The Protein Journal, Vol. 25, No. 1, January 2006 (© 2006)
DOI: 10.1007/s10930-006-0011-7

Prediction of Three Dimensional Structure of Calmodulin

Ke Chen,[1] Jishou Ruan,[1] and Lukasz A. Kurgan[2,3]

Received August 4, 2005

Calmodulin (CaM) is an important human protein, which has multiple structures. Numerous researchers studied the CaM structures in the past, and about 50 different structures in complex with fragments derived from CaM-regulated proteins have been discovered. Discovery and analysis of existing and new CaM structures is difficult due to the inherent complexity, i.e. flexibility of 6 loops and a central linker that constitute part of the CaM structure. The extensive interest in CaM structure analysis and discovery calls for a comprehensive study, which based on the accumulated expertise would design a method for prediction and analysis of future and existing CaM structures. It is also important to find the mechanisms by which the protein adjusts its structure with respect to various factors. To this end, this paper analyzes the known CaM structures and finds four factors that influence CaM structure, which include existence of Ca^{2+} binding, different binding segments, measuring surroundings, and sequence mutation. The degree of influence of specific factors on different structural regions is also investigated. Based on the analysis of the relation between the four factors and the corresponding CaM structure a novel method for prediction of the CaM structure in complex with novel segments, given that the surroundings of the complex, is developed. The developed prediction method is tested on a set aside, newest CaM structure. The prediction results provide useful and accurate information about the structure verifying high quality of the proposed prediction method and performed structural analysis.

KEY WORDS: Calmodulin; Ca^{2+}-binding protein; protein structure; binding segment.

1. INTRODUCTION

Calmodulin (CaM) is a ubiquitous Ca^{2+}-binding protein consisting of 148 residues, which plays important role in the Ca^{2+}-dependent signaling pathways of eukaryotic cells (Klee and Vanaman, 1982). A wide range of physiological processes are mediated by CaM through Ca^{2+}-dependent regulation of target enzymes such as myosin light chain kinase (MLCK), CaM-dependent kinases, protein phosphatase calcineurin, phosphor-diesterase, nitric oxide syntheses, Ca^{2+}-ATPase pumps as well as cytoskeletal structural proteins (Means *et al.*, 1991; Vogel, 1994; James *et al.*, 1995). This broad functionality is manifested in the CaM binding regions of these target proteins, which differ significantly in their amino acid (AA) sequences. Recently, several crystal and NMR structures of CaM in complex with fragments derived from CaM-regulated proteins revealed novel ways for CaM to interact with its targets, including the Ca^{2+}-activated K^{+}-channel and the anthrax exotoxin (Drum *et al.*, 2001, 2002; Schumacher *et al.*, 2001; Shen *et al.*, 2002a, 2004). Analysis of interactions of the CaM with other proteins is based on knowledge and analysis of the

[1] College of Mathematical Sciences and LPMC, Chern Institute of Mathematics and Liuhui Center for Applied Mathematics, Nankai University, Tianjin, 300071, Peoples Republic of China.
[2] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada.
[3] To whom correspondence should be addressed. E-mail: lkurgan@ece.ualberta.ca

Abbreviation: CaM, calmodulin; MLCK, myosin light chain kinase; AA, amino acid; NMR, nuclear magnetic resonance; PDB, Protein Data Bank; 3D, three-dimensional; RMSD, root mean square deviation.

CaM structure. As of October 2004 there were 45 experimental CaM structures in the Protein Data Bank (PDB) (Berman *et al.*, 2000), including Ca^{2+}-free, Ca^{2+}-bound, and bound to various protein segments structures. In contrast with the number of protein segments that can interact with CaM, the number of CaM structures is relatively low. In fact, three-dimensional (3D) CaM structure is still highly adverted and several new structures of CaM complex with novel segments are mensurated and reported every year, see Fig. 1. Since the number of new structures in recent years becomes larger, i.e. since 2000, 18 structures were reported, compared to 27 structures reported between 1988 and 1999, development of a method for prediction and analysis of feature CaM structures would provide invaluable assistance for researchers in this field. To this end, this paper develops a prediction method for the CaM structure in complex with new segments. The development is based on analysis of the known CaM structures and tests the developed method by performing successful 3D structure prediction on an excluded from the development process structure.

A cartoon diagram that shows overall CaM structure together with its binding segments is shown in Fig. 2. The 3D structure consists of (1) a linker, which is shown using the middle 'bridge', (2) N-domain and C-domain, which are shown using 'clouds', and (3) protein segments, which are shown using 'wings'. Additionally, $\oplus$ symbol depicts the Ca^{2+} molecules bound to CaM. Both Ca^{2+} and protein segments are frequently absent in different CaM structures.

Even though the cartoon diagram seems relatively simple, the CaM structure is difficult to analyze and predict due to its flexibility. To further analyze the CaM's structure, its secondary structure is considered. The secondary structure consists of six loops, one central linker (loop), and eight helices, see Fig. 3. The protein includes four so-called EF-hands, which consist of two helices and a loop as its central part and commonly appear in Ca^{2+}-activated proteins.

The known 45 CaM structures of are not always consistent even though some of them are in general congruent. This paper is the first to perform comprehensive and detailed analysis of different CaM structures. The analysis applies a sequence of the following steps:

Analysis of the CaM structure. The analysis is performed based on alignment of different CaM sequences in order to find conserved portions of the sequence, i.e. where no AAs are inserted or obliterated.

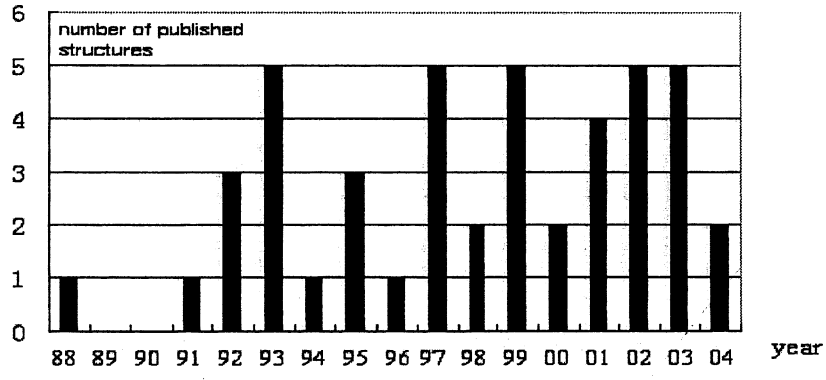


Fig. 1. The number of measured (mensurated) and published in PDB CaM-structures.

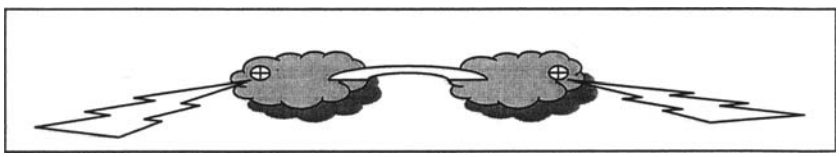
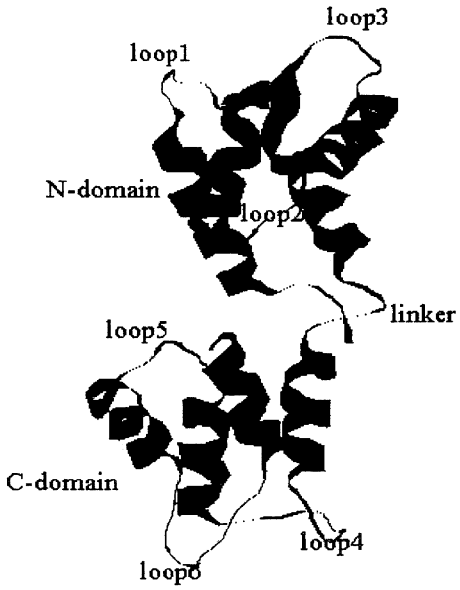


