Primary Structure and Molecular Evolution of Protein. Internal Homology and Intramolecular Evolutionary Tree of Calmodulin

Yôichi IIDA

Department of Chemistry, Faculty of Science, Hokkaido University, Sapporo 060

(Received March 7, 1984)

Synopsis. One notable feature of the primary structure of calmodulin is its internal homology. It can be subdivided into four domains having a similar amino acid sequence. In the present paper, an intramolecular evolutionary tree was constructed with the four domains. The origin and evolution of calmodulin was discussed on the basis of these results.

Origin and evolution of protein is one of the most important problems in biology and biochemistry. Evolutionary trees of typical proteins, such as cytochrome c and globins, have been constructed with a number of biological species.[1] In previous papers,[2,3] we studied molecular evolution of calmodulin (CaM), one of the calcium-binding proteins. It has four calcium-binding sites, and interacts reversibly with Ca^{2+} to form a calmodulin-Ca^{2+} complex.[4] CaM plays an important role in regulating other enzymatic reactions or cellular processes. The complete amino acid sequence of bovine brain CaM was first determined by Watterson *et al.*[5] and by Kasai *et al.*[6] It contains 148 amino acids and has a trimethylated lysin at position 115.

One notable feature of CaM is that it is widely distributed throughout eukaryotes and that the primary structure of CaM is highly conservative against evolution.[2] Another interesting feature is that the primary structure of CaM has internal homology, that is, it can be subdivided into four domains having a similar amino acid sequence. This is shown, for example, in Fig. 1 with bovine brain CaM, where the first domain (residues 8 to 40), the second (residues 44 to 76), the third (residues 81 to 113) and the fourth (residues 117 to 148) are aligned for purposes of comparison. The amino acids which do not belong to the four domains are residues 1 to 7, 41 to 43, 77 to 80 and 114 to 116. These are the linking sequences between the domains. In the following, we only discuss the amino acids belonging to the four domains. In the previous paper,[3] we found by visual inspection that the amino acid sequence homology is greatest when the first domain is aligned with the third, and the second domain with the fourth . In order to explain this fact, a two-step intragenic duplication model (refer to Fig. 4 of Ref. 3) was proposed to the primary amino acid sequence of CaM. In this model, it is assumed that an ancestral one-domain quarter-calmodulin (AQ) underwent an elongation by the first intragenic duplication, which resulted in an ancestral two-domain half-calmodulin (AH). In the period from the birth of AH to the next intragenic duplication (stage 1 in Fig. 4 of Ref. 3), AH underwent evolutionary changes and some replacements of amino acids occurred in it. At the end of stage 1, the second intragenic duplication took place, resulting in an ancestral four-domain calmodulin (AC). In the period from the birth of AC to the present time (stage 2 in Fig. 4 of Ref. 3), AC underwent further evolution and some replacements of amino acids took place.

```
                 8
Domain (1)  -Gln-Ile-Ala-Glu-Phe-Lys-Glu-Ala-Phe-Ser-Leu-Phe-
                44
Domain (2)  -Thr-Glu-Ala-Glu-Leu-Gln-Asp-Met-Ile-Asn-Glu-Val-
                81
Domain (3)  -Ser-Glu-Glu-Glu-Ile-Arg-Glu-Ala-Phe-Arg-Val-Phe-
               117
Domain (4)  -Thr-Asp-Glu-Glu-Val-Asp-Glu-Met-Ile-Arg-Glu-Ala-

                20
      (1)   Asp-Lys-Asp-Gly-Asp-Gly-Thr-Ile-Thr-Thr-Lys-Glu-
                56
      (2)   Asp-Ala-Asp-Gly-Asp-Gly-Thr-Ile-Asp-Phe-Pro-Glu-
                93
      (3)   Asp-Lys-Asp-Gly-Asp-Gly-Tyr-Ile-Ser-Ala-Ala-Glu-
               129
      (4)   Asp-Ile-Asp-Gly-Asp-Gly-Gln-Val-Asn-Tyr-Glu-Glu-

                32
      (1)   Leu-Gly-Thr-Val-Met-Arg-Ser-Leu-Gly-
                68
      (2)   Phe-Leu-Thr-Met-Met-Ala-Arg-Lys-Met-
               105
      (3)   Leu-Arg-His-Val-Met-Thr-Asn-Leu-Gly-
               141
      (4)   Phe-Val-Gln-Met-Met-Thr-Ala-Lys
```

Fig. 1. Comparison and homology of the amino acid sequences of the four domains in bovine brain calmodulin. The first domain (residues 8 to 40), the second (residues 44 to 76), the third (residues 81 to 113) and the fourth (residues 117 to 148) are aligned for purposes of comparison.

Clearly, this model is based on the finding that relations between the first and third domains and between the second and fourth domains are closest. However, these two relations do not necessarily support the above two-step intragenic duplication model. In order to support this model definitely, we should construct intramolecular phylogenetic tree of the four domains. If such a tree is once constructed, we can clearly show how the four domains are related with one another.

