J Mol Evol (1990) 30:522–562

Journal of Molecular Evolution

Evolution of EF-Hand Calcium-Modulated Proteins. I. Relationships Based on Amino Acid Sequences

Nancy D. Moncrief,[1,*] Robert H. Kretsinger,[1] and Morris Goodman[2]

[1] Department of Biology, University of Virginia, Charlottesville, Virginia 22901, USA
[2] Department of Anatomy and Cell Biology, Wayne State University School of Medicine, Detroit, Michigan 48201, USA

Summary. The relationships among 153 EF-hand (calcium-modulated) proteins of known amino acid sequence were determined using the method of maximum parsimony. These proteins can be ordered into 12 distinct subfamilies—calmodulin, troponin C, essential light chain of myosin, regulatory light chain, sarcoplasmic calcium binding protein, calpain, aequorin, *Strongylocentrotus purpuratus* ectodermal protein, calbindin 28 kd, parvalbumin, α-actinin, and S100/intestinal calcium-binding protein. Eight individual proteins—calcineurin B from *Bos*, troponin C from *Astacus*, calcium vector protein from *Branchiostoma*, caltractin from *Chlamydomonas*, cell-division-cycle 31 gene product from *Saccharomyces*, 10-kd calcium-binding protein from *Tetrahymena*, LPS1 eight-domain protein from *Lytechinus*, and calcium-binding protein from *Streptomyces*—are tentatively identified as unique; that is, each may be the sole representative of another subfamily. We present dendrograms showing the relationships among the subfamilies and uniques as well as dendrograms showing relationships within each subfamily.

The EF-hand proteins have been characterized from a broad range of organismal sources, and they have an enormous range of function. This is reflected in the complexity of the dendrograms. At this time we urge caution in assigning a simple scheme of gene duplications to account for the evolution of the 600 EF-hand domains of known sequence.

Key words: Calcium-modulated protein — EF-hand — Maximum parsimony — Calmodulin — Troponin C — Light chains of myosin — Parvalbumin — S100 — Calpain — Calbindin

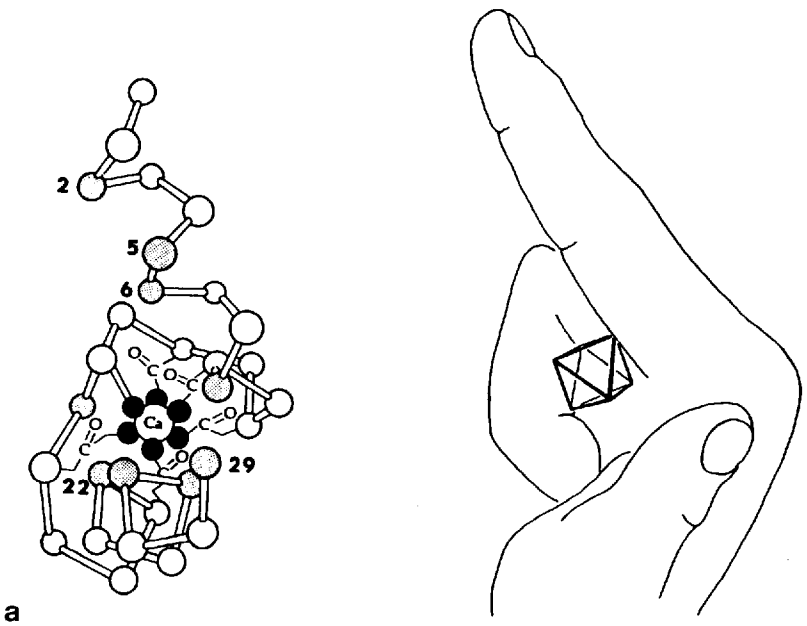
Introduction

Overview and Description of the EF-Hand

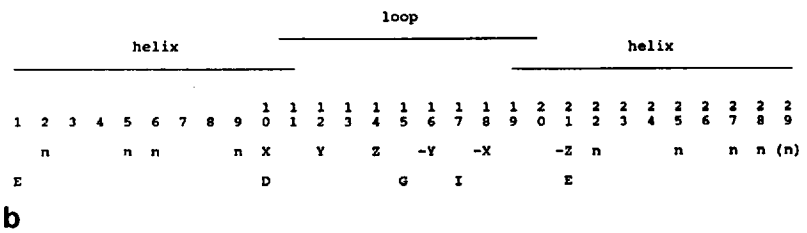
Various extracellular stimuli, including neurotransmitters and hormones, cause calcium to be released from internal stores or permit its entry from outside the cell. In the cytosol calcium functions as a second messenger, coupling the stimulus to cellular responses such as exocytosis, contraction, and enzyme activation. The targets of calcium functioning as a second messenger in the cytosol are calcium-modulated proteins (Kretsinger 1975).

Calcium-modulated proteins in the cytosol bind this messenger calcium, and, in their calci-forms, they are active as enzymes or regulate other enzymes and structural proteins. Most of these calcium-modulated proteins are homologs and contain from two to eight copies of the EF-hand, or calmodulin fold. This basic functional and evolutionarily conserved domain consists of 29 amino acids arranged in a helix, loop, helix conformation, whose functionally important amino acids have been inferred from the crystal structures of parvalbumin (PARV), intestinal calcium-binding protein (ICBP), troponin C (TNC), and calmodulin (CAM). To date, members of this superfamily have been found only within the cytosol or on membranes facing the cytosol; they bind calcium (under cytosolic conditions of ~3 mM free Mg^{2+} ion) with an affinity for calcium [pK_d(Ca)

* *Present address:* Department of Mammalogy, Virginia Museum of Natural History, Martinsville, Virginia 24112, USA
Offprint requests to: R.H. Kretsinger



a



b

Fig 1. **a** The EF-hand, or calmodulin fold, consists of an α-helix (symbolized by the forefinger of a right hand), a loop around the Ca^{2+} ion (represented by the clenched middle finger), and a second α-helix (symbolized by the thumb). Amino acids 1–11 comprise the first α-helix; 19–29 the second. The stipled α-carbons—2, 5, 6, 9, 22, 25, 26, and 29—usually have hydrophobic side chains. They point inward, as does the side chain of residue 17, and interact with the homologous residues of another EF-hand to form a stable hydrophobic core. The calcium ion, when present, is coordinated by an oxygen atom (or by a water molecule bridged to an oxygen atom) of the side chains of residues 10, 12, 14, 18, and 21; the carbonyl oxygen of residue 16 also coordinates calcium. **b** The canonical domain consists of 29 residues in a helix, loop, helix conformation; the Ca^{2+} ion, if bound, is coordinated by six residues, whose positions are approximated by the vertices of an octahedron. Five of these, X, Y, Z, $-$X, and $-$Z, usually have oxygen-containing side chains: Asp (D), Asn (N), Ser (S), Thr (T), Glu (E), or Gln (Q). The oxygen at position 16 ($-$Y) comes from the main chain and can be supplied by any amino acid. As indicated, Asp (D) is usually found at position 10 and Glu (E) is often found at position 21. Gly (G) at position 15 permits a sharp bend ($\Phi = 90°$, $\psi = 0°$) in the calcium-binding loop. Ile (I), Leu (L), or Val (V) at position 17 attaches the loop to the hydrophobic core of the molecule.

~ 6] such that they are calcium free in the unstimulated cell and calcium bound following a pulse of messenger calcium.

The EF-hand was first observed in the crystal structure of PARV (Moews and Kretsinger 1975). PARV contains three calcium-binding domains, designated AB, CD, and EF from N-terminus to C-terminus. The C-terminal, or EF domain, is the domain for which the CAM fold was named, hence the designation EF-hand.

Kretsinger (1987) discussed in detail the characteristics of the EF-hand domain (Fig. 1) and its numerous variations. Some EF-hands do not have the ability to bind calcium; they are easily recognized as homologs, however, because they have most of the charcteristic residues of the EF-hand and because they often occur in tandem with other EF-hands.

The canonical domain consists of 29 residues in an α-helix, calcium-binding loop, α-helix conformation (Fig. 1). The first α-helix frequently begins with Glu at position 1. There is much more variation in sequence and in structure prior to position 1 and less following it; hence its designation as 1. The first α-helix has hydrophobic residues on the side facing the core of the molecule at positions 2, 5, 6, and 9. Similarly, the second α-helix frequently begins with Glu, which also coordinates calcium with the two oxygen atoms of its carboxylate group, at residue 21. Residues 22, 25, and 26 are hydrophobic. Residue 29 at the inside of the end of the second α-helix may or may not be hydrophobic depending on the nearby tertiary structure. Calcium is coordinated by the side chains of five amino acids that can be assigned to the vertices of an octahedron: X, 10; Y, 12; Z, 14; $-$X, 18; and $-$Z, 21. Asp is almost invariant at 10 and Glu at 21, with a broader distribution of Asx, Ser, Thr, and Glx at the other vertices; see reviews by Reid et al. (1981), Kretsinger (1987), and Strynadka and James (1989). Calcium is coordinated by a peptide carbonyl atom at $-$Y 16, and various amino acids are found in this position. Although six amino acids (or in some cases water substituting for a side chain) are involved in calcium coordination, seven oxygen atoms are actually involved, because one carboxylate, usually from Glu at $-$Z, functions as a bidentate ligand. Frequently Gly is found at 15. Ile, Leu, or Val at 17

contribute to the hydrophobic core of the molecule. Most EF-hands occur as members of pairs in which the hydrophobic inside residues form a stable core.

Two variations from the canonical EF-hand are often observed. In the first variation, insertions or deletions occur in domains that are demonstrated or inferred not to bind calcium. The second variation occurs in the first domain of the S100 subfamily; this domain binds calcium with reduced affinity. Only one side chain, that of Glu 21, coordinates calcium. The ligands at vertices X, Y, Z, and −Y are provided by main-chain carbonyl oxygens, and those residues are more variable. In the crystal structure of one member of this subfamily, ICBP (Szebenyi and Moffat 1986), there is a water molecule at the −X vertex. There are two additional amino acids in the first domain of most S100 subfamily proteins; one (12b) is inserted between the X and the Y vertices; a second (16b) occurs between the Z and the −Y vertices.

Historical Perspective

In 1972 Kretsinger proposed that the three domains of PARV resulted from gene triplication and tandem splicing. Weeds and McLachlan (1974) and Collins (1976a,b) subsequently determined the amino acid sequences of myosin light chains and TNC; each of these publications identified four homologous domains. As more proteins were sequenced and their homology recognized, a series of evolutionary studies was performed on members of this protein superfamily.

Prior to the initiation of our study, the most comprehensive analysis that used amino acid sequences to infer evolutionary relationships among these proteins was performed by Baba et al. (1984). They constructed an evolutionary tree relating the 50 calcium-modulated proteins whose amino acid sequences were known at that time. Baba et al. concluded that an interaction between gene duplication and natural selection resulted in the evolution of at least six distinct subfamilies: CAM, TNC, regulatory light chain of myosin (RLC), essential light chain of myosin (ELC), PARV, and the 9-kd ICBP. More recently, Parmentier et al. (1987) analyzed amino acid sequences from six subfamilies, and Perret et al. (1988b) analyzed genomic DNA and amino acid sequences from representatives of eight subfamilies. Each of these studies suggested that a series of tandem gene duplications produced the variation in domain number and protein function that is evident in existing proteins.

Since 1984, when the report of Baba et al. was published, more than a hundred additional amino acid sequences of calcium-modulated proteins have become available. We recently published a preliminary report that included 129 sequences available as of November 1987 (Kretsinger et al. 1988). This study includes 153 amino acid sequences determined by chemical methods or inferred from DNA sequences; genomic DNA (gDNA) sequences currently are available for approximately 25 proteins, and complementary DNA (cDNA) has been sequenced for about 80 of these homologs. Crystal structures have been determined for representatives of four different subfamilies: PARV (Moews and Kretsinger 1975), ICBP (Szebenyi et al. 1981; Szebenyi and Moffat 1986), TNC (Herzberg and James 1985, 1988; Satyshur et al. 1988), and CAM (Babu et al. 1985, 1988; Kretsinger et al. 1986). This is the first in a series of reports that will use amino acid and nucleotide sequences to establish the classification of members of the EF-hand superfamily. From this ordering we hope to gain insights into the structures, functions, and evolution of the EF-hand domain and the proteins that contain it. The purpose of this report is to document our data base of 153 amino acid sequences and references and to present our results of maximum parsimony analyses of the amino acid sequences.

Materials and Methods

Description of Data Base. The first and most fundamental series of tasks completed for this study was to establish a computerized data base of amino acid sequences, to format the sequences for subsequent computations, to proofread the sequences against the primary references, to resolve errors in published sequences, and to compile a comprehensive listing of references. The data base used for this study consists of 153 complete or near complete amino acid sequences available to us either through publication or personal communication as of October 26, 1988 (Appendices I and II). Some of the amino acid sequences were determined directly by chemical means; some were deduced from cDNA or from gDNA sequences. These DNA sequences are being compiled in parallel data bases. Amino acid sequences reported after October 26, 1988, as well as DNA sequences will be included in future analyses and subsequent publications.

We have tried to structure our data bases and this report so that they are of optimal use to our colleagues. We will honor requests (via electronic mail, rhk5i@virginia.edu) for our data base (Appendix I), or requests can be sent to National Biomedical Research Foundation, Georgetown University, 3900 Reservoir Road N.W., Washington, DC 20007, (202) 687–2121, pir-mail@gunbrf.bitnet. This data base comprises 174 amino acid sequences as of December 5, 1989. Every effort was made to assure the accuracy of sequences because this critically affects the analyses. References for all sequences are included in this report. We welcome receipt of new sequences as well as corrections of sequences and references reported here. Several sequences have been corrected by the original investigators, but the corrections have not been published; corrected sequences are indicated by a † in Appendix II.

The following sequences were made available to us, and/or we became aware of them after October 26, 1988: (1) gDNA of PARV from *Homo sapiens* (Berchtold 1988), (2) cDNA of PARV from *Mus musculus* (Zuhlke et al., personal communication), (3)

amino acid sequence of a 15-kd protein from *Hemicentrotus pulcherrimus* (Hosoya et al. 1988), (4) cDNA of calbindin (CLBN) from *M. musculus* (Wood et al. 1988), (5) amino acid sequence of TNC from *Meleagris* sp. (Axelsen, personal communication), (6) amino acid sequence of TNC from *Electrophorus electricus* (Axelsen, personal communication), (7) gDNA of TNC from slow skeletal/cardiac muscle of *Coturnix coturnix* (Maisonpierre and Emerson, personal communication), (8) cDNA of S100 from lung of *Bos taurus* (Glenney et al. 1989), (9) cDNA of p24 thyroid protein from *Canis familiaris* (Lefort et al. 1989), (10) cDNA of α-actinin (ACTN) from skeletal muscle of *Gallus gallus* (Arimura et al. 1988), (11) cDNA of ELC from smooth muscle of *H. sapiens* (Lenz et al. 1989), (12) gDNA of RLC from *Drosophila melanogaster* (Parker et al. 1985), (13) two PARVs from *E. electricus* (Zhu et al. 1985), (14) Spec 2d from *Strongylocentrotus purpuratus* (Hardin and Klein 1987), (15) cardiac TNC from *M. musculus* (Parmacek and Leiden 1989), (16) cardiac TNC from *G. gallus* (Putkey et al. 1987), (17) gDNA, CAM pseudogenes from *Rattus norvegicus* (Nojima 1989), (18) cDNA of CAM genes from *H. sapiens* and *R. norvegicus* (SenGupta et al. 1987).

We have not incorporated these sequences in the calculations reported here because the evaluation of alternate topologies depends on direct comparison of numerical scores for each topology. These scores reflect the number of sequences and provide a quantitative measure of each topology. Incorporation of sequences after we began the analyses would have required recomputation of all topologies examined prior to the addition of those sequences. We will comment on several of these recently available sequences in the local context of subfamilies.

Sequence Nomenclature. For all subfamilies except PARV, domains are numbered sequentially from N-terminus to C-terminus beginning with 1. PARV begins with domain 2, in accordance with the findings of Baba et al. (1984), Epstein et al. (1986), Parmentier et al. (1987), and Perret et al. (1988b). Within the canonical domain, positions are numbered 1–29 sequentially from the N-terminus. Deletions skip the appropriate number(s). Insertions carry the number of the previous canonical residue, followed by b, c, etc. For example, the two inserted residues in the first domain of members of the S100 subfamily are designated 12b and 16b. The canonical residue is itself inferred to be position a; residue 12 is implied to be residue 12a.

The positions between domains and those following the C-terminal domain are numbered sequentially from the N-terminal direction +1, +2, etc. The previous domain may be obvious from context; if not, it can be explicitly stated (e.g., 2+1 and 2+2 refer to the first and second residues, respectively, immediately following domain 2). The positions preceding the N-terminal domains are numbered from the residue immediately preceding position 1 of the domain, starting with −1, then continuing successively toward the N-terminus with −2, −3, etc.

Alignment of Domains. Dendrogram construction is critically dependent on the correct alignment of amino acid and/or nucleic acid sequences. All of the known proteins in the EF-hand superfamily contain two to eight EF-hand domains in a single polypeptide chain. At the outset we assumed, as does any study of this sort, that these EF-hands are homologs. The results of our analyses are consistent with this initial assumption. Kretsinger (1987) described several proteins that contain helix–loop–helix domains that bind calcium and are similar to EF-hands; however, these proteins are not homologs of the EF-hand superfamily and were not included in this study. The domains that bind calcium are relatively easily recognized and aligned using the criteria of Tufty and Kretsinger (1975) and Kretsinger (1987). However, some of the EF-hand proteins have diverged so greatly and harbor so many insertions and deletions that determining alignments was a major undertaking. Several domains in aequorin (AEQ), calpain (CALP), and in the sarcoplasmic calcium-binding proteins (SARCs) possess multiple insertions and/or deletions.

In all cases, except PARV, the domains were aligned with one another by taking the N-terminal domain first and proceeding sequentially toward the C-terminus. In PARV, the domains were treated as 2, 3, and 4. Justification for this ordering was provided by Epstein et al. (1986), who found evidence of the first domain in the 5′-flanking region of cDNA for *Rattus* PARV, as well as analyses by Baba et al. (1984), Parmentier et al. (1987), Perret et al. (1988b), and our own computations (data not shown). Similarly, Desplan et al. (1983a) reported the remnants of a third domain in the 3′ noncoding region of a cDNA for *Rattus* ICBP; however, this interpretation was disputed by Parmentier et al. (1987). Nevertheless, in accordance with findings reported by Baba et al. (1984) and our own analyses (data not shown), we aligned the two domains of these proteins in the S100 subfamily against domains 1 and 2 of the other members of this superfamily.

Alignment of Interdomain Regions. The interdomain regions are readily aligned with homologs from the same subfamily; however, their alignments with interdomain regions from distantly related subfamilies often verge on the arbitrary. As Parmentier et al. (1987) noted,

> Comparisons of sequences that evolved independently for about a billion years is not a trival problem. The linkers joining the calcium-binding domains are very dissimilar in length and sequence, especially between the most distant proteins, so that unambiguous alignment is often impossible to achieve.

Alignments among nodal sequences constructed for the interdomain regions of each subfamily should be more easily accomplished. We plan to use this technique to align interdomains in subsequent analyses; this strategy will allow us to consider computations based on entire sequences in future reports.

Alignments for the domains (Appendix I) are valid within and among all subfamilies. However, alignments for the interdomain regions (not shown) are valid only within subfamilies and have yet to be optimized for comparisons among subfamilies.

Description of Computer Programs. Before describing the construction and interpretation of the dendrograms presented in this report, we will briefly review the general strategy and series of calculations. The suite of programs we used to analyze the calcium-modulated proteins is listed in Fig. 2.

Most of the computations were performed with the programs SWAP1 and SWAP8. These programs use a maximum parsimony algorithm (Moore et al. 1973; Moore 1976; Goodman et al. 1979) that calculates the lowest NR score (nucleotide substitutions that cause amino acid replacements) for each network examined; deletions or insertions are not scored. This algorithm iteratively computes the NR scores for different branching arrangements (topologies); the fundamental difference between the programs SWAP1 and SWAP8 is the way in which successive arrangements are determined for the iterative calculation procedure. SWAP1 examines only nearest-neighbor single-step changes in the network. It calculates the NR score of the three alternative arrangements for each set of four branches that originate from each pair of adjacent internal nodes. For each iteration of SWAP1, during which the three scores for each pair of internal nodes are computed, the arrangement of sequences that results in the lowest NR score is taken as the input network for the next iteration. If a lower-score network is computed, the new arrangement is saved in memory, and the process repeats itself. This branch-swapping procedure continues, rearranging one branch per iteration, until SWAP1 either computes two consecutive networks of equal NR score or a higher score is computed for the next network to be considered. At this point in the analysis more

Name of Program	Input	Form of Output	Next Use of Output
SWAP1	dendrogram, sequences	dendrogram with score	low score dendrogram, input for SWAP1, SWAP8, or LAD
SWAP8	dendrogram, sequences	all dendrograms for 8 external nodes	low score dendrogram(s), input for SWAP1, SWAP8, or LAD
MMD	sequences	distance matrix	input for FTE or UPGMA
UPGMA	distance matrix	dendrogram	input for SWAP1 or SWAP8
FTE	distance matrix	dendrogram	input for SWAP1 or SWAP8
LAD	dendrogram, sequences	branch lengths, ancestral sequences at internal nodes	

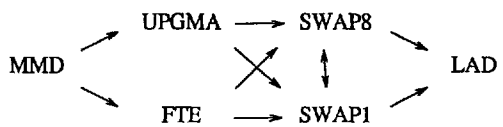


Fig. 2. Suite of programs used for this study. SWAP1, maximum parsimony program that swaps one branch per iteration; SWAP8, maximum parsimony program that computes scores for all trees with eight exterior nodes; MMD, minimum mutation distance; UPGMA, unweighted pair-group method with arithmetic averaging; FTE, Farris tree; LAD, branch lengths and ancestral sequences for a given dendrogram. The diagram at the bottom of the figure illustrates possible routes for input and output.

extensive rearrangements can be explored by rewriting the resultant output dendrogram and submitting the newly altered dendrogram as input to SWAP1. This allows computation of scores for networks that cannot be derived solely from a series of nearest-neighbor single-step changes. SWAP1 is so named because it swaps one branch per iteration. Following this nomenclature, SWAP8 iteratively computes all NR scores for a network that consists of no more than eight external nodes; there are 10,395 possible unrooted arrangements for eight sequences (Moore 1976). The number of possible arrangements for as few as 20 sequences is greater than 10^{20} (Moore 1976), making impractical the calculation of scores for all arrangements of more than eight external nodes. Data sets with more than eight sequences can be accommodated by designating eight subsets of the sequences in the data set; under these conditions, the ancestral node of each subset is treated as a single external node. This sort of brute force calculation will yield the most parsimonious network for a given set of eight or fewer external nodes, although it does not necessarily produce the global minimum network for the entire set of sequences. We alternately used SWAP1 and SWAP8 to examine several hundred thousand networks for the 153 sequences included in this study.

The programs SWAP1 and SWAP8 require a starting tree as input. Usually, SWAP1 performs relatively few branch swaps (less than 30) before two networks of equal or higher score are computed in consecutive iterations and SWAP1 terminates. Therefore, SWAP1 usually does not produce a network that differs greatly from the input dendrogram in topology. This situation might allow the outcome of a set of SWAP1 analyses to be biased by submitting only input networks that represent preconceived notions of the optimum topology. In order to avoid this possible source of bias in our analyses, we employed two different clustering algorithms that construct starting trees using a distance matrix as input. To compute this matrix, we used the program MMD (minimum mutation distance), which calculates distances among all pairs of aligned sequences using the method described by Jukes (1963) and Fitch and Margoliash (1967). The distance matrix was used as input for programs UPGMA and FTE. UPGMA executes the clustering procedure of Sokal and Michener (1958), which is called the unweighted pairgroup method with arithmetic averaging, and constructs a dendrogram. FTE (Farris tree) computes a dendrogram using the distance algorithm described by Farris (1972). These two dendrograms, one constructed by UPGMA, the other by FTE, and the amino acid sequences are then used as input for SWAP1 and/or SWAP8 to construct the dendrogram that requires the fewest changes (lowest score) among all external nodes (extant sequences) and internal nodes (inferred precursors), while accounting for all of the data. Finally, the program LAD (branch lengths and ancestral sequences for a given dendrogram) is used to determine the most probable ancestral sequences at each of the internal nodes as well as to compute the individual branch lengths for a given dendrogram and set of sequences. Figure 2 details the input and output for each of these programs and provides a brief schematic of the order in which they are used. Copies of these programs are available from John Czelusniak, Department of Anatomy and Cell Biology, Wayne State University School of Medicine, 540 E. Canfield Ave., Detroit, Michigan 48201.

Comments on Computational Procedures and Results. There are several major categories of methods for inferring evolutionary relationships among molecular sequences. Felsenstein (1988) presented a lucid description of these methods and reviewed their statistical properties. One way to construct evolutionary trees "... is to count the minimum number of base substitutions that are required.... That tree requiring the fewest changes is preferred" (Felsenstein 1988). This is the method of maximum parsimony (Fitch 1971; Hartigan 1973; Moore et al. 1973; Moore 1976) and was the primary method of analysis used herein.

Distance methods, the second major category, fit a tree to a matrix of pairwise distances between species.... The phy-

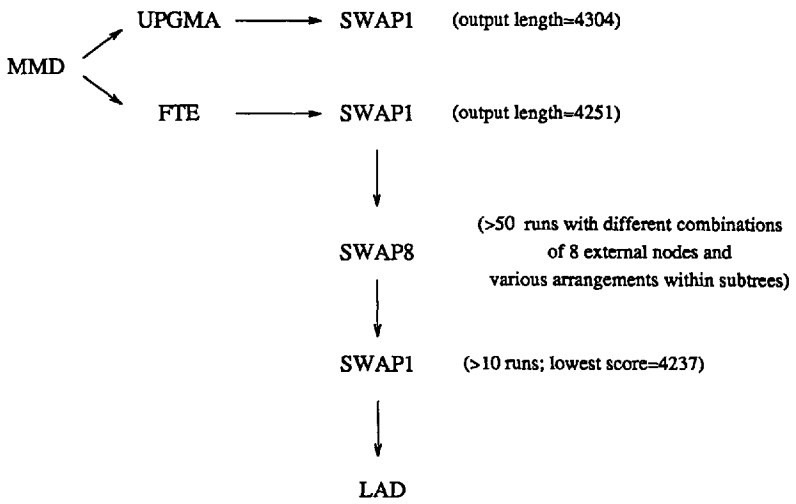


Fig. 3. Order in which programs were used for this study. Lengths are given as NR scores, which represent the minimum number of nucleotide substitutions that could account for amino acid replacement in the bounding nodes for a given connecting branch. Acronyms are as listed for Fig. 2.

> logeny makes a prediction of the distance for each pair as the sum of branch lengths in the path from one [external node] to another through the tree. A measure of goodness of fit of the observed distances to the expected distances is used, and that phylogeny is preferred which minimizes the discrepancy between them as evaluated by this measure. (Felsenstein 1988)

We used distance algorithms in the programs UPGMA and FTE to construct dendrograms, which were then used as input for the maximum parsimony algorithms.

The methods for constructing dendrograms are fairly straightforward. However, Felsenstein (1988) pointed out that "The question of how to obtain confidence intervals and carry out statistical tests [on estimates of phylogeny] is in a relatively primitive state." Felsenstein (1988) concluded that his

> . . . survey of methods for inferring phylogenies and assessing their reliabilities shows that the field is in an incomplete but interesting state. We have a number of different approaches: parsimony, distance matrix methods, and likelihood methods. The assumptions in these methods are only sketchily known—we have hints but little in the way of comprehensive proofs that particular assumptions are required. It is clear from the failings of different methods in particular cases that they all have assumptions; no method allows one to make inferences about evolutionary patterns in a well-justified way without making any assumptions about evolutionary processes.