Fig. 2. Overview of the CaM structure, which consists of a linker (represented by the middle 'bridge'), N-domain and C-domain (represented by the 'clouds'), protein segments (represented by the 'wings'), and Ca^{2+} molecules that are bound to CaM (represented by $\oplus$ symbol).



helix1	EEQIAEFKEAFSLF	6-19
helix2	TKELGTVMRS	29-38
helix3	EAELQDMINEV	45-55
helix4	FPEFLTMMARK	65-75
helix5	EEEIREAFRV	82-91
helix6	AAELRHVMTN	102-111
helix7	DEEVDEMIRE	118-127
helix8	EEFVQMMT	139-146
loop1	DKDGDGTIT	20-28
loop2	LGQNPT	39-44
loop3	DADGNGTID	56-64
loop4	FDKDGNGYIS	92-101
loop5	LGEKLT	112-117
loop6	ADIDGDGQVNY	128-138
linker	MKDTDS	76-81

(a) cartoon diagram of the secondary structure **(b)** AA sequence of the corresponding secondary structure fragments

Fig. 3. Secondary structure and amino acid sequence of the 1CFC CaM protein; helixes and loops are numbered according to the order they appear in the sequence.

To measure the difference between two structures, we adopt root mean square deviation (RMSD) (Lesk, 1986). As a result, we determine which segments of the CaM's secondary structure are flexible, and which are structurally conserved.

Search for factors that control CaM structure. Each of the flexible structural segments is analyzed with respect to different CaM structures. The segments are analyzed and divided into three sets depending of their position with respect to the CaM's domains: (1) loops inside EF-hands, including loops 1, 3, 4 and 6, (2) loops connecting EF-hands, including loop 2 and loop 5, and (3) the central linker. A tree that represents structural relationship between the known CaM structures is developed.

Analysis of shifts in CaM structure. Based on the tree and description of the corresponding structures four factors, which control CaM structure, are found. A detailed analysis concerning how CaM shifts its structure with respect to each of the factors is performed.

Development and testing of the structure prediction method. After learning how each factor controls shifts of the CaM structure, a novel method for prediction of the CaM structure when binding to new segments is proposed and tested.

2. MATERIALS AND METHODS

2.1. Alignment of Different CaM Structures

There are 45 proteins in PDB related to CaM structure, which vary in their sequence information. Sequence alignment is used to find which portions of the sequences do not contain AAs, which are either inserted or obliterated. Super pairwise alignment (Shen *et al*., 2002b) is adopted to align the sequences, see Fig. 4.

The alignment considers the 45 known versions of the CaM sequence and shows that residues at 48 sites are different between different sequences. The differences at 45 sites are due to mutation, and the changes at the remaining three sites are due to missing AAs. The 1AHR protein is missing Thr AA and Asp AA at positions 79 and 80, and the 1DEG protein is missing Glu at position 84. Thus, there are no missing AAs in the two regions: from position 1(Ala) to position 78

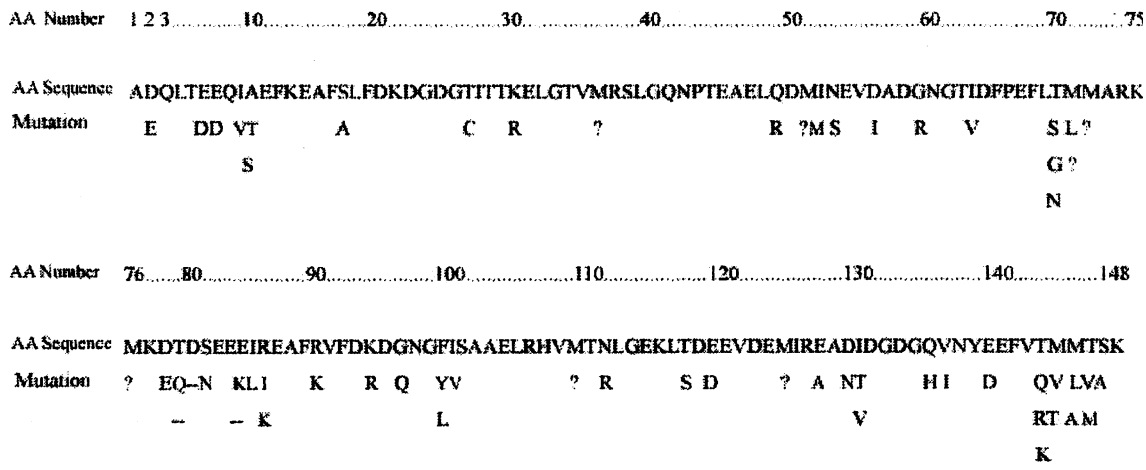


Fig. 4. The multiple alignment result for CaM sequences; the 'AA Number' rows show residue number, the 'AA Sequence' rows show the most probable, based on alignment, AA for each position, and the 'mutation' rows show all possible mutations at each position where '?' stands for the nonstandard AA 'MSE'.

(Asp) and from position 85 (Leu) to position 145 (Met). Therefore the structures within these two regions can be directly compared.

2.2. Determining Flexible and Conserved Segments of the CaM Structure

The flexible and the invariant CaM regions are found using method proposed by Holm and Sander (1993). The CaM sequence is divided into seven residues long segments, i.e. 1st to 7th residues, 2nd to 8th residues, ..., and 142nd to 148th residues. For each of the segments RMSD is computed for different CaM structures and used to judge whether the segment's structure is flexible or conserved (invariant). The average RMSD values are shown in Fig. 5. The x-axis shows the residue position while the y-axis gives the average RMSD between the segment structures, where values for central AA are given. The six loops and the linker are in white color, and the eight helices are in black.

The figure shows seven wave crests, which denote flexible CaM structure regions. More specifically, the seven flexible regions are: position 20 (Asp) to position 28 (Thr), position 39 (Leu) to position 44 (Thr), position 56 (Asp) to position 64 (Asp), position 73 (Ala) to position 85 (lie), position 92 (Phe) to position 101 (Ser), position 112 (Leu) to position 117 (Thr), and position 128 (Ala) to position 138 (Tyr). Coincidently, the seven regions constitute the six loops and the central linker shown in Fig. 3. Additionally, the central linker is much more flexible than other six loops, and it is the most flexible region. For the remaining CaM regions, which include the eight helices, there are no significant differences in the CaM structures. Next, factors that affect the structure of the loops and the linker are discussed.

2.3. Analysis of the Loop Structures

Initial analysis concentrates on nine, among the 45, mensurated in solution CaM structures, i.e. 1CFF, 1MUX, 1CFC, ICED, 1DMO, 1CKK, 1NWD, 2BBM, and 2BBN (Ikura *et al.*, 1992; Kuboniwa *et al.*, 1995; Zhang *et al.*, 1995; Osawa *et al.*, 1998, 1999; Elshorst *et al.*, 1999; Yap *et al.*, 2003; Guo *et al.*, 2004). The last four proteins are structurally conserved, and the maximal RMSD between their multiple structures is less than 3A. The other five proteins are flexible although there are structurally conserved within their N-domain and the C-domain, which is shown using values on diagonal in Table 1.

