

Sequence Motifs Determine Structure and Ca⁺⁺-binding by EF-hand Proteins

Hooman H. Rashidi, Margarethe Bauer,
Joe Patterson, and Douglas W. Smith*

Department of Biology, 0116,
and Center for Molecular Genetics,
University of California, San Diego,
La Jolla, CA 92093, USA

Abstract

Prediction of protein structural and functional characteristics based on specific motif interactions could serve as a powerful tool in many facets of the biological sciences. Such improvements in protein modeling will be instrumental in the enhancement of drug design. A new approach to a sequence description of EF-hand motifs with more than one EF-hand domain is presented here; this permits precise insight into the structural and functional properties of many members of the EF-hand superfamily of calcium-binding proteins. Three separate regular expressions, or signatures, are used to describe an EF-hand motif, and specific relationships must exist between the two sequence motifs for the two neighboring EF-hands in a given calcium-binding domain. Specifically, each of the sequence motifs has a conserved phenylalanine. These two phenylalanine residues are separated by 57±10 amino acid residues but interact closely with each other in the tertiary structure of the calcium-binding domain. Changes in conserved residues in the sequence motifs have been shown experimentally to decrease or eliminate the ability of the protein to bind calcium. This new approach of use of multiple sequence motifs, with motif interrelationships, yields a highly specific and robust tool for the prediction of structural and functional properties of new and novel proteins.

Introduction

The helix-loop-helix EF-hand protein motif, involved in binding of calcium ions, was first discovered in the crystal structure of parvalbumin (Kretsinger and Kockolds, 1973), and has since been identified in numerous other calcium-binding proteins. Many of the known essential calcium-binding proteins belong to a large superfamily of evolutionarily related proteins containing this motif (review: Persechini *et al.*, 1989).

Muscle contraction, nucleotide metabolism, cell cycle control, and signal transduction are examples of the essential cellular processes that are tightly regulated by calcium. Uncontrolled high cellular levels of calcium ions

can activate certain biochemical processes that lead to protein degradation and ultimately to cell death. Alzheimer's disease, Parkinson's disease, Downs syndrome, Acquired Immune Deficiency Syndrome (AIDS), epilepsy, retinopathy, and ischemia are some of the neurodegenerative disorders that involve EF-hand calcium-binding proteins (Hermann and Cox, 1995). The regulatory and buffering role of the calcium-binding proteins is vital to the organism and their potential malfunctions can have pathological consequences. Understanding the structural and functional characteristics of these essential macromolecules could lead to potential therapeutic routes specifically targeting malfunction in one or more of the mediators responsible for the pathogenic event.

EF-hand calcium-binding proteins have been extensively studied (Persechini *et al.*, 1989). EF-hand motifs are typically found in pairs in a given protein and display a cooperative binding of calcium. EF-hand motif protein domains are functionally classified into 1) "regulatory" domains and 2) "structural" or "buffer" domains, according to their conformational response upon binding calcium (Ikura, 1996). The "regulatory" domains are those that undergo larger conformational changes and are found to activate other interacting proteins. In contrast, the "structural" or "buffer" domains exhibit smaller conformational changes and function as a sequestering agent in control of intracellular calcium concentration (Ikura, 1996). EF-hand motifs in buffering domains are generally found to have a higher binding affinity for calcium than do EF-hand motifs in regulatory domains.

Calmodulin, troponin C, and calcineurin B are a few of the essential cellular mediators with multiple EF-hand motifs. Most of the EF-hand calcium-binding motifs are found in pairs in a given EF-hand protein domain. The presence of such neighboring EF-hand motifs enhances the overall binding affinity of the protein for calcium (Waltersson *et al.*, 1993). Positive cooperativity is believed to be responsible for this increased affinity, where binding of calcium to one motif enhances the binding to the neighboring motif. Thus, understanding the intrinsic properties of these motifs and their inter-motif relationships could be instrumental in gaining insight into the structural and functional characteristics of novel EF-hand calcium-binding proteins.

We present here a new approach to a description of EF-hand motifs. In this approach, three separate regular expressions, or signatures, are used to describe an EF-hand motif. Further, specific relationships must exist between the two sequence motifs for the two neighboring EF-hands in a given calcium-binding domain. Although each of the three signatures is less specific than the general PROSITE (Bairoch *et al.*, 1997) EF-hand signature, when coupled with the intermotif relationships, the three signatures create a highly specific and robust tool which precisely describes the structural and functional properties of the calcium-binding domain of the protein.

Received March 2, 1999; revised April 1, 1999; accepted April 1, 1999.
*For correspondence. Email dsmith@ucsd.edu; Tel. 619-534-2620; Fax. 619-534-7108.