We chose maximum parsimony algorithms for this report because of a historical precedence for using parsimony methods to infer phylogenies of calcium-modulated proteins, as well as the availability of suitable computer programs.

Documentation of Procedures Used for These Analyses. For this study we first used MMD to construct a distance matrix for all pairwise comparisons among the 153 amino acid sequences in our data base (Fig. 3). Next, we used the programs UPGMA and FTE to construct two different starting trees, which were then used as input with the amino acid sequences for SWAP1. Thus, SWAP1 was used to produce two different output dendrograms. Of these, the lower NR score (4251) was associated with the dendrogram that used the FTE-generated dendrogram as input. Using the two dendrograms that were output by SWAP1, we determined the compositions of the major subfamilies and a first approximation of their relationships to one another. All sequences within an entity that we have designated as a subfamily were consistently placed with other members of that subfamily. Eight sequences are considered not to be members of the subfamilies; we tentatively refer to them as uniques.

Next, SWAP8 was used extensively to explore optimal arrangements (i.e., to find the lowest score networks) within and among the unique sequences and the 10 subfamilies recognized at that point in our analyses. All 153 sequences were grouped into eight subtrees in various combinations; more than 50 different input arrangements were submitted to SWAP8. Optimal arrangements within each subfamily were tested with at least one run of SWAP8, and as many as seven different input arrangements were submitted to SWAP8 for subfamilies represented by more than 20 individual sequences and to determine optimal positions for each of the uniques. Arrangements that lowered the NR length within each subfamily were then incorporated into a final input tree, which was submitted to SWAP1. The NR score of the lowest length arrangement of sequences determined by this series of analyses was 4237. We will present and interpret this arrangement in subsequent sections.

Identification of Subfamilies. As we discussed, these sequences can be unambiguously aligned (with a few exceptions) because of the distinctive features of the tertiary structure of the EF-hands. Hence, we feel that the domain-only alignments and comparisons are the most valid for defining and comparing subfamilies. The domains seem to be subject to much stricter evolutionary constraints than are the interdomains. Conversely, because some domains (e.g., those of the animal calmodulins) evolve very slowly, the interdomains have accumulated more differences and provide more information for determining relationships among sequences within a subfamily.

For this report we have been more concerned with the general identification and classification of subfamilies. A subfamily consists of all external nodes (extant sequences) distal to a designated internal node of the dendrogram. The choice of the designated internal node, or root, may reflect biochemical, functional, and/or evolutionary relationships; however, given the internal node, the designation of a subfamily is unambiguous. Even so, the definition of subfamily is somewhat arbitrary because some subfamilies may be further subdivided while still satisfying the preceding definition. Therefore, within subfamilies we refer to groups; analogously, all external nodes distal to an internal node designated as a group node are members of that group. This problem of assignment of rank is encountered in all classification schemes. It is important to realize that the assignment of classificatory rank does not alter the structure of the dendrograms we present, only their interpretation.

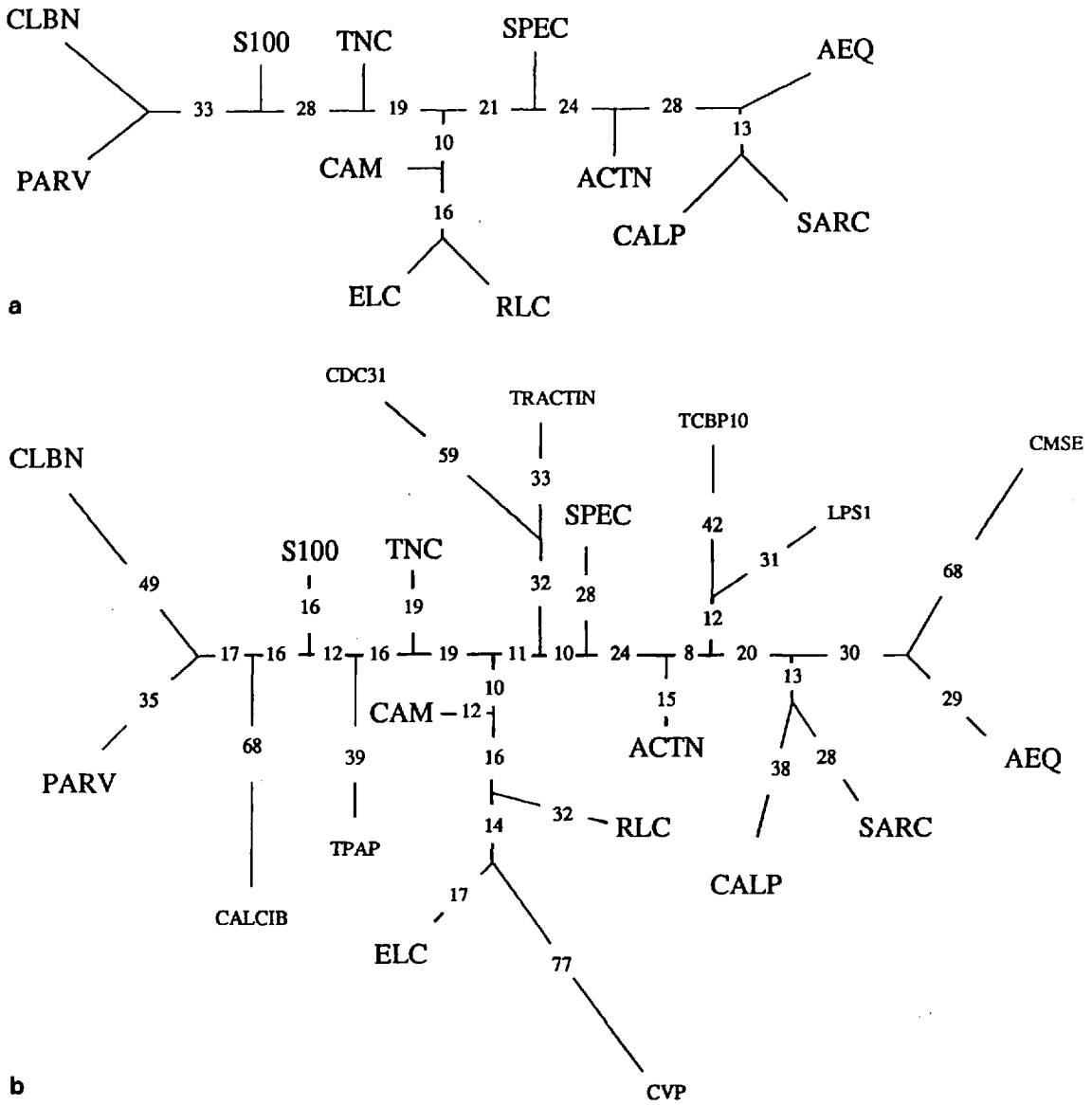


Fig. 4. Relationships among EF-hand homologs; numbers indicate branch lengths (see legend of Fig. 3) and represent relative amounts of divergence among sequences. **a** Relationships among 12 subfamilies: PARV, parvalbumin; CLBN, calbindin; S100, S100 and several other groups of two-domain proteins; TNC, troponin C; CAM, calmodulin; ELC, essential light chain of myosin; RLC, regulatory light chain of myosin; SPEC, *Strongylocentrotus purpuratus* ectodermal protein; ACTN, α-actinin; AEQ, aequorin- and luciferin-binding protein; CALP, calpain; and SARC, sarcoplasmic calcium-binding protein. **b** Relationships among these 12 subfamilies and eight unique proteins; CALCIB, calcineurin B from *Bos*; TPAP, troponin C from *Astacus*; CVP, calcium vector protein from *Branchiostoma*; TRACTIN, caltractin from *Chlamydomonas*; CDC31, cdc31 gene product from *Saccharomyces*; TCBP10, 10-kd calcium-binding protein from *Tetrahymena*; LPS1, eight-domain protein from *Lytechinus*; and CMSE, calcium-binding protein from *Streptomyces*.

We were concerned whether or not relationships within subfamilies differed when only the EF-hand domains were used to construct dendrograms as opposed to when the entire sequences, including the interdomain regions, were used. To address this concern, we constructed a dendrogram for the complete sequences of all proteins classified as part of the CAM subfamily, using complete sequences of representatives from TNCs and both RLCs and ELCs as outgroups. The dendrogram for CAM based on the entire sequence (not shown) is very similar to that based solely on the domains, and relative branch lengths are also similar. Relationships within and among all other subfamilies were determined using only the EF-hand domains. By extrapolation from the results we obtained for CAM, we anticipate few major changes in relationships within subfamilies when complete sequences of all the calcium-modulated proteins are used, a topic of the next paper in this series.

Results

Overview of Relationships among EF-Hand Homologs

Figure 4a shows relationships among the 12 subfamilies of calcium-modulated proteins, and Fig. 4b de-

Table 1. Calcium-binding in subfamilies and unique EF-hand homologs

		Ca^{2+} binding by EF-hand domains							
	Acronym	1	2	3	4	5	6	7	8
Calmodulin	CAM[a,b,c]	+	+	+/?	+/?				
Troponin C	TNC[a,d,e]	+/−	+	+/−	+				
Essential light chain of myosin	ELC[f]	+/−	−	+/−	+/−				
Regulatory light chain of myosin	RLC	+	−	−	−				
Sarcoplasmic calcium-binding protein	SARC	+	+/−	+/−	+/−				
Calpain	CALP*	+	+	−	−				
Aequorin	AEQ	+	−	+	+				
Strongylocentrotus purpuratus ectodermal protein	SPEC*	+	+	+	+/−				
Calbindin	CLBN	+	−	+	+	+	−		
Parvalbumin	PARV[a]		−	+	+				
α-actinin	ACTN	+/−	+/−						
S100	S100[a,g]	+/−	+/−						
Calcineurin B (*Bos*)	CALCIB	+	+	+	+				
Troponin C (*Astacus*)	TPAP	−	+	−	+				
Calcium vector protein (*Branchiostoma*)	CVP	−	−	+	+				
Caltractin (*Chlamydomonas*)	TRACTIN*	+	+	+	+				
CDC31 (*Saccharomyces*)	CDC31*	+	−	−	+				
10-kd protein (*Tetrahymena*)	TCBP10*	+	+						
Eight-domain protein (*Lytechinus*)	LPS1*	+	+	+	+	+	+	+	−
Calcium-binding protein (*Streptomyces*)	CMSE*	+	+	+	+				

Domains that are demonstrated or strongly inferred to bind calcium are indicated by +; those that are demonstrated or strongly inferred not to bind calcium are indicated by −. Homologs for which calcium-binding ability is inferred solely from primary sequence are indicated by *

[a] The four subfamilies for which crystal structures have been solved

[b] Domain 4 of CAM from *Saccharomyces* apparently does not bind calcium

[c] The CAM pseudogenes from rat may not bind calcium in domains 3 and 4

[d] The first domain of cardiac TNC does not bind calcium

[e] Domains 1 and 3 from *Halocynthia* apparently do not bind calcium

[f] Domain 1 as well as domain(s) 3 and/or 4 from *Physarum* may bind calcium

[g] The first domain of S100 has two inserted amino acids

picts relationships among the 12 subfamilies and eight unique EF-hand proteins. These figures are presented as networks rather than dendrograms, because we have not yet determined the position of the root or origin for the superfamily of EF-hand homologs. In contrast, dendrograms for individual subfamilies (Figs. 5–15) are presented as rooted trees, because relationships within each subfamily have been determined relative to all sequences not in that subfamily.

Table 1 details the number of domains in each subfamily and unique proteins and indicates those EF-hands that are demonstrated or inferred to bind calcium. Note that relationships among subfamilies (Fig. 4a) are not determined by the number of domains that characterize each subfamily. PARV, which has three domains, and CLBN, with six domains, are closely related to each other. Next is S100, which has only two domains, then TNC with four domains, then a grouping of four-domain proteins consisting of CAM, ELC, and RLC. These are adjacent to *S. purpuratus* ectodermal protein (SPEC), which has four domains. Next is ACTN, which has only two domains; it is adjacent to another subfamily of four-domain proteins, AEQ. The four-domain subfamilies CALP and SARC are closely related to each other and to AEQ.

Similarly, the unique proteins, which vary in domain number, are distributed throughout the dendrogram shown in Fig. 4b. Calcineurin B (CALCIB), a four-domain homolog, was placed along the branch between S100 (two domains) and the node that joins CLBN (six domains) and PARV (three domains). *Astacus leptodactylus* troponin C (TPAP) (four domains) was placed on the branch between S100 (two domains) and TNC (four domains). Calcium vector protein (CVP) (four domains) clusters with ELC (four domains). Caltractin (TRACTIN) and cell division-cycle 31 gene product (CDC31), both of which have four domains, are closely related to each other; they are placed on the branch between CAM–RLC–ELC–CVP (four domains) and SPEC (four domains). The calcium-modulated protein from *Streptomyces erythraeus* (CMSE) (four domains) is closely related to AEQ (four domains). The *Tetrahymena thermophila* 10-kd calcium-binding protein (TCBP10) (two domains) and the protein from *Lytechinus pictus* (LPS1) (eight domains) are each placed on the branch between ACTN (two domains), and the AEQ–CMSE pairing.

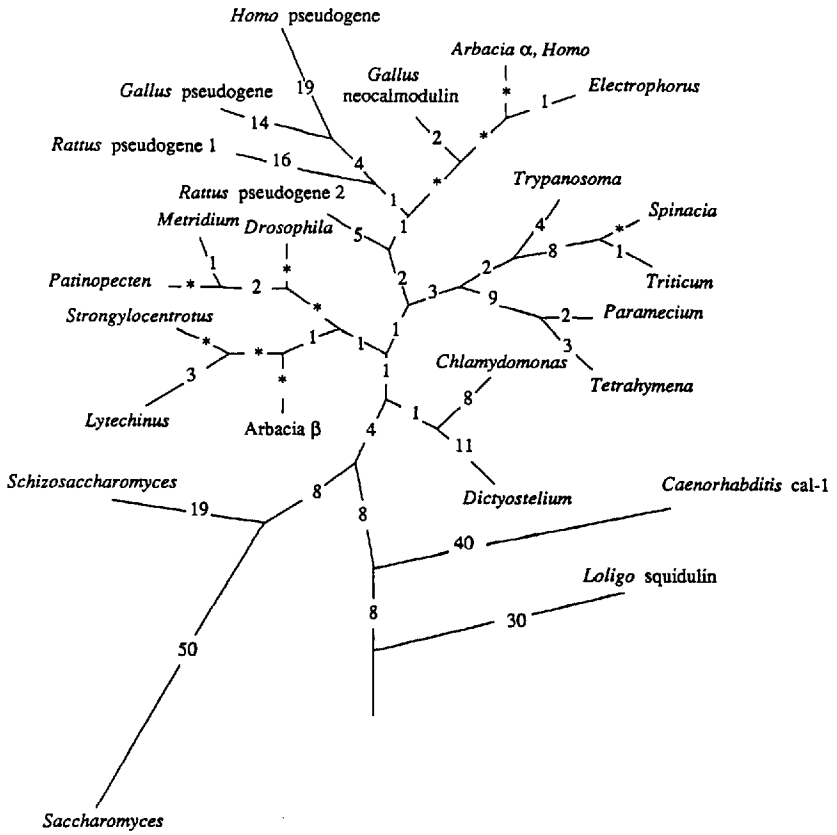


Fig. 5. Relationships among members of the CAM subfamily, which are indicated by the genus names of their organismal sources and other identifiers when necessary. Numbers indicate branch lengths (see legend of Fig. 3) and represent relative amounts of divergence among sequences; * indicates a branch length of zero. CAM from *Bos, Oryctolagus, Rattus, Mus, Gallus,* and *Xenopus* (not shown) is identical to that from *Homo,* as is the α form from *Arbacia.* CAM from *Renilla* (not shown) is identical to *Metridium* CAM. Cal-1 and squidulin do not function as CAM, but are most closely related to CAM.

These relationships and relationships within subfamilies will be discussed in detail in the following sections. Evolutionary, structural, and functional aspects of each subfamily and the unique homologs will also be considered.

CAM

CAM (Fig. 5) is inferred, by extrapolation from numerous studies of animals and several studies of plants, protocists, and fungi, to be found in all cells of all eukaryotes. There is convincing evidence that mammalian CAM activates, in situ, at least 12 and as many as 20 target enzymes or structural proteins. Except for plant NAD kinase, all of these target proteins are from mammals. These targets, so far as sequence information is available, do not appear to be homologous to one another; however, several do contain cationic, amphipathic α-helices that have been demonstrated or inferred to be the site of calmodulin binding. Persechini and Kretsinger (1988) have presented evidence supporting their postulate ". . . that the linker region of the central helix functions as a flexible tether to permit the two lobes of calmodulin to enfold the target helix." This linker region can undergo various deletions, insertions, and substitutions yet still retain its ability to activate several of these targets. The seeming paradox is then presented: How can CAM retain activity with so many changes, yet at the same time be so highly conserved evolutionarily? Two points should be considered. First, natural selection undoubtedly operates on characteristics (e.g., longevity within the cell) in addition to in vitro activation of several enzymes.

The second point is obvious from an examination of the amino sequences of CAM and is especially well illustrated in the dendrogram (Fig. 5): CAM is much less highly conserved outside of the animals. Excluding the sequences of cal-1 and squidulin, which will be discussed below, there are only 91 invariant positions in a total of 148 residues. This number is reduced to 69 invariant positions of one also considers the four pseudogenes from *Homo, Gallus,* and *Rattus.* It is certainly possible that the relative conservation among the animal CAMs (and the lower rate of evolution that is inferred from this conservation) reflects CAM's having acquired more targets in the animals than in the other kingdoms.

The interdomain linker region of the central helix of CAM (and probably of TNC as well) seems to be delicately balanced between two forms: extended α-helix and a currently ill-defined bent form. The inferred nodal sequence of the CAM 2,3 interdomain linker is Met*Lys*Asp*Thr*Asp*Ser*Glu*Glu, the same as observed in the animals. However, the linker of cal-1 from *Caenorhabditis* (Met*Lys*Glu*Thr*Asp*Ser*Glu) is one residue shorter. The dele-

tion mutants of vertebrate CAM engineered by Persechini et al. (1989) still activate several target enzymes; hence, cal-1 may assume some functions resembling those of CAM. In contrast, the next nearest protein is squidulin from *Loligo,* whose linker contains two Pros (Met*Gly*Pro*Thr*Asp*Pro *Glu*Lys). This linker surely is not helical, and the gene duplication inferred at this node generated molecules with quite different structures. The organisms in which cal-1 and squidulin are found, *Caenorhabditis* and *Loligo,* respectively, also have genes that encode CAM (Salvato et al. 1986; Head 1989), indicating that these proteins are indeed closely related to, but distinct from, proteins that are identified as CAM.

TNC

The TNCs from chordates cluster together in one subfamily (Fig. 6). The proteins Wnuk et al. (1986) and Wnuk (1988) identified as TNC from the arthropod *A. leptodactylus* (TPAP) are located between the TNC subfamily and the S100 subfamily. We have classified the TPAPs as uniques (Fig. 4b), although by biochemical criteria they function as TNC. This situation contrasts with the CAM subfamily, in which cal-1 and squidulin are at the base of the CAM dendrogram (Fig. 5), yet they appear to have other functions. Within birds and mammals (and probably all vertebrates), there are two genes encoding TNC, fast skeletal and slow skeletal/cardiac. Gahlmann et al. (1988) reported that

> . . . the fast skeletal muscle TNC gene appears to be expressed exclusively in skeletal muscle. Only the slow TNC gene is expressed in human cardiac muscle. The slow skeletal TNC gene is also expressed in skeletal muscle, and surprisingly, in several human fibroblast cell lines.

The human slow skeletal/cardiac TNC sequence reported by Gahlmann et al. (1988) differs from that reported by Roher et al. (1986) at position 115. Gahlmann et al. reported Asp at this position; Roher et al. reported Glu. Gahlmann et al. suggested that this discrepancy is attributable to a genomic polymorphism, because the differing amino acids (Glu and Asp) are related by a single third base codon transversion. Thus, in contrast to the report of Roher et al., the slow skeletal/cardiac sequences of *Oryctolagus* (which has Asp at position 115; Wilkinson 1980) and *Homo* are identical, and the slow skeletal/cardiac sequence of *Bos,* which has Glu at position 115 (van Eerd and Takahashi 1976), differs from that of *Homo* and *Oryctolagus* at this one position.

The TNC from the body wall muscle of *Halocynthia roretzi,* a protochordate (Takagi and Konishi 1983), was placed at the root of the TNC tree, before the divergence of vertebrate slow skeletal/

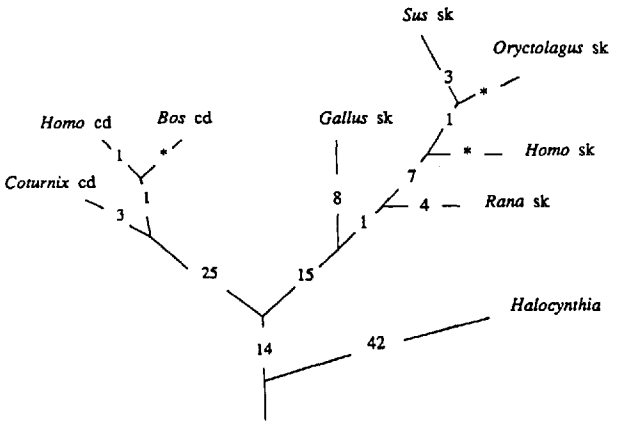


Fig. 6. Relationships among members of the TNC subfamily; sk represents skeletal and cd indicates cardiac isoforms. Numbers indicate branch lengths (see legend of Fig. 3) and represent relative amounts of divergence among sequences; * indicates a branch length of zero. The sequence fragment of skeletal TNC from *Rattus* (not shown) is identical to the corresponding portion from *Oryctolagus,* and the cardiac isoform from *Oryctolagus* (not shown) is identical to that from *Homo.*

cardiac and fast skeletal genes. The body wall muscle of *Halocynthia* is smooth muscle; however, it differs from other smooth muscles in that it is multinucleated and regulated by the troponin–tropomyosin system. The first and third calcium-binding domains of the TNC from *Halocynthia* appear to have lost the ability to bind calcium (Takagi and Konishi 1983).

Myosin Light Chains

Myosin is found in (nearly) all eukaryotic cells in highly regular polymers, as seen in skeletal muscle, or as individual molecules, as in seen in *Acanthamoeba.* It is a heterohexamer consisting of four light chains and two (near) identical heavy chains that have ATPase activity. There are two types of light chains—ELC and RLC; two (nearly) identical ELCs and two RLCs occur per hexamer.

We have followed the nomenclature of the primary references in identifying myosin light chains. Generally, ELCs are designated either 1 and 3 or 1 and 4; RLCs are designated 2. However, the ELCs are sometimes called A1 and A2, and RLCs are sometimes called 2 and 3. These numerals reflect the relative mobilities of the proteins on gel electrophoresis. In our analyses, we have adopted the convention of identifying skeletal ELCs as 1 and 3 or 1 and 4; we denote the skeletal RLCs as 2, never 3.

The ELCs and RLCs are found in muscle and nonmuscle tissues. The amino acid sequences of myosin light chains from brain, cytoplasm, and smooth, atrial, ventricular, and fast skeletal muscle have been shown to differ from each other, and

many are encoded by different genes. However, not all tissues express different light chains, and light chain expression in a particular tissue also depends on the developmental stage of the organism. These attributes make ELCs and RLCs and the genes that encode them excellent model systems for studies of gene expression and developmental regulation; this is reflected by the large number of amino acid, cDNA, and gDNA sequences available for these two subfamilies (Appendix II).

ELC of Myosin

In vertebrates the ELCs (Fig. 7), which are also known as enzymatic or alkali light chains, do not bind calcium, even though they include four recognizable EF-hand domains (Table 1). Invertebrates are characterized by ELCs with four domains, one of which may bind calcium.

Our analyses suggest that a gene duplication event occurred in chordates, producing the two major classes of ELC (fast skeletal and nonmuscle/smooth/cardiac/slow skeletal), after the appearance of vertebrates and before the emergence of birds and mammals. Further duplication events occurred within the nonmuscle/smooth/cardiac/slow skeletal lineage. In mammals, atrial (at) and ventricular (vt) ELCs apparently are encoded by separate genes (represented by *Mus* at and *Homo* vt). In birds, ventricular, atrial, and slow skeletal isoforms (represented by *Gallus* vt) are encoded by the same gene. According to Barton et al. (1988) "The atrial form [of mammals] is also expressed in fetal skeletal and fetal ventricular muscle." In addition Barton et al. suggested that the atrial isoform specific to adults arose during mammalian evolution, because this ELC is not seen in *Gallus* or *Xenopus*. Barton et al. reported that the *Mus* embryonic ELC/adult atrial ELC gene is orthologous to the *Gallus* L23 gene, which is expressed mainly in fetal smooth muscle and persists in the brain from embryonic to adult stages (Kawashima et al. 1987). Our analyses indicate that the proteins encoded by these two genes are indeed very similar. Supporting the suggestion of Barton et al. that the atrium-specific form is characteristic only of mammals is the report of Nakamura et al. (1988) on *Gallus* ventricular ELC:

> The gene for the cardiac alkali light chain [of *Gallus*] has proved to be expressed in ventricular muscle and in atrial and latissimus dorsi muscles, the last of these being characteristic of slow skeletal muscles. In these muscles two kinds of mRNA for the cardiac myosin alkali light chain, identical with those in ventricular muscle, were expressed and their relative amount in each tissue was almost the same as in ventricular muscle.

In addition our analyses reflect the similarity between *Gallus* nonmuscle and smooth muscle ELC, which are the products of a single gene (Nabeshima et al. 1987). Lenz et al. (1989) recently reported that the ". . . alkali light chains of human smooth and nonmuscle myosins are [also] encoded by a single gene." These data suggest that the duplication event that produced this gene lineage occurred before the emergence of birds and mammals. Even so, there appear to be differences between birds and mammals in tissue-specific exon usage in this gene (Lenz et al. 1989). The sequences for *Homo* nonmuscle and smooth muscle ELCs will be included in subsequent analyses.

The fast skeletal lineage of mammals and birds is characterized by the production of two polypeptide chains (named either 1 and 3 or 1 and 4), which are encoded by a single gene (Nabeshima et al. 1984, *Gallus*; Periasamy et al. 1984, *Rattus*; Robert et al. 1984, *Mus*; Strehler et al. 1985, *Rattus*; Seidel et al. 1987, *Homo*). This gene probably generates different mRNAs through differential splicing of the same primary transcript via alternative association of exons into the mature RNA (Robert et al. 1984). According to Parker et al. (1985), the polypeptides are identical

> . . . over their carboxyterminal 141 amino acids. As a result of differential promoter utilization and RNA splicing, [ELC 1] has 42 amino-terminal residues not present in [ELC 3 or 4], and [ELC 3 or 4] has 8 amino-terminal residues not present in [ELC 1].

There are some parallels to this mechanism in invertebrates:

> In *Drosophila*, there is a developmental difference in splicing patterns of the [ELC] gene, resulting in at least two slightly different polypeptides encoded by and expressed by a single gene; [one isoform is produced only in adults; one is produced in both larvae and adults]. (Falkenthal et al. 1985)

However, as Falkenthal et al. pointed out, exon usage by *Drosophila* ELCs and vertebrate fast skeletal ELCs are strikingly different in detail:

> The polypeptide sequence differences between chicken LC1 and LC3 occur exclusively at the amino termini and arise through differential use of transcription initiation sites and splice sites in the 5′ region of the gene. However, identical use is made of exons and introns in the 3′ region of the single gene, in marked contrast to that seen for the *Drosophila* [ELC] gene.

It is interesting that Robert et al. (1984) reported a pseudogene for ELCs 1 and 3 in *Mus domesticus*. This pseudogene is not present in a closely related species, *Mus spretus*, suggesting that it is of very recent origin. The discovery of this pseudogene provides further evidence that complex evolutionary processes are occurring in this subfamily of calcium-binding proteins, underscoring the fact that organismal phylogenies based on molecular phylogenies such as this must be inferred with great caution.