The values give the RMSD between the domains, for instance 1.7–2.3 in the first row and second column is the RMSD between the N-domain of 1CFF and the N-domain of 1MUX. The "—" denotes value given in the upper triangle. The multiple structures of the N-domain of any of the five proteins have a maximal RMSD less than 1.5 A, and for C-domain they have a maximal RMSD less than 2 A. Therefore, the five proteins have virtually the same domain structure. Table 1 also shows that

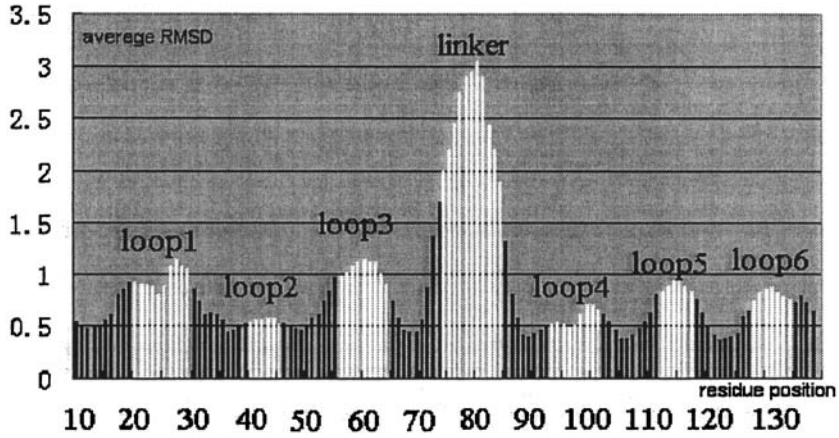


Fig. 5. Average RMSD values for the CaM structures. The *x*-axis shows the residue position while the *y*-axis gives the average RMSD between the segment structures, where values for central AA are given. The six loops and the linker are in white color, and the eight helices are in black.

Table 1. The RMSD Values Computed between N- and C-domains of the Five Proteins Mensurated in Solution

Domain	N-domain					C-domain				
ID	1CFF	1MUX	1CFC	1CFD	1DMO	1CFF	1MUX	1CFC	1CFD	1DMO
1CFF	0.8–1.4	1.7–2.3	3.9–4.2	4.0–4.1	3.9–4.6	0.7–1.0	1.6–2.2	4.8–5.1	4.9–5.1	4.6–5.1
1MUX	–	0.7–1.2	3.8–4.2	3.9–4.1	4.1–4.8	–	0.7–1.3	4.3–4.8	4.4–4.7	3.9–4.8
1CFC	–	–	0.2–0.6	0.3–0.6	0.9–1.6	–	–	0.2–1.3	0.2–0.4	1.7–2.3
1CFD	–	–	–	0	0.9–1.8	–	–	–	0	1.7–2.2
1DMO	–	–	–	–	0.8–1.5	–	–	–	–	0.9–2.0

the structure of the two 1CFF's domains is similar to the structure of 1MUX domains. Also, 1CFC, 1CFD, and 1DMO domains are similar. Based on this discovery, further analysis of similarities between different CaM proteins is performed.

2.4. Analysis of the 3D-structure of the Loops 1, 3, 4 and 6

As showed in Fig. 3, CaM includes four EF-hands, which include loops 1, 3, 4 and 6, respectively, and are denoted as EF-hand$_i$, where $i=1,2,3,4$. In general there are two types of EF-hands in the CaM proteins, see Figs. 6a and b. Since helices in the EF-hands are conserved structurally, as shown in Fig. 5, the 3D-structure difference between the multiple structures of the same EF-hand is most likely caused by the inside loops of the EF-hands. Thus, the analysis concentrates on the individual loops.

The number of residues of EF-hands is much smaller than that of integrated proteins, and RMSD usually increases with the increase of number of residues of the compared structures. Therefore, when comparing two structures the commonly used threshold of 3 A for RMSD criterion should not be adopted to judge if the two structures of EF-hands are similar. Instead, the 1.5 A threshold is used, which results in ability to detect more subtle differences in structure. An example is used to illustrate that two EF-hand structures with RMSD of 1.5 A have little difference in structure. The structure of EF-hand$_1$ in 1CFF (Elshorst *et al.*, 1999), see Fig. 6c, is similar to the structure of EF-hand$_1$ in 1A29 (Vandonselaar *et al.*, 1994), see Fig. 6a. The only difference between the two hands is between positions 19 (Phe) and 22 (Asp), which are a little bend in the helix1 shown in light gray in the corresponding figures.

The EF-hand$_1$ structures from the 45 proteins are divided into two types: A and B using the threshold of 1.5 A. Every pair of structures for which value of the RMSD is greater than 3 A is of different types, while RMSD less than 1.5 A denotes pair of structures of the same type. Based on this

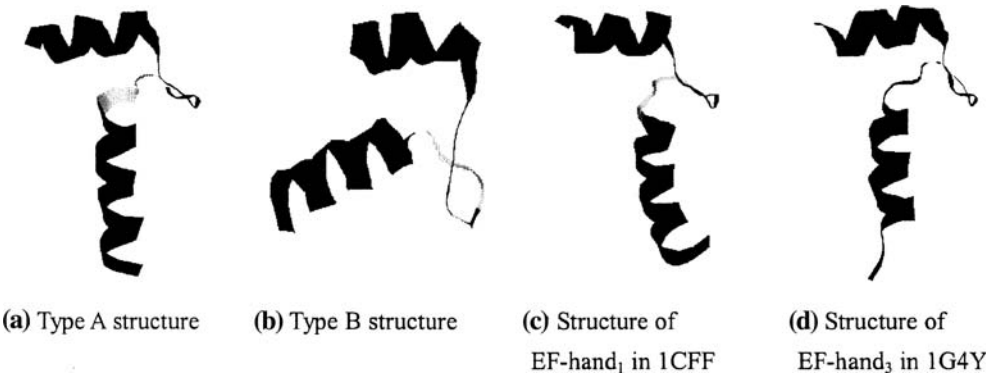


Fig. 6. Typical EF-hand structures; EF-hand consists of two helices and an inter-helical loop, which together form a Ca^{2+} binding site. Types A and B are typical EF-hand structures for CaM, while structures shown in (c) and (d) are relatively rare.

assumption, type A structures are characteristic for 36 proteins including 1A29, 1AHR, 1CDL, 1CDM, 1CKK, 1CLL, 1CLM, 1CM1, 1CM4, 1CTR, 1EXR, 1G4Y, 1GGZ, 1IQ5, 1IQW, 1L7Z, 1LIN, 1MUX, 1MXE, 1NIW, 100J, 1OSA, 1PRW,1QIV, 1QIW, 1QS7, 1QTX, 1RFJ, 1VRK, 3CLN, 4CLN 1CFF, 1DEG, 1NWD, 2BBM and 2BBN (Taylor *et al.*, 1991; Chattopadhyaya *et al.*, 1992; Ikura *et al.*, 1992; Meador *et al.*, 1992; Meador *et al.*, 1993; Rao *et al.*, 1993; Ban, 1994; Cook *et al.*, 1994; Vandonselaar *et al.*, 1994; Tabernero *et al.*, 1997; Wall *et al.*, 1997; Babu *et al.*, 1998; Elshorst *et al.*, 1999; Mirzoeva *et al.*, 1999; Osawa *et al.*, 1999; Harmat *et al.*, 2000; Wilson *et al.*, 2000; Kurokawa *et al.*, 2001; Schumacher *et al.*, 2001; Clapperton *et al.*, 2002; Han *et al.*, 2002; Aoyagi *et al.*, 2003; Fallon *et al.*, 2003; Symersky *et al.*, 2003; Yamauchi *et al.*, 2003; Yap *et al.*, 2003; Matsubara *et al.*, 2004; Yun *et al.*, 2004). On the other hand, type B structures are characteristic for the remaining nine proteins including 1CFC, 1CFD, 1DMO, 1K90, 1K93, 1LVC, 1QX5, 1S26 and 1SK6 (Kuboniwa *et al.*, 1995; Zhang *et al.*, 1995; Drum *et al.*, 2002; Shen *et al.*, 2002a; Guo *et al.*, 2004; Shen *et al.*, 2004; Schumacher *et al.*, 2004). Following the same procedure, classification for EF-hand$_2$ structures gives the same results as for EF-hand$_1$.

