JOURNAL OF STRUCTURAL AND FUNCTIONAL GENOMICS,**1:** 8-14 2000

Calmodulin Target Database

Kyoko L. Yap†, Justin Kim†, Kevin Truong, Marc Sherman, Tao Yuan and Mitsuhiko Ikura *

Division of Molecular and Structural Biology, Ontario Cancer Institute and Department of Medical Biophysics, University of Toronto, 610 University Avenue, Toronto, Ontario, M5G 2M9, Canada

Received July 13, 2000, Accepted October 5, 2000

KEYWORDS **: calcium signaling, EF-hand protein, molecular recognition, protein-protein interaction**

The intracellular calcium sensor protein calmodulin (CaM) interacts with a large number of proteins to regulate their biological functions in response to calcium stimulus. This molecular recognition process is diverse in its mechanism, but can be grouped into several classes based on structural and sequence information. We have developed a web-based database (http://calcium.uhnres.utoronto.ca/ctdb) for this family of proteins containing CaM binding sites or, as we propose to call it herein, CaM recruitment signaling (CRS) motifs. At present the CRS motif found in approximately 180 protein sequences in the databases can be divided into four subclasses, each subclass representing a distinct structural mode of molecular recognition involving CaM. The database can predict a putative CRS location within a given protein sequence, identify the subclass to which it may belong, and structural and biophysical parameters such as hydrophobicity, hydrophobic moment, and propensity for α-helix formation.

INTRODUCTION

Ca^{2+} is one of the most diverse and widespread second messengers in eukaryotic cells, being involved in a number of cellular processes including fertilization, cell division and cell differentiation, as well as long-term potentiation and depression in neurons (reviewed in [1]). As a well-established, common mechanism for intracellular Ca^{2+} signaling, the ubiquitous Ca^{2+} sensor calmodulin (CaM) is recruited by numerous proteins and enzymes, by which these target proteins express Ca^{2+} sensitivity in their biological functions (reviewed in [2,3]). These Ca^{2+}-CaM stimulated proteins include various Ser/Thr protein kinases, protein phosphatase calcineurin, nitric oxide synthases, ion transporters and cytoskeletal proteins.

Previous studies on CaM recruitment signaling (CRS) motifs have shown that many display little, if any, sequence homology to one another. Despite this apparent insensitivity of CaM to target sequence, certain features have appeared consistent in the majority of CRS motifs. The typical CRS is unstructured in absence of CaM, and whereupon binding CaM forms an α-helix. The hydrophobic face of the α-helix interacts with the exposed hydrophobic pockets of CaM, while the overall basic ($\geq +4$) charge is well suited for electrostatic interaction given the acidic nature of CaM (reviewed in [4,5]). There are some possible exceptions, such as MARCKS protein [6] and phosphorylase kinase [7], in which the CaM binding domain in presence of CaM is believed to be in a non-helical or extended conformation. Many CaM-binding proteins also contain PEST (proline, glutamate, serine and threonine-rich) sequences responsible for calpain recognition that are often close to the CRS motif [8].

After solution of the first three-dimensional structures of CaM in complex with a target peptide [9-11] it became possible to more accurately characterize the modes by which CaM might interact with its many targets. Based on these structures and the ever-increasing number of sequences found to bind CaM, consensus sequences based on the positions of interacting hydrophobic and basic residues have been proposed for CRS motifs [5,12,13]. A large number of Ca^{2+}-dependent CRS motifs have been characterized as belonging to either the 1-8-14 or the 1-5-10 class, the numbers indicating the positions of bulky hydrophobic residues that form major interactions with CaM. Rhoads and Friedberg [13] also compiled a large number of IQ (IQxxxRGxxxR) motifs, which bind (or are predicted to bind) CaM in a predominantly Ca^{2+}-independent manner. The oft-tandem IQ motif was in fact first isolated as a binding motif for myosin light chains on the heavy chain before being recognized as having high sequence similarity to the CaM binding

† These authors contributed equally to this work.