Finally, we note that the myosin light chain from

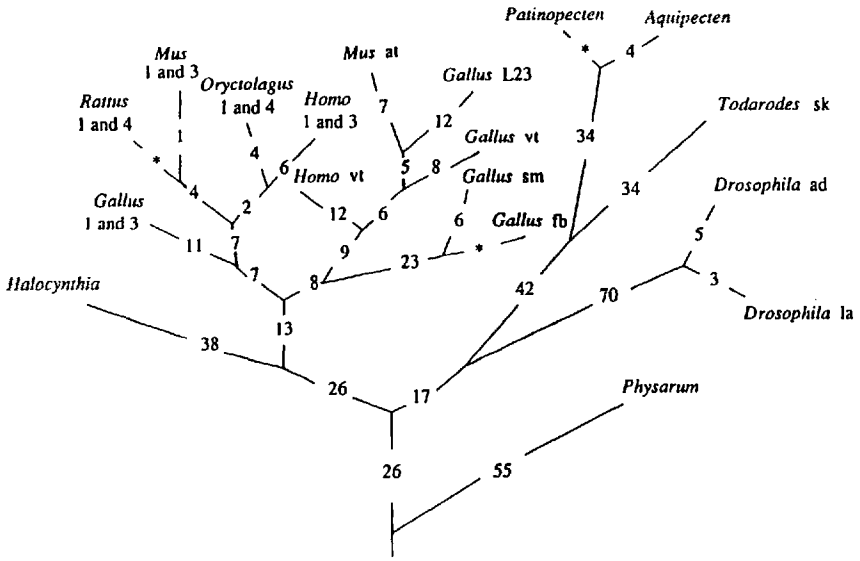


Fig. 7. Relationships among members of the ELC subfamily. Skeletal isoforms of vertebrates are indicated by 1 and 3 and 1 and 4; at indicates atrial; *Homo* vt is ventricular only; *Gallus* vt is atrial, ventricular, and slow skeletal; L23 is expressed embryonically and throughout life; sm indicates smooth muscle; fb indicates fibroblast; sk indicates skeletal; ad indicates adult only; la indicates larval and adult. Numbers indicate branch lengths (see legend of Fig. 3) and represent relative amounts of divergence among sequences; * indicates a branch length of zero.

the slime mold *Physarum polycephalum* was placed with the ELCs in our analyses, confirming the suggestion of Kobayashi et al. (1988a) that it is ". . . akin to alkali light chain." This ELC is encoded by a single gene, and binds at least two equivalents of Ca^{2+} in the absence of magnesium (Kobayashi et al.), probably in domains 3 and/or 4 as well as in domain 1.

RLC of Myosin

The RLCs (Fig. 8) are also known as DTNB (dinitrobenzoic acid-removable) light chains in vertebrates and EDTA (ethylenediaminetetraacetic acid-extractable) light chains in invertebrates. Tanaka et al. (1988) summarized RLC function:

> Contraction of vertebrate striated muscles is regulated by Ca^{2+} through the tropomyosin–troponin system, contraction of molluscan muscles is regulated through the light chains of myosin. The RLC or EDTA-light chain is responsible for specific Ca^{2+}-binding and Ca^{2+}-dependent Mg-ATPase activity of molluscan myosin. Vertebrate muscle myosins also possess light chains analogous to the molluscan RLC, such as gizzard [smooth muscle] 20 kD light chain, skeletal-DTNB light chain, and cardiac L2 light chain. The 20 kD light chain as well as molluscan RLC [can] confer Ca-sensitivity to scallop myosin Mg-ATPase activity, but the DTNB and L2 light chains [do] not. RLC's can be classified [into] three types according to their physiological roles: 1) contraction of molluscan muscle regulated through Ca^{2+} binding, 2) contraction of gizzard [smooth] muscle regulated by phosphorylation of RLC, and 3) contraction of vertebrate striated muscle regulated by troponin–tropomyosin.

Clearly, our dendrogram (Fig. 8) reflects these physiological roles. It is possible that more than one gene encoding RLCs was present in the ancestor of chordates and molluscs. One of these genes may have encoded a protein that regulated muscle contraction in the ancestor of molluscs and chordates, as reflected by the placement of *Gallus* smooth muscle RLC with the molluscan RLCs. The other gene soon coded for a protein that does not regulate muscle contraction, as reflected by the placement of *Halocynthia* RLC with vertebrate cardiac and skeletal RLCs. Coincident with these evolutionary events, a troponin system evolved to regulate muscle contraction in the ancestor to chordates. The comments of Takagi et al. (1986b) are interesting in light of this interpretation: "The body wall muscle of ascidian is morphologically smooth muscle, and the actin–myosin interaction is regulated by troponin. Thus ascidian smooth muscle is the first example of smooth muscle found to be regulated by troponin." Yet another gene duplication event occurred after the protochordate–vertebrate divergence, resulting in skeletal and cardiac isoforms of RLC in vertebrates.

As mentioned earlier, vertebrate smooth muscle RLCs, represented in our dendrogram by a sequence from *Gallus* gizzard, apparently are involved in calcium-linked regulation of muscle contraction, differentiating them from vertebrate skeletal and cardiac muscle. Thus, as Tanaka et al. (1988) reported, ". . . sequence homology between molluscan and gizzard RLCs . . . is higher than that between molluscan and vertebrate striated RLCs." The observations of Tanaka et al. accord well with our findings and interpretation of this dendrogram.

In addition, our dendrogram reflects the findings of Nudel et al. (1984), who stated that ". . . the predicted amino acid sequence [of rat skeletal RLC] is identical to that from rabbit except that the rat sequence lacks one of two Gly residues located at positions 12 and 13 in the rabbit sequence." Further, our dendrogram suggests that the gene duplication event that produced the two types of smooth muscle RLC in *Patinopecten* is a very recent event.

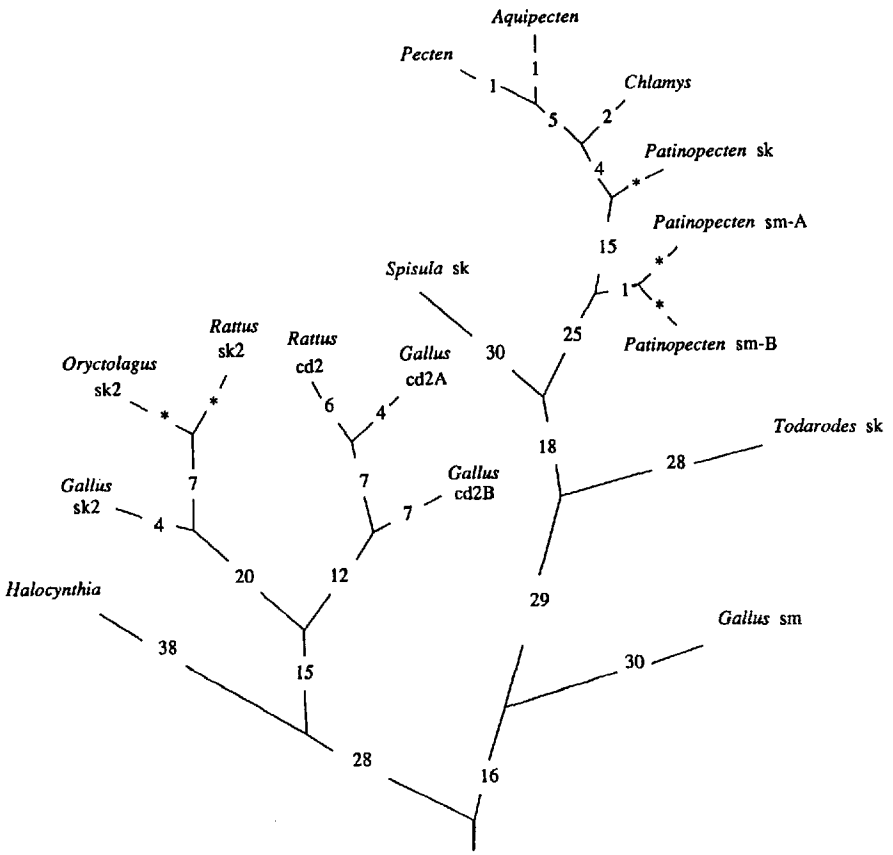


Fig. 8. Relationships among members of the RLC subfamily. Skeletal isoforms of vertebrates are indicated by sk2; skeletal isoforms of invertebrates are indicated by sk; cd2 indicates cardiac; sm indicates smooth muscle. Numbers indicate branch lengths (see legend of Fig. 3) and represent relative amounts of divergence among sequences; * indicates a branch length of zero.

In contrast to ELCs, in which mammalian ventricular and atrial isoforms result from different genes, the cardiac RLCs of mammals apparently are encoded by a single gene (Kumar et al. 1986; Henderson et al. 1988). However, there are two types of cardiac RLCs in *Gallus* (A and B, as indicated in our data base, Appendix II, and in Fig. 8), and according to Matsuda et al. (1981a), there are two types of RLC in cardiac muscle myosin in *Oryctolagus, Homo, Papio,* and *Canis.*

The sequence of RLC from *Drosophila* (Parker et al. 1985) was not included in this study because we were not aware of it prior to October 26, 1988. Future analyses will include this sequence. Its placement in our dendrogram should be especially interesting in light of the comments of Parker et al. (1985): "[There] are unexpected similarities [among the amino-terminal] sequence of the *Drosophila* MLC-2 protein [and] vertebrate myosin alkali light chains."

SARC

The SARCs (Fig. 9) were sequenced from several invertebrate species in a search for an analog or homolog of PARV, which has been found only in vertebrates. Like PARV, SARC seems to be more abundant in fast muscles, but no functional relationship has been established from this distribution. To date, SARCs have been sequenced from representatives of the Mollusca, Annelida, Arthropoda, and Chordata (Protochordata). Relationships among these sequences generally accord with commonly accepted theories of organismal phylogeny: the arthropods *A. leptodactylus* and *Penaeus* sp. group together, the annelids *Nereis diversicolor* and *Perinereis vancaurica* cluster together, and the order of divergence is Mollusca, Annelida, Arthropoda, and Chordata.

The fast rate of evolution of SARCs indicates that they are probably not very specific in their function and is reflected in their divergent relationships (Fig. 9). Takagi et al. (1986a) noted that "Antibodies raised against SARC from amphioxus does not crossreact with SARC's from crayfish or sandworm and has free N-terminal Gly, also in contrast to SARC's from the three other phyla of non-vertebrates." Takagi et al. further reported that "SARC's from crustaceans occur as homo- or heterodimers, a pecularity not shared by SARC's of other phyla."

Our analyses indicate that the SARC from *Patinopecten* is the most divergent of these sequences. In fact, it was placed closer to CALP of vertebrates than to the other SARC sequences, all of which are from nonvertebrate taxa. The uniqueness of the SARC from *Patinopecten* in comparison with the

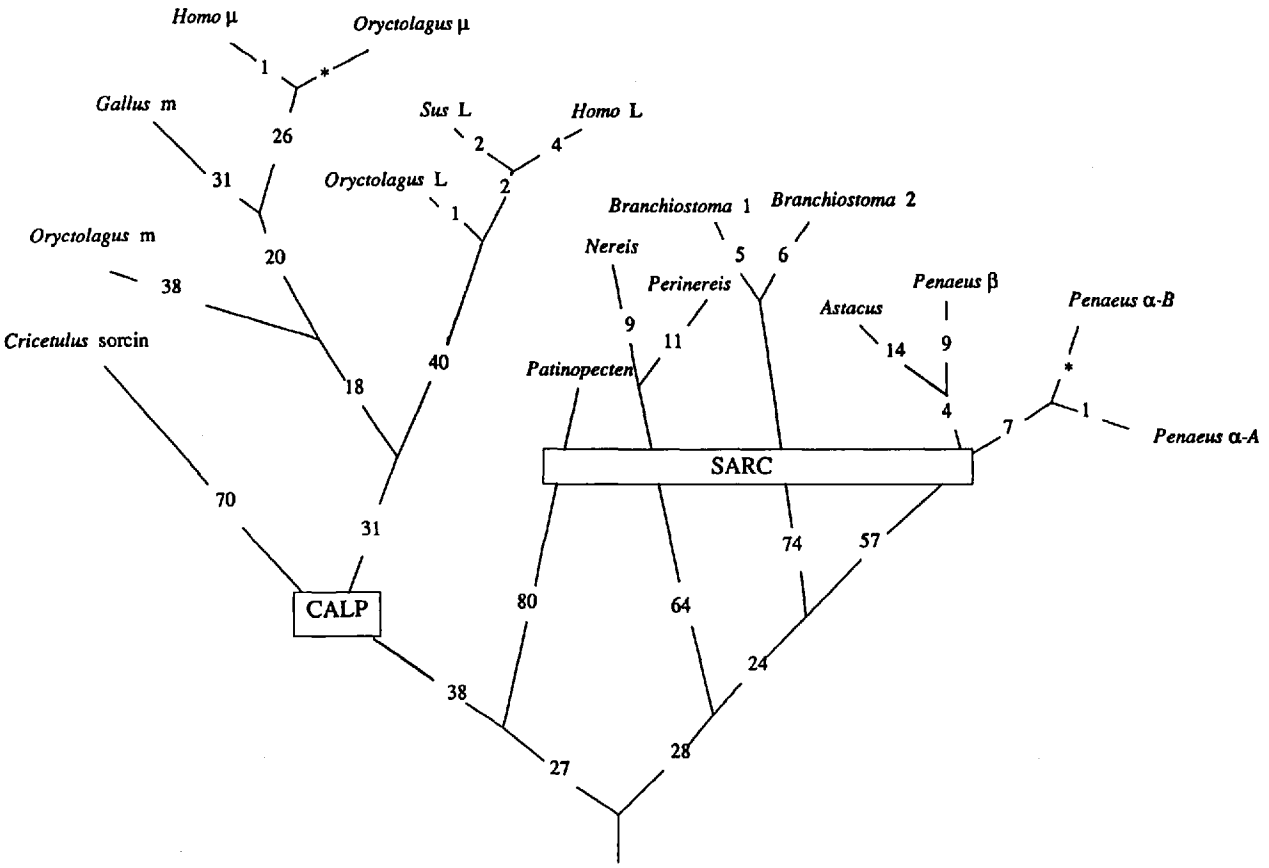


Fig. 9. Relationships among members of the CALP and SARC subfamilies. CALP light chains are indicated by L; the two types of CALP heavy chains are indicated by m and μ. Numbers indicate branch lengths (see legend of Fig. 3) and represent relative amounts of divergence among sequences; * indicates a branch length of zero. Sorcin, from multidrug-resistant *Cricetulus* cell lines, is a member of the CALP subfamily. The sequence from *Patinopecten* is considered a SARC, even though it is placed on the branch leading to the CALP subfamily.

other SARCs is readily seen in Table 2. The calcium-binding domains of SARC from *Patinopecten* follow a 1, 3 pattern of inferred or confirmed calcium binding; SARCs from other species follow two patterns, both different from the SARC of *Patinopecten*: 1, 2, 3 and 1, 3, 4. As indicated in Table 2, several domains that have putative calcium-binding side chains also have characteristics that might alter their conformations and calcium affinities: + [−] helix E nonstandard; + (−) not Gly at position 17; and + {−} helix F incomplete.

CALP

The CALPs (Fig. 9) are calcium-dependent, neutral, intracellular thiol proteases that have four EF-hands and are found ubiquitously in tissues of higher animals. According to Ohno et al. (1984), in *Gallus*

> ... Ca^{2+}-protease (80 kDa subunit) consists of 4 functional domains. Domain II is homologous to thiol protease like papain, and domain IV to calcium-binding protein like calmodulin [which in turn consists of four EF-hand domains]. The origins and functions of domains I and III are, however, unknown. Thus, Ca^{2+}-protease probably arose through the

Table 2. Patterns of calcium-binding in members of the SARC subfamily

Protein	Domain 1	2	3	4
SARC *Penaeus α-A*	+ [−]	+	+ {−}	−
SARC *Penaeus α-B*	+ [−]	+	+ {−}	−
SARC *Astacus*	+ [−]	+	+ {−}	−
SARC *Penaeus β*	+ [−]	+	+ {−}	−
SARC *Branchiostoma* 1	+	+ (−)	+	−
SARC *Branchiostoma* 2	+	+ (−)	+	−
SARC *Perinereis*	+	−	+ (−)	+
SARC *Nereis*	+	−	+ (−)	+
SARC *Patinopecten*	+	−	+ (−)	−
AEQ	+	−	+	+
CALP	+	+	−	−
CMSE	+	+	+	+

Data from the AEQ and CALP subfamilies and the unique homolog CMSE are provided for comparison. SARCs are identified by genus. + indicates that the domain is inferred to bind calcium; − indicates that it is not; + [−] indicates that helix E is nonstandard; + (−) indicates that position 17 is not Gly; + {−} indicates that helix F is incomplete

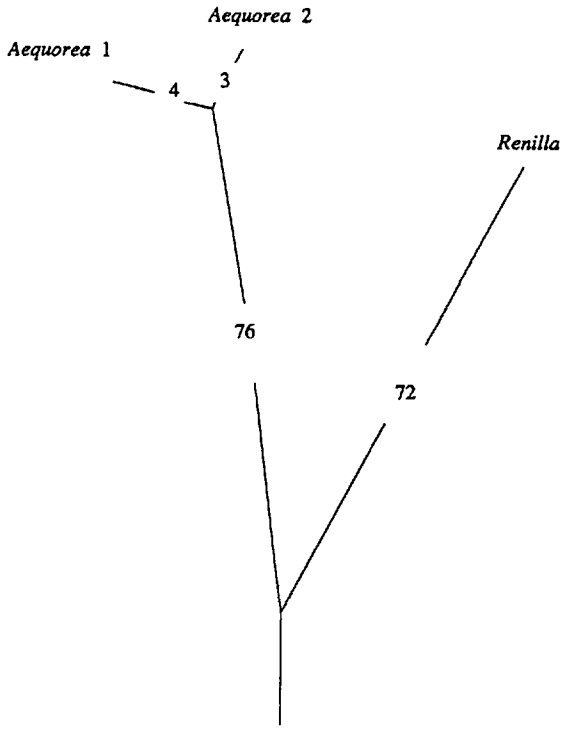


Fig. 10. Relationships among members of the AEQ subfamily. Numbers indicate branch lengths (see legend of Fig. 3) and represent relative amounts of divergence among sequences. Aequorin itself is represented by *Aequorea*; *Renilla* represents the related protein, luciferin-binding protein.

fusion of genes encoding polypeptides with completely different functions.

Aoki et al. (1986) observed that

. . . CALP's are composed of two subunits, a large [or heavy] (catalytic M_r 80,000) one and a small [or light] (regulatory M_r 28,000) one. Two types of CALP heavy chains exist in mammals, μCALP and mCALP, which respectively require micromolar and millimolar Ca^{2+} for their activity. As the small subunit is common to both [mammalian] CALP's, the Ca^{2+} requirement [=sensitivity] of CALP is determined apparently by the large subunit. Only one molecular species of CALP, with an intermediate Ca^{2+} sensitivity, has been found in chicken.

From our analyses we can infer that the small (light, L) and large (heavy, μ and m) subunits shared a common precursor, and that the gene duplication event that produced these forms occurred prior to the emergence of birds and mammals. The event that produced the μ and m heavy chains probably also occurred before the emergence of mammals and birds, after the gene duplication that produced small and large subunits. Because the N-terminal structures of heavy and light CALP subunits "are totally different, [the N-terminal portions of these two chains] may be derived from different evolutionary origins," suggesting two independent fusion events (Emori et al. 1986a).

Our results contradict the suggestion of Emori et al. (1986b) that the μ type CALP of *Oryctolagus* is more similar to *Gallus* m type CALP than it is to m of *Oryctolagus*. Perhaps *Gallus* has lost the μ gene, perhaps this gene has just not been detected in *Gallus* yet, or perhaps *Gallus* m may be an evolutionarily intermediate form; the calcium sensitivity of *Gallus* m is intermediate between that of mammalian m and μ (Emori et al.). If *Gallus* CALP is designated μ instead of m, this would resolve the discrepancy.

The node that includes the protein called sorcin, which is from the multidrug-resistant *Cricetulus griseus* ovary cell line CH5C5 (Van der Bliek et al. 1986), is proximal to the sequences identified as CALPs. We tentatively consider sorcin to be a member of the CALP subfamily, in agreement with statements made by Van der Bliek et al. The function of sorcin and its contribution to multidrug resistance are unknown; however, probes for *Cricetulus* sorcin hybridize to *Homo* and *Mus* DNA (Van der Bliek et al.).

AEQ

The AEQ subfamily includes the bioluminescence protein aequorin and the luciferin-binding protein from the cnidarian coelenterates *Aequorea victoria* and *Renilla reniformis,* respectively (Fig. 10). The bioluminescence systems of these organisms are especially interesting due to the presence of energy transfer systems, which are well characterized biochemically (Cormier 1978). Sequence heterogeneity exists in AEQ within single organisms; Charbonneau et al. (1985) reported that there are at least three isotypes. Prasher et al. (1987) demonstrated multiple isotypes in a single *Aequorea* circumoral ring; so, the fact that ". . . aequorin was isolated from several hundred thousand individual jellyfish (*Aequorea victoria*) . . . during collections conducted over a period of several years . . ." (Charbonneau et al.) probably is not responsible for the heterogeneity. Prasher et al. (1987) reported that ". . . five aequorin cDNA's have been compared and shown to code for three aequorin isoforms," and noted that ". . . this variation can be explained either by a multigene family or by alternative splicing of a primary mRNA transcript from multiple exons comprising the aequorin gene. A definitive explanation must wait until genomic clones are available."

Like CALP, the second domain of AEQ is distinctive: ". . . [evidence presented] suggests splicing of DNA coding for the [luciferin and oxygen binding sites] into the site previously occupied by the second domain of a four-domain protein" (Charbonneau et al. 1985). Charbonneau et al. noted that "One other precedent exists for an enzyme with both EF-hand domains and an identified enzymatic activity within

the same polypeptide chain. This is the Ca(II)-dependent protease calpain." They suggested that "In the case of aequorin a similar splicing event could have formed the catalytic site, or the active site could have diverged from one EF hand within a four-domain precursor."

Thus, within both aequorin and luciferin-binding protein three domains are canonical EF-hands, but domain 2 bears little resemblance and would not be identified as an EF-hand if it were not flanked by canonical domains 1 and 3. The position of AEQ relative to the other subfamilies (Fig. 4) is not changed when only domains 1, 3, and 4 are considered (data not shown).

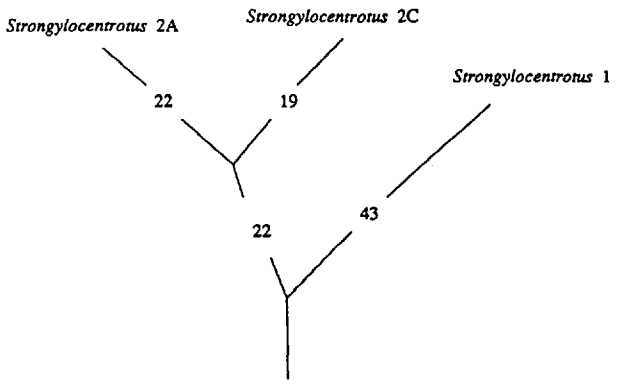


Fig. 11. Relationships among members of the SPEC subfamily. As the name implies, all members of this subfamily are from *Strongylocentrotus*. Numbers indicate branch lengths (see legend of Fig. 3) and represent relative amounts of divergence among sequences.

SPEC

The SPEC cDNAs (Fig. 11) were originally derived from clones of mRNAs extracted from the aboral ectoderm of *S. purpuratus* during embryogenesis and larval development. According to Carpenter et al. (1984), ". . . they are not a sea urchin equivalent of any other known protein sequence." SPECs have four EF-hands.

Tomlinson and Klein (1989) commented that

> In the *S. purpuratus* genome there is a single Spec 1 gene and six or seven related Spec 2 genes (Hardin et al. 1985; 1988). Four of these genes, Spec 1, Spec 2a, Spec 2c and Spec 2d, have been cloned and studied in considerable detail. . . . While the members of the Spec gene family display identical cell type specificity, Spec messages are individually regulated: Spec 1 mRNA begins to accumulate at the early blastula stage and is most abundant; Spec 2a/Spec 2c mRNAs begin accumulating at the late blastula–early gastrula stage and reach about one-half the levels of Spec 1; and Spec 2d mRNA accumulates mostly during gastrula and pluteus stages, with levels reaching only 2% of Spec 1 (Hardin et al. 1988).

The function of the Spec proteins remains unknown.

An attempt to find homologs of these proteins in the distantly related sea urchin *Lytechinus pictus* (Xiang et al. 1988) produced cDNA clones for an eight-domain protein, LPS1. Interestingly, LPS1 did not cluster with the SPECs in our analyses (Fig. 4b); its placement will be discussed in more detail in the section on unique homologs.

Two published sequences were not included in these analyses. One is from SPEC 2d, published by Hardin and Klein (1987). Its placement in future analyses should be quite enlightening; Hardin and Klein note that SPEC 2d is "the most divergent member of the family." Also, we have tentatively classified a 15-kd protein from *H. pulcherrimus* as belonging to this subfamily, following the suggestion of Hosoya et al. (1988) that "It was most homologous to Spec proteins." Subsequent analyses will address the relationship of these two sequences to those represented in Fig. 11.

CLBN

CLBNs (Fig. 12) are six-domain homologs whose synthesis is dependent on vitamin D-derived hormones. The placement of CLBN relative to the other subfamilies was determined using the first four N-terminal domains of CLBN and all four domains of the calretinin fragment. The arrangement of sequences among the CLBNs was based on comparisons of all six domains and is in agreement with organismal phylogeny.

Because the 28-kd CLBN was first obtained from chicken intestine, these molecules are sometimes referred to as avian calbindin. This distinguishes them from the ICBPs that are two-domain homologs that belong to the S100 subfamily. ICBPs are sometimes called 9-kd calbindins and are also known as mammalian calbindins because they have not been sequenced from chicken (or any other bird). We recommend that the name CLBN be reserved for the 28-kd six-domain molecule, and that the modifier avian be dropped, because this protein is found in vertebrate classes other than Aves.

According to Fullmer and Wasserman (1987),

> The primary structure [of CLBN] shows six homologous regions of sequences based on the EF-hand concept of calcium binding, four of which are predicted to actually bind calcium [loops 2 and 6 are predicted not to bind calcium from structural evidence, Table 1]. Aside from these regions, there is no overall structural identity or apparent similarity with the mammalian calbindins (9 kDa) [=ICBP], calmodulin, or troponin C.

Parmentier et al. (1987) also suggested that these proteins are distinct members of the EF-hand superfamily:

> Comparisons of calcium-binding domains from various proteins suggested that all members of the troponin C superfamily derive from a common two-domained [precursor], but that duplications leading to calbindin and to the four-

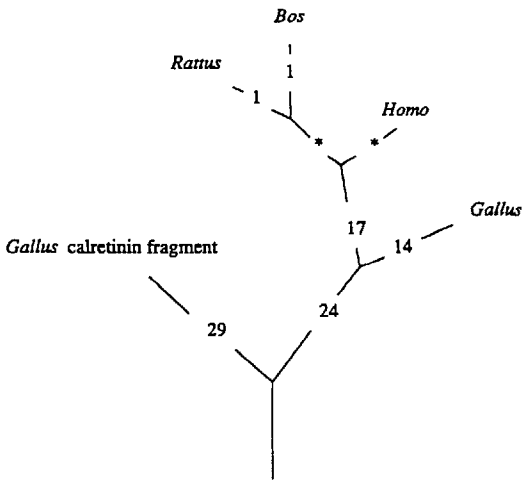


Fig. 12. Relationships among members of the CLBN subfamily. Numbers indicate branch lengths (see legend of Fig. 3) and represent relative amounts of divergence among sequences; * indicates a branch length of zero. Calretinin, represented by a four-domain fragment, is a member of this subfamily.

domained calcium-binding proteins took place independently on different branches of the evolutionary tree.