The structure types for EF-hand$_3$ and EF-hand$_4$, which are also classified into types A and B, are assigned to different sets of proteins. Type B is characteristic for 1CFC, 1CFD, 1DMO, and 1QX5 proteins, while for the remaining proteins type A is characteristic. The consistency of the results for the EF-hand$_1$ and EF-hand$_2$ pair, and the EF-hand$_3$ and EF-hand$_4$ pair is due to the overall CaM structure, where the corresponding pairs belong to the same domains.

All type A EF-hand structures bind to Ca^{2+}, while type B structures are Ca^{2+}-free, except the EF-hand$_3$ and EF-hand$_4$ in the 1G4Y protein. The 1G4Y protein is the only crystal and the Ca^{2+}-free structure. Its EF-hands have a RMSD of only 1.6 A from type A, which is surely not induced by Ca^{2+}. Therefore, we believe that crystal surrounding can change and stabilize the structure of the EF-hands to a certain degree. Despite this exception, obviously Ca^{2+} binding is the main factor in determining EF-hand structure. The next section concentrates on analysis of the remaining two loops.

2.5. Analysis of the 3D-structure of the Loops 2 and 5

The loop 2, which is located between position 39 (Leu) and 44 (Thr), is characterized by limited flexibility when compared with other loops. This can be observed in Fig. 5 where loop 2 has the lowest RMSD values. The flexibility of loop 2 is further studied by calculating the RMSD between the structures of segment from position 29 (Thr) to position 55 (Val), which include helix 2, loop 2 and helix 3. The short length of loop 2 does not allow for using it for calculating the RMSD and thus the adjacent helix 2 and helix 3, which are structure conserved, are added. Thus, the resulting difference between the structures of the entire segment is caused by the loop 2.

The RMSD values between the structures of the selected segment for the 45 CaM structures range between 0.1 and 1.4 A. Two structures of about 30 residues with RMSD of 1.5 A between them are not significantly different, for instance

compare type A structure in Fig. 6a and EF-hand$_1$ in 1CFF in Fig. 6c. Therefore we conclude that the loop 2 has limited flexibility with regard to its surroundings or binding of a segment, which are characteristic factors for the considered CaM structures.

The flexibility of the loop 5 is analyzed in the analogous way by selecting the segment from position 102 (Ala) to position 127 (Glu), which include helix 6, loop 5 and helix 7. The difference between the structures in this segment is caused by the loop 5 since again both helices are structurally conserved.

The results are different from the result for loop 2. First, 1QX5 protein has a loop 5 structure, which is different from the loop 5 structures of the other proteins. This structure, see Fig. 7b, forms a RMSD of at least 4.7 A from other structures. The reason for this difference is that the 1QX5 includes two CaM structures, which form a dimer by embedding the C-domains. Fig. 5 shows that the loop 5 is more flexible than the loop 2. After excluding the 1QX5, RMSD values between loop 5's structures vary between 0.2 and 2.2 A, while the RMSD for loop 2 is in the 0.1–1.4 A range. The largest value of 2.2 A for the loop 2 is for the 1A29 and 1DMO proteins, which are shown in Fig. 7a and c respectively. The two loop 5 structures are similar despite the relatively larger RMSD value and therefore we conclude that the loop 5 also has limited flexibility if the 1QX5 protein is excluded.

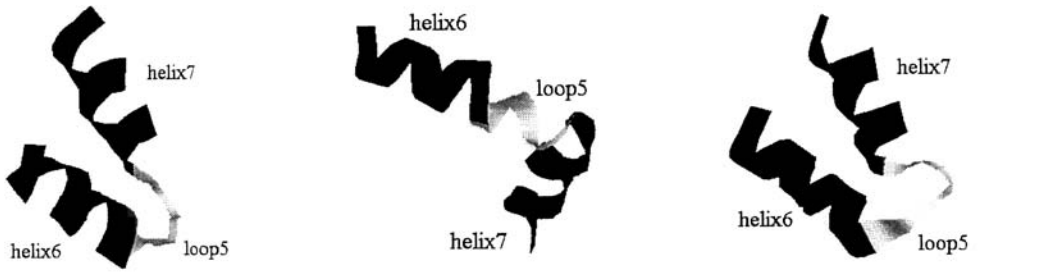
2.6. Analysis of the Central Linker Structure

The 1CDL, 1EXR, 1AHR, 1CFF, 1DEG, 1G4Y, 1MUX, and 1NWD proteins share a similar structure in both of the domains, and thus the unique significant difference between these proteins is the link style of the central linker. They cover the eight typical types of linker structure in CaM complexes. The eight types of structures for the above proteins are shown in Fig. 8.

Next, structure of the linker for the eight proteins is analyzed, and three factors that affect the structure of the linker are introduced:

X-ray crystal surroundings constitute the first factor that affects the linker structure, especially when CaM binds to no segment. The linker of CaM, which binds to no segment and is mensurated by X-ray diffraction, is a long helix, see Fig. 8b. This is caused by crystal packing (Barbato *et al.*, 1992; Wall *et al.*, 1997), and not the linker's intrinsic structure. At the same time, the linker of CaM, which binds to no segment and in mensurated in solution, e.g. in the 1CFC and 1DMO proteins, is flexible and folds into many shapes. At the same time, when linker is stabilized by the binding segment, such as when CaM binds to a helix-like segment, its crystal structure and solution structure are similar and both form structure shown in Fig. 8a. In fact, 21 out of the 45 CaM structures are in complex with helix-like segment and form this structure. They include 1A29, 1CDL, 1CDM, 1CM1, 1CM4, 1CTR, 1IQ5, 1IWQ,1L7Z, 1LIN, 1MXE, 1PRW, 1NIW, 1QIV, 1QIW, 1QS7, 1QTX and 1VRK, which are crystal structures, and 1CKK, 2BBM and 2BBN, which are solution structures.

Binding segments are the second factor that results in different linker structure. Figures 8a, e, f and h show CaM structure that binds to different segments. The linkers of these four structures are stabilized by the bound segments, and have visibly



(a) the structure of segment from ALA102 to GLU127 in the 1A29 **(b)** the structure of segment from ALA102 to GLU127 in the 1QX5 **(c)** the structure of segment from ALA102 to GLU127 in the 1DMO

Fig. 7. Typical loop 5 structures; (a), (b), and (c) show structures of segment from position 102 (ALA) to position 127 (GLU) in 1A29, 1QX5 and 1DMO respectively; this segment includes helix 6, loop 5 and helix 7.