* To whom correspondence should be addressed.
Tel: +1 416 946 2025 Fax: +1 416 946 2055/6529 Email: mikura@uhnres.utoronto.ca

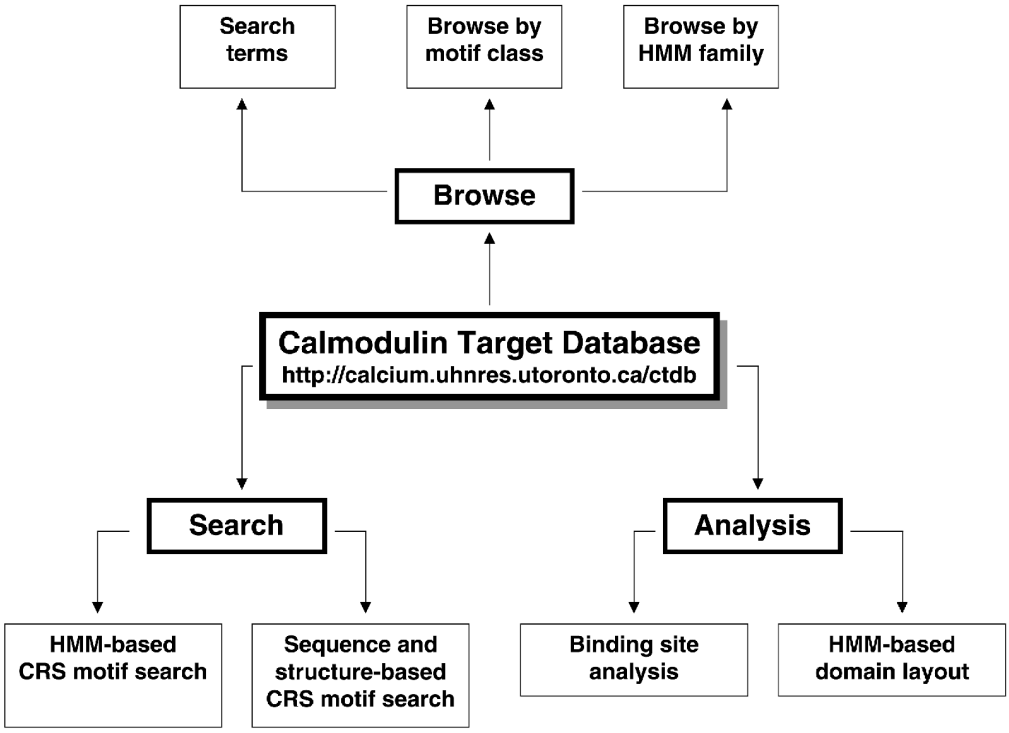

Figure 1. Structure of http://calcium.uhnres.utoronto.ca/ctdb. The website has three key features: a browse function which enables the user to search the database by keyword, browse sequences by motif class or browse CaM target proteins by HMM family; a search function in which a query protein or genomic sequence is analyzed for potential CRS motifs, either by comparing to HMMs or by calculating and comparing sequence-based biophysical and structural parameters; and an analysis function, which can calculate these parameters for a query sequence or present a functional domain layout as predicted by HMMs for a query protein.

motifs of neuromodulin and neurogranin [14]. Despite the ability to categorize many known CRS motifs into these families, there remain a number of atypical sequences and much biochemical evidence that there are other modes of interaction between CaM and its targets (*e.g.* [6,7,15-18]).

To date, at least 180 unique CRS motifs have been identified. As more protein and genomic sequences are determined, there is no doubt that the number of CaM binding proteins will increase. To manage and make sense of this large amount of data, it is beneficial to assemble a database containing these target protein sequences and any associated biochemical, biophysical and structural parameters. By dissecting this information one may be able to better classify the large number of known CRS motifs and thus further define the characteristics of commonly occurring CRS motifs. Based on this characterization, prediction of CaM-binding functionality attributable to a particular sequence within a protein should also be possible. This database should be valuable in probing the relationship between different CaM binding proteins, and in understanding the manner by which CaM can bind and activate such diverse targets.