Supporting this finding, Wilson et al. (1988) reported that

> There are ten introns [in CLBNs], most of which do not fall at homologous positions, neither with respect to the sixfold repeating structure of the calbindin protein, nor with respect to previously sequenced genes for calmodulin and other calcium-binding proteins.

This and other theories that pertain to the evolution of the various subfamilies of EF-hand proteins are reviewed in the discussion.

Fullmer and Wasserman (1987) predicted that loops 2 and 6 do not bind calcium; this would result in pairings of domains that do not bind calcium and those that do: 1 and 2, and 5 and 6. They noted that

> Kretsinger and Barry (1975) proposed that EF-hands are arranged in pairs. [It] is clear [from ICBP] that an altered and a normal Ca-binding loop can pair [but,] it is not known if a non-Ca-binding loop can pair with a Ca-binding. However, this has been predicted for cardiac [cardiac/slow skeletal] TNC loops I and II. It is predicted that the secondary structure of 28-kDa calbindin-D is significantly different from other proteins of this class, which bind four calcium atoms.

Parmentier et al. (1987) observed

> If the only function of calbindin were to bind calcium, one would expect, according to the neutral theory of evolution, that . . . inactive domains would exhibit much higher evolutionary rates than the active ones . . . [therefore] the selective pressure exerted on calbindin is not restricted to its ability to bind calcium. . . . It could, for instance, interact in a regulatory way with other proteins, or use its two degenerated calcium-binding domains for other important, albeit still undefined functions.

Our analyses agree with the statement by Wilson et al. (1988), which indicates that CLBN and calretinin are quite similar, "The introns are in the same positions in the calretinin and calbindin genes." In fact, Rogers (1987) states that "Both genes date from before the divergence of chicks from mammals."

PARV

The PARVs (Fig. 13) are, in many ways, the best-characterized calcium-binding proteins. They were the first purified (Henrotte 1952), the first of known amino acid sequence (Pechére et al. 1971), the first of known crystal structure (Kretsinger et al. 1971), and the first for which multiple isoforms were recognized in a single individual organism (Kretsinger 1972). They are inferred to function as a kinetic buffer of calcium in fast-twitch muscle (Gillis et al. 1982), perhaps increasing the rate of muscle relaxation. PARVs are also present in brain, bone, and several endocrine tissues; however, no function has been inferred for PARV in these tissues (Epstein et al. 1986).

Two types of PARV, α and β, are generally recognized. In Fig. 13 we indicate the nomenclature adopted by the authors of the primary-sequence references. There are at least 11 residues characteristically different between α and β forms of PARV. These 11 residues are not invariant within either subgroup, but we do suggest that scoring these positions will usually distinguish between the two forms. The 11 positions and characteristic amino acids are as follows: positions 1–8, α Met, β Ile; domain 2, position 4, α Lys, β Ala; position 9, α Phe, β Cys; position 21, α Lys, β Thr; position 26, α Leu, β Lys; position 2+1, α Lys, β Ala; domain 3, position 7, α His, β Gly or Lys; position 25, α Ile, β Phe; position 3+3, α Glu, β Ser or Asp; domain 4, position 5, α Leu, β Phe; position 11, α Lys, β Ser.

We can infer from the currently available sequences that at least two forms of PARV existed in the ancestor of vertebrates, because both α and β types have been sequenced from fishes, amphibians, and mammals. The α form has not been sequenced from a reptile, and neither type has been sequenced from a bird. This may reflect an evolutionary loss of specific isoforms of PARV in these lineages, or it may reflect the incompleteness of our data base at this time.

No protein identified as a β PARV in the primary reference has been sequenced from a mammal. However, the tumor protein identified as oncomodulin by Gillen et al. (1987) always clusters with the β PARVs in our analyses; therefore, oncomodulin is clearly a PARV. Mutus et al. (1985) reported that oncomodulin, but not PARV, can activate cyclic nucleotide phosphodiesterase (as does CAM), albeit

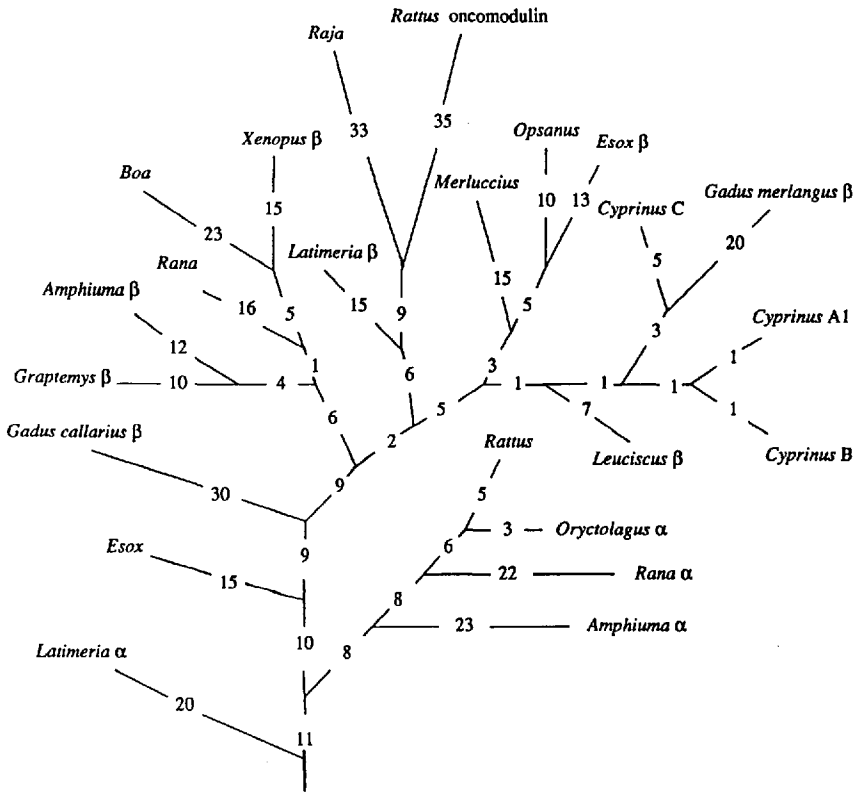


Fig. 13. Relationships among members of the PARV subfamily. Designations for α and β isoforms are presented as specified in the primary references. Numbers indicate branch lengths (see legend of Fig. 3) and represent relative amounts of divergence among sequences.

at 25 times higher concentration. If *Rattus* has no other β form, oncomodulin can be considered to be the β form of PARV in mammals; on the other hand, oncomodulin may be another isoform of PARV. If this is the case, the gene encoding the β form may have been deleted or inactivated in *Rattus* and other mammals. Additional sequences are required to resolve these questions. There are two PARV sequences from mammals that were not included in these analyses. According to Berchtold (1988), sequences of *Homo* (Berchtold) and *Mus* (Zuhlke et al., personal communication) PARVs ". . . are very similar to the other mammalian parvalbumins," suggesting that they may be α PARV.

Isoforms of PARV other than α and β may indeed exist. Note the sequence of *Esox* near the base of the β branch. If this sequence is considered to be α, then our analyses would indicate a separate evolutionary origin of α in *Esox*. A simpler explanation is that two types of PARV coexist within *Esox*; one can be assigned to the β group, the other belongs to an as yet unnamed group. The occurrence of more than two isoforms of PARV is also supported by the fact that the sequences designated β from *Gadus merlangus* and *Gadus callarias* are not placed together; they are, in fact, quite divergent from each other and are probably not orthologs. If these sequences did represent orthologous gene products, we would expect them to be very near (if not adjacent) to each other, because they were obtained from two species in the same genus. This situation contrasts with the multiple isoforms that were recovered from a single individual of the species *Cyprinus carpio*. These sequences cluster very near each other, reflecting a close evolutionary relationship that probably resulted from a very recent polyploidization of the entire genome of this species and other cyprinids.

Clearly, PARVs are inappropriate for inferring organismal phylogenies (Fig. 13); it is difficult to detect which of the numerous isoforms are orthologs. As Maeda et al. (1984) observed, ". . . varying rates of amino acid replacement, much homoplasy, considerable gene duplication, plus complicated lineages make the set of parvalbumin sequences unsuitable for systematic study of the origin of the tetrapods and other higher-taxa divergence. . . ."

ACTN

ACTN (Fig. 14) is a component of the Z-line in skeletal muscle and can be divided into three distinct regions: (1) the N-terminal 240 amino acids probably represent the actin-binding domain, (2) amino acids 270–740 contain four repeats of a spectrin-like sequence, and (3) the C-terminal sequence contains two EF-hand domains (Baron et al. 1987). The cDNA sequence reported by Baron et al. is from a chick embryo fibroblast library. The deduced protein contains two EF-hands that probably do not

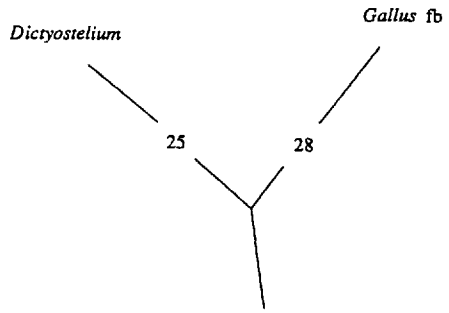


Fig. 14. Relationships among members of the ACTN subfamily; fb indicates fibroblast. Numbers indicate branch lengths (see Fig. 3) and represent relative amounts of divergence among sequences.

bind calcium. The cDNA hybridizes to only one gene in chicken; Baron et al. concluded that "results of Northern blots are consistent with the view that smooth and skeletal muscle α-actinins are encoded by separate genes, which are considerably divergent." Arimura et al. (1988) presented data that agree with this statement. We will include the recently published sequence of ACTN from chicken skeletal muscle (Arimura et al. 1988) in future analyses.

According to Baron et al. (1987),

> Distinct isoforms of α-actinins have been isolated from different tissues, and even from the same tissue. [A] major functional difference [is] found between muscle α-actinins, which are not Ca^{2+}-sensitive in their binding to actin, and non-muscle α-actinins, which are [Ca^{2+} sensitive]. [The cDNA from chicken embryo fibroblasts] matches all the protein sequence data we have directly obtained from chick smooth muscle α-actinin, [although it is] not clear why fibroblasts should be expressing a muscle-type α-actinin.

The

> α-actinin from ... the slime mould *Dictyostelium discoideum* [clusters with the chicken smooth muscle ACTN and] carries two ... EF-hand structures at the C terminus [whose] calcium-binding loops contain all necessary liganding oxygens and most likely form the structural basis for the calcium sensitivity of strictly calcium-regulated non-muscle α-actinins. (Noegel et al. 1987)

Noegel et al. (1987) present

> ... the first complete sequences of a nonmuscle α-actinin [which is] completely inhibited by calcium; the alignment scores [from comparisons with typical members of the CAM, TNC, ELC, and PARV subfamilies] suggest that α-actinin assumed the EF-hand structure very early in evolution, most likely before the separation of slime moulds and higher plants.

S100

All of the two-domain proteins except the α-actinins and the unique calcium-binding protein from *Tetrahymena,* which will be discussed below, group together in a single subfamily, which for brevity we call S100 (Fig. 15). The first domain, as seen in the crystal structure of bovine ICBP (Szebenyi and Moffat 1986) is unique to the S100 subfamily. It has two extra amino acids, whose sites are designated 12b and 16b to indicate insertions after canonical positions and 12 and 16. The calcium ion is coordinated by four carbonyl oxygen atoms from residues Ala (10 X), Glu (12 Y), Asp (14 Z), and Gln (16 −Y), by the side chain of Glu (21 −Z), and by a water molecule bridged to the side chain of Ser (18 −X). We infer that all known members of the S100 subfamily (except the p11 proteins) share this ICBP-hand in the first domain; the second domain (except that of the p11 proteins) binds calcium with a canonical EF-hand. At this time functions are not known for any of the members of the S100 subfamily; however, the characteristics of these proteins are so diverse that several different functions must exist. Nonetheless, we believe that some unifying concepts for this subfamily will emerge as more data are collected.

In comparing patterns of nucleotide and amino acid sequence evolution, the S100s provide an interesting contrast to the other subfamilies of EF-hand proteins. For example, in the CAMs of animals (exclusive of pseudogenes), there are only five known differences in amino acid sequence (Fig. 5). Furthermore, the amino acid sequence of the α form of CAM from *Arbacia* is identical to that found in seven vertebrates (*Homo, Oryctolagus, Rattus, Mus, Bos, Gallus,* and *Xenopus*). Yet in these same CAMs, the third base of equivalent triplets has diverged to randomness in the 6×10^8 years since the divergence of the ancestor of echinoderms and vertebrates. In contrast the S100s have been sequenced from only a limited number of mammals (*Bos, Sus, Rattus,* and *Homo*), yet they have a broader distribution of amino acid sequences. However, as noted by Saris et al. (1987),

> ... these genes show a greater conservation at the nucleotide level than at the amino acid level. This coupled with the high degree of conservation between the 3' untranslated sequences of bovine and murine p11 indicates that the nucleotide sequence of this gene family has been under strong selective pressures independent of that required to maintain product function.

In a subsequent paper, we will describe in detail a parallel study of evolution in EF-hand homologs based on analyses of cDNA and gDNA sequences.

There are four easily recognized groups of proteins in the S100 subfamily. We assume that the conformations of these proteins are very similar to that of ICBP (Szebenyi and Moffat 1986). We will discuss each of these in turn, beginning with S100 itself. The name S100 was originally coined to describe the solubility in 100% saturated $(NH_4)_2SO_4$, at pH 7 of this group of dimeric proteins (reviewed by Donato 1986 and by Baudier 1988). To date,

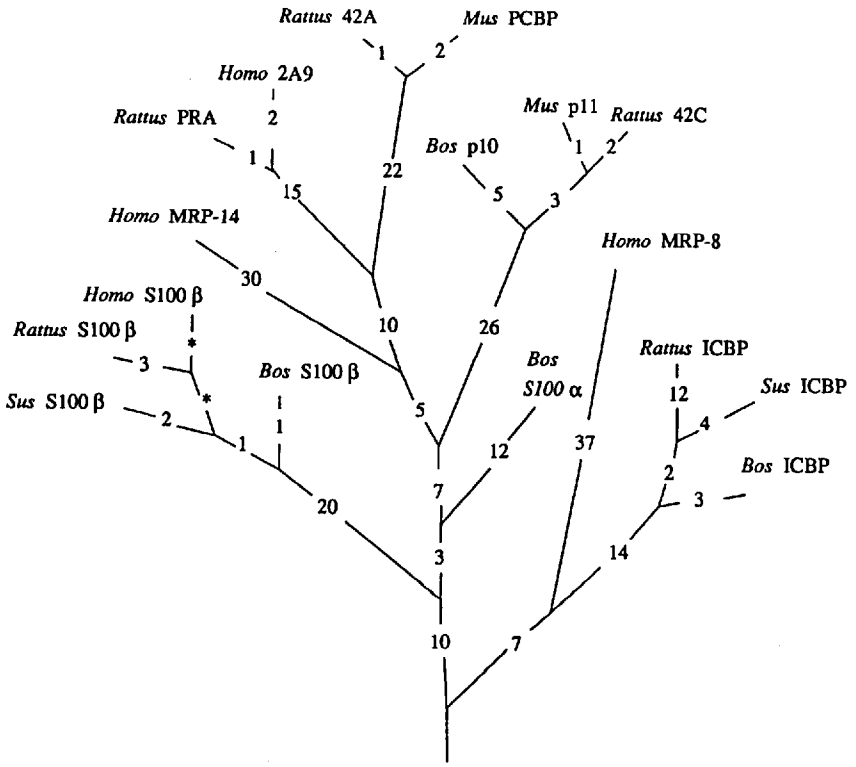


Fig. 15. Relationships among members of the S100 subfamily; both genus name of the organismal source and protein designation are given because a standard nomenclature has not been established. The sequence of p11 from *Sus* (not shown) is identical to the sequence of p10 from *Bos*. Numbers indicate branch lengths (see legend of Fig. 3) and represent relative amounts of divergence among sequences; * indicates a branch length of zero.

they have been found only in mammals and are present in high concentrations in nervous tissue, as well as in cardiac tissue and adipocytes. There are two distinct classes of subunits, as confirmed by our analyses, α and β (Fig. 15). The dimeric proteins are called either a_0 (α, α), a (α, β), or b (β, β).

The subunits appear freely interchangeable, and the concentrations of a_0, a, and b within any cell appear to reflect the binomial distribution expected for random association of α and β subunits. S100a and S100b are found in glial cells; $S100a_0$ is enriched in neurons. Donato (1986) noted that ". . . rat brain S-100 is mostly, if not exclusively, S-100β." Zimmer and Van Eldik (1989) found that the content of S100, primarily as the $\beta\beta$ form, increases fourfold during differentiation of C6 glioma cells. In addition to tissues of ectodermal origin, S100 is found in heart, primarily as $\alpha\alpha$ (Kato and Kimura 1985), and in adipose tissue, primarily as b (Hidaka et al. 1983), as well as in trachea and skin. The β subunit contains two Cys and the α subunit contains one Cys; however, the subunits within the dimer are not crosslinked by disulfide bonds as the protein is extracted from these tissues. In contrast Kligman and Marshak (1985) reported that disulfide-crosslinked S100β isolated from bovine brain has neurite extension activity when assayed with a primary culture of neurons at low density in serum-free medium. Donato (1986) noted the value of S100 to clinical diagnoses: ". . . the presence of S-100 in the cerebrospinal fluid has been taken as an index of cell injury in the nervous system."

The function(s) of S100 remains to be established; no catalytic activity has been assigned to it. $S100a_0$ stimulates the basal (Mg^{2+}-activated) adenylate cyclase of skeletal muscle (Fano et al. 1989); in contrast S100b inhibits this activity. S100 binds to unassembled tubulin (Donato 1988), thereby preventing polymerization into microtubules or driving the equilibrium toward depolymerization. Hagiwara et al. (1988) reported that S100 binds to p36, as does its homolog p11, thereby inhibiting phosphorylation of its tyrosine(s) by p60sarc kinase and by p130fps kinase. Zimmer and Van Eldik (1989 and previous work cited therein) showed that S100β activates fructose-1,6-bisphosphate aldolase C. Numerous studies (reviewed by Donato 1986 and by Baudier 1988) show that S100 interacts with synthetic and natural membranes.

In addition to its dimeric nature, S100 is distinguished from the other groups within its subfamily by its calcium- and zinc-binding characteristics. As anticipated from its sequence it binds four Ca^{2+} ions per dimer (i.e., one per domain). The affinity for Ca^{2+} is moderate, $pK_d(Ca) \sim 5.0$, with no other salt present; however, at physiological KCl, 120 mM, the affinity is reduced to $pK_d \sim 3.0$ for both a_0 and b forms (Baudier 1988). Calcium affinity is increased, possibly to a physiologically significant range, in the presence of lipid vesicles (Zolese et al.

1988); and in turn S100b affinity for cardiolipin vesicles is increased by calcium. They estimate from circular dichroism measurements that the α-helix content decreases from 33% to 28% upon binding calcium, to 25% upon interaction with vesicles, and to 9% with calcium and cardiolipin. The sensitivity of calcium coordination is further illustrated by the increase in $pK_d(Ca^{2+})$ to ~6 upon alkylation of Cys 85α or of Cys 84β with the thiol-specific probe Bimane (Zolese et al. 1988). Furthermore, S100b is unique, even relative to $S100a_0$ or other groups within its subfamily, in that it binds four equivalents of zinc with $pK_d(Zn)$ ~7.5. This zinc binding also reduces the amount of α-helix and it increases the affinity of S100b for calcium. Baudier et al. (1986) noted "... that S100b protein becomes highly hydrophobic upon Zn^{2+} binding whereas S100a and S100a' are not affected." They exploited this effect to retard S100b selectively in phenyl-Sepharose chromatography. The change in conformation associated with calcium binding has several manifestations. Baudier and Gerard (1986) concluded that "... only the α-subunit exposes hydrophobic domains to solvent in the presence of calcium and that cysteine residues exposed upon Ca^{2+} binding to S100 correspond to Cys 85α and Cys 84β." They further concluded that "At acidic pH, or in the presence of calcium, aromatic residues are exposed to solvent and the quaternary structure becomes less stable." How these in vitro characteristics are related to the function of S100 remain to be deciphered; however, the inferred evolution of zinc-binding ability is especially interesting. The β subunit has two Cys and five His; the α subunit has one Cys and two His. The coordination of zinc by S100 is not understood, but zinc-binding regions usually have four ligands with the sulfur of Cys and one of the nitrogen atoms of the His side chain being the favored ligands of proteins.

The second group within the S100 subfamily (Fig. 15), ICBP, is especially abundant in the duodenum and placenta of mammals and is also found in mammalian skin and kidney. ICBP is nearly absent in rachitic animals; its synthesis is dependent upon vitamin D and precedes increased calcium transport. The usual isotype is 78 residues long. Wasserman and Taylor (1966) suggested that ICBP be called calbindin D 9 kd to distinguish it from calbindin D 28 kd, whose synthesis is also dependent upon vitamin D. According to our analyses (results not shown), none of the six domains of CLBN are closely related to ICBP and especially not to the singular domain 1. As seen in the refined crystal structure (Szebenyi and Moffat 1986), domain 2 of ICBP has the canonical EF-hand structure and is essentially superimposable on the canonical EF-hands of PARV, TNC, and CAM. Both domains of ICBP bind Ca^{2+} ions weakly (pK_d ~5). Although the function of ICBP remains unknown, Kretsinger et al. (1982) presented a model in which ICBP facilitated the diffusion of calcium across the cytosol of epithelial cells. Feher et al. (1989) have confirmed that ICBP facilitates calcium diffusion in vitro, and in fact, it does so better than does CAM, which has a higher affinity for calcium.

The third group of proteins in this subfamily, the p11 proteins, are subunits of calpactin. Calpactin is a heterotetramer consisting of two heavy chains 36,000 kd and two light chains 11,000 kd. It is found in brain, spleen, and thymus and is present in high concentrations in kidney, intestine, and lung. Although its function remains unknown, calpactin has been shown to interact with phospholipid, actin, and nonerythroid spectrin in a calcium-dependent manner. Calpactin I heavy chain is the name gaining favor for a protein variously designated p34, p36, p39, lipocortin II, 36-kd calelectrin, 33-kd lymphocyte Ca-binding protein, 33-kd calcimedin, and chromobindin 8 (Klee 1988). Most cells contain a molar excess of heavy chain relative to light chain, hence the suggestion that the light chain, which we will call p11, may regulate the association of monomeric heavy chain into an active heterotetrameric form. The heavy chain is a major in vivo substrate of retroviral and growth factor receptor protein–tyrosine kinases, as well as of protein kinase C. The calcium-binding site has not been identified. The heavy chain is not an EF-hand homolog; however, the isolated p11, which is a homolog, does not bind calcium with high affinity. The amino acid sequences of the second domains of all members of the S100 subfamily except the p10, p11, 42C grouping indicate canonical structures capable of binding calcium. All three p11s have Cys at the Y vertex, instead of the Asp or Asn almost always found at Y in domains of demonstrated calcium binding. All three also have Ser at −Z instead of the usual Glu. Although the sulfur of Cys could coordinate a Ca^{2+} ion weakly, and the long side chain and usual bidentate coordination of Glu at −Z is replaced by bridging waters, this pair of substitutions is certainly consistent with p11's not binding calcium in its second domain. Similarly, the sequences of all of the first domains in the S100 subfamily (except those of the p11s) honor sequence precedents set by ICBP. The calcium coordination here provides fewer guidelines because the X, Y, Z, and −Y oxygen atoms are carbonyl oxygens from the main chain. The important point is that, whereas the singular ICBP domain 1 has two amino acids inserted relative to the canonical EF-hand for a total of 31 residues, the p11s have 28 residues; that is, one amino acid is deleted from the p11s relative to the 29 residues of the canonical hand. Again this is con-

sistent with an inability to bind calcium; however the calcium coordination by calpactin is left unresolved. The heavy chain has no EF-hand or obvious calcium coordination site, and p11 cannot bind calcium. The calcium affinity is low and best demonstrated in the presence of lipid. Perhaps the putative calcium-binding site(s) is at an interface and is comprised of several components. Saris et al. (1987) mapped the p11 gene of *Mus* to chromosome 3, near Gbp-1. By comparison of linkage groups they predict that the *Homo* p11 maps either to chromosome 1p or to 4q.

The fourth group of S100 proteins are called calcyclin, 42A, 42C, and PCBP (Fig. 15). Calabretta et al. (1986) identified, by differential screening of *Homo* cDNA libraries, a cDNA (2A9) encoding an uncharacterized protein called calcyclin. According to Calabretta et al., ". . . the mRNA corresponding to the 2A9 cDNA is not detectable in G_0 cells," and ". . . 2A9 reaches its peak of expression in mid-G_1. Ferrari et al. (1987) subsequently isolated the gene encoding this protein from a *Homo* genomic library. According to Ferrari et al., "The calcyclin gene is a unique copy gene and has 3 exons . . . [and it] has been localized to the long arm of human chromosome 1, near the ski oncogene." Murphy et al. (1988) used ". . . antisera raised against partially purified rabbit mammary gland prolactin receptor" to isolate ". . . a cDNA from a T-47D human breast cancer cell line . . ." and called it prolactin receptor-associated protein (PRA). This cDNA has the same coding sequence as 2A9. The function of calcyclin remains to be established.

Masiakowski and Shooter (1988) differentially probed a cDNA library from *Rattus* pheochromocytoma PC12 cells with DNA from naive PC12 cells and PC12 cells exposed to nerve growth factor (NGF) for seven days. Two mRNAs detected by these cDNAs encode two proteins called 42A and 42C. The second, 42C, is almost identical to p11 from *Mus*. The finding of its induction by NGF is consistent with the suggestion that 42C regulates the assembly of the heavy subunit of calpactin into the heterotetramer, $p36_2/p11_2$. The former, 42A, is different from 42C and identical to *Rattus* p9Ka, whose cDNA was identified in the myoepithelial cells and smooth muscle cells of the normal *Rattus* mammary gland (Barraclough et al. 1988). It is ". . . 15-fold more abundant in the myoepithelial-like cells than in the parental cuboidal epithelial stem cells." Jackson-Grusby et al. (1987) characterized an mRNA, called 18A2, that increases in cultured *Mus* fibroblasts following addition of serum. In whole mice, the highest concentrations of 18A2 mRNA were found in uterus and placenta; it is not detectable in the placenta following the 10th day of pregnancy. Its corresponding protein was called placental calcium-binding protein (PCBP) and is very similar in amino acid sequence to 42A from *Rattus*.

The remaining two members of the S100 subfamily, MRP-8 and MRP-14 (Fig. 15), were isolated from macrophages by Odink et al. (1987). They then synthesized probes from partial amino acid sequences of these proteins, isolated, and sequenced the corresponding cDNAs. The original isolation involved an antibody directed against *Homo* macrophage migration inhibitory factor, MIF. Therefore, Odink et al. designated these sequences MIF-related proteins, MRP. The larger, MRP-14, has 114 amino acids; there are 28 amino acids C-terminal to the second EF-hand domain. Both MRP-14 and MRP-8 are found in granulocytes, monocytes, and macrophages. Odink et al. noted that ". . . In acutely inflamed tissues macrophages can express MRP-14 but not MRP-8. . . ." The MRP-8 cDNA corresponds to the cDNA related to the cystic fibrosis antigen (CFA) after a minor correction in the CFA sequence (Dorin et al. 1987) is made. Subsequently, Bruggen et al. (1988) found that concentrations of MRP-14, but not MRP-8, are significantly elevated in the plasmas of obligate heterozygotes and about a hundred times higher in homozygotes. As in other members of this subfamily, the functions of MRP-14 and of MRP-8 remain unknown. Interestingly, MRP-14 and MRP-8 are not closely related; MRP-14 is more like PRA/2A9 and 42A/PCBP, and MRP-8 is the closest relative of the ICBPs.