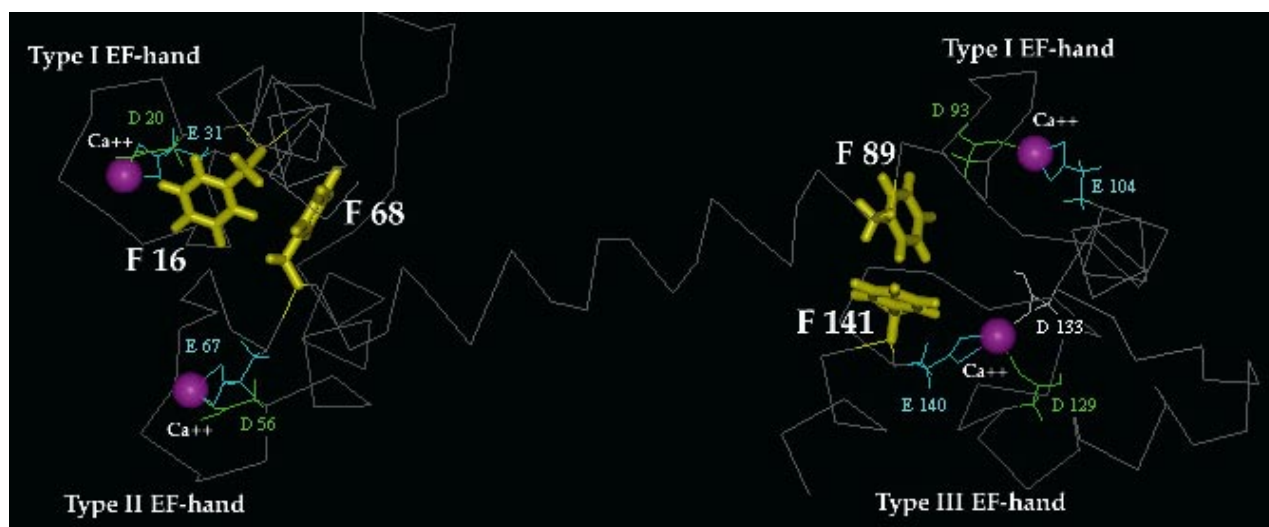


Figure 1. Tertiary Structure of Human Calmodulin

The alpha-carbon backbone of the complete molecule, using coordinates for PDB (Sussman *et al.*, 1998) entry 1CLL, is shown. Stick structures are shown for the calcium-coordinating conserved acidic residues, and a solid stick structure is shown for the two conserved phenylalanine residues, one in each of the type I-II (N-terminus) and type I-III (C-terminus) sequence motifs. Bound calcium ions are shown as spheres. The pi-stacking interaction between these paired phenylalanine residues is apparent. Residue numbers for human calmodulin are indicated. A colour version of this figure is available on our Web site: <http://www.jmmb.net>

Results

Three EF-hand Motifs and Three Relationships Between Them

Many calcium-binding proteins contain EF-hand motifs as pairs of motifs in a single domain of the protein. For example, calmodulins typically contain two such EF-hand protein domains, at each end of the protein, with two EF-hand motifs present in each domain (see Figure 1), whereas parvalbumins contain a single EF-hand protein

1. Presence of the Sequence Motifs I, II and III in Pairs

I **F**-X(3)-D-X(10)-[ED]-[LIVMFYWAGC]

AND

II [LIVMFYWAGC]-X(3)-D-X(3)-{D}-X(6)-[ED]-**F**

OR

III [LIVMFYWAGC]-X(3)-D-X(3)-D-X(6)-[ED]-**F**

2. Order of the Sequence Motifs

Motif I always followed by Motif II or III

3. Separation of the Sequence Motifs

The 2 highly conserved phenylalanine (**F**) residues at the N-terminus of motif I and the C-terminus of motif II or III, respectively, must be 57 ± 10 residues apart.

Figure 2. Sequence Motifs of EF-hand Proteins

Three sequence motifs and three relationships between them which together define a distinct class of EF-hand proteins. Motifs II and III are identical except for position 9; this position has any amino acid except aspartic acid in motif II but requires an aspartic acid in motif III.

domain. Three distinct EF-hand sequence motifs, termed types I, II, and III, present in EF-hand calcium-binding proteins containing such pairs of EF-hand motifs in an EF-hand protein domain are identified here. The order in which these three motifs are found in the EF-hand protein domain, their amino acid content, and the number of residues between two key phenylalanine residues, one in each pair of these motifs, provide definitive structural and functional information about the EF-hand motifs in these molecules. These three sequence motifs and the key relationships between these motifs for structure and function information are shown in Figure 2. Each EF-hand in a given identified Ca-binding protein contains at most one of the three motifs and the motifs do not overlap each other.

The pair of motifs present in a given EF-hand protein domain are found to be either a type I – type II pair or a type I – type III pair. Further, the EF-hand motif pair type I – type II is found preferentially at the N-terminus and the pair type I – type III is found mainly at the C-terminus of Ca-binding proteins. However, the presence alone of the two motifs in a given EF-hand protein domain is insufficient to yield maximal information. Two additional relationships are required. These relationships are: 1) type I motif always precedes a type II or type III motif, and 2) the two highly conserved phenylalanine residues, one in each of the type I and type II motifs or in each of the type I and type III motifs, are separated by 57 ± 10 residues in the polypeptide chain. The predictive value of the three motifs for molecular structure and function is optimized when all three requirements are satisfied.

The conserved residues in each of the three motifs serve specific functional and structural roles. The acidic aspartate and glutamate residues function to bind calcium ions while the phenylalanine and hydrophobic residues serve as structural moieties (Martin *et al.*, 1992; Waltersson *et al.*, 1993; George *et al.*, 1996). Each of the three motifs contains a highly conserved phenylalanine residue (Figure

2, bold). This phenylalanine is always found at the N terminus of the type I motif and at the C terminus of the type II and type III motifs. The main structural interaction of the type I EF-hand motif with either the type II or type III motif is via direct interaction between the ring moieties of these highly conserved phenylalanine residues (see Figure 1). Presence of these phenylalanine residues is the key feature distinguishing the EF-hand proteins identified using these motifs and the relationships described here between these motifs. Upon calcium binding, the phenylalanine residues move closer to each other and their ring moieties assume a pi-stacking conformation. The perpendicular pi-stacking conformation of these residues is shown in Figure 1.

Relationship of the PROSITE EF-hand Signature to EF-hand Sequence Motifs I, II, and III

The PROSITE (Bairoch *et al.*, 1997) EF-hand signature (PS00018) is a common predictive tool used to identify novel proteins as potential EF-hand calcium-binding proteins, as well as to identify their potential metal binding sites. Although the general application of this and other regular expression signatures is very useful in identifying potential functional protein candidates, information present in these signatures is often imprecise, and structural and functional insight gained is minimal. A comparison of the PROSITE EF-hand signature with those of sequence motifs I, II, and III is shown in Figure 3. Although each motif by itself (I, II, or III) is less restrictive than that of the PROSITE EF-hand signature, when the three motifs *per se* are combined with the required relationships between them (Figure 2), a robust and powerful tool is created that provides highly specific structural and functional information and eliminates uncertainties present in application of the PROSITE signature. By itself, application of the PROSITE signature to new proteins results in ~ 20% false positive prediction of proteins that bind calcium. In addition, this signature *per se* provides very little insight into the tertiary

structure and functional properties of the residues involved. By contrast, application of the three sequence motifs described here, coupled with their three relationships, leads to NO false positive predictions of Ca-binding proteins and provides considerable structural and functional information relevant to calcium-binding potential and binding affinity. However, the PROSITE signature is able to detect some EF-hand proteins not found using the EF-hand motifs and their relationships described here. Thus, the gain in specificity results in some loss in sensitivity.