MATERIALS AND METHODS

176 sequences demonstrated to bind CaM were culled from literature published up to July 2000 (*e.g.* from [13]; a complete list of references can be found at http://calcium.uhnres.utoronto.ca/ctdb in the 'General' section). Several biochemical/biophysical parameters (average hydrophobicity, average hydrophobic moment, average propensity for α-helix formation) were calculated as described elsewhere [19-23]. Structural information was derived from the three-dimensional structures of CaM in complex with peptides derived from skeletal myosin light chain kinase (MLCK) [9], smooth muscle MLCK [10], Ca^{2+}-CaM dependent protein kinase IIα [11], Ca^{2+}-dependent CaM kinase kinase [24] and plasma membrane Ca^{2+} pump [25]. Based on calculated parameters, residue and peptide charge and sequence homology, most sequences were classified into four major motif families comprising ten subfamilies, while the remainder either belonged to one

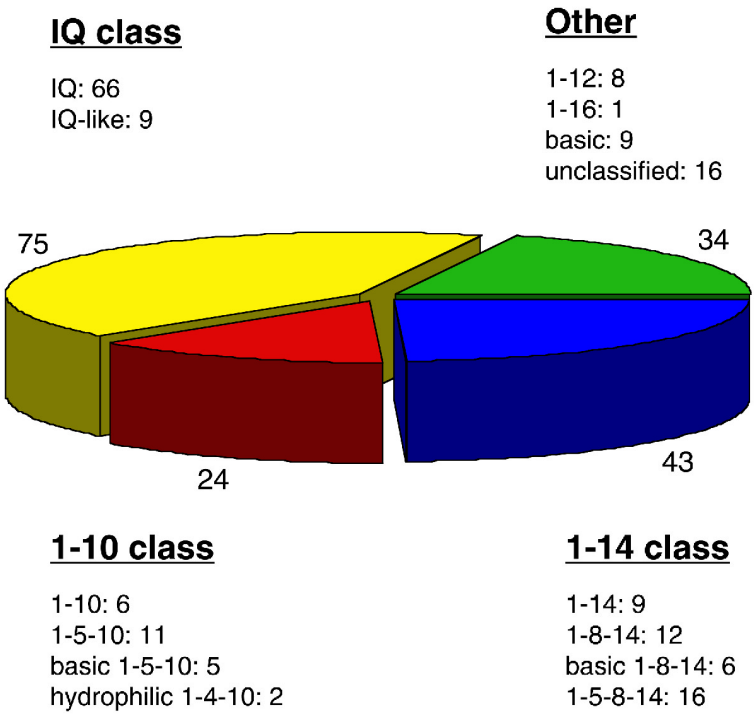

Figure 2. Distribution of CRS motifs in classes and subclasses as organized in the Calmodulin Target Database as of July 2000.

of three less common motif classes or were unclassifiable. 1260 sequences homologous to those already determined to bind CaM were obtained using BLAST [26] on the SWALL database with default parameters (WU-blastp program with blosum62 matrix). While it may be speculated that many of these sequences bind CaM, they are stored separately from the sequences experimentally shown to bind CaM.

To map domain layouts of CaM target proteins, PFAM [27] and SMART [28] were used. PFAM and SMART are both collections of hidden Markov models (HMMs) used to find frequently occurring functional domains in proteins. HMMs are considerably more powerful than BLAST or FASTA searches since they are created from multiple sequence alignments and thus incorporate multiple sets of information [29]. HMMs were also built from alignments of CRS motifs and are used as a first step in classifying query proteins into particular motifs.

Putative CRS motifs are identified first by searching for sequence homology to any entry in the database. If no sequence homology is found in any region along the query sequence, it is checked against motif-based HMMs and numerous criteria (such as presence of proline residues, distribution of hydrophobicity and basic charge and similarity to typical database parameters) which allow a score to be assigned to individual residues via a 20-residue 'sequence-walk' routine. Scores for the query sequence are returned, from which potential CaM binding sites are easily identified.

The database is stored using Oracle and its browse, search and analysis features (see below) were coded in JavaScript and Perl.

RESULTS AND DISCUSSION

There is no single consensus sequence for CaM binding domains. However, the CRS motif is typically hydrophobic and basic in nature, and consists of 15-30 amino acid residues which tend to form an α-helix. An increasing number of CaM target proteins reported in literature and the diversity in sequences for CaM binding prompted the need to organize all available data, enabling an easier comparison of new CaM binding sequences with known ones. To this end, we have developed a web-based CaM target database (http://calcium.uhnres.utoronto.ca/ctdb) (Fig. 1). The database contains almost all published CRS sequences and over 1200 possible homologues collected from the non-redundant SWALL database (SWISSPROT, TREMBL and TrEMBL New Databases) [30].