It is especially intriguing that the members of the S100 subfamily are quite similar in terms of sequence; yet, as the diversity of names and acronyms would imply, their functions or relationships to one another remain unknown. It is probably just as well to retain these alpha-numeric designations now; however, we anticipate a more rational and meaningful system of designations as we acquire more information about their function(s) and evolution.

Unique Homologs

CALCIB

CALCIB is a protein phosphatase present in mammalian brain. It is composed of two subunits, termed A (61,000 kd), which interacts with CAM, and B (15,000 kd), which binds four Ca^{2+} ions per molecule with affinities in the micromolar range (Aitken et al. 1984). Our assignment of the B subunit, CALCIB (Fig. 4b), as unique accords with remarks by Aitken et al.: ". . . the B subunit is a new member of this family of Ca^{2+}-binding proteins," and

> It is worth noting that there is more inter-protein homology between the Ca^{2+}-binding loops that are in the equivalent positions in the B subunit and calmodulin sequences, than intra-protein homology between the four Ca^{2+}-binding loops

of the B subunit itself. This suggests that the gene duplication events [producing CALCIB] predated divergence in the family of Ca^{2+}-binding proteins.

Also, CALCIB is unusual because it is N-myristilated; the N-termini of most EF-hand proteins are blocked, usually by acetylation.

TPAP

Two proteins from the crayfish *A. leptodactylus* were identified by Wnuk et al. (1986) as TNC because they

provide Ca^{2+} sensitivity for the actinomyosin ATPase in the presence of two other troponin subunits (TnI and TnT) and tropomyosin. Both [isoforms from crayfish] restore Ca^{2+} sensitivity to skinned rabbit adductor (fast-twitch) fibres, devoid of endogenous TnC by extraction with an EDTA solution, [although in both proteins] domains I and III have lost their ability to bind Ca^{2+}.

Although the TPAPs function as TNC and are biochemically quite similar to the other TNCs (Wnuk 1988), they are placed outside the TNC subfamily by our analyses (Fig. 4b). This situation provides an interesting contrast with squidulin and cal-1, which are discussed with CAM; they are placed within, but near the base of the CAM subfamily (Fig. 5).

We derived the acronym TPAP from troponin C, *Astacus pontastacus*. After submittal of this manuscript, J. Cox (personal communication) stated that *A. leptodactylus* is the name of the crayfish from which these proteins were obtained.

CVP

CVP (Fig. 4b) is an 18-kd calcium-binding protein from amphioxus (*Branchiostoma lanceolatum*) muscle. It has four identifiable EF-hand domains; domains 3 and 4 probably bind calcium. Domain 3 contains two ϵ-*N*-trimethyllysine residues in the α-helices flanking the calcium-binding loop (Kobayashi et al. 1987). CVP interacts with a 36-kd protein, which may be a target analogous to those of calmodulin (Cox 1990); in other properties, CVP resembles TNC. CVP contains a disulfide link between Cys-16 (domain 1, position 2) and Cys-78 (domain 2, position 26). With only slight distortion, this S–S bond could be accommodated in the crystal structures of PARV, TNC, or CAM. Kobayashi et al. (1987) noted that the inferred stabilization may obviate the need for calcium binding in domains 1 and 2.

According to Cox (1986), "CVP does not substitute for calmodulin in a specific enzyme assay nor for troponin C in restoring Ca^{2+} sensitivity to skinned muscle fibers." Kobayashi et al. (1987) recognized the uniqueness of this protein,

the amino acid sequence . . . indicates that CVP is directly derived from the four-domain [precursor] but displays some very unusual characteristics . . . the structure of the Ca^{2+}-binding protein found in amphioxus muscle is chimeric . . . half strongly resembles calmodulin and troponin C, which explains its capacity for protein–protein interaction. The other half is reminiscent of the abortive Ca^{2+}-binding domains which occur in parvalbumins and sarcoplasmic Ca^{2+}-binding proteins.

TRACTIN

TRACTIN (Fig. 4b) is a basal body-associated calcium-binding protein that is encoded by a single copy gene in the unicellular flagellated green alga *Chlamydomonas reinhardtii*. It is 20 kd and has four EF-hand domains, all of which appear to bind calcium. "In mitotic cells [TRACTIN] was specifically associated with the poles of the mitotic spindle at the sites of the duplicated basal body complexes" (Huang et al. 1988b). According to Huang et al. (1988a), it is ". . . a component of calcium-sensitive contractile fibers that link the basal bodies of the complex to each other and the complex as a whole to the nucleus." In addition, Huang et al. (1988a) observed that

The deduced amino acid sequence of [TRACTIN] shows a strong sequence relatedness with . . . the deduced amino acid sequence of the yeast CDC31 gene product required for spindle pole body duplication. . . . The association of these sequence-related proteins to microtubule-organizing centers of divergent structure suggests that the proteins may be functionally related.

The close relationship of amino acid sequences from CDC31 and TRACTIN is confirmed by our analyses (Fig. 4b).

CDC31

The CDC31 gene (Fig. 4b) of *Saccharomyces cerevisiae* encodes a four-domain protein that has at least two binding sites for Ca^{2+} in domains 1 and 4. It is distinct from CAM; another yeast gene that is more closely related to CAM from animals (Fig. 5) has been cloned and sequenced (Davis et al. 1986). According to Baum et al. (1986) the CDC31 gene product ". . . performs a specialized role in spindle pole body duplication under the regulation of Ca^{2+} fluxes coordinated with the cell cycle. Such fluxes would regulate other essential processes by interaction with other Ca^{2+}-binding proteins. . . ." Baum et al. further note that ". . . it seems likely that the CDC31 product . . . acts by regulation of another protein . . ." and speculate that the CDC31 product may be localized at the site of the spindle pole body. The location of this protein and the proposal by Baum et al. that "Ca^{2+} fluxes within the yeast cell play a key role in . . . the organization of microtubule arrays . . ." accord well with the comments of Huang et al. (1988a) regarding functional similarities between CDC31 and TRACTIN.

TCBP10

TCBP10 is a 10-kd calcium-binding protein present in the cilia and cell body of the ciliated protozoan *T. thermophila* (Kobayashi et al. 1988b). It contains two EF-hand domains, both of which appear to bind calcium. Although its function has not been demonstrated, TCBP10 was discovered during Kobayashi et al.'s (1988b) studies of calcium-dependent ciliary reversal in *Tetrahymena.* Our analyses (Fig. 4b) confirm the statement of Kobayashi et al. (1988b) that "TCBP-10 has a unique primary structure as compared with the other calcium-binding proteins known so far."

Shortly after the submittal of this manuscript we learned that TCBP10 is, in fact, the degraded product of TCBP25, a 25-kd calcium-binding protein that contains four EF-hands (Takagi, personal communication). We have corrected this error in our data base and will address the relationships of TCBP25 in subsequent reports.

LPS1

Xiang et al. (1988) isolated cDNA clones of LPS1 from *L. pictus* in the course of their studies of SPEC proteins. They observed that

> The sequence of LpS1 reveals the presence of eight EF-hand domains, which share structural homology with the Spec1 or Spec2 EF-hands; however, little else in the protein sequence is conserved. The results support the hypothesis that the LpS1 gene arose from a duplication of a [precursory] Spec gene and that the overall structural features of the Spec family of proteins are more conserved than the amino acid sequences.

Apparently all domains except domain 8 of LPS1 bind calcium.

For our analyses we treated LPS1 as two separate four-domain sequences; otherwise, domains 7 and 8 would not have been aligned with any other sequence, and domains 5 and 6 would have been aligned with only the two C-terminal domains of CLBN. The two separate four-domain sequences of LPS1 always clustered together, supporting the view that this eight-domain protein is the result of a recent duplication event, which occurred after divergence of these genes of *Lytechinus* from the SPEC genes of *Strongylocentrotus.* Because SPEC and LPS1 are some distance apart in the dendrogram (Fig. 4b), we cannot be certain that they perform the same function in *Strongylocentrotus* and *Lytechinus.* Interestingly, the two-domain protein from *Tetrahymena,* TCBP10, clusters with LPS1 in our analyses. This affinity is based on similarities between the two domains of TCBP10 and domains 1, 2 and 5, 6 of LPS1, because as noted above, we included LPS1 as two four-domain sequences.

CMSE

Swan et al. (1987) cloned and sequenced a gene that encodes a calcium-binding protein in the gram-positive bacterium that they identified as *S. erythraeus.* They inferred four EF-hands, all of which appear to bind calcium, and further stated that

> The EF-hand motif may have arisen in an ancient protein before the divergence of the eukaryotes and prokaryotes. The similarity of the bacterial protein to CAM rather than to other known EF-hand proteins supports the view that the eukaryotic EF-hand superfamily has diverged from a common calmodulin-like precursor with four calcium-binding domains. Less likely alternatives are that convergent evolution, or later gene transfer from eukaryote to prokaryote may be involved.

These points regarding the structural similarities of CMSE to CAM are well taken; however, our analyses support the subsequent interpretation of Cox and Bairoch (1988) that CMSE is more similar in structure to the SARCs (and by our computations, the AEQs).

Discussion

Relationships among Subfamilies

The relationships among calcium-modulated proteins presented in Fig. 4 were calculated using only the EF-hand domains, and the relative positions of subfamilies and uniques that have unequal numbers of domains are determined by only those domains that overlap in our alignment (Table 1 and Appendix I). Thus, the placement of the two-domain proteins S100, TCBP10, and ACTN relative to other homologs depends only on domains 1 and 2 of the other proteins. Similarly, the relationship of PARV to other calcium-modulated proteins is based on only domains 2, 3, and 4 of those homologs. This situation may be ameliorated by analyses that compare individual domains and pairs of domains to one another (see Directions for Future Research). As discussed in the Materials and Methods, inclusion of sequences from interdomain regions is more likely to resolve ambiguities in relationships within subfamilies than among subfamilies and uniques.

The domains of all proteins are numbered 1–6 from N-terminus to C-terminus, with two exceptions. The three domains of PARV are numbered 2, 3, and 4, in accordance with earlier analyses (Baba et al. 1984; Epstein et al. 1986; Parmentier et al. 1987; Perret et al. 1988b) and our own analyses (results not shown). The unique LPS1 contains eight domains (Table 1). As discussed earlier, we treated it as two four-domain sequences, each numbered 1, 2, 3, and 4 (Appendix I). Each of these facts must be considered when interpreting global relationships among all EF-hand homologs (Fig. 4).

Summary of Previous Evolutionary Analyses of Calcium-Modulated Proteins

The series of publications that used maximum parsimony methods to examine evolution in the EF-hand family include reports by Goodman and Pechére (1977), Goodman et al. (1979), Goodman (1980), Baba et al. (1984), and Kretsinger et al. (1988). Each of these studies examined a progressively greater number of amino acid sequences. This allowed more conclusive statements to be made about relationships among and within subfamilies of EF-hand proteins.

Goodman et al. (1979) analyzed 25 amino acid sequences and reported that there are

> . . . six major present-day lineages [of calcium-modulated proteins], three of which—calcium dependent modulator protein [=CAM], heart and skeletal muscle troponin Cs, and alkali light chains of myosin [=ELC]—were found to share a closer kinship with one another than with the other lineages. Similarly, parvalbumins and regulatory light chains of myosin were depicted as more closely related, whereas the branch of intestinal calcium-binding protein proved to have the most distant separation. The . . . common ancestor of these six lineages [was] a four domain protein; . . . parvalbumins evolved by deletion of domain I [and] intestinal calcium-binding protein evolved by deletion of domains III and IV. . . . [Our] results suggested that tandem duplication in a precursor gene caused a primordial one-domain polypeptide . . . to double and then quadruple in size to a four-domain (I–II–III–IV) protein with domain I genetically closer to III and II to IV.

Baba et al. (1984) examined relationships among the 50 amino acid sequences of EF-hand proteins available at that time. They reasserted the findings of Goodman et al. (1979):

> . . . this phylogenetic reconstruction shows all 50 proteins as originating from a one-domain polypeptide about 36–40 amino acid residues long with the central 12-residue Ca^{2+}-binding site followed by two tandem duplications to become first a two-domain protein then a four-domain calmodulin-like protein in which each domain had the central Ca^{2+}-binding loop. . . . In the . . . ancestral sequence for this primal four-domain protein, domains I and III (the protein's first and third quadrants) appear to have descended from the N-terminal half of the earlier two-domain protein, whereas domains II and IV descended from the C-terminal half. Subsequent duplications in the basal eukaryotes and later prevertebrate metazoans may then have produced the ancestral loci for major protein branches of the calmodulin family.

Their depiction of relationships among subfamilies also reiterated the earlier work of Goodman et al. (1979). CAM, TNC, and ELC were found to share a closer kinship with one another than with the other lineages; PARV and RLC were depicted as closely related; and ICBP and S100 were placed at the root of the tree.

Parmentier et al. (1987) used the program Fast P (Lipman and Pearson 1985) to examine relationships among individual domains from representatives of six subfamilies. Their analyses indicated that

> . . . calmodulin, troponin C and myosin catalytic light chains share a common four-domained [precursor;] . . . the last two domains of parvalbumin are more homologous [to domains 3 and 4 of calmodulin; and] . . . it is obvious that the four-domained [precursor] derived from a single domain by two successive duplications. . . .

Therefore, Parmentier et al. concluded that ". . . all members of the troponin C superfamily derive from a common two-domained ancestor, but that duplications leading to calbindin and to the four-domained calcium-binding proteins took place independently on different branches of the evolutionary tree." According to Parmentier et al. ". . . the two-domained proteins more probably derived from the two-domained [precursor] of calmodulin," rather than being derived from the N-terminal half of a four-domain precursor as Goodman et al. (1979) and Baba et al. (1984) had proposed. Thus, Parmentier et al. suggested the following scenario. There was a single-domain precursor, which duplicated to form the two-domain precursor of all calcium-modulated proteins. This two-domain protein was the immediate precursor of the S100 subfamily and underwent a triplication to give rise to CLBN. A duplication of the two-domain protein produced the four-domain precursor of CAM, TNC, ELC, RLC, CALCIB, and (after deletion of one domain) PARV. Parmentier et al. did not examine representatives of CALP, SARC, AEQ, or most of the uniques included in our study, so they did not comment on the evolutionary origins of these homologs.

Perret et al. (1988b) compared and analyzed the structures of the recently described genes coding for CAM, SPEC, ELC, RLC, PARV, ICBP, and CALP. This analysis of gene structure revealed several highly conserved residues; additionally, the codon for one of these is interrupted by an intron. Perret et al. speculated

> that the four-site primordial ancestor gene produced by two successive duplications had introns that flanked each domain . . . and would have permitted duplication. . . . The distinct branches within this family of proteins can be interrelated if it is assumed that divergence took place by remodeling the structure of this ancestral gene and principally loss and insertion of introns.

They recognized five major evolutionary lineages of calcium-modulated proteins. The first contains CAM, SPEC, ELC, and PARV; the second consists solely of RLC, whose ". . . gene structure is very different from that of the [CAM, ELC, SPEC 1, and PARV] branches. . . ."

The third evolutionary lineage recognized by Perret et al. (1988b) consists of ICBP and other members of the S100 subfamily, which share a modified

first domain. According to Perret et al., S100s could be

> . . . derived directly from the two-site primordial ancestor PA2 or indirectly from the four-site ancestor [PA4] by loss of two sites . . . all sites II of these proteins are more similar to calmodulin site IV than to [CAM] site II. In the Kanehisa alignment analysis, a site (e.g., site I) is very often more similar to its true homologue (I) than to its duplicated homologue (III). Derivation from PA4 would involve a loss of sites I and II rather than III and IV of the ancestral gene. This would also explain the loss of the ancestral intron located just after the initiation codon, which is highly conserved in the other lines in this superfamily.

The fourth lineage recognized by Perret et al. (1988b) is CALP; they presented evidence that

> . . . the first three sites in the calpains correspond to the last three sites of [CAM]. . . . The present intron–exon structure of the calcium-binding domain in calpains could be explained by the loss of site I of the ancestral four-site primordial gene and addition of a fourth domain [by exon shuffling], together with a number of genomic rearrangements.

The fifth gene lineage is CLBN, whose gene structure has not been determined. They stated that

> . . . internal similarities among either sites I, III, and IV [sic] or among II, IV, and VI . . . are in favor of a triplication. Examination of the linker regions following sites II, IV, and VI shows a high degree of conservation that may also have resulted from triplication . . . [; CLBN] seems to be the protein that has undergone the greatest change from its evolutionary origin. Its genomic structure and the reconstruction of the genetic tree using the method of maximum parsimony will be of interest.

Perret et al. (1988b) concluded that

> . . . the evolutionary scenario allows explanation of the discrepancy between the intramolecular similarities observed in all of the members of this family of homologous proteins, the exon-shuffling model, and the actual structure of the genes encoding these proteins. The ancestral gene, produced by two successive duplications, would consist of at least four exons. The four exons coding for the four calcium-binding subdomains would be separated by three vestigial introns from the two successive duplications. The different lineages of this family would have evolved by remodeling the structure of the ancestral gene by different genomic rearrangements and, in particular, loss and insertion of introns.

The most comprehensive treatment of these proteins to date (Kretsinger et al. 1988) described our preliminary findings based on 129 amino acid sequences. In that report, we recognized 10 subfamilies: S100, PARV, CAM, TNC, RLC, ELC, AEQ, SARC, CALP, and CLBN. In addition, we identified six proteins as unique: SPEC, CDC31, TRACTIN, CALCIB, CVP, and CMSE. Analyses performed for this study included the 129 sequences from our preliminary study (Kretsinger et al. 1988) as well as 24 sequences that were not available at the time of our previous work.

Summary of Analyses Reported Herein

From results presented in this report, we conclude that there are at least 12 subfamilies of EF-hand proteins and eight unique homologs. The unique sequences might reflect an unusual function, a recent origin, a rapid evolution, or simply an inadequate sampling of tissues and organisms. For whatever reason, these eight sequences seem to be quite different from the 12 recognized subfamilies.

There are many very diverse calcium-modulated proteins distributed throughout the fungi, protocists, plants, and animals. Additionally, CMSE was obtained from a prokaryote. The question of whether *Streptomyces* acquired a eukaryotic gene or whether CMSE evolved from a precursor in the common ancestor of eubacteria and eukaryotes is important; the answer may lend insight into the question of whether or not prokaryotes use calcium as a cytosolic messenger.

Of the 12 subfamilies, only ELC, CAM, and ACTN have representatives outside Animalia. To date, S100, PARV, CALP, and CLBN have been found only in vertebrates; AEQ, SPEC, and SARC have been found only in nonvertebrate animals. The unique proteins come from a mammal (CALCIB), a crustacean (TPAP), a protochordate (CVP), two protocists (TRACTIN, TCBP10), a yeast (CDC31), an echinoderm (LPS1), and a prokaryote (CMSE).

Directions for Future Research

In a most parsimonious scheme, one previously suggested by several groups (Weeds and McLachlan 1974; Collins 1976a,b; Barker et al. 1977; Goodman and Pechère 1977; Watterson et al. 1980; Iida 1982) there would be a single EF-hand domain precursor in a species ancestral to the eukaryotes. Its encoding gene would duplicate, thereby generating a two-domain protein. A subsequent duplication would produce a four-domain protein that would be the precursor to all existing four-domain members of the homolog family and, by deletion of either one or two domains, to all three- and two-domain proteins, respectively. In this, and similar, evolutionary schemes, we would expect dendrograms constructed using only domains 1 and 2 to be, within statistical limits, congruent with those constructed using domains 3 and 4, as well as with those constructed using each domain independently. It is difficult to estimate the actual noise associated with the resulting reduced lengths of sequence under comparison in each of these instances. However, it is obvious that the (quasi) random nature of mutation events will have a greater effect as one reduces the length of sequence under examination from approximately 116 amino acids (representing four do-

mains) to only 29 residues (constituting a single domain) and in the extreme case, a quarter domain of 7 amino acids.

Such a comparison of individual domains and pairs of domains may resolve conflicting opinions about the origin of several subfamilies. For example, Perret et al. (1988b) concluded that S100 is descended from the C-terminal half of a four-domain precursor; according to Baba et al. (1984), S100 evolved from the N-terminal half of a four-domain precursor. The analyses of Parmentier et al. (1987) indicated that S100 is a direct descendent of the two-domain precursor from which all calcium-modulated proteins arose. Analyses involving S100 are complicated by the fact that most members of the S100 subfamily have an altered domain 1.

The origin of CLBN is also debatable; several scenarios are possible. Parmentier et al. (1987) and Perret et al. (1988b) concluded that triplication of the two-domain precursor resulted in the six-domain precursor of CLBN. Our analyses of pairs of domains (results not shown) appear to corroborate their view. Alternatively, CLBN could be the product of duplications or quadruplications followed by deletions.

A series of deletions and splicing events almost certainly occurred in the AEQ and SARC subfamilies. Similar rearrangements may have produced CALP, as Perret et al. (1988b) suggested.

An evaluation of pairs of domains and of single domains is the subject of the next publication in this series. Preliminary results from analyses we have already performed indicate overall similarity, but incomplete congruence within and among several subfamilies. This does not invalidate the results we present in this report; however, we caution that some counterintuitive relationships among entire molecules reported herein may reflect different evolutionary histories for individual domains, or regions, within a single molecule. The optimal classification and evolutionary interpretation of EF-hand homologs perhaps should be based on domains compared independently of one another as opposed to comparisons of intact molecules.

Conclusions

This is the most comprehensive, fully documented treatment of the EF-hand homologs to date. More than two-thirds of the sequences included in this study were published since the 1984 study of Baba et al., and we have added more than 20 sequences since our preliminary findings were published (Kretsinger et al. 1988). In addition to the subfamilies identified by Baba et al. (1984), this evolutionary analysis includes amino acid sequences of CLBN, SPEC, ACTN, CALP, AEQ, SARC, as well as CALCIB, TPAP, CVP, TRACTIN, CDC31, TCBP10, LPS1, and CMSE. Thus, our findings indicate that most calcium-modulated proteins can be ordered into at least 12 subfamilies and that at least eight proteins are unique.

There is still much to learn about these proteins, however. For example, relationships in Fig. 4 do not reflect a systematic gain or loss of calcium-binding ability in the EF-hand domains of these proteins. Enabling or disabling mutations surely occurred numerous times during evolution.

The diversity and distribution of known calcium-modulated proteins indicate a broad range of functions and ancient gene duplication events. However, using presently available data, we cannot confidently establish the root for the network depicting relationships among subfamilies of calcium-modulated proteins (Fig. 4). One technique that would allow us to determine the position of the root for this network involves including representatives of the closest relatives (the sister group) of these proteins. Unfortunately, calcium-modulated proteins are of such ancient origin that identifying their sister group of molecules is not possible. Indeed, simply determining the most divergent subfamily of EF-hand homologs, which would also allow us to infer the position of the root of relationships among subfamilies, may prove to be a formidable task. In contrast, we can infer the position of the root among sequences within each subfamily, because relationships within subfamilies are determined relative to all sequences outside that subfamily.

Because we cannot establish a root for inferring relationships among subfamilies within the superfamily as a whole, we cannot suggest the order and timing of genetic events that produced the known EF-hand homologs. However, we are relatively confident that the evolutionary history of this group of molecules is more complex than is suggested by the most parsimonious scenario. In this scenario, two tandem duplication events would have produced a four-domain protein that was the precursor to all members of this superfamily. Although we think that such a scenario does not adequately explain the observed variation in domain number and function, it would be premature at this time to offer other hypotheses. Before doing so, we must analyze further the relationships among individual domains and pairs of domains as well as the cDNA and gDNA data. Analyses such as these are necessary before we can favor a single scenario that accounts for evolution of all EF-hand homologs.

Acknowledgments. This study was supported by NASA grant NAGW-1233 to R.H.K. and NSF grant BSR 8607202 to M.G. J. Czelusniak provided the computer programs. We thank R.

Cordaro for his invaluable assistance in solving hardware and software problems.

