Maldanado F, Barker JE, Bacquet R, Villafranca JE. Crystal structures of human calcineurin and the human FKBP12-FK506-calcineurin complex. Nature 1995;378:641–644.
81. Garcin ED, Bruns CM, Lloyd SJ, Hosfield DJ, Tiso M, Gachhui R, Stuehr DJ, Tainer JA, Getzoff ED. Structural basis for isozyme-specific regulation of electron transfer in nitric-oxide synthase. J Biol Chem 2004;279(36):37918–37927.
82. Dhe-Paganon S, Ottinger EA, Nolte RT, Eck MJ, Shoelson SE. Crystal structure of the pleckstrin homology-phosphotyrosine binding (PH-PTB) targeting region of insulin receptor substrate 1. Proc Natl Acad Sci USA 1999;96(15):8378–8383.
83. Holbourn KP, Sutton JM, Evans HR, Shone CC, Acharya KR. Molecular recognition of an ADP-ribosylating Clostridium botulinum C3 exoenzyme by RalA GTPase. Proc Natl Acad Sci USA 2005;102(15):5357–5362.
84. Dawson R, Muller L, Dehner A, Klein C, Kessler H, Buchner J. The N-terminal domain of p53 is natively unfolded. J Mol Biol 2003;332(5):1131–1141.
85. Kussie PH, Gorina S, Marechal V, Elenbaas B, Moreau J, Levine AJ, Pavletich NP. Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain [comment]. Science 1996;274(5289):948–953.
86. Bell S, Klein C, Muller L, Hansen S, Buchner J. p53 contains

large unstructured regions in its native state. J Mol Biol 2002; 322(5):917–927.

87. Rustandi RR, Baldisseri DM, Weber DJ. Structure of the negative regulatory domain of p53 bound to S100B(bb). Nat Struct Biol 2000;7(7):570–574.

88. Fuxreiter M, Simon I, Friedrich P, Tompa P. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. J Mol Biol 2004;338(5):1015–1026.

89. Garner E, Romero P, Dunker AK, Brown C, Obradovic Z. Predicting binding regions within disordered proteins. Genome Inform Ser Workshop Genome Inform 1999;10:41–50.

90. Bourhis JM, Johansson K, Receveur-Brechot V, Oldfield CJ, Dunker KA, Canard B, Longhi S. The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner. Virus Res 2004;99(2):157–167.

91. Callaghan AJ, Aurikko JP, Ilag LL, Gunter Grossmann J, Chandran V, Kuhnel K, Poljak L, Carpousis AJ, Robinson CV, Symmons MF, Luisi BF. Studies of the RNA degradosome-organizing domain of the Escherichia coli ribonuclease RNase E. J Mol Biol 2004;340(5):965–979.

92. Kingston RL, Hamel DJ, Gay LS, Dahlquist FW, Matthews BW. Structural basis for the attachment of a paramyxoviral polymerase to its template. Proc Natl Acad Sci USA 2004;101(22):8301–8306.

93. Oldfield CJ, Cheng Y, Cortese MS, Romero PR, Uversky VN, Dunker AK. Structural disorder-to-order transitions in proteins: predicting alpha-helical molecular recognition elements. Biochemistry 2005. Submitted for publication.

94. Spolar RS, Record II MT. Coupling of local folding to site-specific binding of proteins to DNA. Science 1994;263:777–784.

95. Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. Curr Opin Struct Biol 2002;12(1):54–60.

96. Demchenko AP. Recognition between flexible protein molecules: induced and assisted folding. J Mol Recognit 2001;14(1):42–61.

97. Uversky VN, Oldfield CJ, Dunker AK. Showing your ID: intrinsic disorder as an ID for regcognition, regulation, and cell signaling. J Mol Recogntion 2005;18:343–384.

98. Fischer E. Einfluss der configuration auf die wirkung der enzyme. Ber Dt Chem Ges 1894;27:2985–2993.

99. Koshland DE Jr. Application of a theory of enzyme specificity to protein synthesis. Proc Natl Acad Sci USA 1958;44:98–104.

100. Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M. Automated analysis of interatomic contacts in proteins. Bioinformatics 1999;15(4):327–332.

101. Coureux PD, Sweeney HL, Houdusse A. Three myosin V structures delineate essential features of chemo-mechanical transduction. EMBO J 2004;23(23):4527–4537.