Conserved Residues in the Three Motifs: EF-hand Domain Structure and Calcium Binding

The relationship between the conserved residues in the three sequence motifs and their role in EF-hand domain structure and Ca-binding was determined from examination of 25 EF-hand Ca-binding protein domains whose tertiary structures are known. These 25 structures are all of those present in the SCOP database (Murzin *et al.*, 1995) as of July, 1998 (Table 1). The three sequence motifs each span the loop found between two alpha helices traditionally called the E- and F-helices (Figure 3). A 12 amino acid stretch, found between the E- and F-helices in EF-hand calcium binding proteins, contains the essential coordinating residues that are involved in calcium binding, namely, those found at positions 1, 3, 5, 9, and 12 (Figure 3). These include the conserved aspartate and glutamate residues of motifs I, II, and III at positions 1, 5, and 12. For the type I EF-hand motif, the aspartate residue at position 1 is always found at the beginning of the loop, while the next conserved acidic residue, aspartate or glutamate, is at the C-terminal end of the EF-hand sequence motif just within helix F, at position +3 of the F-helix (position 12 in Figure 3). The type II EF-hand motif begins with aspartate at position 1 at the beginning of the loop after the E-helix, followed by the acidic residue, aspartate or glutamate, at the C-terminal end of the motif at positions +3 of the F-helix. The type III motif is similar to the type II motif, but

	E-Helix				Calcium Binding Loop										F-Helix					
Position in Motif	<u>-4</u>	<u>-3</u>	<u>-2</u>	<u>-1</u>	1	2	3	4	5	6	7	8	9	10	11	12	13			
Consensus pattern (PROSITE)					D-x-[DNS]-{ILVFW}-[DENSTG]-[DNQHRK]-{GP}-[LIVMC]-[DENQSTAGC]-x-x-[ED]-[LIVMFYW]															
Sequence Motif I	F	-x	-x	-x	D	-x	x	-	x	-	x	-	x	-	x	-x	-x	[ED]	-	[B]
Sequence Motif II	[B]	-x	-x	-x	D	-x	x	-	{D}	-	x	-	x	-	x	-x	-x	[ED]	-	F
Sequence Motif III	[B]	-x	-x	-x	D	-x	x	-	D	-	x	-	x	-	x	-x	-x	[ED]	-	F

Figure 3. Comparison of the PROSITE EF-hand Signature with the Three EF-hand Motifs. Positions of amino acids in the motifs are numbered relative to the PROSITE signature sequence. Standard syntax is used for regular expression nomenclature. Note that the first four residues of the three EF-hand motifs are not defined in the PROSITE signature. Key residues of the motifs (see text) are shown in bold. Amino acid positions that are involved in calcium-binding are boxed, and positions that are part of the structure stabilizing the EF-hand hydrophobic core are underlined. Residues present in the E and F alpha helices, versus those in the intervening loop, are indicated by the cartoon. The symbol B is used to designate the hydrophobic amino acids L, I, V, M, F, Y, W, A, G, and C.

Table 1. Phe-Phe C α Carbon Atom Distances in the Paired Sequence Motifs Present in EF-hand Ca-binding Proteins

PDB code ^a	N-terminal protein domain			C-terminal protein domain		
	Motif ^b relations	Calcium ^c bound	Phe-Phe ^d distance (Å)	Motif ^b relations	Calcium ^c bound	Phe-Phe ^d distance (Å)
4cln	+	+	7.24	+	+	7.61
1cll	+	+	7.28	+	+	7.53
1osa	+	+	7.29	+	+	7.31
1lin	+	+	7.13	+	+	7.21
1cmf	-	-	N/A ^e	+	-	8.22
1cfc	+	-	8.02	+	-	9.75
1pva	-	-	N/A	+	+	6.86
4cpv	-	-	N/A	+	+	6.95
5pal	-	-	N/A	+	+	7.13
1rtf	-	-	N/A	+	+	7.18
1rro	-	-	N/A	+	+	7.38
1tco	+	+	7.72	+	+	7.16
5tnc	+	-	8.93	+	+	7.49
1top	+	-	8.79	+	+	7.44
1tnq	+	+	7.45	-	-	N/A
1tnw	+	+	7.09	+	+	7.01

^aProtein Databank code for the protein (see Experimental Procedures).

^bMotif relations are as defined in Figure 1 and are satisfied (+) or not (-).

^cMolecular structure was determined with calcium bound (+) or not (-).

^dDistance, in Angstroms, between the two highly-conserved phenylalanine residues in motif I and motif II or III.

^eN/A: Calcium-binding domain not present in this domain of the protein.

differs by presence of an additional conserved aspartate residue at position 5 of the calcium-binding loop (Figure 3). The acidic residues at positions 1 and 12 in all three motifs always directly coordinate calcium ions, as does an aspartate when found at position 5. Residues at positions 3 and 9 mainly coordinate calcium ions indirectly through a water molecule (Martin *et al.*, 1992; Waltersson *et al.*, 1993). The presence of the additional aspartate residue at position 5 in the type III motif can account for the higher calcium affinity observed in the C-terminal domain of Ca-binding proteins (Martin *et al.*, 1992; George *et al.*, 1996).

The highly conserved phenylalanine residues are found at the N-terminal end of the type I motif, mainly at position -4 within the E-helix, and at the C-terminal ends of the type II and III motifs, mainly at position +4 of the F-helix (position 13 in Figure 3). The two phenylalanine residues in each member of a pair of motifs (type I – type II or type I – type III) function by interacting directly to bring the two motifs together to form the globular structure typical of an EF-hand protein domain (see Figure 1). In addition, this globular structure is stabilized by hydrophobic interactions between the conserved hydrophobic residue at position 13 in the F-helix of a type I motif (Figure 3) and a variety of nonconserved hydrophobic residues found in different EF-hand proteins at positions -1, -4, -5, and -8 of the E-helix, position 8 of the loop, and positions +7, +8, and +11 of the F-helix (Corson *et al.*, 1986; Sekharudu and Sundaralingam, 1988; Houdusse *et al.*, 1997). The presence of these position-specific residues in the sequence motif pairs described here yields highly reliable information with respect to prediction of structure and calcium-binding function in novel, putative calcium-binding proteins.