References

Aitken A, Klee CB, Cohen P (1984) The structure of the B subunit of calcineurin. Eur J Biochem 139:663–671

Aoki K, Imajoh S, Ohno S, Emori Y, Koike M, Kosaki G, Suzuki K (1986) Complete amino acid sequence of the large subunit of the low-Ca^{2+}-requiring form of human Ca^{2+}-activated neutral protease (μCANP) deduced from its cDNA sequence. FEBS Lett 205:313–317

Arimura C, Suzuki T, Yanagisawa M, Imamura M, Hamada Y, Masaki T (1988) Primary structure of chicken skeletal muscle and fibroblast α-actinins deduced from cDNA sequences. Eur J Biochem 177:649–655

Baba ML, Goodman M, Berger-Cohn J, Demaille JG, Matsuda G (1984) The early adaptive evolution of calmodulin. Mol Biol Evol 1:442–455

Babu YS, Sack JS, Greenhough TJ, Bugg CE, Means AR, Cook WJ (1985) Three-dimensional structure of calmodulin. Nature 316:37–40

Babu YS, Bugg CE, Cook WJ (1988) Structure of calmodulin refined at 2.2 Å resolution. J Mol Biol 204:191–204

Barker WC, Ketcham LK, Dayhoff MO (1977) Evolutionary relationships among calcium-binding proteins. In: Wasserman RH, Corradino R, Carafoli E, Kretsinger RH, MacLennan D, Siegel F (eds) Calcium-binding proteins and calcium function. North-Holland, New York, pp 73–75

Baron MD, Davidson MD, Jones P, Critchley DR (1987) The sequence of chick α-actinin reveals homologies to spectrin and calmodulin. J Biol Chem 262:17623–17629

Barraclough R, Savin J, Dube SK, Rudland PS (1987) Molecular cloning and sequence of the gene for p9Ka. A cultured myoepithelial cell protein with strong homology to S-100, a calcium-binding protein. J Mol Biol 198:13–20

Barraclough R, Kimbell R, Rudland PS (1988) The identification of a normal rat gene located close to a gene for the potential myoepithelial cell calcium-binding protein, p9Ka. J Biol Chem 263:14597–14600

Barton PJR, Robert B, Cohen A, Garner I, Sassoon D, Weydert A, Buckingham ME (1988) Structure and sequence of the myosin alkali light chain gene expressed in adult cardiac atria and fetal striated muscle. J Biol Chem 263:12669–12676

Baudier J (1988) S100 proteins: structure and calcium binding properties. In: Gerday C, Gillis R, Bolis L (eds) Calcium and calcium binding proteins: molecular and functional aspects. Springer-Verlag, New York, pp 102–113

Baudier J, Gerard D (1986) Ions binding to S100 proteins. II. Conformational studies and calcium-induced conformational changes in S100$\alpha\alpha$ protein: the effect of acidic pH and calcium incubation on subunit exchange in S100a ($\alpha\beta$) proteins. J Biol Chem 261:8204–8212

Baudier J, Glasser N, Gerard D (1986) Ions binding to S100 proteins. I. Calcium- and zinc-binding properties of bovine brain S100($\alpha\alpha$), S100a($\alpha\beta$), and S100b($\beta\beta$) protein: Zn^{2+} regulates Ca^{2+} binding on S100b protein. J Biol Chem 261:8192–8203

Baum P, Furlong C, Byers B (1986) Yeast gene required for spindle pole body duplication: homology of its product with Ca^{2+}-binding proteins. Proc Natl Acad Sci USA 83:5512–5516

Bender PK, Dedman JR, Emerson CP Jr (1988) The abundance of calmodulin mRNAs is regulated in phosphorylase kinase-deficient skeletal muscle. J Biol Chem 263:9733–9737

Berchtold MW (1988) Structural organization of the human parvalbumin gene. In: Hidaka H (ed) Ca^{2+} protein signaling. Plenum Press, New York (in press)

Berchtold MW, Heizmann CW, Wilson KJ (1982) Primary structure of parvalbumin from rat skeletal muscle. Eur J Biochem 127:381–389

Berchtold MW, Epstein P, Beaudet AL, Payne ME, Heizmann CW, Means AR (1987) Structural organization and chromosomal assignment of the parvalbumin gene. J Biol Chem 262:8696–8701

Bruggen J, Tarcsay L, Cerletti N, Odink K, Rutishauser M, Hollander G, Sorg C (1988) The molecular nature of the cystic fibrosis antigen. Nature 331:570

Calabretta B, Battini R, Kaczmarek L, de Riel JK, Baserga R (1986) Molecular cloning of the cDNA for a growth factor-inducible gene with strong homology to S-100, a calcium-binding protein. J Biol Chem 261:12628–12632

Capony J-P, Ryden L, Demaille J, Pechére J-F (1973) The primary structure of the major parvalbumin from hake muscle. Overlapping peptides obtained with chemical and enzymatic methods. The complete amino acid sequence. Eur J Biochem 32:97–108

Capony J-P, DeMaille J, Pina C, Pechére J-F (1975) The amino-acid sequence of the most acidic major parvalbumin from frog muscle. Eur J Biochem 56:215–227

Capony J-P, Pina C, Pechére J-F (1976) Parvalbumin from rabbit muscle: isolation and primary structure. Eur J Biochem 70:123–135

Carpenter CD, Bruskin AM, Hardin PE, Keast MJ, Anstrom J, Tyner AL, Brandhorst BP, Klein WH (1984) Novel proteins belonging to the troponin C superfamily are encoded by a set of mRNAs in sea urchin embryos. Cell 36:663–671

Charbonneau H, Walsh KA, McCann RO, Prendergast FG, Cormier MJ, Vanaman TC (1985) Amino acid sequence of the calcium-dependent photoprotein aequorin. Biochemistry 24: 6762–6771

Chen Q, Taljanidisz J, Sarkar S, Tao T, Gergely J (1988) Cloning, sequencing and expression of a full-length rabbit fast skeletal troponin-C cDNA. FEBS Lett 228:22–26

Chien Y-H, Dawid IB (1984) Isolation and characterization of calmodulin genes from *Xenopus laevis*. Mol Cell Biol 4:507–513

Coffee CJ, Bradshaw RA (1973a) Carp muscle calcium-binding protein. I. Characterization of the tryptic peptides and the complete amino acid sequence of component B. J Biol Chem 248:3305–3312

Coffee CJ, Bradshaw RA (1973b) Erratum. J Biol Chem 248: 6576

Coffee CJ, Bradshaw RA, Kretsinger RH (1974) The coordination of calcium ions by carp muscle calcium binding proteins A, B and C. Adv Exp Med Biol 48:211–233

Collins, JH (1976a) Structure and evolution of troponin C and related proteins. In: Calcium in biological systems, Soc Exp Biol Symp XXX, Cambridge University Press, London, pp 303–334

Collins JH (1976b) Homology of myosin DTNB light chain with alkali light chains, troponin C and parvalbumin. Nature 259:699–700

Collins JH, Greaser ML, Potter JD, Horn MJ (1977) Determination of the amino acid sequence of troponin C from rabbit skeletal muscle. J Biol Chem 252:6356–6362

Collins JH, Jakes R, Kendrick-Jones J, Leszyk J, Barouch W, Theibert JL, Spiegel J, Szent-Györgyi AG (1986) Amino acid sequence of myosin essential light chain from the scallop *Aquipecten irradians*. Biochemistry 25:7651–7656

Collins JH, Cox JA, Theibert JL (1988) Amino acid sequence of a sarcoplasmic calcium-binding protein from the sandworm *Nereis diversicolor*. J Biol Chem 263:15378–15381

Cormier MJ (1978) Applications of *Renilla* bioluminescence. Methods Enzymol 57:237–244

Cox JA (1986) Isolation and characterization of a new M_r 18,000 protein with calcium vector properties in amphioxus muscle and identification of its endogenous target protein. J Biol Chem 261:13173–13178

Cox JA (1990) Calcium vector protein and sarcoplasmic calcium binding proteins from invertebrate muscle. In: Dedman JR, Smith VL (eds) Stimulus–response coupling: the role of intracellular calcium. Telford Press, West Caldwell NJ (in press)

Cox JA, Bairoch A (1988) Sequence similarities in calcium-binding proteins. Nature 331:491–492

Davis TN, Urdea MS, Masiarz FR, Thorner J (1986) Isolation of the yeast calmodulin gene: calmodulin is an essential protein. Cell 47:423–431

Dedman JR, Jackson RL, Schreiber WE, Means AR (1978) Sequence homology of the Ca^{2+}-dependent regulator of cyclic nucleotide phosphodiesterase from rat testis with other Ca^{2+}-binding proteins. J Biol Chem 253:343–346

Desplan C, Heidmann O, Lillie JW, Auffray C, Thomasset M (1983a) Sequence of rat intestinal vitamin D-dependent calcium-binding protein derived from a cDNA clone: evolutionary implications. J Biol Chem 258:13502–13505

Desplan C, Thomasset M, Moukhtar M (1983b) Synthesis, molecular cloning, and restriction analysis of DNA complementary to vitamin D-dependent calcium-binding protein mRNA from rat duodenum. J Biol Chem 258:2762–2765

Donato R (1986) S-100 proteins. Cell Calcium 7:123–145

Donato R (1988) Calcium-independent, pH-regulated effects of S-100 proteins on assembly–disassembly of brain microtubule protein in vitro. J Biol Chem 263:106–110

Dorin JR, Novak M, Hill RE, Brock DJH, Secher DS, van Heyningen V (1987) A clue to the basic defect in cystic fibrosis from cloning the CF antigen gene. Nature 326:614–617

Elsayed S, Bennich H (1975) The primary structure of allergen M from cod. Scan J Immunol 4:203–208

Emori Y, Kawasaki H, Imajoh S, Kawashima S, Suzuki K (1986a) Isolation and sequence analysis of cDNA clones for the small subunit of rabbit calcium-dependent protease. J Biol Chem 261:9472–9476

Emori Y, Kawasaki H, Sugihara H, Imajoh S, Kawashima S, Suzuki K (1986b) Isolation and sequence analyses of cDNA clones for the large subunits of two isozymes of rabbit calcium-dependent protease. J Biol Chem 261:9465–9471

Enfield DL, Ericsson LH, Blum HE, Fischer EH, Neurath H (1975) Amino-acid sequence of parvalbumin from rabbit skeletal muscle. Proc Natl Acad Sci USA 72:1309–1313

Epstein P, Means AR, Berchtold MW (1986) Isolation of the rat parvalbumin gene and full length cDNA. J Biol Chem 261: 5886–5891

Falkenthal S, Parker VP, Mattox WW, Davidson N (1984) *Drosophila melanogaster* has only one myosin alkali light-chain gene which encodes a protein with considerable amino acid sequence homology to chicken myosin alkali light chains. Mol Cell Biol 4:956–965

Falkenthal S, Parker VP, Davidson N (1985) Developmental variations in the splicing patterns of transcripts from the *Drosophila* gene encoding myosin alkali light chain result in different carboxyl-terminal amino acid sequences. Proc Natl Acad Sci USA 82:449–453

Fano G, Angelella P, Mariggio D, Aisa MC, Giambanco I, Donato R (1989) S-100α_0 protein stimulates the basal (Mg^{2+}-activated) adenylate cyclase activity associated with skeletal muscle membranes. FEBS Lett (in press)

Farris JS (1972) Estimating phylogenetic trees from distance matrices. Am Nat 106:645–668

Feher JJ, Fullmer CS, Fritzsch GK (1989) Comparison of the enhanced steady-state diffusion of calcium by calbindin-D 9k and calmodulin: possible importance in intestinal calcium absorption. Cell Calcium 10:189–203

Felsenstein J (1988) Phylogenies from molecular sequences: inference and reliability. Annu Rev Genet 22:521–565

Ferrari S, Calabretta B, de Riel JK, Battini R, Ghezzo F, Lauret E, Griffin C, Emanuel BS, Gurrieri F, Baserga R (1987) Structural and functional analysis of a growth-related gene, the human calcyclin. J Biol Chem 262:8325–8332

Fischer R, Koller M, Flura M, Mathews S, Strehler-Page M-A, Krebs J, Penniston JT, Carafoli E, Strehler EE (1988) Multiple divergent mRNAs code for a single human calmodulin. J Biol Chem 263:17055–17062

Fitch WM (1971) Toward defining the course of evolution: minimum change for a specific tree topology. Syst Zool 20: 406–416

Fitch WM, Margoliash E (1967) The construction of phylogenetic trees—a generally applicable method utilizing estimates of the mutation distance obtained from cytochrome C sequences. Science 155:279–284

Floyd EE, Gong Z, Brandhorst BP, Klein WH (1986) Calmodulin gene expression during sea urchin development: persistence of a prevalent maternal protein. Dev Biol 113:501–511

Frank G, Weeds AG (1974) The amino-acid sequence of the alkali light chains of rabbit skeletal-muscle myosin. Eur J Biochem 44:317–334

Frankenne F, Joassin L, Gerday C (1973) The amino acid sequence of the pike (*Esox lucius*) parvalbumin III. FEBS Lett 35:145–147

Fullmer CS, Wasserman RH (1981) The amino acid sequence of bovine intestinal calcium-binding protein. J Biol Chem 256:5669–5674

Fullmer CS, Wasserman RH (1987) Chicken intestinal 28-kilodalton calbindin-D: complete amino acid sequence and structural considerations. Proc Natl Acad Sci USA 84:4772–4776

Gahlmann R, Wade R, Gunning P, Kedes L (1988) Differential expression of slow and fast skeletal muscle troponin C. Slow skeletal muscle troponin C is expressed in human fibroblasts. J Mol Biol 201:379–391

Garfinkel LI, Periasamy M, Nadal-Ginard B (1982) Cloning and characterization of cDNA sequences corresponding to myosin light chains 1, 2, and 3, troponin-C, troponin-T, α-tropomyosin, and α-actin. J Biol Chem 257:11078–11086

Gerday C (1976) The primary structure of the parvalbumin II of pike (*Esox lucius*). Eur J Biochem 70:305–318

Gerday C (1988) Soluble calcium binding proteins in vertebrate and invertebrate muscles. In: Gerday C, Bolis L, Gilles R (eds) Calcium and calcium binding proteins: molecular and functional aspects. Springer-Verlag, Berlin, pp 23–39

Gerday C, Collin S, Pironi A (1978) Phylogenetic relationships between Cyprinidae parvalbumins. II. The amino acid sequence of the parvalbumin V of chub (*Leuciscus cephalus* L.). Comp Biochem Physiol 61B:451–457

Gerke V, Weber K (1985) The regulatory chain in the p36-kd substrate complex of viral tyrosine-specific protein kinases is related in sequence to the S-100 protein of glial cells. EMBO J 4:2917–2920

Gillen MF, Banville D, Rutledge RG, Narang S, Seligy VL, Whitfield JF, MacManus JP (1987) A complete complementary DNA for the oncodevelopmental calcium-binding protein, oncomodulin. J Biol Chem 262:5308–5312

Gillis J-M, Thomason DB, Le Fevre J, Kretsinger RH (1982) Parvalbumin and muscle relaxation: a computer simulation study. J Muscle Res Cell Motil 3:377–398

Glenney JR Jr, Tack BF (1985) Amino-terminal sequence of p36 and associated p10: identification of the site of tyrosine phosphorylation and homology with S-100. Proc Natl Acad Sci USA 82:7884–7888

Glenney JR Jr, Kindy MS, Zokas L (1989) Isolation of a new

member of the S100 protein family: amino acid sequence, tissue, and subcellular distribution. J Cell Biol 108:569–578

Goodman M (1980) Molecular evolution of the calmodulin family. In: Siegel FL, Carafoli E, Kretsinger RH, MacLennan DH, Wasserman RH (eds) Calcium-binding proteins: structure and function. Elsevier North-Holland, New York, pp 347–354

Goodman M, Pechére J-F (1977) The evolution of muscular parvalbumins investigated by the maximum parsimony method. J Mol Evol 9:131–158

Goodman M, Pechére J-F, Haiech J, DeMaille JG (1979) Evolutionary diversification of structure and function in the family of intracellular calcium-binding proteins. J Mol Evol 13: 331–352

Goodwin EB, Szent-Györgyi AG, Leinwand LA (1987) Cloning and characterization of the scallop essential and regulatory myosin light chain cDNAs. J Biol Chem 262:11052–11056

Grand RJA, Perry SV (1978) Amino acid sequence of the troponin C-like protein (modulator protein) from bovine uterus. FEBS Lett 92:137–142

Grand RJA, Shenolikar S, Cohen P (1981) The amino acid sequence of the δ subunit (calmodulin) of rabbit skeletal muscle phosphorylase kinase. Eur J Biochem 113:359–367

Hagiwara M, Achiai M, Owada K, Tanaka T, Hidaka H (1988) Modulation of tyrosine phosphorylation of p36 and other substrates by the S-100 protein. J Biol Chem 263:6438–6441

Hardin PE, Klein WH (1987) Unusual sequence conservation in the 5′ and 3′ untranslated regions of the sea urchin Spec mRNAs. J Mol Evol 25:126–133

Hardin PE, Angerer LM, Hardin SH, Angerer RC, Klein WH (1988) Spec 2 genes in *Strongylocentrotus purpuratus*. Structure and differential expression in embryonic aboral ectoderm cells. J Mol Biol 202:417–431

Hardin SH, Carpenter CD, Hardin PE, Bruskin AM, Klein WH (1985) Structure of the Spec1 gene encoding a major calcium-binding protein in the embryonic ectoderm of the sea urchin, *Strongylocentrotus purpuratus*. J Mol Biol 186:243–255

Hardin SH, Keast MJ, Hardin PE, Klein WH (1987) Use of consensus oligonucleotides for detecting and isolating nucleic acids encoding calicum binding domains of the troponin C superfamily. Biochemistry 26:3518–3523

Hardy DO, Bender PK, Kretsinger RH (1987) Two calmodulin genes are expressed in *Arbacia punctulata*. An ancient gene duplication is indicated. J Mol Biol 198:223–227

Hartigan JA (1973) Minimum mutation fits to a given tree. Biometrics 29:53–65

Head JF (1989) Amino acid sequence of a low molecular weight, high affinity calcium-binding protein from the optic lobe of the squid *Loligo pealei*. J Biol Chem 264:7202–7209

Henderson SA, Xu Y-C, Chien KR (1988) Nucleotide sequence of full length cDNAs encoding rat cardiac myosin light chain-2. Nucleic Acids Res 16:4722

Henrotte JG (1952) A crystalline constitutent from myogen of carp muscle. Nature 169:968–969

Herzberg O, James MNG (1985) Structure of the calcium regulatory protein troponin-C at 2.8 Å resolution. Nature 313: 653–659

Herzberg O, James MNG (1988) Refined crystal structure of troponin C from turkey skeletal muscle at 2.0 Å resolution. J Mol Biol 203:761–779

Hidaka H, Endo T, Kawamoto S, Yamada E, Umekawa H, Tanabe K, Hara K (1983) Purification and characterization of adipose tissue S-100b protein. J Biol Chem 258:2705–2709

Hoffmann E, Shi QW, Floroff M, Mickle DAG, Wu T-W, Olley PM, Jackowski G (1988) Molecular cloning and complete nucleotide sequence of a human ventricular myosin light chain 1. Nucleic Acids Res 16:2353

Hofmann T, Kawakami M, Hitchman AJW, Harrison JE, Dorrington KJ (1979) The amino acid sequence of porcine intestinal calcium-binding protein. Can J Biochem 57:737–748

Hosoya H, Takagi T, Mabuchi I, Iwaasa H, Sakai H, Hiramoto Y, Konishi K (1988) The amino acid sequence, immunofluorescence and microinjection studies on the 15 kDa calcium-binding protein from sea urchin egg. Cell Struct Funct 13:525–532

Huang B, Mengersen A, Lee VD (1988a) Molecular cloning of cDNA for caltractin, a basal body-associated Ca^{2+}-binding protein: homology in its protein sequence with calmodulin and the yeast CDC31 gene product. J Cell Biol 107:133–140

Huang B, Watterson DM, Lee VD, Schibler MJ (1988b) Purification and characterization of a basal body-associated Ca^{2+}-binding protein. J Cell Biol 107:121–131

Huang W-Y, Cohn DV, Hamilton JW (1975) Calcium-binding protein of bovine intestine. The complete amino acid sequence. J Biol Chem 250:7647–7655

Hunziker W (1986) The 28-kDa vitamin D-dependent calcium-binding protein has a six-domain structure. Proc Natl Acad Sci USA 83:7578–7582

Iida Y (1982) Molecular evolution of protein: internal homology in the amino acid sequence of calmodulin. J Mol Biol 159:167–177

Inouye S, Noguchi M, Sakaki Y, Takagi Y, Miyata T, Iwanaga S, Miyata T, Tsuji FI (1985) Cloning and sequence analysis of cDNA for the luminescent protein aequorin. Proc Natl Acad Sci USA 82:3154–3158

Isobe T, Okuyama T (1978) The amino-acid sequence of S-100 protein (PAP I-b protein) and its relation to the calcium-binding proteins. Eur J Biochem 89:379–388

Isobe T, Okuyama T (1981) The amino-acid sequence of the α subunit in bovine brain S-100a protein. Eur J Biochem 116: 79–86

Jackson-Grusby LL, Swiergiel J, Linzer DIH (1987) A growth-related mRNA in cultured mouse cells encodes a placental calcium-binding protein. Nucleic Acids Res 15:6677–6690

Jamieson GA Jr, Bronson DD, Schachat FH, Vanaman TC (1980) Structure and function relationships among calmodulins and troponin C-like proteins from divergent eukaryotic organisms. Ann NY Acad Sci 356:1–13

Jauregui-Adell J, Pechére J-F (1978) Parvalbumins from coelacanth muscle. III. Amino acid sequence of the major component. Biochim Biophys Acta 536:275–282

Jauregui-Adell J, Pechére J-F, Briand G, Richet C, DeMaille JG (1982) Amino-acid sequence of an α-parvalbumin, pI = 4.88, from frog skeletal muscle. Eur J Biochem 123:337–345

Jauregui-Adell J, Wnuk W, Cox JA (1989) Complete amino acid sequence of the sarcoplasmic calcium-binding protein (SCP-I) from crayfish (*Astacus leptodactilus* [sic]). FEBS Lett 243:209–212

Jensen R, Marshak DR, Anderson C, Lukas TJ, Watterson DM (1985) Characterization of human brain S100 protein fraction: amino acid sequence of S100β. J Neurochem 45:700–705

Joassin L, Gerday C (1977) The amino acid sequence of the major parvalbumin of the whiting (*Gadus merlangus*). Comp Biochem Physiol 57B:159–161

Jukes TH (1963) Some recent advances in studies of the transcription of the genetic message. Adv Biol Med Phys 9:1–41

Kasai H, Kato Y, Isobe T, Kawasaki H, Okuyama T (1980) Determination of the complete amino acid sequence of calmodulin (phenylalanine-rich acidic protein II) from bovine brain. Biomed Res 1:248–264

Kato K, Kimura S (1985) $S100a_0$ ($\alpha\alpha$) protein is mainly located in the heart and striated muscles. Biochim Biophys Acta 842: 146–150

Kawashima M, Nabeshima Y, Obinata T, Fujii-Kuriyama Y (1987) A common myosin light chain is expressed in chicken embryonic, skeletal, cardiac, and smooth muscles and in brain

continuously from embryo to adult. J Biol Chem 262:14408–14414

Kay BK, Shah AJ, Halstead WE (1987) Expression of the Ca^{2+}-binding protein, parvalbumin, during embryonic development of the frog, *Xenopus laevis*. J Cell Biol 104:841–847

Kendrick-Jones J, Jakes R (1977) Myosin-linked regulation: a chemical approach. In: Riecker G, Weber A, Goodwin J (eds) Myocardial failure. International Boehringer Mannheim Symposium. Springer-Verlag, Berlin, pp 28–40

Klee CB (1988) Ca^{2+}-dependent phospholipid- (and membrane-) binding proteins. Biochemistry 27:6645–6653

Kligman D, Marshak DR (1985) Purification and characterization of a neurite extension factor from bovine brain. Proc Natl Acad Sci USA 82:7136–7139

Kobayashi T, Takasaki Y, Takagi T, Konishi K (1984) The amino acid sequence of sarcoplasmic calcium-binding protein obtained from sandworm, *Perinereis vancaurica tetradentata*. Eur J Biochem 144:401–408

Kobayashi T, Takagi T, Konishi K, Cox JA (1987) The primary structure of a new Mr 18,000 calcium vector protein from amphioxus. J Biol Chem 262:2613–2623

Kobayashi T, Takagi T, Konishi K, Hamada Y, Kawaguchi M, Kohama K (1988a) Amino acid sequence of the calcium-binding light chain of myosin from the lower eukaryote, *Physarum polycephalum*. J Biol Chem 263:305–313

Kobayashi T, Takagi T, Konishi K, Ohnishi K, Watanabe Y (1988b) Amino acid sequence of a calcium-binding protein (TCBP-10) from *Tetrahymena*. Eur J Biochem 174:579–584

Koller M, Strehler EE (1988) Characterization of an intronless human calmodulin-like pseudogene. FEBS Lett 239:121–129

Kretsinger RH (1972) Gene triplication deduced from the tertiary structure of a muscle calcium binding protein. Nature New Biol 240:85–88

Kretsinger RH (1975) Hypothesis: calcium modulated proteins contain EF-hands. In: Carafoli E (ed) Calcium transport in contraction and secretion. North-Holland Publishing, Amsterdam, p 469

Kretsinger RH (1987) Calcium coordination and the calmodulin fold: divergent versus convergent evolution. Cold Spring Harbor Symp Quant Biol 52:499–510

Kretsinger RH, Barry CD (1975) The predicted structure of the calcium binding component of troponin. Biochim Biophys Acta 405:40–52

Kretsinger RH, Hardy DO (1987) Duplication of calmodulin gene inferred to occur prior to echinoderm, chordate divergence. In: Yagi K, Miyazak T (eds) Calcium signal and cell response. Springer-Verlag, Berlin, pp 157–163

Kretsinger RH, Nockolds CE, Coffee CJ, Bradshaw RA (1971) The structure of a calcium binding protein from carp muscle. Cold Spring Harbor Symp Quant Biol 36:217–220

Kretsinger RH, Mann JE, Simmonds JG (1982) Model of facilitated diffusion of calcium by the intestinal calcium binding protein. In: Normal AW, Schaefer K, Herrath DV, Grigoleit H-G (eds) Proceedings of the fifth workshop on vitamin D: chemical, biochemical and clinical endocrinology of calcium metabolism. de Gruyter, New York, pp 233–248

Kretsinger RH, Rudnick SE, Weissman LJ (1986) Crystal structure of calmodulin. J Inorg Biochem 28:289–302

Kretsinger RH, Moncrief ND, Goodman M, Czelusniak J (1988) Homology of calcium-modulated proteins: their evolutionary and functional relationships. In: Morad M, Nayler W, Kazda S, Schramm M (eds) The calcium channel: structure, function and implications. Springer-Verlag, New York, pp 16–35

Kumar CC, Cribbs L, Delaney P, Chien KR, Siddiqui MAQ (1986) Heart myosin light chain 2 gene. Nucleotide sequence of full length cDNA and expression in normal and hypertensive rat. J Biol Chem 261:2866–2872

Kumar R, Wieben E, Beecher SJ (1989) The molecular cloning of the complementary deoxyribonucleic acid for bovine vitamin D-dependent calcium-binding protein: structure of the full-length protein and evidence for homologies with other calcium-binding proteins of the troponin-C superfamily of proteins. Mol Endocrinol 3:427–432

Kuwano R, Usui H, Maeda T, Fukui T, Yamanari N, Ohtsuka E, Ikehara M, Takahashi Y (1984) Molecular cloning and the complete nucleotide sequence of cDNA to mRNA for S-100 protein of rat brain. Nucleic Acids Res 12:7455–7465

Lagace L, Chandra T, Woo SLC, Means AR (1983) Identification of multiple species of calmodulin messenger RNA using a full length complementary DNA. J Biol Chem 258:1684–1688

Lefort A, Lecocq R, Libert F, Lamy F, Swillens S, Vassart G, Dumont JE (1989) Cloning and sequencing of a calcium-binding protein regulated by cyclic AMP in the thyroid. EMBO J 8:111–116

Lenz S, Lohse P, Seidel U, Arnold H-H (1989) The alkali light chains of human smooth and nonmuscle myosins are encoded by a single gene. Tissue specific expression by alternative splicing pathways. J Biol Chem 264:9009–9015

Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227:1435–1441

Lorkin PA, Lehmann H (1983) Malignant hyperthermia in pigs: a search for abnormalities in Ca^{2+} binding proteins. FEBS Lett 153:81–87

Lukas TJ, Iverson DB, Schleicher M, Watterson DM (1984) Structural characterization of a higher plant calmodulin *Spinacia oleracea*. Plant Physiol 75:788–795

Lukas TJ, Wiggins ME, Watterson DM (1985) Amino acid sequence of a novel calmodulin from the unicellular alga *Chlamydomonas*. Plant Physiol 78:477–483

MacManus JP, Watson DC, Yaguchi M (1983) The complete amino acid sequence of oncomodulin—a parvalbumin-like calcium-binding protein from Morris hepatoma 5123tc. Eur J Biochem 136:9–17

MacManus JP, Watson DC, Yaguchi M (1986) The purification and complete amino acid sequence of the 9000-Mr Ca^{2+}-binding protein from rat placenta. Identity with the vitamin D-dependent intestinal Ca^{2+}-binding protein. Biochem J 235: 585–595

Maeda N, Zhu D, Fitch WM (1984) Amino acid sequences of lower vertebrate parvalbumins and their evolution: parvalbumins of boa, turtle, and salamander. Mol Biol Evol 1:473–488

Maisonpierre PC, Hastings KEM, Emerson CP Jr (1987) The cloning and the codon and amino acid sequence of the quail slow/cardiac troponin C cDNA. Methods Enzymol 139:326–337

Maita T, Umegane T, Kato Y, Matsuda G (1980) Amino-acid sequence of the L-1 light chain of chicken cardiac-muscle myosin. Eur J Biochem 107:565–575

Maita T, Chen J-I, Matsuda G (1981a) Amino-acid sequence of the 20000-molecular-weight light chain of chicken gizzard-muscle myosin. Eur J Biochem 117:417–424

Maita T, Umegane T, Matsuda G (1981b) Amino-acid sequence of the L-4 light chain of chicken skeletal-muscle myosin. Eur J Biochem 114:45–49

Maita T, Konno K, Ojima T, Matsuda G (1984) Amino acid sequences of the regulatory light chains of striated adductor muscle myosins from Ezo giant scallop and Akazara scallop. J Biochem 95:167–177

Maita T, Konno K, Maruta S, Norisue H, Matsuda G (1987a) Amino acid sequence of the essential light chain of adductor muscle myosin from Ezo giant scallop, *Patinopecten yessoensis*. J Biochem 102:1141–1149

Maita T, Tanaka H, Konno K, Matsuda G (1987b) Amino acid sequence of the regulatory light chain of squid mantle muscle myosin. J Biochem 102:1151–1157

Mangelsdorf DJ, Komm BS, McDonnell DP, Pike JW, Haussler

MR (1987) Immunoselection of cDNAs to avian intestinal calcium binding protein 28K and a novel calmodulin-like protein: assessment of mRNA regulation by the vitamin D hormone. Biochemistry 26:8332–8338

Marshak DR, Clarke M, Roberts DM, Watterson DM (1984) Structural and functional properties of calmodulin from the eukaryotic microorganism *Dictyostelium discoideum*. Biochemistry 23:2891–2899