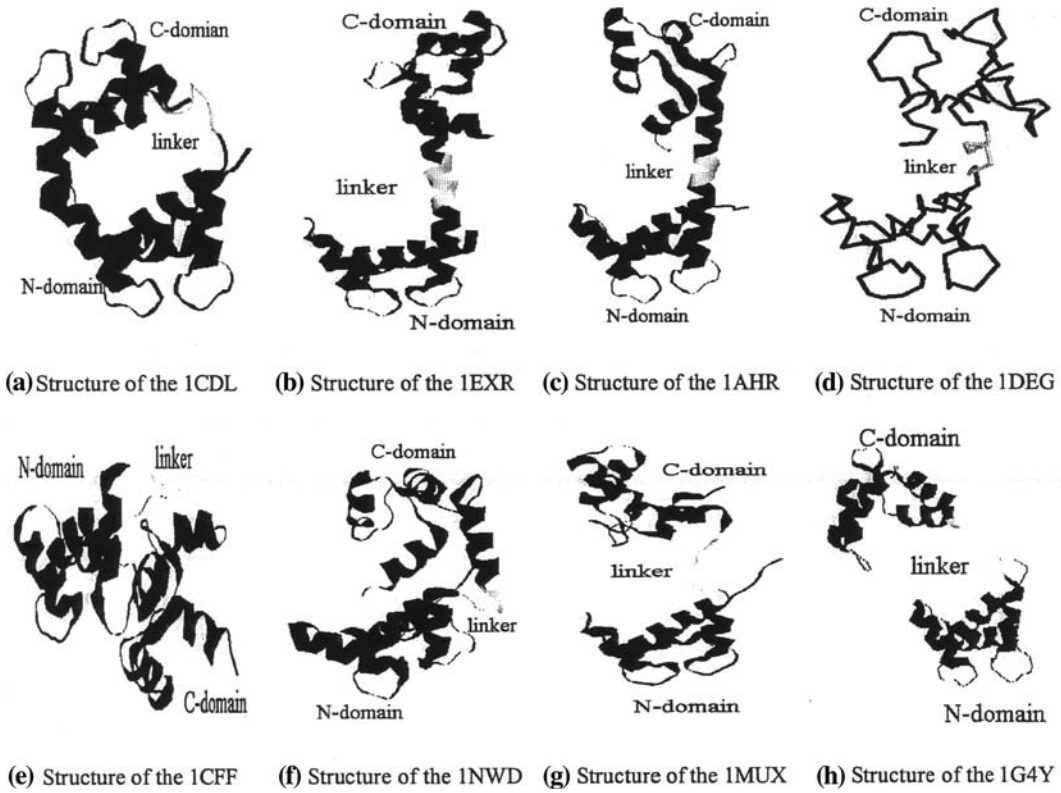


Fig. 8. Eight typical linkers of the CaM-structure; N- and C- domains are in black, and linker is in light gray; PDB only provides main chain of 1DEG, and thus only the main chain structure is shown.

different structure. As discussed in 1, after binding to a segment, the difference in linker structure is not caused the mensurating surroundings, i.e. in crystal or in solution. Instead, we believe that different segments and different binding style, such as binding region or number of the segments involved in the interaction, are the reasons for the difference of the linker structure.

Sequence mutations are the third factor. Missing linker residues affect CaM structure. Although 1AHR and 1DEG proteins share the same surroundings with the proteins in 1EXR, see Fig. 8b–d, they form different structures. The difference is due to missing AA Thr at position 79 and AA Asp at position 80 for the 1AHR, and missing AA Glu at position 84 for the 1DEG Missing residues do not change the secondary structure of the linker. The linker is still helical for both 1AHR and 1DEG, but the missing residues induce that the corresponding long helices are incomplete leading to different structures.

Further details about eight types of linker structures and their binding information are summarized in Table 2.

In short, the analysis shows that only the four loops, i.e. loops 1, 3, 4 and 6, and the linker are flexible are thus affect the resulting CaM structure. The remaining two loops and all helices are structurally conserved. The main factor that controls loop structure is Ca^{2+} binding, while for the linker the factors are binding segments, measuring surroundings, and sequence mutation. Next, the results of the above analysis are tied with respect to the specific CaM structures and presented in an easy to understand manner.

2.7. Analysis of the Structural Relation between the 45 Known CaM Structures

An RMSD distance tree is created based on computations independent of the results shown in Sections 2.3 and 2.4. The tree consists of three levels and visualizes relation between different CaM's structures in an easy to understand manner. Proteins in the same node of the third (lowest) level have similar overall structure. Proteins in the same

Table 2. Detailed Description of the Eight Types of the Linker Structure

PDB protein ID	Binding segment (BS)	BS length	BS conformation	Binding CaM region	Binding type	CaM conformation	Linker style	Reason for special link
1CDL	smMLCK	20	Helix	11–19, 32–39, 71–74, 89–92, 105–114, 124–128, 144–145	Domain-BS-domain	Collapsed	Loop	The most common linker
1EXR	None	–	–	–	–	Extended	Helix	Crystal packing
1AHR	None	–	–	–	–	Extended	Helix	Crystal packing and mutation at linker
1DEG	None	–	–	–	–	Extended	Helix	Crystal packing and mutation at linker
1NWD	Dimeric GAD domain	28	Helix	32–39, 48–55, 71–72, 85–92, 105–112, 124–128, 144–145	Domain-BS-BS-domain	Collapsed	Loop	Special dimeric BS and special binding model
1MUX	W-7 Peptide	–	Small molecule	92, 100, 105, 109, 124, 128, 136, 141,144, 145	Domain-BS	Extended	–	Special BS and special binding model
1CFF	C20W	20	Helix	100, 105, 124, 125, 128, 136, 144	Domain-BS	Extended	–	Special BS and special binding model
1G4Y	CAMBD in SK channel	82	Two helices	12, 19, 35, 36, 39, 68, 71, 72, 111–116	Domain-BS-domain	Collapsed	Loop	Special BS and special binding model

The binding type column shows how the binding segment (BS) binds. It can bind to one of the domains, to both domains, or it can be dimeric and in this case every monomer binds to one of the domains. The CaM conformation column has two values: "extended" stands for the case when both of the CaM domains are separated far away, and "collapsed" stands for the case when the two domains are drawn together the BS forming a 'cavity'.

branch of the second level have similar structure in both of the CaM domains. Finally, proteins in the same branch of the first level have similar structure in one of the two domains. To create the tree, for any pair of the CaM's structures (denoted as a and b), we define

$$\begin{aligned} d_N &= \mathrm{RMSD}(domainN_a, domainN_b) \\ d_C &= \mathrm{RMSD}(domainC_a, domainC_b) \\ d &= \mathrm{RMSD}(structure_a, structure_b) \end{aligned}$$

Vector $V_{ab} = (d_N, d_C, d)$ describes the difference between two structures. In general, if the RMSD between two structures is less than 2.5 A, the two structures are assumed similar (Lesk *et al.*, 1986; Holm and Sander, 1993). Therefore, if max (d_N, d_C, d) < 2.5 A than a and b structures are assumed similar and grouped in the same branch of the tree. The tree is constructed in the following manner:

Compute the RMSD vector V_{ab} between every two CaM structures.

If two structures are similar, that is if max (d_N, d_C, d) < 2.5 A, they are grouped in the same branch on the lowest (third) tree level.

If two branches obtained in the second step have similar structures in both N-domain and C-domain, that is if max $(d_N, d_C) < 2.5$ A, they are grouped under the same branch on the second tree level.

If the two branches obtained in the third step have similar structure in N-domain or C-domain, that is $d_N < 2.5$ A or $d_C < 2.5$ A, they are grouped under the same branch on the first tree level. The distance tree for the 45 known CaM structures is shown in Fig. 9.

Following, the found clusters (groups) of structurally similar CaM structures are described: The largest leaf node A contains 21 proteins. They have similar structure in spite of different surroundings and binding segments. These proteins bind to a segment, which interacts with both of the CaM domains to form a cave-like shape with the segment in its center.

The nine proteins included in the leaf node B have the same structure, which is different from the structure of proteins from node A. These proteins share the same surroundings, i.e. they are mensurated by X-ray, they do not bind to a segment, and their both domains are bound to Ca^{2+}. Their central linker forms a long helix and the two domains are separated relatively far away.