Phe-Phe Distance and Calcium Binding

Analysis of EF-hand calcium-binding proteins whose structure is known yielded two important structural relationships between the highly conserved phenylalanine residues present in the EF-hand motifs and the three dimensional fold of these motifs. The two phenylalanine residues have a pi stacking conformation in their calcium-bound state (see Figure 1) and their C-alpha carbon distances are quite conserved (Figure 4; also see Table

1). In the state with no calcium bound (apo state), the distance between the C-alpha carbons of the conserved phenylalanine residues in the two motifs is 8-10 Å, while in the calcium bound state they are 6-8 Å apart (Figure 4 and Table 1). In the apo state, the C-alpha carbon distance between the two conserved phenylalanine residues is

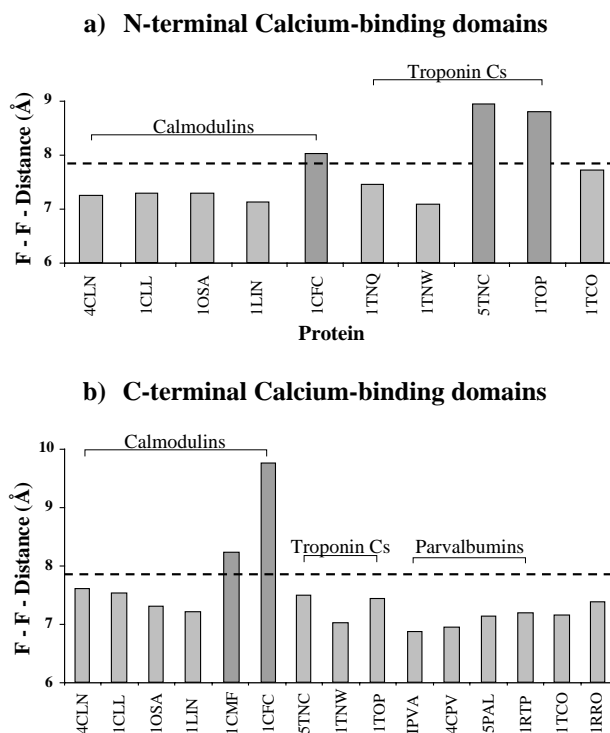


Figure 4. Phenylalanine-Phenylalanine C α Carbon Distances for Ca-Binding Proteins of Known Structure

A, N-terminal Ca-binding domains. B, C-terminal Ca-binding domains. Each protein is designated by its Protein DataBase (PDB) name (Sussman *et al.*, 1998). Dashed line, threshold distance of 7.8 Å below which calcium binding activity is present. proteins with distances less than the threshold distance. proteins with distances greater than the threshold distance. Proteins are grouped by type of Ca-binding protein.

Table 2. Types of Ca-Binding Proteins that Satisfy the Three Sequence Motif Requirements

A. N-terminal domain of proteins		
Type of Calcium-binding protein ^a	I-II motif pairs per database entries ^c	Phe-Phe separation (res) ^d
Alpha Actinin	1 / 3	51
Calcineurin B	7 / 7	47
Calcineurin B (isoform 2)	2 / 2	47
Ca-Dependent Protein Kinase	5 / 7	51
Calcyphosine	3 / 3	51
Calmodulin	47 / 48	52
Caltractin	13 / 14	51
Fimbrin	2 / 3	55
Hippocalcin	2 / 2	62
Neurocalcin	4 / 4	62
Neuronal Calcium Sensor	5 / 6	62
Nucleobindin	3 / 3	67
Plastin	5 / 5	55
Sorcin	1 / 3	52
Troponin C (Skeletal)	7 / 7	51
Visinin-like Protein	5 / 6	62
Others ^b	11 / 11	47 – 57

B. C-terminal domain of proteins		
Type of Calcium-binding protein ^a	I-III motif pairs per database entries ^c	Phe-Phe separation (res) ^d
Calcineurin B	7 / 7	56
Calcineurin B (isoform 2)	2 / 2	56
Ca-Dependent Protein Kinase	5 / 7	50
Calmodulin	47 / 48	51
Caltractin	4 / 14	51
DNA BP NEFA (Precursor)	2 / 2	67
cGMP Activating Protein	4 / 4	59
Hippocalcin	2 / 2	63
Neurocalcin	4 / 4	63
Neuronal Calcium Sensor	6 / 6	63
Nucleobindin	3 / 3	67
Oncomodulin	3 / 3	54
Parvalbumin	32 / 33	54
Sarcoplasmic Ca Protein	2 / 3	49
Troponin C (Cardiac)	4 / 5	51
Troponin C (Skeletal)	7 / 7	51
Visinin-like Protein	6 / 6	65
Others ^b	19 / 19	51 – 63

^aTypes of calcium-binding proteins present in SWISSPROT 36.0 / TrEMBL (Bairoch and Apweiler, 1999).

^bProtein types with only one available representative sequence.

^cNumber of proteins with either motif I - motif II, or motif I - motif III, pairs per number of database entries for a given group of proteins.

^dNumber of residues between the highly conserved phenylalanine residues in motif I and motif II or III.

above a threshold distance of 7.8 Å (dashed line fill, Figures 4a and 4b) and this C-alpha carbon distance is less than the 7.8 Å threshold when calcium is bound (grey fill, Figures 4a and 4b). Calcium binding is accompanied by energetically coupled conformational changes (Fujimori *et al.*, 1990). The presence of the EF-hand motifs and their position-specific interactions will be a powerful tool for prediction of structural and functional properties of novel calcium-binding proteins yet to be identified.

Calcium-Binding Proteins with These Motifs and Relationships

A search of SWISS-PROT 36.0 / TrEMBL (Bairoch and Apweiler, 1999) with the PROSITE EF-hand signature PS00018 yielded 537 proteins. Of these, PROSITE 15.0 (Bairoch *et al.*, 1997) correctly identified 411 and missed 25 of these, and additionally identified 25 proteins of unknown Ca-binding ability and incorrectly identified 100 non-EF-hand proteins. The BLOCKS database 10.1 (Henikoff and Henikoff, 1994) lists 413 under block

BL00018. Of the 537 proteins found in this SWISS-PROT / TrEMBL search, 242 contain a single EF-hand motif, and hence are excluded from the analysis presented here. Of the remaining 295 proteins, we found 188 (64%) to satisfy the three sequence motif requirements described here (see Figure 2). These 188 Ca-binding proteins, of over 20 different types from different species, are shown in Table 2. Only proteins that are true EF-hand proteins were identified (no false positives). All of those identified proteins for which structural data is available and which satisfy the three relationships between these sequence motifs have been shown experimentally to bind calcium. Sequence data are available for more than one protein in most of the protein types, usually from different species. Those types of Ca-binding proteins for which only one sequence is known are grouped together in Table 2 under "Others". The number of residues found between the conserved phenylalanine residues in the two sequence motifs of a given Ca-binding domain, either N-terminal (Table 2A) or C-terminal (Table 2B), are all found to be within 10 residues of 57 residues. This is the basis for relationship 3 between the pairs of sequence motifs (see Figure 2). Remarkably, the number of residues between the two phenylalanine residues for all proteins in a given type is precisely the same. This conservation of residue distance in the protein sequence between two conserved residues is of high predictive value, and is a predictive parameter absent when a single, general motif signature is used.