Masiakowski P, Shooter EM (1988) Nerve growth factor induces the genes for two proteins related to a family of calcium-binding proteins in PC12 cells. Proc Natl Acad Sci USA 85: 1277–1281

Matsuda G, Maita T, Suzuyama Y, Setoguchi M, Umegane T (1977a) Amino acid sequence of the L-2 light chain of rabbit skeletal muscle myosin. J Biochem 81:809–811

Matsuda G, Suzuyama Y, Maita T, Umegane T (1977b) The L-2 light chain of chicken skeletal muscle myosin. FEBS Lett 84:53–56

Matsuda G, Maita T, Suzuyama Y, Setoguchi M, Umegane T (1978) The amino acid sequences of the tryptic, chymotryptic, and peptic peptides from the L-2 light chain of rabbit skeletal muscle myosin. Hoppe-Seyler's Z Physiol Chem 359: 629–640

Matsuda G, Maita T, Kato Y, Chen J-I, Umegane T (1981a) Amino acid sequences of the cardiac L-2A, L-2B, and gizzard 17000-M_r light chains of chicken muscle myosin. FEBS Lett 135:232–236

Matsuda G, Maita T, Umegane T (1981b) The primary structure of L-1 light chain of chicken fast skeletal muscle myosin and its genetic implication. FEBS Lett 126:111–113

Miyake S, Emori Y, Suzuki K (1986) Gene organization of human calcium-activated neutral protease. Nucleic Acids Res 14:8805–8817

Miyanishi T, Maita T, Morita F, Kondo S, Matsuda G (1985) Amino acid sequences of the two kinds of regulatory light chains of adductor smooth muscle myosin from *Patinopecten yessoensis*. J Biochem 97:541–551

Moews PG, Kretsinger RH (1975) Refinement of the structure of carp muscle calcium-binding parvalbumin by model building and difference Fourier analysis. J Mol Biol 91:201–228

Moore GW (1976) Proof for the maximum parsimony ("Red King") algorithm. In: Goodman M, Tashian RE (eds) Molecular anthropology. Plenum Press, New York, p 117

Moore GW, Barnabas J, Goodman M (1973) A method for constructing maximum parsimony ancestral amino acid sequences on a given network. J Theor Biol 38:459–485

Murphy LC, Murphy LJ, Tsuyuki D, Duckworth ML, Shiu RPC (1988) Cloning and characterization of a cDNA encoding a highly conserved, putative calcium binding protein, identified by an anti-prolactin receptor antiserum. J Biol Chem 263: 2397–2401

Mutus B, Karuppiah N, Sharma RK, MacManus JP (1985) The differential stimulation of brain and heart cyclic-AMP phosphodiesterase by oncomodulin. Biochem Biophys Res Commun 131:500–506

Nabeshima Y, Fujii-Kuriyama Y, Muramatsu M, Ogata K (1982) Molecular cloning and nucleotide sequences of the complementary DNAs to chicken skeletal muscle myosin alkali light chain mRNAs. Nucleic Acids Res 10:6099–6110

Nabeshima Y, Fujii-Kuriyama Y, Muramatsu M, Ogata K (1984) Alternative transcription and two modes of splicing result in two myosin light chains from one gene. Nature 308: 333–338

Nabeshima Y, Nabeshima Y-I, Nonomura Y, Fujii-Kuriyama Y (1987) Nonmuscle and smooth muscle myosin light chain mRNAs are generated from a single gene by the tissue-specific alternative RNA splicing. J Biol Chem 262:10608–10612

Nakamura S, Nabeshima Y-I, Kobayashi H, Nabeshima Y, Nonomura Y, Fujii-Kuriyama Y (1988) Single chicken cardiac myosin alkali light-chain gene generates two different mRNAs by alternative splicing of a complex exon. J Mol Biol 203: 895–904

Noegel A, Witke W, Schleicher M (1987) Calcium-sensitive non-muscle α-actinin contains EF-hand structures and highly conserved regions. FEBS Lett 221:391–396

Nojima H (1989) Structural organization of multiple rat calmodulin genes. J Mol Biol 208:269–282

Nojima H, Sokabe H (1986) Structure of rat calmodulin processed genes with implications for a mRNA-mediated process of insertion. J Mol Biol 190:391–400

Nojima H, Kishi K, Sokabe H (1987) Multiple calmodulin mRNA species are derived from two distinct genes. Mol Cell Biol 7:1873–1880

Nudel U, Calvo JM, Shani M, Levy Z (1984) The nucleotide sequence of a rat myosin light chain 2 gene. Nucleic Acids Res 12:7175–7186

Odink K, Cerletti N, Bruggen J, Clerc RG, Tarcsay L, Zwadlo G, Gerhards G, Schlegel R, Sorg C (1987) Two calcium-binding proteins in infiltrate macrophages of rheumatoid arthritis. Nature 330:80–82

Ohno S, Emori Y, Imajoh S, Kawasaki H, Kisaragi M, Suzuki K (1984) Evolutionary origin of a calcium-dependent protease by fusion of genes for a thiol protease and a calcium-binding protein? Nature 312:566–570

Ohno S, Emori Y, Suzuki K (1986) Nucleotide sequence of a cDNA coding for the small subunit of human calcium-dependent protease. Nucleic Acids Res 14:5559

Parker VP, Falkenthal S, Davidson N (1985) Characterization of the myosin light-chain-2 gene of *Drosophila melanogaster*. Mol Cell Biol 5:3058–3068

Parmacek MS, Leiden JM (1989) Structure and expression of the murine slow/cardiac troponin C gene. J Biol Chem 264: 13217–13225

Parmentier M, Lawson DEM, Vassart G (1987) Human 27-kDa calbindin complementary DNA sequence. Evolutionary and functional implications. Eur J Biochem 170:207–215

Pearson RB, Jakes R, John M, Kendrick-Jones J, Kemp BE (1986) Phosphorylation site sequence of smooth muscle myosin light chain (M_r = 20000). FEBS Lett 168:108–112

Pechére J-F, Capony JP, Ryden L, DeMaille J (1971) The amino acid sequence of the major parvalbumin from hake muscle. Biochem Biophys Res Commun 43:1106–1111

Pechére J-F, Capony J-P, DeMaille J (1973) Evolutionary aspects of the structure of muscular parvalbumins. Syst Zool 22:533–548

Pechére J-F, Rochat H, Ferraz C (1978) Parvalbumins from coelacanth muscle. II. Amino acid sequence of the two less acidic components. Biochim Biophys Acta 536:269–274

Periasamy M, Strehler EE, Garfinkel LI, Gubits RM, Ruiz-Opazo N, Nadal-Ginard B (1984) Fast skeletal muscle myosin light chains 1 and 3 are produced from a single gene by a combined process of differential RNA transcription and splicing. J Biol Chem 259:13595–13604

Perret C, Lomri N, Gouhier N, Auffray C, Thomasset M (1988a) The rat vitamin-D-dependent calcium-binding protein (9-kDa CaBP) gene. Complete nucleotide sequence and structural organization. Eur J Biochem 172:43–51

Perret C, Lomri N, Thomasset M (1988b) Evolution of the EF-hand calcium-binding protein family: evidence for exon shuffling and intron insertion. J Mol Evol 27:351–364

Persechini A, Kretsinger RH (1988) The central helix of calmodulin functions as a flexible tether. J Biol Chem 263:12175–12178

Persechini A, Blumenthal DK, Jarrett HW, Klee CB, Hardy DO, Kretsinger RH (1989) The effects of deletions in the central helix of calmodulin on enzyme activation and peptide binding. J Biol Chem 264:8052–8058

Prasher DC, McCann RO, Longiaru M, Cormier MJ (1987)

Sequence comparisons of complementary DNAs encoding aequorin isotypes. Biochemistry 26:1326–1332

Putkey JA, Ts'ui KF, Tanaka T, Lagace L, Stein JP, Lai EC, Means AR (1983) Chicken calmodulin genes. A species comparison of cDNA sequences and isolation of a genomic clone. J Biol Chem 258:11864–11870

Putkey JA, Carroll SL, Means AR (1987) The nontranscribed chicken calmodulin pseudogene cross-hybridizes with mRNA from the slow-muscle troponin C gene. Mol Cell Biol 7:1549–1553

Reid RE, Gariépy J, Saund AK, Hodges RS (1981) Calcium-induced protein folding. Structure–affinity relationships in synthetic analogs of the helix-loop-helix calcium binding unit. J Biol Chem 256:2742–2751

Reinach FC, Karlsson R (1988) Cloning, expression, and site-directed mutagenesis of chicken skeletal muscle troponin C. J Biol Chem 263:2371–2376

Robert B, Daubas P, Akimenko M-A, Cohen A, Garner I, Guenet J-L, Buckingham M (1984) A single locus in the mouse encodes both myosin light chains 1 and 3, a second locus corresponds to a related pseudogene. Cell 39:129–140

Rogers JH (1987) Calretinin: a gene for a novel calcium-binding protein expressed principally in neurons. J Cell Biol 105: 1343–1353

Roher A, Lieska N, Spitz W (1986) The amino acid sequence of human cardiac troponin-C. Muscle Nerve 9:73–77

Romero-Herrera AE, Castillo O, Lehmann H (1976) Human skeletal muscle proteins. The primary structure of troponin C. J Mol Evol 8:251–270

Sakihama T, Kakidani H, Zenita K, Yymoto N, Kikuchi T, Sasaki T, Kannagi R, Nakanishi S, Ohmori M, Takio K, Titani K, Murachi T (1985) A putative Ca^{2+}-binding protein: structure of the light subunit of porcine calpain elucidated by molecular cloning and protein sequence analysis. Proc Natl Acad Sci USA 82:6075–6079

Salvato M, Sulston J, Albertson D, Brenner S (1986) A novel calmodulin-like gene from the nematode *Caenorhabditis elegans*. J Mol Biol 190:281–290

Saris CJ, Kristensen T, D'Eustachio P, Hicks LJ, Noonan DJ, Hunter T, Tack BF (1987) cDNA sequence and tissue distribution of the mRNA for bovine and murine p11, the S100-related light chain of the protein-tyrosine kinase substrate p36 (calpactin I). J Biol Chem 262:10663–10671

Sasagawa T, Ericsson LH, Walsh KA, Schreiber WE, Fischer EH, Titani K (1982) Complete amino acid sequence of human brain calmodulin. Biochemistry 21:2565–2569

Satyshur KA, Rao ST, Pyzalska D, Drendel W, Greaser M, Sundaralingam M (1988) Refined structure of chicken skeletal muscle troponin C in the two-calcium state at 2-Å resolution. J Biol Chem 263:1628–1647

Schaefer WH, Hinrichsen RD, Burgess-Cassler A, Kung C, Blair IA, Watterson DM (1987a) A mutant *Paramecium* with a defective calcium-dependent potassium conductance has an altered calmodulin: a nonlethal selective alteration in calmodulin regulation. Proc Natl Acad Sci USA 84:3931–3935

Schaefer WH, Lukas TJ, Blair IA, Schultz JE, Watterson DM (1987b) Amino acid sequence of a novel calmodulin from *Paramecium tetraurelia* that contains dimethyllysine in the first domain. J Biol Chem 262:1025–1029

Seidel U, Bober E, Winter B, Lenz S, Lohse P, Arnold HH (1987) The complete nucleotide sequences of cDNA clones coding for human myosin light chains 1 and 3. Nucleic Acids Res 15:4989

SenGupta B, Friedberg F, Detera-Wadleigh SD (1987) Molecular analysis of human and rat calmodulin complementary DNA clones. J Biol Chem 262:16663–16670

Sherbany AA, Parent AS, Brosius J (1987) Rat calmodulin cDNA. DNA 6:267–272

Simmen RCM, Tanaka T, Ts'ui KF, Putkey JA, Scott MJ, Lai EC, Means AR (1985) The structural organization of the chicken calmodulin gene. J Biol Chem 260:907–912

Smith VL, Doyle KE, Maune JF, Munjaal RP, Beckingham K (1987) Structure and sequence of the *Drosophila melanogaster* calmodulin gene. J Mol Biol 196:471–485

Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kansas Sci Bull 38:1409–1438

Stein JP, Munjaal RP, Lagace L, Lai EC, O'Malley BW, Means AR (1983) Tissue-specific expression of a chicken calmodulin pseudogene lacking intervening sequences. Proc Natl Acad Sci USA 80:6485–6489

Strehler EE, Periasamy M, Strehler-Page M-A, Nadal-Ginard B (1985) Myosin light chain 1 and 3 gene has two structurally distinct and differentially regulated promotors evolving at different rates. Mol Cell Biol 5:3168–3182

Strynadka NCJ, James MNG (1989) Crystal structures of the helix-loop-helix calcium-binding proteins. Annu Rev Biochem 58:951–998

Suzuyama Y, Umegane T, Maita T, Matsuda G (1980) The amino acid sequence of the L-2 light chain of chicken skeletal muscle myosin. Hoppe-Seyler's Z Physiol Chem 361:119–127

Swan DG, Hale RS, Dhillon N, Leadlay PF (1987) A bacterial calcium-binding protein homologous to calmodulin. Nature 329:84–85

Szebenyi DME, Moffat K (1986) The refined structure of vitamin D-dependent calcium-binding protein from bovine intestine. Molecular details, ion binding, and implications for the structure of other calcium-binding proteins. J Biol Chem 261:8761–8777

Szebenyi DME, Obendorf SK, Moffat K (1981) Structure of vitamin D-dependent calcium-binding protein from bovine intestine. Nature 294:327–332

Takagi T, Konishi K (1983) Amino acid sequence of troponin C obtained from ascidian (*Halocynthia roretzi*) body wall muscle. J Biochem 94:1753–1760

Takagi T, Konishi K (1984a) Amino acid sequence of the β chain of sarcoplasmic calicum binding protein (SCP) obtained from shrimp tail muscle. J Biochem 96:59–67

Takagi T, Konishi K (1984b) Amino acid sequence of α chain of sarcoplasmic calcium binding protein obtained from shrimp tail muscle. J Biochem 95:1603–1615

Takagi T, Nemoto T, Konishi K, Yazawa M, Yagi K (1980) The amino acid sequence of the calmodulin obtained from sea anemone (*Metridium senile*) muscle. Biochem Biophys Res Commun 96:377–381

Takagi T, Kobayashi T, Konishi K (1984) Amino-acid sequence of sarcoplasmic calcium-binding protein from scallop (*Patinopecten yessoensis*) adductor striated muscle. Biochim Biophys Acta 787:252–257

Takagi T, Konishi K, Cox JA (1986a) Amino acid sequence of two sarcoplasmic calcium-binding proteins from the protochordate amphioxus. Biochemistry 25:3585–3592

Takagi T, Kudoh S, Konishi K (1986b) The amino acid sequence of ascidian (*Halocynthia roretzi*) myosin light chains. Biochim Biophys Acta 874:318–325

Takagi T, Nojiri M, Konishi K, Maruyama K, Nonomura Y (1986c) Amino acid sequence of vitamin D-dependent calcium-binding protein from bovine cerebellum. FEBS Lett 201: 41–45

Takeda T, Yamamoto M (1987) Analysis and *in vivo* disruption of the gene coding for calmodulin in *Schizosaccharomyces pombe*. Proc Natl Acad Sci USA 84:3580–3584

Tanaka H, Maita T, Ojima T, Nishita K, Matsuda G (1988) Amino acid sequence of the regulatory light chain of clam foot muscle myosin. J Biochem 103:572–580

Thatcher DR, Pechére J-F (1977) The amino-acid sequence of

the major parvalbumin from thornback-ray muscle. Eur J Biochem 75:121–132

Thomasset M, Desplan C, Warembourg M, Perret C (1986) Vitamin-D dependent 9 kDa calcium-binding protein gene: cDNA cloning, mRNA distribution and regulation. Biochimie 68:935–940

Toda H, Yazawa M, Kondo K, Honma T, Narita K, Yagi K (1981) Amino acid sequence of calmodulin from scallop (*Patinopecten*) adductor muscle. J Biochem 90:1493–1505

Toda H, Yazawa M, Sakiyama F, Yagi K (1985) Amino acid sequence of calmodulin from wheat germ. J Biochem 98:833–842

Tomlinson CR, Klein WH (1989) Temporal and spatial transcriptional regulation of the aboral ectoderm-specific Spec genes during sea urchin embryogenesis. Mol Rep Dev (in press)

Tschudi C, Young AS, Ruben L, Patton CL, Richards FF (1985) Calmodulin genes in trypanosomes are tandemly repeated and produce multiple mRNAs with a common 5′ leader sequence. Proc Natl Acad Sci USA 82:3998–4002

Tufty RM, Kretsinger RH (1975) Troponin and parvalbumin calcium binding regions predicted in myosin light chain and T7 lysozyme. Science 187:167–169

Umegane T, Maita T, Matsuda G (1982) Amino-acid sequence of the L-1 light chain of chicken fast skeletal-muscle myosin. Hoppe-Seyler's Z Physiol Chem 363:1321–1330

Vanaman TC, Sharief F (1979) Structural properties of calmodulin from divergent eucaryotic organisms. Fed Proc 38: 788

Van der Bliek AM, Meyers MB, Biedler JL, Hes E, Borst P (1986) A 22-kd protein (sorcin/V19) encoded by an amplified gene in multidrug-resistant cells, is homologous to the calcium-binding light chain of calpain. EMBO J 5:3201–3208

van Eerd J-P, Takahashi K (1976) Determination of the complete amino acid sequence of bovine cardiac troponin C. Biochemistry 15:1171–1180

van Eerd J-P, Capony J-P, Ferraz C, Pechére J-F (1978) The amino-acid sequence of troponin C from frog skeletal muscle. Eur J Biochem 91:231–242

Varghese S, Lee S, Huang Y-C, Christakos S (1988) Analysis of rat vitamin D-dependent calbindin-D28K gene expression. J Biol Chem 263:9776–9784

Wasserman RH, Taylor AN (1966) Vitamin D_3-induced calcium-binding protein in chick intestinal mucosa. Science 152: 791–793

Watanabe B, Maita T, Konno K, Matsuda G (1986) Amino-acid sequence of LC-1 light chain of squid mantle muscle myosin. Biol Chem Hoppe-Seyler 367:1025–1032

Watterson DM, Sharief F, Vanaman TC (1980) The complete amino acid sequence of the Ca^{2+}-dependent modulator protein (calmodulin) of bovine brain. J Biol Chem 255:962–975

Weeds AG, McLachlan AD (1974) Structural homology of myosin alkali light chains, troponin C and carp calcium binding protein. Nature 252:646–649

Wilkinson JM (1976) The amino acid sequence of troponin C from chicken skeletal muscle. FEBS Lett 70:254–256

Wilkinson JM (1980) Troponin C from rabbit slow skeletal and cardiac muscle is the product of a single gene. Eur J Biochem 103:179–188

Wilson PW, Harding M, Lawson DEM (1985) Putative amino acid sequences of chick calcium-binding protein deduced from a complementary DNA sequence. Nucleic Acids Res 13:8867–8881

Wilson PW, Rogers J, Harding M, Pohl V, Pattyn G, Lawson DEM (1988) Structure of chick chromosomal genes for calbindin and calretinin. J Mol Biol 200:615–625

Wnuk W (1988) Calcium binding to troponin C and the regulation of muscle contraction. In: Gerday C, Bolis L, Gilles R (eds) Calcium and calcium binding proteins: molecular and functional aspects. Springer-Verlag, Berlin, pp 44–68

Wnuk W, Schoechlin M, Kobayashi T, Takagi T, Konishi K, Hoar PE, Kerrick WGL (1986) Two isoforms of troponin C from crayfish. Their characterization and a comparison of their primary structure with the tertiary structure of skeletal troponin C. J Muscle Res Cell Motil 7:67–68

Wood TL, Kobayashi Y, Frantz G, Varghese S, Christakos S, Tobin AJ (1988) Molecular cloning of mammalian 28,000 M_r vitamin D-dependent calcium binding protein (calbindin-D 28K): expression of calbindin-D28K RNAs in rodent brain and kidney. DNA 7:585–593

Xiang M, Bedard P-A, Wessel G, Filion M, Brandhorst B, Klein WH (1988) Tandem duplication and divergence of a sea urchin protein belonging to the troponin C superfamily. J Biol Chem 263:17173–17180

Yamakuni T, Kuwano R, Odani S, Miki N, Yamaguchi Y, Takahashi Y (1986) Nucleotide sequence of cDNA to mRNA for a cerebellar Ca-binding protein, spot 35 protein. Nucleic Acids Res 14:6768

Yamanaka MK, Saugstad JA, Hanson-Painton O, McCarthy BJ, Tobin SL (1987) Structure and expression of the *Drosophila* calmodulin gene. Nucleic Acids Res 15:3335–3348

Yazawa M, Yagi K, Toda H, Kondo K, Narita K, Yamazaki R, Sobue K, Kakiuchi S, Nagao S, Nozawa Y (1981) The amino acid sequence of the *Tetrahymena* calmodulin which specifically interacts with guanylate cyclase. Biochem Biophys Res Commun 99:1051–1057

Zhu D-X, Maeda N, Fitch WM (1985) Amino acid sequences of two parvalbumins from electric eel (*Electrophorus electricus*). Sci Sin 28B:926–941

Zimmer DB, Van Eldik LJ (1989) Analysis of the calcium-modulated proteins, calmodulin, and their target proteins during C6 glioma cell differentiation. J Cell Biol 108:141–151

Zolese G, Tangorra A, Curatola G, Giambanco I (1988) Interaction of S100-b protein with cardiolipin vesicles as monitored by electron spin resonance, pyrene fluorescence and circular dichroism. Cell Calcium 9:149–157

Zot AS, Potter JD, Straus WL (1987) Isolation and sequence of a cDNA clone for rabbit fast skeletal muscle troponin C. Homology with calmodulin and parvalbumin. J Biol Chem 262:15418–15421

Received October 9, 1989/Revised and accepted December 20, 1989

Appendix I. The data base: representative amino acid sequences for available calcium-modulated proteins, illustrating the alignment of EF-hand domains used in this study

0	0001	1	CAMHS	E FKEAFS LF D K D GDGT I TTKELGTVMRSL
0	0016	1	CAMSC	E FKEAFA LF D K D NNGS I SSSELATVMRSL
0	0055	1	CALCIB	R LGKRFK KL D L D NSGS L SVEEFMSLP EL
0	0056	1	CAL1CE	E FREAFM MF D K D GNGT I STKELGIAMRSL
0	0057	1	SPEC1	E FKRRFK NK D T D KSKS I TAEELGEFFKST
0	0058	1	TRACTIN	E IREAFD LF D T D GSGT I DAKELKVAMRAL
0	0061	1	LPS1A	EALKQEFKDNY D T N KDGT V SCAELVKLMNWT
0	0062	1	LPS1B	EYYKNEFE KF D K N GDGS L TTAEMSEFM SK
0	0063	1	QUIDLN	E IKDAFD MF D I D GDGQ I TSKELRSVMKSL
0	0064	1	aACTDD	E FKACFS HF D K D NDNK L NRLEFSSCLKSI
0	0070	1	CMSE	R LKKRFD RW D F D GNGA L ERADFEKEAQHI
0	0080	1	TPHUCS	E FKAAFD MF D A D GGGD I SVKELGTVMRML
0	0150	1	CDC31	E IYEAFS LF D M N NDGF L DYHELKVAMKAL
0	0151	1	TPHR	Q FRAAFD IF VAD A KDGT I SSKELGKVMKML
0	0152	1	TPAP1	A LQKAFD SF D T D SKGF I TPETVGIILRMM
0	0176	1	MOPP1	Q IQECFQ IF D K D NDGK V SIEELGSALRSL
0	0178	1	MORBLD	E FKEAFT VI D Q N RDGI I DKEDLRDTFAAM
0	0185	1	MOSWLE	E MKEAFS MI D V D RDGF V SKDDIKAISEQL
0	0188	1	CVP	E CMKIFD IF D R N AEN IAPVSDTMDMLTKL
0	0198	1	MOSWLD	D LKDVFE LF D FWDG RDGA V DAFKLGDVCRCL
0	0207	1	MOHSA1	E FKEAFL LF D S T GDSK I ILSQVGDVLRAL
0	0263	1	SCBPND	KTYFNR I D F D KDGA I TRMDFESMAERF
0	0289	1	CALBNHS	Q FFEIWL HF D A D GSGY L EGKELQNLIQEL
0	0318	1	CALPNOC	S CRSMVN LM D R D GNGK L GLVEFNILWNRI
0	0320	1	AEQAV1	R HKHMFN FL D V N HNGK I SLDEMVYKASDI
0	0358	1	PVNESA	
0	0374	1	PVNESB	
0	0375	1	ONC	
0	0440	1	KLBOI	E LKGIFE KY A A K EGDPNQL SKEELKLLLQTE
0	0444	1	BCBOIB	A LIDVFH QY S G R EGDKHKL KKSELKELINNE
0	0448	1	2A9	L LVAIFH KY S G R EGDKHTL SKKELKELIQKE
0	0450	1	P10BT	T MMFTFH KF A GDKGYL TKEDLRVLMEKE
0	0459	1	TCBP10	DVARRLFK RY D K D GS GQL QDDEIAGLLKDT
0	0001	2	CAMHS	E LQDMIN EV D A D GNGT I DFPEFLTMMARK
0	0016	2	CAMSC	E VNDLMN EI D V D GNHQ I EFSEFLALMSRQ
0	0055	2	CALCIB	L VQRVID IF D T D GNGE V DFKEFIEGVSQF
0	0056	2	CAL1CE	E ILEMIN EV D I D GNGQ I EFPEFCVMMKRM
0	0057	2	SPEC1	Q IDKMIS DV D T D ESGT I DFSEMLMGIAEQ
0	0058	2	TRACTIN	E IKKMIS EI D K D GSGT I DFEEFLTMMTAK
0	0061	2	LPS1A	Q N IIA RL D V N SDGH M QFDEFILYM EG
0	0062	2	LPS1B	E IEYLIS RV D L N DDGR V QFNEFFMHL DG
0	0063	2	QUIDLN	E LEEMIR EV D T D GNGT I EYAEFVEMMAKQ
0	0064	2	aACTDD	Q LNQVIS KI D T D GNGT I SFEEFIDYMVSS
0	0070	2	CMSE	L FDYLAK EA G V G SDGS L TEEQFIRVTENL
0	0080	2	TPHUCS	E LDAIIE EV D E D GSGT I DFEEFLVMMVRQ
0	0150	2	CDC31	E ILDLID EY D S E GRHL M LYDDFYIVMGEK
0	0151	2	TPHR	D LQEMIE EV D I D GSGT I DFEEFCLMMYRQ
0	0152	2	TPAP1	H LQQVIS ET D E D GSGE I EFEEFAELAAKF
0	0176	2	MOPP1	E LNTIKG E LNAKEF DLATFKTVYRKP
0	0178	2	MORBLD	E LDAMM K E A SGP I NFTVFLTMFGEK
0	0185	2	MOSWLE	E LTAML K E A PGP L NFTMFLS DK
0	0188	2	CVP	TEAIMK EARG P K GDKKNI GPEEWLTLCSKW
0	0198	2	MOSWLD	D VFA VGGTH K M G EKS L PFEEFLPAYEGL
0	0207	2	MOHSA1	E VRKVLGNPS N E E LNAKKI EFEQFLPMMQAI
0	0263	2	SCBPND	DNFLTAVAG G K GIDE T TFINSM
0	0289	2	CALBNHS	E MKTFVD QY G Q R DDGK I GIVELAHVLPTE
0	0318	2	CALPNOC	N YLAIFR KF D L D KSGS M SAYEMRMAIESA
0	0320	2	AEQAV1	K RHKDAV EA F F G GAGM K YGVETDWPAYIE
0	0358	2	PVNESA	D INKAIH AF K A G EA F DFKKF VHLLGL
0	0374	2	PVNESB	D IEAALS SV K A A ES F NYKTF FTKCGL