Guided by three-dimensional structural information on CaM-target complexes, we categorize the CRS motifs into three major classes plus one class of less common motifs (Fig. 2) ,each class possessing a distinct mode of CaM interaction involving bulky

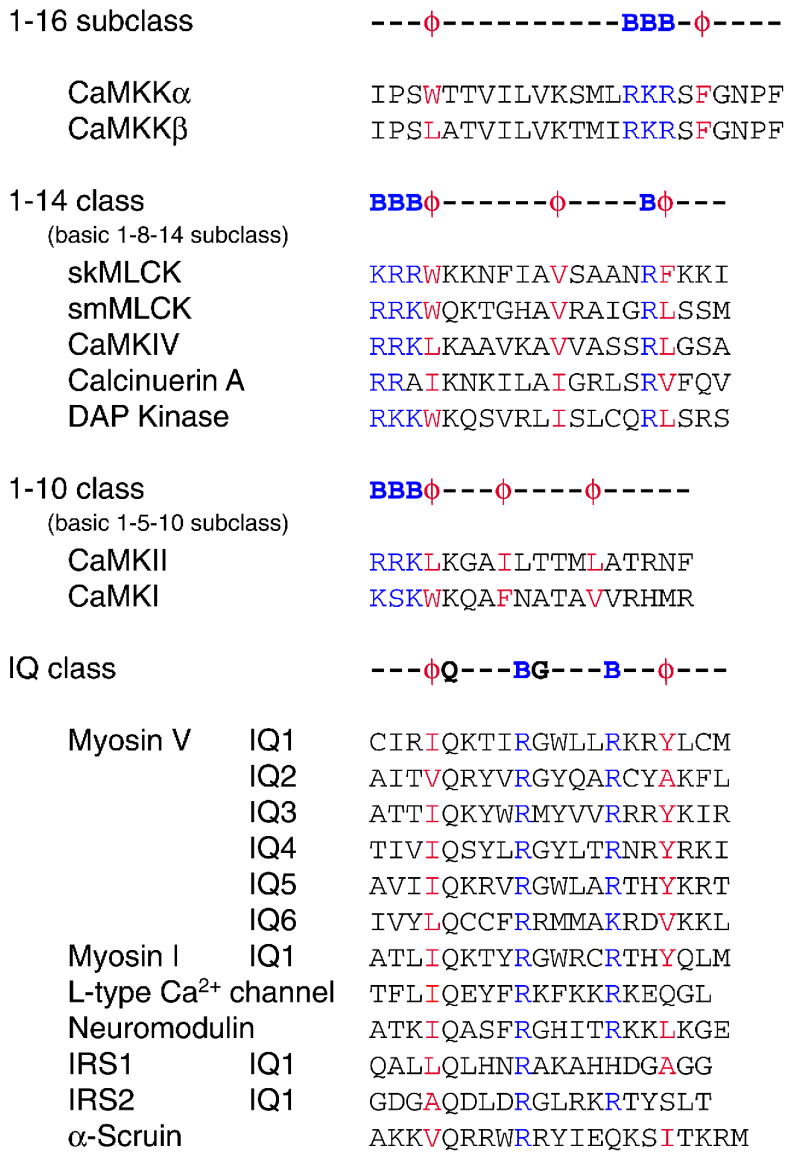

Figure 3. Sequence alignment of CaM-binding domains. There are three major classes and several subclasses of CRS motifs in the CaM target database. Representative members of these major classes and one subclass are shown. ϕ in the consensus sequence indicates a hydrophobic residue, B indicates a basic residue. The 1-16 subclass thus far contains only one protein. The 1-14 class [13] comprises the 1-14, 1-8-14, basic 1-8-14 and 1-5-8-14 subclasses, where each number represents the presence of a hydrophobic residue; for the basic 1-8-14 motif, three basic residues precede the first hydrophobic residue. The 1-10 class [13] consists of the 1-10, 1-5-10, basic 1-5-10 and hydrophilic 1-4-10 subclasses. The latter motif has three hydrophilic residues following the hydrophobic residue at the tenth position. The IQ class [14] consists of both complete and incomplete IQ motifs, the latter so-called for the absence of the second basic residue. Following are SWISSPROT/TrEMBL accession numbers and starting residues of the aligned sequences. 1-16 class: rat Ca^{2+}-calmodulin-dependent kinase kinase isoform α (CaMKKα), Q64572: 441; rat CaMKKβ O88831: 477. Note that contrary to the 1-14 and 1-10 classes, the trio of basic residues appears at the C-terminus of the CRS motif. This is proposed to influence the orientation of the peptide in the CaM-CaMKK complex structure, which is of opposite polarity to that seen in the CaM-MLCK and CaM-CaMKII complex structures [24]. 1-14 class, basic 1-8-14 subclass: rabbit skeletal muscle myosin light chain kinase (skMLCK), P07313: 577; rabbit smooth muscle myosin light chain kinase (smMLCK), P29294/Q28729: 974; mouse Ca^{2+}-calmodulin-dependent kinase IV (CaMKIV), P08414/Q61381: 319; *Neurospora crassa* calcineurin subunit A, Q05681: 406; human death-associated protein (DAP) kinase I, P53355: 302. 1-10 class, basic 1-5-10 subclass: rat Ca^{2+}-calmodulin-dependent kinase II (CaMKII), Q94608: 297; rat Ca^{2+}-calmodulin-dependent kinase I (CaMKI), Q63450/Q63084: 300. IQ class: mouse dilute non-muscle myosin V heavy chain (all six tandem IQ motif repeats shown), Q99104: 770, 793, 818, 841, 866, 889; bovine (brush border) myosin I heavy chain-like protein (first of two IQ motifs shown), P10568: 701; rabbit L-type Ca^{2+}-channel subunit α_{1c}, P15381/Q99243: 1651; bovine neuromodulin, P06836: 35; mouse insulin receptor substrate 1 (IRS1) (first of two IQ motifs shown), P35569: 108; mouse insulin receptor substrate 2 (IRS2) (first of two IQ motifs shown), P81122: 557; crab α-scruin, Q25390: 434.