The six, i.e. 1G4Y, 1CFF, 1MUX, 1AHR, 1DEG, and 1NWD, proteins, which form individual

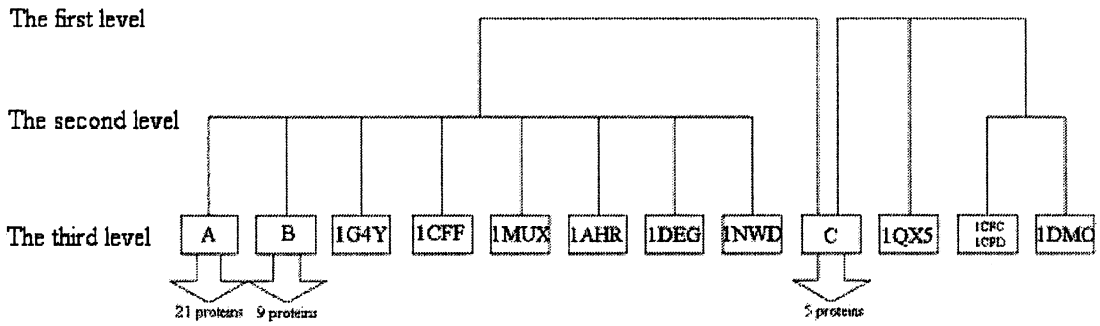


Fig. 9. RMSD distance tree for the 45 known CaM structures; leaf node A includes 1A29, 1CDL, 1CDM, 1CKK, 1CM1, 1CM4, 1CTR, 1IQ5, 1IWQ, 1L7Z, 1LIN, 1MXE, 1PRW, 1NIW, 1QIV, 1QIW, 1QS7, 1QTX, 1VRK, 2BBM, and 2BBN proteins; leaf node B includes 1CLL, 1CLM, 1EXR, 1GGZ, 1OOJ, 1OSA, 1RFJ, 3CLN, and 4CLN proteins; leaf node C includes 1K90, 1K93, 1LVC, 1S26, and 1SK6 proteins.

leafs and are grouped together with nodes A and B in the second tree level, have the same structure in both domains as proteins in A and B. The only difference is in the linker structure, see Section 2.4.

The five proteins included in the leaf node C have the same structure in one of the domains when compared with structures clustered in A, B and the above six proteins. The difference in the other domain is due to structure of loops 1 and 3, which in case of these five proteins are of type B, while the six proteins and proteins in A and B have the loops of type A, see Section 2.3.1.

The 1QX5 protein, which is grouped on the first tree level with proteins in node C and 1DMO, 1CFC, and 1CFD, includes two CaM structures, which form a dimer by embedding the C-domains (see Section 2.3.2). Therefore, it has similar structure with some other CaM structures only in one of the domains.

The 1CFC and 1CFD proteins are structurally similar. Both domains of the 1CFC, 1CFD, and 1DMO proteins are similar and thus are clustered on the second tree level, see Sections 2.3 and 2.3.1.

Finally, 1QX5, 1CFC, 1CFD, and 1DMO proteins have similar structure in one of the domains, and are different in both domains than proteins in A, B, C, and the six proteins. This is due to their loop 1, 3, 4 and 6 structures (Section 2.3.1). The former four proteins have structure of loops 4 and 6 of type B, while the latter proteins have type A. The four proteins and proteins in C have type B structure of loops 1 and 3, while the proteins in A, B and the six proteins have type A. This results in differences in both of the CaM domains.

In summary, the tree reveals structural relation between different CaM structures in a convenient and consistent way with respect to the analysis performed in Sections 2.3 and 2.4.

3. RESULTS

3.1. Analysis of Shift Patterns in CaM Structure

The six loops and the central linker, which are part of a CaM molecule, adjust their structure corresponding to various factors including the change of surrounding or replacement of binding segments. Based on our analysis, different degrees of flexibility were assigned to each of the above structures. Our investigation shows that four factors have impact on changes in CaM 3D-structure: binding of Ca^{2+}, binding segment together with its binding type, X-ray surroundings and sequence mutation at the linker. The degree of flexibility and the degree to which each of the factors impacts the structure is estimated and summarized in Table 3. The table uses keywords 'essential', which means that a factor controls the structure, 'no impact' which means that a factor has no impact on the structure, and 'very little', 'little', 'large', and 'very large', which mean that a factor has insignificant, small, large and substantial impact on the structure respectively.

The discovered facts about the shifts in CaM structure are summarized below:

Binding of Ca^{2+} can change and stabilize the 3D-structure of loops innate EF-hands. The structure of Ca^{2+}-conjoint EF-hand, see Fig. 6a, is evidently different from the structure of Ca^{2+}-free EF-hand, see Fig. 6b. At the same time Ca^{2+} binding brings virtually no change to the structures of loops 2, 5 and the linker, see the second row in Table 3.

Table 3. The Degree of Flexibility of the Six Loops and the Linker, and the Degree by Which their Structure is Affected by the Four Factors

Structure Flexibility		Loop 1 Large	Loop 2 Little	Loop 3 Large	Loop 4 Large	Loop 5 Little	Loop 6 Large	Linker Very large
Structure affecting factors	Binding of Ca^{2+}	Essential	No impact	Essential	Essential	No impact	Essential	No impact
	Binding segment and type	Very little	Very little	Very little	Very little	Very little	Very little	Essential
	X-ray crystal surrounding	Little	Very little	Little	Little	Little	Little	Essential
	Mutation at linker	Very little	Very little	Very little	Very little	Very little	Very little	Essential

Binding segments mainly affect structure of the CaM linker, Different binding types, such as different number of binding segments and binding region, result in different structures, see the third row in Table 3.

Crystal surroundings also mainly affect structure of the linker, i.e. the long-helix linker structure is not its intrinsic structure, but it is due to the crystal surrounding, see the fourth row in Table 3. Crystal surrounding can also affect, to some degree, structure of the loops inside EF-hands, see Fig. 6a and d.

Finally, replacement of residues in the CaM domains results in virtually no change in the CaM structure, while missing residues in the linker result in different linker structures. For instance, for 1AHR and 1DEG proteins missing residues in the linker result in change in relative orientation between the two CaM domains, see the fifth row in Table 3.

3.2. Prediction Method

Prediction of the CaM structure in complex with a novel binding-segment(s) requires knowledge of the information related to the four factors, i.e. we need to know if Ca^{2+} are bound, if there are any missing residues in the linker, the binding type with the binding-segment, and more precisely the binding region and the number of the binding segments. Based on this information, the crystal or solution structure of the predicted CaM–segment complex can be found using the following two step procedure:

Determining structure of loops in EF-hands, including loops 1, 3, 4 and 6.

If a given EF-hand contains Ca^{2+}, its structure is of type A, otherwise the structure is of type B. Given the EF-hand structure, the structure of the loop is automatically determined.

Determine the structure of the linker.

Three factors play crucial role to determine the structure. When ordered from the most important to the least important, they include binding segment and other binding information, crystal surrounding or solution, and missing residues at the linker.

The remaining CaM segments, which include loops 2 and 5, and helices are structurally conserved. More specifically, loop 2 is structurally conserved, while loop 5 is also structurally conserved except for 1QX5 protein, see Section 2.3.2. Since this protein is a CaM dimer, for prediction of single CaM structures, the predicted structure for these loops is the same as structure in the existing CaM complexes.

Based on the analysis given in Section 2.4, the possible linker structures include eight cases, see Fig. 10. The upper four cases include complexes where there are no binding segments, while the lower four cases with the binding segments.

Figure 10e, f, g and h show linker structure when a typical long-helix segment binds to CaM. They can be different and should be accordingly replaced when other segments bind to CaM. The crystal surroundings affect the linker structure mainly when no segment binds to CaM, see Fig. 10b and d, and have marginal effect on the linker structure when some segments bind to CaM, see Fig. 10f and h. Missing residues at the linker do not affect its secondary structure, but only elongate or shorten it, see Fig. 10c, d, g and h. Finally, when no segment binds to CaM and it is mensurated in solution than the linker is flexible, see Fig. 10a and c. Based on the information about the above factors, the predicted linker structure is selected based on the eight cases.