Motif sequences present in representative EF-hand proteins from a variety of different species are shown in Table 3. The number of functional calcium binding sites in these proteins, as based on experimental evidence, is shown for each protein. All of the EF-hand motifs that have mutations in any of the conserved residues of sequence motifs I, II, or III are known to be nonfunctional in their ability to bind calcium while those EF-hand motifs that possess the conserved residues within the sequence motifs are known to bind calcium. When the Ca-binding conserved residues (aspartates and glutamates) are altered, calcium-binding is dramatically decreased or, in most cases, eliminated, whereas when the structural residues (phenylalanines) are altered, calcium is still bound but with decreased affinity (Starovasnik *et al.*, 1992; Chandra *et al.*, 1994). The examples chosen are mainly from different organisms and represent different types of calcium-binding proteins.

Although all of the proteins in Table 2 contain Ca-binding domains that completely meet the EF-hand motif sequence requirements and satisfy the three relationships between the motifs, some proteins contain additional EF-hand motifs that do not satisfy the three relationships. This is true for 11 of the 188 proteins, as shown explicitly in Table 3. In the type III motif from *S. cerevisiae* calmodulin (CALM_YEAST), all three of the essential acidic calcium-binding residues have been substituted with non-acidic residues. This motif would be expected to be unable to bind calcium, and yeast calmodulin is known to bind three calcium ions rather than four (Brockhoff *et al.*, 1992). Several of these proteins have mutations at positions of conserved aspartates, including the N-terminal domain of APLC_APLCA, BTV3_BETVE, CATR_NAEGR, SCP_NERDI, TPCC_COTJA, and TPCC_RABIT, and the C-terminal domain of CATR_NAEGR, CAYP_CANFA, and TPCC_COTJA. The conserved aspartate in motif II of

Table 3. Sequence Motifs in EF-hands of Representative Calcium-binding Proteins

A. N-terminal protein domain						
SWISS-PROT name ^a	Sequence Motif I ^b		Sequence Motif II ^b		Calcium binding sites ^c	
APLC_APLCA	21	FTEAEIKQWHKGF RKDC	36	68 FN VFDENKDGFI SFG EF	84	1
BTV3_BETVE	45	FDLFDKNSDGIITV DEL	60	81 VKS F TREGNIGLQ F EDF	97	1
CABO_LOLPE	16	FDMFDIDGGQITS KEL	31	52 IRE V TDGNGTIE Y A EF	68	2
CALB_NAEGR	34	FKKLDKDGNGTISK DEF	49	66 ISIF D ENGDSGVNF K EF	82	2
CALM_DICDI	17	FSLFDKDGSGSIT TKEL	33	54 INE V DADGNNGNID F PEF	70	2
CALM_YEAST	17	FALFDKDNNGSIS SSEL	33	53 MNE I DVDGNHQIE F SEF	69	2
CATR_NAEGR	38	FDLFDMDGSGKID ADEL	54	73 MIS G IDNGSGKID F NDF	89	1
CAYP_CANFA	30	FRRLDRDRSRSLD SREL	45	66 CRR W DRDGSGLD L EEF	82	2
CDP1_ORYSA	383	FKAMDTKNSRVV TGEL	399	418 ME A ADDTTSTIN W EEFI	434	1
CDP2_ORYSA	394	FTNMDTDNSGTI TYEEL	410	430 ME A ADV D GNGSID V VEF	446	2
CDP3_ORYSA	389	FKAVDTKNSRVV TGEL	405	425 ME A AH D N N V T IHY E EF	441	1
PLSL_MOUSE	18	FAKVDTDGNGYI SCNEL	33	58 MAT G DLDDGKIS F DEF	74	2
SCP_NERDI	12	FNRIDFDKDGAI TRMDF	27	44 EH A KV L M D S L T G V W DNF	60	1
TPCC_COTJA	24	FDIFVLGAEDGCI STKE	40	61 I D E V D E D G S G T V D F DQF	77	0
TPCC_RABIT	24	FDIFVLGAEDGCI STKE	40	65 E L Q E M I D E V D E D G S G T	76	0
TPCS_RANES	26	FDMFDTDGGDI STKEL	42	62 I E V D E D G S G T I D F E EF	78	2
B. C-terminal protein domain						
SWISS-PROT name ^a	Sequence Motif I ^b		Sequence Motif III ^b		Calcium binding sites ^c	
APLC_APLCA	104	FRLYDLDNDGFI TRDEL	120	152 FQ V M D K N K D D K L T F D EF	168	2
BTV3_BETVE	139	FKVFD E D E D G D G Y I S A R E L	155	177 I V S V D S N R D G R V D F F E F	193	2
CA22_RAT	118	FRLYDLDKDKI SRDEL	134	159 I Q E A D Q D G S A I S F T E F	175	2
CABO_LOLPE	89	FRVFDKDNGLI TAAEL	105	126 I R E A D I D G D G M V N Y E E F	142	2
CALB_NAEGR	103	FKVYDIDG D G Y I S N G E L	119	144 I L E A D E D G D G K I S F E E F	160	2
CALM_DICDI	91	FKVFDKIDG N G Y I S A E L	107	127 I R E A D L D G D G Q V N Y D E F	143	2
CALM_YEAST	90	FKVFDKNGDGLI SAAEL	106	125 M L R E V S D G S G E I N I Q Q F	141	1
CATR_NAEGR	110	FRLF E D E D S G F I T F A N L	126	146 I E E A D R S N Q Q I S K E D F	156	1
CAYP_CANFA	102	FAKLDRSGD G V V T V D D L	118	146 D N F D S S E K D G Q V T L A E F	162	1
CDP1_ORYSA	453	FTYFDK D G S G F I T V D K L	469	487 I L E V D Q N N D G Q I D Y A E F	503	1
CDP2_ORYSA	466	FQYFDK D K N S G F I T R D E L	482	501 I S E V D T D N D G R I N Y E E F	517	2
CDP3_ORYSA	461	FTYFDK D G S G Y I T V D E L	477	495 I S E V D Q N N D G Q I D Y A E F	511	2
GCAP_BOVIN	95	FKLYD V D G N G C I D R D E L	111	139 F S K I D V N G D G E L S L E E F	155	2
NCS1_CAEEL	104	FKLYD L D Q D G F I T R N E M	120	152 F R M D K N N D A Q L T L E E F	168	2
NECD_BOVIN	104	FSMYD L D G N Y I S K A E M	120	152 F R Q M D T N R D G K L S L E E F	168	2
NUBN_HUMAN	249	FILHD I N S D G V L D E Q E L	265	301 M K N V D T N Q D R L V T L E E F	317	2
ONCO_HUMAN	47	FRFID N D Q S G Y L D E E E L	63	86 M A A A D N D G D G K I G A E E F	102	2
PLSI_HUMAN	20	FNKID I D N S G V S D Y E L	36	60 L S V A D S N K D G K I S F E E F	76	2
PRVA_MACFU	47	FHILDK D K S G F I E E D E L	63	86 M A A G D K D G D G K I G V D E F	102	2
PRVB_MERMR	47	FVFID Q D K S G F I E E D E L	63	86 L K A G D S D G D G A I G V E E W	101	2
SCP_NERDI	100	FRAVDT N E D N N I S R D E Y	116	134 L D A I D T N D G L L S L E E F	150	2
TPCC_COTJA	101	FRMFD K N A D G Y I D L E E L	117	137 M K D G K N N D G R I D Y D E F	153	1
TPCC_RABIT	101	FRMFD K N A D G Y I D L E L	117	137 M K D G K N N D G R I D Y D E F	152	2
TPCS_RANES	102	FRIFD K N A D G Y I D S E L	118	138 M K D G K N N D G K I D F D E F	154	2
VIS3_RAT	104	FSMYD L D G N Y I S R S E M	120	152 F R Q M D T N D G K L S L E E F	168	2