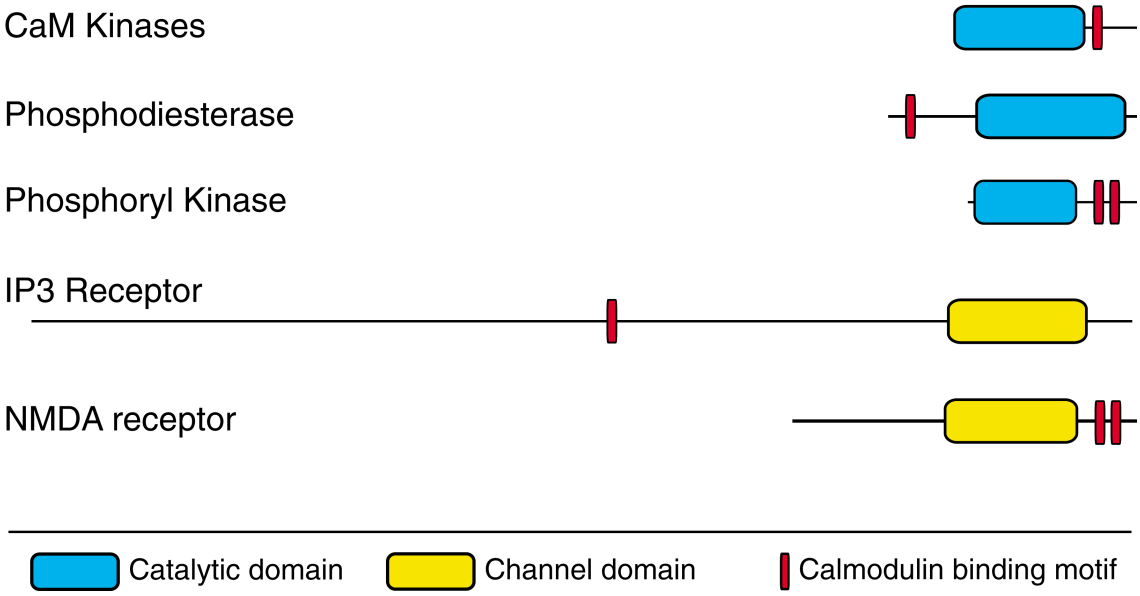

Figure 4. Domain organization of CaM target proteins. CaM recruitment signaling (CRS) motifs vary in number of occurrences and position relative to catalytic or channel domains. CaM kinase sequences contain a CRS motif immediately C terminal to the catalytic domain, whereas phosphodieterases contain the motif N terminal to the catalytic domain. Phosphoryl kinase contains two CRS repeats. IP_3 and NMDA receptors are both ligand-gated Ca^{2+} channels that contain a CRS motif N terminal and C terminal to the channel domain, respectively. Some proteins, such as ras-related proteins, contain CRS motifs within the catalytic domains.

hydrophobic residues in the target (Fig. 3). The 1-14 class [13] was seen in two similar structures [9,10] of CaM complexed with skeletal or smooth muscle myosin light chain kinase (MLCK). In this class, two key hydrophobic residues are positioned 14 residues apart (W582 and F595 in rat skeletal muscle MLCK; W1734 and L1747 in chicken smooth muscle MLCK), each interacting with the hydrophobic cavity of the C-terminal and N-terminal domain of CaM, respectively. The shortest distance (10 residues) between the two hydrophobic residues (L300 and L309) was identified in the structure of CaM in complex with CaM kinase II [11], representing the 1-10 class [13]. Although a three-dimensional CaM complex structure is yet to be determined, the IQ motif [14] should represent a unique third class, as all proteins of this class contain an Ile-Gln or equivalent sequence and, unlike other CRS sequences, many IQ sequences bind Ca^{2+}-free CaM. As one of the less common motifs, the 1-16 subclass is characterized by the structure of CaM in complex with a CaM kinase kinase peptide [24], in which the two key hydrophobic residues Trp444 and Phe459 are placed 16 residues apart, interacting with the N-terminal and C-terminal domain of CaM, respectively, in opposite direction to that observed for the 1-14 class. This novel interaction, which has not been observed in any other known CaM target protein thus far, may constitute a fourth major motif class as more such CRS motifs are discovered.