3.3. Test of the Developed Prediction Method

CaM complex, which is not among the 45 proteins studied in this paper, was used to verify

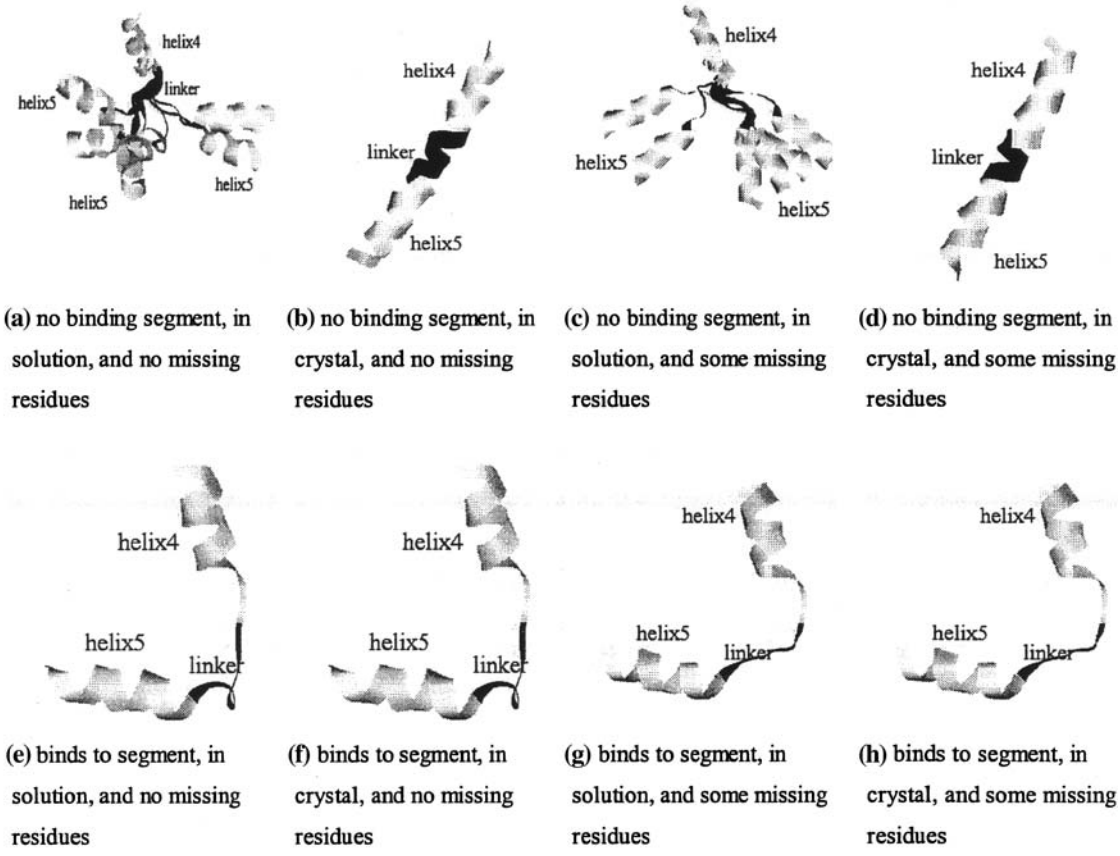


Fig. 10. The eight possible CaM linker structures.

the quality of the developed prediction method by performing structure prediction. The CaM complex is a fragment of the olfactory CNG channel. The complex has been recently mensurated by NMR, and its PDB ID is 1SY9 (Contessa *et al.*, 2005).

First, the required information is gathered:

The four BF-hands of this complex contain Ca^{2+}.

The binding segment is a monomer, and the sequence is "QQRRGGFRRIARLVGVL REWAYRNFR". Its secondary structure is a long helix. Similar to type A structure, the helix binds to both of the CaM domains.

It is a solution structure.

No residues are missing at the linker.

Next, structure of flexible segments is predicted. Based on point 1 the four 4 EF-hands adopt structure of type A. The liker structure is assumed to be as in Fig. 10e based on points 2, 3, and 4. Therefore, the predicted structure of the 1SY9 protein is given in Fig. 11a. The predicted structure is very similar to the experimentally mensurated structure, i.e. the RMSD between the predicted and the mensurated structures is 2.3 A.

The predicted structure is different in two aspects when compared with the mensurated structure. First, region from 1 (Ala) to 5 (Thr) are flexible and has no assigned secondary structure. Thus, due to intrinsic flexibility it is difficult to precisely estimate the structure of this region. Second, the predicted linker structure is slightly different since the 1S9Y binds to a novel segment, which is different from all binding segments in the known 45 CaM structures. At the same time, visually, the difference is very small and the result is still acceptable. Finally, based on the analyzing the four factors, the 1SY9 structure was predicted to be of type A, which fully agrees with the mensurated result. This shows high quality and practical usefulness of the performed analysis and the proposed prediction method.

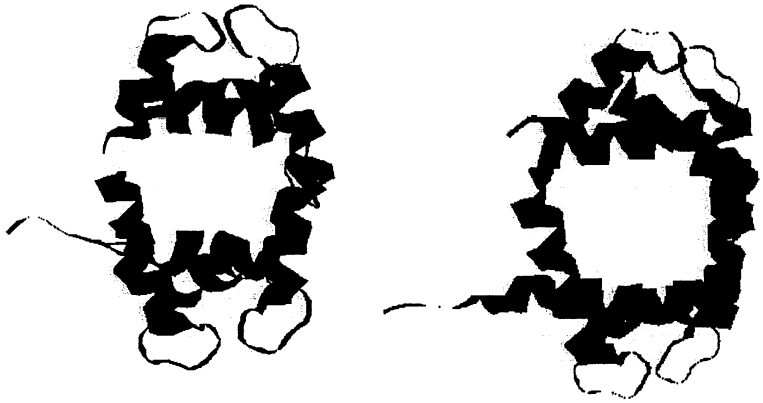


Fig. 11. Comparison between the predicted and mensurated structures of the 1SY9 CaM complex.

4. SUMMARY AND CONCLUSIONS

This paper is the first to perform comprehensive study of the calmodulin structures and to propose prediction method for the future CaM structures in complex. The 45 already mensurated CaM structures is used to perform detailed analysis of their structure. To study the structure, the CaM sequence is divided into structure-conserved regions, which include helices, and structure-flexible regions, which include loops and the linker. For each of the flexible regions, four factors that can affect their structure are determined and their impact on the structure shifts is described. Based on this information, a novel method for prediction of the CaM structure in complex with novel binding segment is proposed. The proposed method is evaluated using a CaM complex, which was the most recently mensurated and was not considered during the design of the prediction method. Given information about the four factors, the method was used to perform structure prediction, and the result was carefully verified to be virtually the same as the mensurated structure.

The high quality of the predicted structure validates results of our study, including discovered relationships between the four factors and loop and linker structures, and the developed prediction method. The presented information is of high practical usefulness not only due to the high quality, but also most importantly due to increasing interest in CaM complexes, which are analyzed and mensurated in increasing numbers over the last couple of years.

ACKNOWLEDGEMENTS

Dr. Ruan and Mr. Chen would like to thank the MITACS Network of Centers of Excellence, the Natural Science Council of China (grant no. 10271061) and the Tianjin–Nankai Universities Joint Program (grant no. 90208022). Dr. Kurgan's work was partially founded by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

REFERENCES

Aoyagi, M., Arvai, A. S., Tainer, J. A., and Getzoff, E. D. (2003). *EMBO J.* **22**(4)**:** 766–775.

Babu, Y. S., Bugg, C. E., and Cook, W. J. (1988). *J. Mol. Biol.* **204**(1)**:** 191–204.