^aThe SWISS-PROT names are given; see Experimental Procedures for more information.

^bMotif sequences, bracketed by their start and stop positions, are shown. Residues at key positions are shown in bold when they agree with the relationships given here (see Figure 2).

^cThe number of functional calcium-binding sites in the protein.

caltractin from *Naegleria gruberi* (CATR_NAEGR) is not in the correct motif position. Even the mutation from an aspartate to a glutamate in the N-terminal domain motif I of aplycalcin from the Californian Sea Hare (APLC_APLCA) is not tolerated. Since the aspartate residue is directly involved in coordinating the calcium ion, its substitution by a residue with a larger sidechain may perturb the coordination geometry, and thus will most likely prevent or dramatically decrease the calcium-binding affinity of this motif. There are three isoforms of calcium-dependent protein kinase from *Oryza sativa* (rice). One of them (CDP2_ORYSA) appears to be experimentally fully functional, while the other two (CDP1_ORYSA and CDP3_ORYSA) have lost their functional type II motifs and hence are likely to be less sensitive to calcium. In beta parvalbumin from whiting fish (PRVB_MERMR), a tryptophan is substituted for the conserved phenylalanine residue in the type III motif, as is also true for skeletal troponin C from the European eel (GenPept 633784; Francois *et al.*, 1993). This substitution most likely decreases the affinity of this motif for the binding of calcium. Skeletal troponin C from both Japanese quail (TPCC_COTJA) and rabbit (TPCC_RABIT) have multiple

essential mutations in both EF-hand motifs in the N-terminal domain, and neither of these proteins binds calcium in this domain.

Structure of Human Calmodulin

Human calmodulin is an example of a calcium-binding protein whose structure is known and which has sequence motifs that satisfy the three relationships in both N-terminal and C-terminal Ca-binding domains. In the structure for human calmodulin (Figure 1), the two conserved phenylalanine residues, one each from sequence motifs I and II, and I and III, for each of the two Ca-binding domains interact directly with each other. This interaction is a pi-electron electrostatic interaction, with the ring plane of one residue in each pair found nearly perpendicular to the ring plane of the other residue. This is the preferred orientation of the ring structures in such pi-stacking interactions. The structural interactions of the conserved calcium-coordinating acidic residues, all aspartates for this protein, with the calcium ions are also shown. The resulting structure of a calcium-binding domain is a globular structure, containing two or fewer bound calcium ions.

Discussion

In this study a new classification scheme based on inter-motif relationships is presented for EF-hand calcium-binding proteins with multiple pairs of EF-hand motifs. Its key criteria, positional dependency of one of the EF-hand sequence motifs with respect to its neighboring motif, and distinct conformational characteristics associated with the interaction of some of the strictly conserved motif residues, for example, Phe-Phe interactions, between two motif neighbors, not only allow assignment of novel proteins to the superfamily of EF-hand calcium-binding proteins but also provide highly reliable information regarding structural and functional features, for example, ligand binding affinity, of such novel proteins.

In addition, this classification approach allows for the first time predictions of the calcium-binding affinity of EF-hand domains in a novel protein by simply inspecting the sequence of the protein and applying the relationship criteria reported in this study (Figure 2). Extensive experimental evidence supports the presence of different types of EF-hand motifs within one protein molecule exhibiting variable affinities for calcium, for example, in calmodulin, troponin C, and other common EF-hand calcium-binding proteins with multiple calcium-binding sites. Many of these proteins contain both N-terminal and C-terminal globular calcium-binding domains, each of which consists of two structurally interacting EF-hand motifs. Though being structurally similar, these globular domains, for example, in troponin C, have been shown to respond quite differently to binding of calcium ions (Potter and Gergely, 1975; Collins *et al.*, 1991; Houdusse *et al.*, 1997) and also exhibit different measures of affinity for calcium. Generally, the N-terminal domain of these molecules shows lower affinity for calcium ions than does the C-terminal domain.

The observed differences in ligand binding affinity of the two protein domains of troponin C (Potter and Gergely, 1975), and other two-domain Ca-binding proteins considered here, can be accounted for by application of the motif relationship criteria presented here. Within the N-terminal domain of these proteins, an EF-hand motif characterized by a type I sequence motif is followed by a second EF-hand motif characterized by a type II sequence motif. However, at the C-terminal domain, a type I sequence EF-hand motif is followed by a type III sequence EF-hand motif. The type III sequence EF-hand motif is characterized by presence of an acidic residue PAIR at positions 1 and 5 (Figure 3, see also Table 3). Such an acidic pair has been experimentally shown (George *et al.*, 1996) to account for increased calcium ion affinity of the respective binding site. Thus, absence of the acidic residue pair in the N-terminal EF-hand domain of these proteins (sequence motifs type I and II) provides a rationale for the observed lower calcium-binding affinity associated with this domain.