There have been known two approaches to construct such an evolutionary tree;[1] one is common ancestor method and the other, matrix method. If the evolutionary rate of protein is slow enough, the two methods reach the same conclusion. In the present paper, we construct intramolecular evolutionary tree of the four domains of bovine brain CaM by the use of the common ancenstor method. "Maximum parsimony method" is used for this purpose. It is generally known that codons of m RNA are translated into amino acids of protein. Any codon is composed of a series of three nucleotides and is expressed by "the genetic code". Therefore, the observed difference of amino acid residues of CaM between the domains (see Fig. 1) is attributable to the difference of nucleotides between the corresponding codons. Referring to the

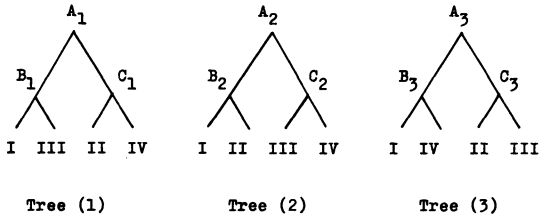

Fig. 2. Three possible intramolecular evolutionary trees (1), (2), and (3) for the four domains of bovine brain calmodulin. The first domain, the second, the third and the fourth are indicated by I, II, III, and IV, respectively. See Fig. 1 for the amino acid sequences of those domains. In each tree, we assume amino acid sequences of common ancestral domains (A_i), (B_i), and (C_i), (i=1, 2 or 3), so that the total number of nucleotide substitutions within the tree can be minimized (the maximum parsimony method). For further details, see the text.

genetic code, we assume that the difference of amino acids is given by the least number of nucleotide substitutions of the codons. In order to examine the above intragenic duplication model, we assume three evolutionary trees, as shown in Fig. 2. The tree (1) represents the evolutionary change, where the first and third domains were diverged from the common ancestral domain (B_1), while the second and fourth domains, from the common ancestor (C_1). The common ancestors (B_1) and (C_1) had been diverged from the common ancestral domain (A_1). Clearly, the tree (1) represents faithfully the evolutionary change of the above-mentioned intragenic duplication model. That is, the common ancestor (A_1) corresponds to the former or latter half of AH, since at the birth of AH it was formed by duplication of AQ. In a similar way, the common ancestor (B_1) corresponds to the first or third domain of AC, while the common ancestor (C_1), to the second or fourth domain of AC (see Fig. 4 of Ref. 3).

In addition to the tree (1), we further consider two possible evolutionary trees (2) and (3). The tree (2) assumes that the first and second domains were diverged from the common ancestral domain (B_2), while the third and fourth domains, from the common ancestor (C_2). These two common ancestors had been diverged from the common ancestral domain (A_2). In this tree, the relations between the first and second domains and between the third and fourth domains are closest. In a similar way, the tree (3) assumes that the first and fourth domains came from the common ancestral domain (B_3), while the second and third domains, from the common ancestor (C_3). Both (B_3) and (C_3) had come from the common ancestral domain (A_3). In order to determine which tree is most probable, we use the maximum parsimony method. In each tree, we assume amino acid sequences of common ancestral domains (A_i), (B_i), and (C_i), (i=1, 2 or 3), so that the total number of nucleotide substitutions within the tree can be minimized. In order to do this, we note a lack of one amino acid of the fourth domain at position 149 (see Fig. 1). The corresponding amino acid residues of the first, second and third domains are Gly at position 40, Met at position 76 and Gly at position 113, respectively. These three amino acid residues are excluded from our counting, because we cannot compare the amino acid residues in all four domains. Taking into account the remaining amino acid residues in the four domains, we calculated the minimum number of nucleotide substitutions of the trees (1), (2), and (3) as 65, 72, and 72, respectively. Therefore, the maximum parsimony method rejects the evolutionary trees (2) and (3) and suggests that the evolutionary tree (1) is most probable. This conclusion confirms that the homologies between the first and third domains and between the second and fourth domains are greatest, and supports strongly our previous two-step intragenic duplication model for the primary amino acid sequence of CaM.

References

1) See, for example, M. O. Dayhoff, "Atlas of Protein Sequence and Structure," National Biomedical Research Foundation, Washington, D. C. (1972), Vol. 5.
2) Y. Iida, *Bull. Chem. Soc. Jpn.*, **55**, 2683 (1982).
3) Y. Iida, *J. Mol. Biology*, **159**, 167 (1982).
4) N. Y. Cheung, *Science*, **207**, 19 (1980).
5) D. M. Watterson, F. Sharief, and T. C. Vanaman, *J. Biol. Chem.*, **255**, 962 (1980).
6) H. Kasai, Y. Kato, T. Isobe, H. Kawasaki, and T. Okuyama, *Biomed. Res.*, **1**, 248 (1980).