Given the diversity in function of the numerous known CaM binding proteins, it is unsurprising that the functional domain layout, specifically the relative location of the CRS motif within the amino acid sequence, also varies (Fig. 4). Generally, proteins that share similar functionality often exhibit similar domain layouts. In many protein kinases, the CRS motif is found proximate (within 9~30 residues) to the C-terminus of the catalytic domain. CRS motifs in integral membrane proteins are typically immediately C-terminal to the last predicted transmembrane domain, or close to the cytoplasmic C-terminus. Cytoplasmic proteins that anchor to the plasma membrane, such as A kinase anchor proteins and those of the band 4.1 family, contain a CaM binding domain at the N-terminus. CRS motifs occur immediately N-terminal to the helix-loop-helix DNA binding domain (*e.g.* in rat transcription factor 12 and Kappa-E2-binding factor) and C-terminal to myosin head domains.

Unlike CaM-dependent protein kinases, many non-protein kinases such as phosphotidyl inositol-3 kinase and phosphodiesterase contain CRS motifs N-terminal to the catalytic domain (separated by 70 residues in the former case). Similarly, the CRS motif in inositol 1,4,5-triphosphate (IP_3) receptor is N-terminal to the channel domain. In contrast, phosphorylase kinase and N-methyl-D-aspartate (NMDA) receptor contains two adjacent CRS domains C-terminal to the catalytic or channel domain. In proteins containing two catalytic domains, such as phosphofructokinase and nitric oxide synthase, the two domains often flank the CRS motif.

The lack of sequence homology in CRS motifs suggests that many of these proteins are evolutionary unrelated to each other. However, the mechanisms underlying CaM binding and its regulatory effects on target functions may be similar in some, if not many cases, as CaM is one of the most conserved proteins in eukaryotes [31]. In the case of protein kinases, the CaM binding directly influences an inhibitory

interaction between the regulatory domain containing the CRS motif and the active site within the kinase domain, thereby enhancing the catalytic activity of the enzyme. This can be achieved by the 1-10, 1-14 or 1-16 class interaction of CaM with CaM-dependent kinases.