Ban, C. (1994). *Acta Crystallogr. D Biol. Crystallogr.* **50**(1)**:** 50–63.

Barbato, G., Ikura, M., Kay, L. E., Pastor, R. W., and Bax, A. (1992). *Biochemistry* **31:** 5269–5278.

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:** 235–242.

Chattopadhyaya, R., Meador, W. E., Means, A. R., and Quiocho, F. A. (1992). *J. Mol. Biol.* **228**(4)**:** 1177–1192.

Clapperton, J. A., Martin, S. R., Smerdon, S. J., Gamblin, S. J., and Bayley, P. M. (2002). *Biochemistry* **41**(50)**:** 14669–14679.

Contessa, G. M., Orsale, M., Melino, S., Torre, V., Paci, M., Desideri, A., and Cicero, D. O. (2005). *J. Biomol NMR* **31**(3)**:** 185–199.

Cook, W. J., Walter, L. J., and Walter, M. R. (1994). *Biochemistry* **33**(51)**:** 15259–15265.

Drum, C. L., Shen, Y., Rice, P. A., Bohm, A., and Tang, W. J. (2001). *Acta Crystallogr. D* **57:** 1881–1884.

Drum, C. L., Yan, S.-Z., Bard, J., Slien, Y.-Q., Lu, D., Soelaiman, S., Grabarek, Z., Bohm, A., and Tang, W.-J. (2002). *Nature* **415:** 396–402.

Elshorst, B., Hennig, M., Forsterling, H., Diener, A., Maurer, M., Schulte, P., Schwalbe, H., Griesinger, C., Krebs, J., Schmid, H., Vorherr, T., and Carafoli, E. (1999). *Biochemistry* **38**(38)**:** 12320–12332.

Guo, Q., Shen, Y., Zhukovskaya, N. L., Florian, J., and Tang, W. J. (2004). *J. Biol. Chem.* **279**(28)**:** 29427–29435.
Han, B. G., Han, M., Sui, H., Yaswen, P., Walian, P. J., and Jap, B. K. (2002). *FEBS Lett.* **521**(1–3)**:** 24–30.
Harmat, V., Bocskei, Z., Naray-Szabo, G., Bata, I., Csutor, A. S., Hermecz, I., Aranyi, P., Szabo, B., Liliom, K., Vertessy, B. G., and Ovadi, J. (2000). *J. Mol. Biol.* **297**(3)**:** 747–755.
Holm, L., and Sander, C. (1993). *J. Mol. Biol.* **233**(1)**:** 123–138.
Ikura, M., Clore, G. M., Gronenborn, A. M., Zhu, G., Klee, C. B., and Bax, A. (1992). *Science* **256**(5057)**:** 632–638.
James, P., Vorherr, T., and Carafoli, E. (1995). *Science* **20:** 38–42.
Fallon, J. L., and Quiocho, F. A. (2003). *Structure* **11:** 1303–1307.
Klee, C. B., and Vanaman, T. C. (1982). *Adv. Protein Chem.* **35:** 312–321.
Kuboniwa, H., Tjandra, N., Grzesiek, S., Ren, H., Klee, C. B., and Bax, A. (1995). *Nat. Struct. Biol.* **2**(9)**:** 768–776.
Kurokawa, H., Osawa, M., Kurihara, H., Katayama, N., Tokumitsu, H., Swindells, M. B., Kainosho, M., and Ikura, M. (2001). *J. Mol. Biol.* **312**(1)**:** 59–68.
Lesk, A. M. (1986) *Acta Crystallogr. A* **42:** 110–113.
Matsubara, M., Nakatsu, T., Kato, H., and Taniguchi, H. (2004). *EMBO J.* **23**(4)**:** 712–718.
Meador, W. E., Means, A. R., and Quiocho, F. A. (1992). *Science* **257**(5074)**:** 1251–1255.
Meador, W. E., Means, A. R., and Quiocho, F. A. (1993). *Science* **262**(5140)**:** 1718–1721.
Means, A. R., VanBerkum, M. F. A., Bagchi, I., Lu, K. P., and Rasmussen, C. D. (1991). *Pharmac. Ther.* **50:** 255–270.
Mirzoeva, S., Weigand, S., Lukas, T. J., Shuvalova, L., Anderson, W. F., and Watterson, D. M. (1999). *Biochemistry* **38**(13)**:** 3936–3947.
Osawa, M., Swindells, M. B., Tanikawa, J., Tanaka, T., and Mase, T, et al., (1998). *J. Mol. Biol.* **276**(1)**:** 165–176.
Osawa, M., Tokumitsu, H., Swindells, M. B., Kurihara, H., Orita, M., Shibanuma, T., Furuya, T., and Ikura, M. (1999). *Nat. Struct. Biol.* **6**(9)**:** 819–824.
Rao, S. T., Wu, S., Satyshur, K. A., Ling, K. Y., Kung, C., and Sundaralingam, M. (1993). *Protein Sci.* **2**(3)**:** 436–447.
Schumacher, M. A., Rivard, A. F., Bachinger, H. P., and Adelman, J. P. (2001). *Nature* **410:** 1120–1124.
Schumacher, M. A., Crum, M., and Miller, M. C. (2004). *Structure (Camb)* **12**(5)**:** 849–860.
Shen, Y., Lee, Y. S., Soelaiman, S., Bergson, P., Lu, D., Chen, A., Beckingham, K., Grabarek, Z., Mrksich, M., and Tang, W.-J. (2002a). *EMBO J.* **21**(24)**:** 6721–6732.
Shen, S. Y., Yang, J., Yao, A., and Hwang, P. I. (2002b). *J. Comput. Biol.* **9**(3)**:** 477–486.
Shen, Y., Guo, Q., Zhukovskaya, N. L., Drum, C. L., Bohm, A., and Tang, W. J. (2004). *Biochem. Biophys. Res. Commun.* **317**(2)**:** 309–314.
Symersky, J., Lin, G., Li, S., Qiu, S., Carson, M., Schormann, N., and Luo, M. (2003). *Proteins* **53**(4)**:** 947–949.
Tabernero, L., Taylor, D. A., Chandross, R. J., VanBerkum, M. F., Means, A. R., Quiocho, F. A., and Sack, J. S. (1997). *Structure* **5**(5)**:** 613–622.
Taylor, D. A., Sack, J. S., Maune, J. F., Beckingham, K., and Quiocho, F. A. (1991). *J. Biol. Chem.* **266**(32)**:** 21375–21380.
Vandonselaar, M., Hickie, R. A., Quail, J. W., and Delbaere, L. T. (1994). *Nat. Struct. Biol.* **1**(11)**:** 795–801.
Vogel, H. J. (1994). *Biochem. Cell Biol.* **2:** 357–376.
Wall, M. E., Clarage, J. B., and Phillips, G. N. (1997). *Structure* **5**(12)**:** 1599–1612.
Wilson, M. A., and Brunger, A. T. (2000). *J. Mol. Biol.* **301**(5)**:** 1237–1256.
Yamauchi, E., Nakatsu, T., Matsubara, M., Kato, H., and Taniguchi, H. (2003). *Nat. Struct. Biol.* **10**(3)**:** 226–231.
Yap, K. L., Yuan, T., Mal, T. K., Vogel, H. J., and Ikura, M. (2003). *J. Mol. Biol.* **328**(1)**:** 193–204.
Yun, C. H., Bai, J., Sun, D. Y., Cui, D. F., Chang, W. R., and Liang, D. C. (2004). *Acta Crystallogr. D Biol. Crystallogr.* **60**(7)**:** 1214–1219.
Zhang, M., Tanaka, T., and Ikura, M. (1995). *Nat. Struct. Biol.* **2**(9)**:** 758–767.