Another example further illustrates the consistency of experimental evidence concerning calcium-binding affinity and calcium-binding specificity of EF-hand calcium-binding proteins with the functional predictions according to the motif relationship requirements presented here. Typically, troponin C molecules consist of three (cardiac isoform) or four (skeletal isoform) EF-hand calcium-binding motifs (Houdusse *et al.*, 1997). Troponin C molecules from five invertebrates (sea squirt, cray fish, lobster, barnacle, and

horseshoe crab) have been shown to exhibit different calcium-binding properties from those found in skeletal and cardiac tissue of vertebrates (Takagi *et al.*, 1994). The invertebrate proteins are found to bind only two calcium ions as opposed to the three or four expected to be bound, and indeed are bound, by the vertebrate troponin C molecules. Furthermore, the two functional calcium-binding sites (II, N-terminal, and IV, C-terminal) of the invertebrate molecules have both been shown to exhibit a low to moderate affinity for calcium-binding (10^5 M^{-1}). Two of the four calcium-binding sites (I and III) of these molecules are believed to be no longer functional due to mutations in essential calcium-coordinating residues. According to the second of the motif relationship requirements (Figure 2), the remaining functional calcium-binding sites II and IV should be EF-hand motifs characterized by a type II sequence motif (lower affinity motif) and a type III sequence motif (higher affinity, C-terminal specific), respectively. However, the EF-hand motif of site IV (C-terminal) in the invertebrate troponin C molecules is found to contain instead of a type III sequence motif a lower affinity type II sequence motif (asp replaced by ser at position 5; see Figure 3). This change can explain the low to moderate affinity for calcium-binding observed for both N- and C-terminal domains of these molecules.

Typically, the higher affinity C-terminal domain of troponin C molecules is less specific for calcium ions and can bind either calcium ions or magnesium ions, while the lower affinity N-terminal domain motifs are specific for calcium ions. In the case of the five invertebrate proteins mentioned above, both N- and C-terminal domains are found to be highly specific towards calcium ions (Wnuk, 1989; Collins *et al.*, 1991).

The three sequence motifs and their relationships as described here identify 188 of the 295 calcium-binding proteins in Swissprot 36.0 containing two or more EF-hand motifs. Examination of the EF-hand sequences in the remaining 107 proteins shows that 88 have amino acid changes at key hydrophobic positions. The remaining 19 proteins contain a different number of residues between the conserved phenylalanine residues than 57 ± 10 . Of the 88 proteins with key amino acid changes, 44 have a hydrophobic residue substituted for the conserved phenylalanine in both type I and type II (or III) sequence motifs, 21 have such a substitution in the type II (or III) sequence motif only, and 13 have such a substitution in the type I sequence motif only. Of these 88 proteins, 25 have a tryptophan substituted for the phenylalanine. 10 other proteins have a nonhydrophobic residue substituted for the conserved phenylalanine. New motifs and relationships are needed to define the structural and calcium-binding functional properties of these proteins (work in progress).

The approach of assigning structural and functional features to novel calcium-binding proteins introduced here shows considerably higher specificity (no false positives found) than the signature approach used by PROSITE (Bairoch *et al.*, 1997) and BLOCKS (Henikoff and Henikoff, 1994), but sacrifices sensitivity. This difference in sensitivity will decrease with the discovery of classes of Ca-binding proteins obeying modified relationships to those described here. Nevertheless, the high sensitivity approach typical of PROSITE signatures complements the high specificity approach presented here, and both methodologies are useful.

The sequence motif trends and relationships observed in EF-hand calcium-binding proteins are most likely not a unique feature found only in calcium-binding proteins but rather similar trends are expected to be found in other metal binding proteins or proteins with two or more specific motifs. Hence, identifying other motifs and inter-motif relationships in other classes of metal-binding proteins could enhance ability to predict structural and functional, e.g. ligand-binding affinity, properties of such molecules, thereby underscoring the importance of the approach presented here as a powerful predictive tool.

Experimental Procedures

BLASTP (Altschul *et al.*, 1990) was initially used to find homologues to human calmodulin, to determine highly conserved residues in EF-hand motifs, and to distinguish one EF-hand from another. The positions of the highly conserved phenylalanine residues were so delineated, resulting in the type I versus type II or type III EF-hand motifs. The GCG (Genetics Computer Group, Inc.; Devereux *et al.*, 1984) program FIND was then used to find all proteins in SWISS-PROT 36.0 (Bairoch and Apweiler, 1999) containing the type I motif. Among these proteins, FIND was then used to find those proteins containing either a type II motif or a type III motif. Sequences of the resulting proteins were then examined manually and Microsoft EXCEL was used to determine the number of residues present between the two conserved phenylalanine residues (see Figure 3). Tertiary structural properties of the calcium binding domains of Ca-binding proteins whose structure is known were determined using the MSI (Molecular Simulations, Inc.) programs HOMOLOGY and VIEWER and the MOE (Molecular Operating Environment) software package of the Chemical Computing Group, Inc. Distances between C-alpha carbon atoms were determined using HOMOLOGY. PDB, the Protein Databank (Sussman *et al.*, 1998), protein names (Figure 4, Table 1), with common organism and protein name, are: 4cln, *D. melanogaster* calmodulin; 1cll, human calmodulin; 1osa, paramoecium calmodulin; 1lin, bovine calmodulin; 1cmf, bovine calmodulin, apo form; 1cfc, frog calmodulin; 1pva, pike alpha parvalbumin; 4cpv, carp beta parvalbumin; 5pal, shark alpha parvalbumin; 1rtp, rat alpha parvalbumin; 1rro, rat oncomodulin; 1tco, bovine calcineurin B; 5tnc, turkey troponin C; 1top, chick troponin C; 1tnq, chick troponin C, N-terminal domain; and 1tnw, chick troponin C, NMR structure. SWISSPROT 36.0 (Bairoch and Apweiler, 1999) protein names (Table 3), with accession number and common organism and protein name, are: APLC_APLCA, Q16981, aplycalcin from californian sea hare; BTV3_BETVE, P43187, allergen Bet V III from white birch; CA22_RAT, Q62877, calcium binding protein P22 from rat; CABO_LOLPE, P14533, squidilin from longfin squid; CALB_NAEGR, P42322, calcineurin beta subunit from *Naegleria gruberi*; CALM_DICDI, P02599, calmodulin from *Dictyostelium discoideum*; CALM_YEAST, P06787, calmodulin from baker's yeast; CATR_NAEGR, P53441, caltractin from *Naegleria gruberi*; CAYP_CANFA, P10463, calcyphosine from dog; CDP1_ORYSA, P53682, CDP2_ORYSA, P53683, CDP3_ORYSA, P53684, calcium dependent protein kinases, isoforms 1, 2, and 11, respectively, from rice; GCAP_BOVIN, P46065, bovine guanylate cyclase activating protein 1; NCS1_CAEEL, P36608, neuronal calcium sensor from *Caenorhabditis elegans*; NECD_BOVIN, P29554, bovine neurocalcin; NUBN_HUMAN, Q02818, human nucleobindin precursor; ONCO_HUMAN, P32930, human oncomodulin; PLSI_HUMAN, Q14651, human I-plastin; PRVA_MACFU, P80050, alpha parvalbumin from japanese macaque; PRVB_MERMR, P02621, beta parvalbumin from whiting; SCP_NERDI, P04571, sarcoplasmic calcium binding protein from sandworm; TPCC_COTJA, P05936, cardiac troponin C from japanese quail; TPCS_RANES, P02589, skeletal troponin C from frog; and VIS3_RAT, P35333, visinin-like protein 3 from rat.