The diversity in domain layouts in which a CRS may be found and the lack of an easily identifiable, single consensus sequence renders the task of determining the location of a CRS motif in a protein sequence rather daunting to the biologist. To address this, the web-based database has a search tool that allows one to identify putative CRS motif sequences within the amino acid sequence of a query protein, based either on HMM analysis or on sequence similarity to any entry in the CRS library. If in the latter case no homology is found in the query sequence, the program employs biophysical and structural criteria to predict a putative CRS motif. Parameters used in the criteria include average hydrophobicity (Kyte-Doolittle values), average hydrophobic moment and average propensity for α-helix formation. A score of probability, ranging from 0 (unlikely) to 9 (very likely), is calculated per amino acid, and typically a stretch of 10-20 residues with scores of 7-9 represents a putative CRS motif. Preliminary tests have shown that the search program correctly identifies new CRS motifs with approximately 80% accuracy, but also yields 'false positives' (putative CRS motifs that do not bind CaM) 1/3 of the time (data not shown). With the incorporation of additional tools, such as HMMs, and additional characterization of currently known and new CRS motifs, we expect to improve the accuracy of the search program.

Fold-based protein structure databases such as CATH [32], DALI [33], and SCOP [34] have proved to be extremely useful for protein research. Assembly of a database for the family of CaM binding proteins represents a new way of reorganizing numerous sequence and structural information that are available in primary databases [35,36]. In contrast to the fold-based structural databases, the CaM target database is not sorted by common folds. It is assembled according to common functional properties – the type and nature of the CaM binding activity of proteins. The CaM binding protein family comprises a large number of proteins with different folds, different domain architectures and diverse biological functions. Classification of those CaM binding proteins into subgroups based on structural criteria may help unveil possible structure-function relationships that could not be identified by conventional classification or analysis.

A secondary database, such as this one, provides a tool to reconstruct a huge volume of database information on sequence and structure into a new dimension. The new dimension can be defined according to similarity in interaction partners and ligands (or pharmaceutical agents), biochemical function, and physiological roles. Along with the fold-based databases, such new databases should help produce biologically useful information that is otherwise hidden or impossible to extract easily. The present database targeted for the CaM binding protein family is one such tool.

CONCLUDING REMARKS

An emerging scheme in CaM-target recognition is that CaM recognizes various CRS motif sequences in different ways, by virtue of the structural plasticity stemming from a flexible domain-domain linker, helix-helix movement and side chain rearrangement [37,38]. The classification of numerous CRS motif sequences into distinct classes may help to understand how various Ca^{2+}-CaM dependent proteins are related to each other in terms of their CaM-dependent regulatory mechanisms. Furthermore, understanding structures and binding properties of CaM-target recognition enables the design of molecular tools that can be used for *in vitro* and *in vivo* studies. Excellent examples include a protein expression and purification system based on a CRS peptide-tagged fusion [39] and an intracellular Ca^{2+} indicator based on a CaM-CRS peptide hybrid and green fluorescent protein (GFP), so-called cameleons [40]. Finally, the present study represents a new way of reorganizing the sequence and structure database information according to a functional property, in this case CaM binding activity.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Medical Research Council of Canada. MI is a Howard Hughes Medical Institute International Scholar and an MRCC Scientist.

REFERENCES

1. Williams, R.J.P. (1999) in: *Calcium as a cellular regulator*, pp. 3-629 (Carafoli, E. and Klee, C.B., Eds.) Oxford University Press, Oxford.
2. James, P., Vorherr, T. and Carafoli, E. (1995) *Trends Biochem Sci* 20, 38-42.
3. Zhang, M. and Yuan, T. (1998) *Biochem Cell Biol* 76, 313-23.
4. O'Neil, K.T. and DeGrado, W.F. (1990) *Trends Biochem Sci* 15, 59-64.
5. Crivici, A. and Ikura, M. (1995) *Annu Rev Biophys Biomol Struct* 24, 85-116.
6. Porumb, T., Crivici, A., Blackshear, P.J. and Ikura, M. (1997) *Eur Biophys J* 25, 239-47.
7. Dasgupta, M., Honeycutt, T. and Blumenthal, D.K. (1989) *J Biol Chem* 264, 17156-63.
8. Barnes, J.A. and Gomes, A.V. (1995) *Mol Cell Biochem* 149-150, 17-27.
9. Ikura, M., Clore, G.M., Gronenborn, A.M., Zhu, G., Klee, C.B. and Bax, A. (1992) *Science* 256, 632-8.
10. Meador, W.E., Means, A.R. and Quiocho, F.A. (1992) *Science* 257, 1251-5.