Appendix I. Continued

0	0375	2	ONC	D IAAALQ EC Q D P DT F EPQKF FQTSGL
0	0440	2	KLBOI	T LDELFE EL D K N GDGE V SFEEFQVLVKKI
0	0444	2	BCBOIB	V VDKVME TL D S D GDGE C DFQEFMAFVAMI
0	0448	2	2A9	E IARLME DL D R N KDQE V NFQEYVTFLGAL
0	0450	2	P10BT	A VDKIMK DL D Q C RDGK V GFQSFFSLIAGL
0	0459	2	TCBP10	D VKIWLQ MA D T N SDGS V SLEEYEDLIIKS
0	0001	3	CAMHS	E IREAFR VF D K D GNGY I SAAELRHVMTNL
0	0016	3	CAMSC	E LLEAFK VF D K N GDGL I SAAELKHVLTSI
0	0055	3	CALCIB	K LRFAFR IY D M D KDGY I SNGELFQVLKMM
0	0056	3	CAL1CE	M IREAFR VF D K D GNGV I TAQEFRYFMVHM
0	0057	3	SPEC1	H YTKAFD DM D K D GNGS L SPQELREALSAS
0	0058	3	TRACTIN	E ILKAFR LF D D D NSGT I TIKDLRRVAKEL
0	0061	3	LPS1A	DEIKQMFD DL D K D GNGR I SPDELNKGVREI
0	0062	3	LPS1B	H IKQQFM AI D K D KNGK I SPEEMVFGITKI
0	0063	3	QUIDLN	E MREAFR VF D K D GNGL I TAAELRQVMANF
0	0064	3	aACTDD	
0	0070	3	CMSE	V VKGTWG MC D K N ADGQ I NADEFAAWLTAL
0	0080	3	TPHUCS	E LAECFR IF D R N ADGY I DPEELAEIFRAS
0	0150	3	CDC31	E IKRAFQ LF D D D HIGK I SIKNLRRVAKEL
0	0151	3	TPHR	E LSEAFR LF D L D GDG I G DELKAALDGT
0	0152	3	TPAP1	E LKEAFR IY D R G GDGY I TTQVLREILKEL
0	0176	3	MOPP1	E MLDAFR AL D K E GNGT I QEAELRQLLLNL
0	0178	3	MORBLD	DVITGAFK VL D P E GKGT I KKQFLEELLTTQ
0	0185	3	MOSWLE	T IRNAFA MF D E Q ENKK L NIEYIKDLLEDM
0	0188	3	CVP	E ILRAFK VF D A N GDGV I DFDEFKFIMQKV
0	0198	3	MOSWLD	D YMEAFK TF D R E GQGF I SGAELRHVLTAL
0	0207	3	MOHSA1	D FVEGLR VF D K E GNGT V MGAELRHGLATL
0	0263	3	SCBPND	P LPLFFR AV D T N EDNN I SRDEYGIFFGML
0	0289	3	CALBNHS	E FMKTWR KY D T D HSGF I ETEELKNFLKDL
0	0318	3	CALPNOC	K LYELII TR Y S E PDLA V DFDNFVCCLVRL
0	0320	3	AEQAV1	W GDALFD IV D K D QNGA I TLDEWKAYTKAA
0	0358	3	PVNESA	D VTKAFH IL D K D RSGY I EEEELQLILKGF
0	0374	3	PVNESB	Q VKKVFD IL D Q D KSGY I EEDELQLFLKNF
0	0375	3	ONC	Q VKDIFR FI D N D QSGY L DGDELKYFLQKF
0	0440	3	KLBOI	
0	0444	3	BCBOIB	
0	0448	3	2A9	
0	0450	3	P10BT	
0	0459	3	TCBP10	
0	0001	4	CAMHS	E VDEMIR EA D I D GDGQ V NYEEFVQMMTAK
0	0016	4	CAMSC	E VDDMLR EV S D GSGE I NIQQF AALLSK
0	0055	4	CALCIB	I VDKTII NA D K D GDGR I SFEEFSAVVGGL
0	0056	4	CAL1CE	E VDEMIK EV D V D GDGE I DYEEFVKMMSNQ
0	0057	4	SPEC1	K IKAIIQ KA D A N KDGK I DREEFMKLIKSC
0	0058	4	TRACTIN	E LQEMIA EA D R N DDNE I DEDEFIRIMKKT
0	0061	4	LPS1A	M ANKLIQ EA D K D GDGH V NMEEFFDTLVVK
0	0062	4	LPS1B	E VAKLIK ES SFE D DDGY I NFNEFVNRF
0	0063	4	QUIDLN	E ISEMIR EA D I D GDGM V NYEEFVKMMTPK
0	0064	4	aACTDD	
0	0070	4	CMSE	E AAEAFN QV D T N GNGE L SLDELLTAVRDF
0	0080	4	TPHUCS	E IESLMK DG D K N NDGR I DFDEFLKMMEGV
0	0150	4	CDC31	E LRAMIE EF D L D GDGE I NENEFIAICTDS
0	0151	4	TPHR	E VDEMMA DG D K N HDSQ I DYEEWVTMMKFV
0	0152	4	TPAP1	N LDEIIE EI D E D GSGT I DFMEFMKMMTG
0	0176	4	MOPP1	E VEELMK EV S V S GDGA I NYESFVDMLVTG
0	0178	4	MORBLD	E IKNMWA AF P P D VGGN V DYKNICYVITHG
0	0185	4	MOSWLE	E MR MTF KE A P V EGGK F DYVKFTAMIK G
0	0188	4	CVP	E VEEAMK EA D E D GNGV I DIPEFMDLIKSK
0	0198	4	MOSWLD	D EIISLT DL Q E D LEGN V KYEDFVKKVMAG

Appendix I. Continued

0	0207	4	MOHSA1	E VEALM AG Q E D SNGC I NYEAFVKHIMSI
0	0263	4	SCBPND	M APASFD AI D T N NDGL L SLEEFVIAGSDF
0	0289	4	CALBNHS	Y TDLMLK LF D S N NDGK L ELTEMARLLPVQ
0	0318	4	CALPNOC	T MFRFFK TL D T D LDGV V TFDLFKWLQLTM
0	0320	4	AEQAV1	D CEETFR VC D I D ESGQ L DVDEMTRQHLGF
0	0358	4	PVNESA	E TKDLLI KG D K D GDGK I GVDEFTSLVAES
0	0374	4	PVNESB	E TKAFLE AG D S D GDGK I GVDEFQALVR S
0	0375	4	ONC	E TKSLMD AA D N D GDGK I GADEFQEMVH S
0	0440	4	KLBOI	
0	0444	4	BCBOIB	
0	0448	4	2A9	
0	0450	4	P10BT	
0	0459	4	TCBP10	

Appendix II. The data base: references for the sequences of calcium-modulated proteins used in this study

OTU no.	ID	Molecule	Species	Common name	Reference
1	CAMHS%	Calmodulin	*Homo sapiens*	Human	Sasagawa et al. 1982; Fischer et al. 1988
		Calmodulin	*Oryctolagus cuniculus*	Rabbit	Grand et al. 1981
		Calmodulin	*Bos taurus*	Cow	Grand and Perry 1978; Kasai et al. 1980; Watterson et al. 1980
	1%#	Calmodulin-1	*Rattus norvegicus*	Rat	Dedman et al. 1978; Nojima and Sokabe 1986; Nojima et al. 1987; Sherbany et al. 1987
	2%#	Calmodulin-2	*Rattus norvegicus*	Rat	Dedman et al. 1978; Nojima and Sokabe 1986; Nojima et al. 1987; Sherbany et al. 1987
	%	Calmodulin-1	*Mus musculus*	Mouse	Bender et al. 1988
	%	Calmodulin-2	*Mus musculus*	Mouse	Bender et al. 1988
	1%	Calmodulin	*Gallus gallus*	Chicken	Putkey et al. 1983
	2#	Calmodulin	*Gallus gallus*	Chicken	Simmen et al. 1985
	%	Calmodulin	*Xenopus laevis*	African clawed frog	Chien and Dawid 1984
2	CAMEE%	Calmodulin	*Electrophorus electricus*	Electric eel	Lagace et al. 1983
3	CAMAPA%	Calmodulin-α*	*Arbacia punctulata*	Sea urchin	Hardy et al. 1987; Kretsinger and Hardy 1987
4	CAMAPB%	Calmodulin-β*	*Arbacia punctulata*	Sea urchin	Hardy et al. 1987; Kretsinger and Hardy 1987
5	CAMSP%	Calmodulin*	*Strongylocentrotus purpuratus*	Sea urchin	Floyd et al. 1986
6	CAMLP#	Calmodulin *	*Lytechinus pictus*	Sea urchin	Hardin et al. 1987
7	CAMDM%#	Calmodulin	*Drosophila melanogaster*	Fruit fly	Smith et al. 1987; Yamanaka et al. 1987
9	CAMT	Calmodulin	*Triticum aestivum*	Wheat	Toda et al. 1985
10	CAMSO	Calmodulin	*Spinacia oleracea*	Spinach	Lukas et al. 1984
11	CAMTP	Calmodulin	*Tetrahymena pyriformis*	Ciliate	Yazawa et al. 1981
12	CAMTBG#	Calmodulin	*Trypanosoma brucei gambiense*	Flagellate	Tschudi et al. 1985
13	CAMPT	Calmodulin	*Paramecium tetraurelia*	Ciliate	Schaefer et al. 1987a,b
14	CAMCR	Calmodulin	*Chlamydomonas reinhardtii*	Flagellate	Lukas et al. 1985
15	CAMDD	Calmodulin	*Dictyostelium discoideum*	Cellular slime mold	Marshak et al. 1984
16	CAMSC#	Calmodulin	*Saccharomyces cerevisiae*	Yeast	Davis et al. 1986

Appendix II. Continued

OTU no.	ID	Molecule	Species	Common name	Reference
17	CAMSCP#	Calmodulin	*Schizosaccharomyces pombe*	Yeast	Takeda and Yamamoto 1987
23	CAMP	Calmodulin	*Patinopecten* sp.	Scallop	Toda et al. 1981
24	CAMMS	Calmodulin	*Metridium senile*	Sea anemone	Takagi et al. 1980
		Calmodulin	*Renilla reniformis*	Sea pansy	Vanaman and Sharief 1979; Jamieson et al. 1980
25	CAMNGG%	Neocalmodulin*	*Gallus gallus*	Chicken	Mangelsdorf et al. 1987
40	CCM1CH#	CAM pseudogene	*Gallus gallus*	Chicken	Stein et al. 1983
41	hSC8RN#	CAM pseudogene 1	*Rattus norvegicus*	Rat	Nojima and Sokabe 1986
42	hSC9RN#	CAM pseudogene 2	*Rattus norvegicus*	Rat	Nojima and Sokabe 1986
43	hGH6HS#	CAM pseudogene	*Homo sapiens*	Human	Koller and Strehler 1988
55	CALCIB	Calcineurin-B	*Bos taurus*	Cow	Aitken et al. 1984
56	CAL1CE#	Cal-1-gene	*Caenorhabditis elegans*	Nematode	Salvato et al. 1986
57	SPEC1%	Spec1	*Strongylocentrotus purpuratus*	Sea urchin	Carpenter et al. 1984; Hardin et al. 1985; Xiang et al. 1988†
58	TRACTIN%	Caltractin	*Chlamydomonas reinhardtii*	Flagellate	Huang et al. 1988a
59	SPEC2A%	Spec2 A	*Strongylocentrotus purpuratus*	Sea urchin	Carpenter et al. 1984
60	SPEC2C%	Spec2 C	*Strongylocentrotus purpuratus*	Sea urchin	Carpenter et al. 1984
61	LPS1A%	Lps1 (doms1–4)	*Lytechinus pictus*	Sea urchin	Xiang et al. 1988
62	LPS1B%	Lps1 (doms5–8)	*Lytechinus pictus*	Sea urchin	Xiang et al. 1988
63	QUIDLN	Squidulin	*Loligo pealei*	Squid	Head 1989
64	aACTDD%	α-actinin	*Dictyostelium discoideum*	Cellular slime mold	Noegel et al. 1987
65	aACTGG%	α-actinin-fb	*Gallus gallus*	Chicken	Baron et al. 1987
70	CMSE#	Bacterial-CAM	*Streptomyces erythraeus*	Actinobacterium	Swan et al. 1987
80	TPHUCS%	Troponin-C-sk	*Homo sapiens*	Human	Romero-Herrera et al. 1976; Gahlmann et al. 1988†
81	TPPGCS	Troponin-C-sk	*Sus scrofa*	Pig	Lorkin and Lehmann 1983
82	TPRBCS%	Troponin-C-sk	*Oryctolagus cuniculus*	Rabbit	Collins et al. 1977; Zot et al. 1987; Chen et al. 1988
		Troponin-C-sk*	*Rattus norvegicus*	Rat	Garfinkel et al. 1982
83	TPCHCS%	Troponin-C-sk	*Gallus gallus*	Chicken	Wilkinson 1976; Reinach and Karlsson 1988
84	TPFGCS	Troponin-C-sk	*Rana esculenta*	Frog	van Eerd et al. 1978
120	TPHUCC%	Troponin-C-cd	*Homo sapiens*	Human	Roher et al. 1986; Gahlmann et al. 1988†
		Troponin-C-cd	*Oryctolagus cuniculus*	Rabbit	Wilkinson 1980
122	TPBOCC	Troponin-C-cd	*Bos taurus*	Cow	van Eerd and Takahashi 1976
123	TPQLCC%	Troponin-C-cd	*Coturnix coturnix*	Quail	Maisonpierre et al. 1987
150	CDC31#	Cell-div.-con.-prot.	*Saccharomyces cerevisiae*	Yeast	Baum et al. 1986
151	TPHR	Troponin-C	*Halocynthia roretzi*	Ascidian	Takagi and Konishi 1983
152	TPAP1	Troponin-C-1	*Astacus leptodactylus*	Crayfish	Wnuk 1988
153	TPAP2	Troponin-C-2	*Astacus leptodactylus*	Crayfish	Wnuk 1988
165	MORTA1%#	Myosin-L1-ELC-sk	*Rattus norvegicus*	Rat	Garfinkel et al. 1982; Periasamy et al. 1984; Strehler et al. 1985
166	MORTA2%#	Myosin-L4-ELC-sk	*Rattus norvegicus*	Rat	Garfinkel et al. 1982; Periasamy et al. 1984; Strehler et al. 1985
167	MORBLA	Myosin-L1-ELC-sk	*Oryctolagus cuniculus*	Rabbit	Frank and Weeds 1974; Collins 1976b
168	MOCHLA%#	Myosin-L1-ELC-sk	*Gallus gallus*	Chicken	Matsuda et al. 1981b; Nabeshima et al. 1982, 1984; Umegane et al. 1982

Appendix II. Continued

OTU no.	ID	Molecule	Species	Common name	Reference
169	MOCHLC%	Myosin-L1-ELC-vt	*Gallus gallus*	Chicken	Maita et al. 1980; Nakamura et al. 1988
170	MOCHG2%	Myosin-LGII-ELC-sm	*Gallus gallus*	Chicken	Matsuda et al. 1981a; Nabeshima et al. 1987
174	MOCHCE%	Myosin-L23-ELC-em	*Gallus gallus*	Chicken	Kawashima et al. 1987
175	MOCHG1	Myosin-LGI-RLC-sm	*Gallus gallus*	Chicken	Maita et al. 1981a; Pearson et al. 1986†
176	MOPP1	Myosin ELC	*Physarum polycephalum*	Plasmodial slime mold	Kobayashi et al. 1988a
177	MOHR	Myosin-ELC-body wall	*Halocynthia roretzi*	Ascidian	Takagi et al. 1986b
178	MORBLD	Myosin-L2-RLC-sk	*Oryctolagus cuniculus*	Rabbit	Collins 1976b; Matsuda et al. 1977a, 1978†
179	MOCHLS	Myosin-L2-RLC-sk	*Gallus gallus*	Chicken	Matsuda et al. 1977b; Suzuyama et al. 1980
181	MOCHAC	Myosin-L2A-RLC-cd	*Gallus gallus*	Chicken	Matsuda et al. 1981a
182	MOCHBC	Myosin-L2B-RLC-cd	*Gallus gallus*	Chicken	Matsuda et al. 1981a
183	MOSWLA	Myosin-RLCa-sm	*Patinopecten yessoensis*	Scallop	Miyanishi et al. 1985
184	MOSWLB	Myosin-RLCb-sm	*Patinopecten yessoensis*	Scallop	Miyanishi et al. 1985
185	MOSWLE	Myosin-EDTA-RLC	*Pecten maximus*	Scallop	Kendrick-Jones and Jakes 1977†
186	MOAI%	Myosin-ELC	*Aquipecten irradians*	Scallop	Collins et al. 1986; Goodwin et al. 1987
187	MODM1%#	Myosin-ELC-sk-la	*Drosophila melanogaster*	Fruitfly	Falkenthal et al. 1984, 1985†
188	CVP	Ca-vector-protein	*Branchiostoma lanceolatum*	Amphioxus	Kobayashi et al. 1987
189	MOCHL4%#	Myosin-LC3-ELC-sk	*Gallus gallus*	Chicken	Maita et al. 1981b; Nabeshima et al. 1982, 1984
192	MORTL2#	Myosin-L2-RLC	*Rattus norvegicus*	Rat	Garfinkel et al. 1982; Nudel et al. 1984†
193	MOHSCC%	Myosin-L1-ELC-vt	*Homo sapiens*	Human	Hoffmann et al. 1988
197	MOSWLC	Myosin-RLC-sk	*Chlamys nipponensis*	Scallop	Maita et al. 1984
198	MOSWLD	Myosin-ELC-sk	*Patinopecten yessoensis*	Scallop	Maita et al. 1987a
199	MOSWLF	Myosin-ELC-sk	*Todarodes pacificus*	Squid	Watanabe et al. 1986
200	MOSWLG	Myosin-RLC-sk	*Patinopecten yessoensis*	Scallop	Maita et al. 1984
201	MOSWLH	Myosin-RLC-sk	*Todarodes pacificus*	Squid	Maita et al. 1987b
202	MOSWLJ	Myosin-RLC-sk	*Spisula sachalinensis*	Surf clam	Tanaka et al. 1988
203	MORBL4	Myosin-ELC-L4-sk	*Oryctolagus cuniculus*	Rabbit	Frank and Weeds 1974
204	MOHR2	Myosin-RLC-body wall	*Halocynthia roretzi*	Ascidian	Takagi et al. 1986b
205	MOMMA1#	Myosin-ELC-L1-sk	*Mus musculus*	Mouse	Robert et al. 1984
206	MOMMA2#	Myosin-ELC-L3-sk	*Mus musculus*	Mouse	Robert et al. 1984
207	MOHSA1%	Myosin-ELC-L1-sk	*Homo sapiens*	Human	Seidel et al. 1987
208	MOHSA2%	Myosin-ELC-L3-sk	*Homo sapiens*	Human	Seidel et al. 1987
209	MOCHF1%	Myosin-fb	*Gallus gallus*	Chicken	Nabeshima et al. 1987
210	MODM2#	Myosin-ELC-sk-ad	*Drosophila melanogaster*	Fruitfly	Falkenthal et al. 1985
211	MORTCR%	Myosin-L2-RLC-cd	*Rattus norvegicus*	Rat	Kumar et al. 1986; Henderson et al. 1988†
212	MOMMCC#	Myosin-ELC-at	*Mus musculus*	Mouse	Barton et al. 1988
213	MOAI2%	Myosin-RLC	*Aquipecten irradians*	Scallop	Goodwin et al. 1987
255	SCBPBL1	SARC1	*Branchiostoma lanceolatum*	Amphioxus	Takagi et al. 1986a†

Appendix II. Continued

OTU no.	ID	Molecule	Species	Common name	Reference
256	SCBPBL2	SARC2	*Branchiostoma lanceolatum*	Amphioxus	Takagi et al. 1986a†
257	SCBPCF	SARC	*Astacus leptodactylus*	Crayfish	Jauregui-Adell et al. 1989
258	SCBPPV	SARC	*Perinereis vancaurica*	Sandworm	Kobayashi et al. 1984
259	SCBPPY	SARC	*Patinopecten yessoensis*	Scallop	Takagi et al. 1984
260	SCBPPB	SARC-β	*Penaeus* sp.	Brine shrimp	Takagi and Konishi 1984a
261	SCBPPAA	SARC-α-A	*Penaeus* sp.	Brine shrimp	Takagi and Konishi 1984b
262	SCBPPAB	SARC-α-B	*Penaeus* sp.	Brine shrimp	Takagi and Konishi 1984b
263	SCBPND	SARC	*Nereis diversicolor*	Sandworm	Collins et al. 1988
285	CALBNGG%#	Calbindin	*Gallus gallus*	Chicken	Wilson et al. 1985, 1988; Hunziker 1986; Fullmer and Wasserman 1987
286	CALBNRN%	Calbindin	*Rattus norvegicus*	Rat	Yamakuni et al. 1986; Varghese et al. 1988
287	CALRTGG%#	Calretinin*	*Gallus gallus*	Chicken	Rogers 1987; Wilson et al. 1988
288	CALBNBT	Calbindin	*Bos taurus*	Cow	Takagi et al. 1986c
289	CALBNHS%	Calbindin	*Homo sapiens*	Human	Parmentier et al. 1987
315	CALPLOC%	Calpain-light-30K	*Oryctolagus cuniculus*	Rabbit	Emori et al. 1986a
316	CALPLSS%	Calpain-light	*Sus scrofa*	Pig	Sakihama et al. 1985
317	CALPMOC%	Calpain-heavy-m	*Oryctolagus cuniculus*	Rabbit	Emori et al. 1986b
318	CALPNOC%	Calpain-heavy-μ	*Oryctolagus cuniculus*	Rabbit	Emori et al. 1986b
319	CALPHGG%	Calpain-heavy-m	*Gallus gallus*	Chicken	Ohno et al. 1984
320	AEQAV1%	Aequorin-1	*Aequorea victoria*	Jellyfish	Charbonneau et al. 1985; Inouye et al. 1985; Prasher et al. 1987
321	AEQAV2%	Aequorin-2	*Aequorea victoria*	Jellyfish	Charbonneau et al. 1985; Inouye et al. 1985; Prasher et al. 1987
322	CALPLHS%	Calpain-light	*Homo sapiens*	Human	Miyake et al. 1986; Ohno et al. 1986
323	LBP	Luciferin-binding protein	*Renilla reniformis*	Sea pansy	Charbonneau et al., unpublished
324	SORC%	Sorcin	*Cricetulus griseus*	Chinese hamster	Van der Bliek et al. 1986
325	CALPNHS%	Calpain-heavy-μ	*Homo sapiens*	Human	Aoki et al. 1986
355	PVLA1	Parvalbumin-α	*Latimeria chalumnae*	Coelacanth	Pechére et al. 1973, 1978
356	PVPK3	Parvalbumin	*Esox lucius*	Pike	Frankenne et al. 1973†
357	PVFGA	Parvalbumin-α	*Rana esculenta*	Frog	Jauregui-Adell et al. 1982
358	PVNESA	Parvalbumin-α	*Amphiuma means*	Two-toed amphiuma	Maeda et al. 1984
359	PVRTA%#	Parvalbumin	*Rattus norvegicus*	Rat	Berchtold et al. 1982, 1987; Epstein et al. 1986
361	PVRB	Parvalbumin-α	*Oryctolagus cuniculus*	Rabbit	Enfield et al. 1975; Capony et al. 1976
362	PVCAB1	Parvalbumin-C	*Cyprinus carpio*	Carp	Coffee et al. 1974
363	PVCAB2	Parvalbumin-B	*Cyprinus carpio*	Carp	Coffee and Bradshaw 1973a,b†; Coffee et al. 1974
364	PVCD	Parvalbumin	*Gadus callarias*	Cod	Elsayed and Bennich 1975
365	PVHK	Parvalbumin	*Merluccius merluccius*	Hake	Capony et al. 1973
366	PVLA2	Parvalbumin-β	*Latimeria chalumnae*	Coelacanth	Pechére et al. 1973; Jauregui-Adell and Pechére 1978
367	PVPK2	Parvalbumin-β	*Esox lucius*	Pike	Gerday 1976
368	PVCHUB	Parvalbumin-β	*Leuciscus cephalus*	Chub	Gerday et al. 1978
369	PVRYC	Parvalbumin	*Raja clavata*	Thornback ray	Thatcher and Pechére 1977
370	PVWI	Parvalbumin-β	*Gadus merlangus*	Whiting	Joassin and Gerday 1977
371	PVFGB	Parvalbumin	*Rana esculenta*	Frog	Capony et al. 1975
372	PVSNBB	Parvalbumin-β	*Boa constrictor*	Boa constrictor	Maeda et al. 1984
373	PVTTMB	Parvalbumin-β	*Graptemys geographica*	Map turtle	Maeda et al. 1984
374	PVNESB	Parvalbumin-β	*Amphiuma means*	Two-toed amphiuma	Maeda et al. 1984
375	ONC%	Oncomodulin	*Rattus norvegicus*	Rat	MacManus et al. 1983; Gillen et al. 1987†

Appendix II. Continued

OTU no.	ID	Molecule	Species	Common name	Reference
376	PVXLB%	Parvalbumin-β	*Xenopus laevis*	African clawed frog	Kay et al. 1987
377	PVTF	Parvalbumin	*Opsanus tau*	Toadfish	Gerday 1988
378	PVCAB3	Parvalbumin-A1	*Cyprinus carpio*	Carp	Coffee et al. 1974
440	KLBOI%	ICBP	*Bos taurus*	Cow	Huang et al. 1975; Fullmer and Wasserman 1981†; Kumar et al. 1989†
441	KLPGI	ICBP	*Sus scrofa*	Pig	Hofmann et al. 1979
442	KLRTI%#	ICBP	*Rattus norvegicus*	Rat	Desplan et al. 1983a,b; MacManus et al. 1986; Thomasset et al. 1986; Perret et al. 1988a
443	BCBOIA	S-100-α	*Bos taurus*	Cow	Isobe and Okuyama 1981
444	BCBOIB	S-100-β	*Bos taurus*	Cow	Isobe and Okuyama 1978
446	MRP-8%	CF-antigen	*Homo sapiens*	Human	Dorin et al. 1987; Odink et al. 1987†
447	MRP-14%	CF-antigen	*Homo sapiens*	Human	Odink et al. 1987
448	2A9%#	2A9/calcyclin	*Homo sapiens*	Human	Calabretta et al. 1986; Ferrari et al. 1987; Murphy et al. 1988
450	P10BT%	p10	*Bos taurus*	Cow	Glenney et al. 1985; Saris et al. 1987†
			Sus scrofa	Pig	Gerke and Weber 1985
451	42A%#	42A/p9ka	*Rattus norvegicus*	Rat	Barraclough et al. 1987, 1988; Masiakowski and Shooter 1988
452	42C%	42C	*Rattus norvegicus*	Rat	Masiakowski and Shooter 1988
453	PCBPMM%	Placental-CaBP/ 18A2	*Mus musculus*	Mouse	Jackson-Grusby et al. 1987
454	BCHS	S-100-β	*Homo sapiens*	Human	Jensen et al. 1985
455	BCRN%	S-100-β	*Rattus norvegicus*	Rat	Kuwano et al. 1984
456	BCSS	S-100-β	*Sus scrofa*	Pig	Kuwano et al. 1984
457	P11MM%	p11	*Mus musculus*	Mouse	Saris et al. 1987
458	PRARN%	PRA/calcyclin	*Rattus norvegicus*	Rat	Murphy et al. 1988
459	TCBP10	TCBP-10	*Tetrahymena thermophila*	Ciliate	Kobayashi et al. 1988b

† indicates a corrected sequence
% indicates amino acid sequence inferred from cDNA sequence
# indicates amino acid sequence inferred from gDNA sequence
* indicates sequence fragment