Acknowledgements

This work was supported by NIH program grant HG00904 to DWS as part of the *Dictyostelium* Developmental Gene Program. We thank Russell Doolittle and Michael Gribskov for helpful comments on the manuscript.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Molec. Biol.* 215: 403-410.
 Bairoch, A., and Apweiler, R. 1999. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.* 27: 49-54.

Bairoch, A., Bucher, P., and Hofmann, K. 1997. The PROSITE database, its status in 1997. *Nucleic Acids Res.* 25: 217-221.
 Bockerhoff, S.E., Edmonds, C.G., and Davis, T.N. 1992. Structural analysis of wild-type and mutant yeast calmodulins by limited proteolysis and electrospray ionization mass spectrometry. *Protein Sci.* 1: 504-516.
 Chandra, M., McCubbin, W.D., Oikawa, K., Kay, C.M., and Smillie, L.B. 1994. Ca⁺⁺, Mg⁺⁺, and troponin I inhibitory peptide binding to a Phe-154 to Trp mutant of chicken skeletal muscle troponin C. *Biochem.* 33: 2961-2969.
 Collins, J.H., Theibert, J.L., Francois, J.-M., Ashley, C.C., and Potter, J.D. 1991. Amino acid sequences and Ca⁺⁺-binding properties of two isoforms of barnacle troponin C. *Biochem.* 30: 702-707.
 Corson, D.C., Williams, T.C., Kay, L.E., and Sykes, B.D. 1986. 1H NMR spectroscopic studies of calcium-binding proteins. 1. Stepwise proteolysis of the C-terminal alpha-helix of a helix-loop-helix binding domain. *Biochem.* 25: 1817-1826.
 Devereux, J., Haerberli, P., and Smithies, O. 1984. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* 12: 387-395.
 Francois, J.-M., Gerday, C., Prendergast, F.G., and Potter, J.D. 1993. Determination of the Ca⁺⁺ and Mg⁺⁺ affinity constants of troponin C from eel skeletal muscle and positioning of the single tryptophan in the primary structure. *J. Muscle Res. Cell Motil.* 14: 585-593.
 Fujimori, K., Sorenson, M., Herzberg, O., Moul, J., and Reinach, F.C. 1990. Probing the calcium-induced conformational transition of troponin C with site-directed mutants. *Nature.* 345: 182-184.
 George, S.E., Su, Z., Fan, D., Wang, S., and Johnson, D.J. 1996. The fourth EF-hand of calmodulin and its helix-loop-helix components: impact on calcium binding and enzyme activation. *Biochem.* 35: 8307-8313.
 Henikoff, S., and Henikoff, J.G. 1994. Protein family classification based on searching a database of blocks. *Genomics.* 19: 97-107.
 Hermann, A., and Cox, J.A. 1995. Sarcoplasmic calcium-binding protein. *Comp. Biochem. Physiol. B Biochem. Molec. Biol.* 111: 337-345.
 Houdusse, A., Love, M.L., Dominguez, R., Grabarek, Z., and Cohen, C. 1997. Structures of four Ca⁺⁺-bound troponin C at 2.0 Å resolution: further insights into the Ca⁺⁺-switch in the calmodulin superfamily. *Structure.* 5: 1695-1711.
 Ikura, M. 1996. Calcium binding and conformational response in EF-hand proteins. *Trends Biochem. Sci.* 21: 14-17.
 Kretsinger, R.H., and Kockolds, C.E. 1973. Carp muscle calcium-binding protein. II. Structure determination and general description. *J. Biol. Chem.* 248: 3313-3326.
 Martin, S.R., Maune, J.F., Beckingham, K., and Bayley, P.M. 1992. Stopped-flow studies of calcium dissociation from calcium-binding-site mutants of *Drosophila melanogaster* calmodulin. *Eur. J. Biochem.* 205: 1107-1114.
 Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Molec. Biol.* 247: 536-540.
 Persechini, A., Moncrief, N.D., and Kretsinger, R.H. 1989. The EF-hand family of calcium-modulated proteins. *Trends Neurosci.* 12: 462-467.
 Potter, J.D., and Gergely, J. 1975. The calcium and magnesium binding sites on troponin and their role in the regulation of myofibrillar adenosine triphosphatase. *J. Biol. Chem.* 250: 4628-4633.
 Sekharudu, Y.C., and Sundaralingam, M. 1988. A structure-function relationship for the calcium affinities of regulatory proteins containing 'EF-hand' pairs. *Protein Eng.* 2: 139-146.
 Starovasnik, M.A., Su, D.R., Beckingham, K., and Klevit, R.E. 1992. A series of point mutations reveal interactions between the calcium-binding sites of calmodulin. *Protein Sci.* 1: 245-253.
 Sussman, J.L., Lin, D., Jiang, J., Manning, N.O., Prilusky, J., Ritter, O., and Abola, E.E. 1998. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta. Crystallogr. D Biol. Crystallogr.* 54: 1078-1084.
 Takagi, T., Petrova, T., Comte, M., Kuster, T., Heizmann, C.W., and Cox, J.A. 1994. Characterization and primary structure of amphioxus troponin C. *Eur. J. Biochem.* 221: 537-546.
 Waltersson, Y., Linse, S., Brodin, P., and Grundstrom, T. 1993. Mutational effects on the cooperativity of Ca⁺⁺ binding in calmodulin. *Biochem.* 31: 7866-7871.
 Wnuk, W. 1989. Resolution and calcium-binding properties of the two major isoforms of troponin C from crayfish. *J. Biol. Chem.* 264: 18240-18246.