11. Meador, W.E., Means, A.R. and Quiocho, F.A. (1993) *Science* 262, 1718-21.
12. Afshar, M., Caves, L.S., Guimard, L., Hubbard, R.E., Calas, B., Grassy, G. and Haiech, J. (1994) *J Mol Biol* 244, 554-71.
13. Rhoads, A.R. and Friedberg, F. (1997) *Faseb J* 11, 331-40.
14. Cheney, R.E. and Mooseker, M.S. (1992) *Curr Opin Cell Biol* 4, 27-35.
15. Ruan, J., Xie, Q., Hutchinson, N., Cho, H., Wolfe, G.C. and Nathan, C. (1996) *J Biol Chem* 271, 22679-86.
16. Nevalainen, L.T., Aoyama, T., Ikura, M., Crivici, A., Yan, H., Chua, N.H. and Nairn, A.C. (1997) *Biochem J* 321, 107-15.
17. Yuan, T. and Vogel, H.J. (1998) *J Biol Chem* 273, 30328-35.
18. Onions, J., Hermann, S. and Grundstrom, T. (2000) *Biochemistry* 39, 4366-74.
19. Schiffer, M. and Edmundson, A.B. (1967) *Biophys J* 7, 121-135.
20. Kyte, J. and Doolittle, R.F. (1982) *J Mol Biol* 157, 105-32.
21. Eisenberg, D., Weiss, R.M. and Terwilliger, T.C. (1982) *Nature* 299, 371-4.
22. Eisenberg, D., Schwarz, E., Komaromy, M. and Wall, R. (1984) *J Mol Biol* 179, 125-42.
23. Erickson-Viitanen, S. and DeGrado, W.F. (1987) *Methods Enzymol* 139, 455-78.
24. Osawa, M., Tokumitsu, H., Swindells, M.B., Kurihara, H., Orita, M., Shibanuma, T., Furuya, T. and Ikura, M. (1999) *Nature Struct Biol* 6, 819-24.
25. Elshorst, B. et al. (1999) *Biochemistry* 38, 12320-32.
26. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) *Nucleic Acids Res* 25, 3389-402.
27. Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. and Sonnhammer, E.L. (2000) *Nucleic Acids Res* 28, 263-6.
28. Schultz, J., Milpetz, F., Bork, P. and Ponting, C.P. (1998) *Proc Natl Acad Sci U S A* 95, 5857-64.
29. Eddy, S.R. (1998) *Bioinformatics* 14, 755-63.
30. Bairoch, A. and Apweiler, R. (2000) *Nucleic Acids Res* 28, 45-8.
31. Copley, R.R., Schultz, J., Ponting, C.P. and Bork, P. (1999) *Curr Opin Struct Biol* 9, 408-15.
32. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B. and Thornton, J.M. (1997) *Structure* 5, 1093-108.
33. Holm, L. and Sander, C. (1996) *Science* 273, 595-603.
34. Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) *J Mol Biol* 247, 536-40.
35. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) *Nucleic Acids Res* 28, 235-42.
36. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and Wheeler, D.L. (2000) *Nucleic Acids Res* 28, 15-8.
37. Vogel, H.J. and Zhang, M. (1995) *Mol Cell Biochem* 149-150, 3-15.
38. Yuan, T., Yap, K.L. and Ikura, M. (2000) in: *Topics in Biological Inorganic Chemistry: Calcium Homeostasis*, Vol. 3 (Carafoli, E. and Krebs, J., Eds.) Springer-Verlag, Heidelberg.
39. Felts, K., Wyborski, D., Bauer, J. and Vaillancourt, P. (1999) *Strategies* 12, 24-27.
40. Miyawaki, A., Llopis, J., Heim, R., McCaffery, J.M., Adams, J.A., Ikura, M. and Tsien, R.Y. (1997) *Nature* 388, 882-7.