DNA Sequence and Intramolecular Evolutionary Tree of Calmodulin Gene. Application of Maximum Likelihood Method

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

(Received September 19, 1984)

DNA sequences of calmodulin genes taken from chicken and electric eel were examined to investigate the molecular evolution of its protein. One notable feature of the amino acid coding region of the gene is that 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 DNA sequences of the four domains.

Origin and evolution of protein is one of the most important problems in biology and biochemistry. In previous papers,[1–4] we studied the 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.[5] CaM plays an important role in regulating various enzymatic reactions or cellular processes. It contains 148 amino acids and has a trimethylated lysine at position 115. One notable feature of CaM is that it is widely distributed throughout eukaryotes and that the primary amino acid sequence is highly conservative against evolution. 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. It is also found that the amino acid sequence homology is greatest when the first domain (residues 8 to 40) is aligned with the third (residues 81 to 113), and the second domain (residues 44 to 76) is aligned with the fourth (residues 117 to 148). In order to explain this fact, a two-step intragenic duplication model was proposed to the primary amino acid sequence of CaM.[2] 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. 2), 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. 2), AC underwent further evolution and some replacements of amino acids took place.

DNA Sequence of Calmodulin Gene

Previously, using the data on the primary amino acid sequences, we attempted to confirm the above two-step intragenic duplication model.[3] Using the common 'ancestor method together with the maximum parsimony technique, we constructed an intramolecular evolutionary tree of the four domains of CaM. At that time, however, only the primary amino acid sequences were available for such purposes, and the observed difference of amino acid residues of CaM between the domains was attributed to the difference of nucleotides between the corresponding codons. Referring to "the genetic code," we assumed that the difference of amino acids was given by the least number of nucleotide substitutions of the codons.

Recently, Lagâce *et al.* isolated m RNA of CaM from the electroplax of electric eel (*Electrophorus electricus*), and its cDNA (DNA complementary to m RNA) has been cloned and sequenced.[6] Moreover, Putkey *et al.* isolated the m RNA from chicken brain and determined the primary nucleotide sequence of its cDNA.[7] For the two biological species, the nucleotide sequences of the amino acid coding regions were aligned and compared in Fig. 1 of Ref. 4. Comparison of the corresponding codons revealed that the situation is more complex than has been expected, that is, the difference of the corresponding codons was not always given by the least number of nucleotide substitutions of codons. Corrections for multiple and revertant changes at homologous sites appear to be important to estimate the evolutionary distances between homologous sequences. It is also found that nucleotide substitutions have occurred most frequently at the third positions of codons and that those substitutions are synonymous with respect to the kinds of the amino acid residues. The synonymous nucleotide substitutions have also occurred at the first positions of codons; such mutations are only allowed in the leucine and arginine residues.

In view of these results, we should examine whether the actual DNA sequences of CaM gene may lead to the same intramolecular evolutionary tree as was given previously by the primary amino acid sequences or not. This was done with the cDNA sequences of the chicken and eel CaM genes. In Fig. 1, we show the nucleotide sequences of the four domains of the genes. For each of the biological species, we construct an intramolecular evolutionary tree of the four domains by using "maximum likelihood method."

Intramolecular Evolutionary Tree and Maximum Likelihood Approach

In order to construct an evolutionary tree in terms of DNA sequence data, several approaches are known, such as maximum parsimony method to estimate the c DNA Sequence of Chicken Calmodulin

Domain 1	CAG ATT GCA GAA TTC AAA GAA GCT TTT TCA CTA TTT GAC AAG GAT GGT
	GAT GGT ACT ATA ACT ACA AAG GAG TTG GGG ACT GTG ATG AGA TCA CTT
Domain 2	ACA GAA GCA GAA TTA CAG GAC ATG ATC AAT GAA GTA GAC GCT GAT GGC
	AAT GGC ACA ATT GAC TTC CCA GAG TTT CTG ACA ATG ATG GCA AGA AAA
Domain 3	AGC GAA GAA GAA ATT AGA GAA GCG TTC CGT GTG TTT GAC AAG GAT GGT
	AAT GGT TAC ATT AGT GCT GCA GAA CTT CGT CAT GTG ATG ACA AAT CTT
Domain 4	ACA GAT GAA GAA GTT GAT GAA ATG ATT AGG GAA GCA GAC ATT GAT GGT
	GAT GGT CAA GTA AAC TAT GAA GAG TTT GTA CAG ATG ATG ACA GCG AAG

c DNA Sequence of Electric-Eel Calmodulin

Domain 1	CAG ATT GCT GAG TTC AAG GAG GCG TTT TCC CTC TTT GAC AAA GAT GGT
	GAC GGC ACC ATC ACC ACC AAA GAG CTG GGT ACT GTG ATG CGC TCT CTG
Domain 2	ACC GAG GCA GAG CTG CAG GAC ATG ATC AAT GAA GTG GAT GCT GAC GGC
	AAT GGA ACA ATA GAC TTC CCG GAG TTC CTG ACC ATG ATG GCC AAG AAA
Domain 3	AGT GAA GAA GAG ATC CGA GAA GCC TTC CGA GTT TTT GAC AAG GAC GGT
	AAT GGC TAC ATC AGT GCA GCC GAG TTG CGA CAT GTC ATG ACT AAC TTG
Domain 4	ACG GAC GAG GAG GTG GAT GAG ATG ATC CGA GAG GCC GAC ATC GAT GGC
	GAC GGC CAG GTG AAC TAT GAA GAG TTC GTG CAA ATG ATG ACT GCA AAG

Fig. 1. The nucleotide sequences of the four domains of the calmodulin (CaM) genes of chicken and eel. For each domain, 96 nucleotides were taken from the c DNA sequence data. See the text and Refs. 6 and 7.

common ancestor sequence and matrix method to estimate pairwise similarity of the sequences.[8] However, most data on DNA sequences involve moderate to large amount of change, such as multiple and revertant changes. In such cases, the maximum parsimony method may fail to construct a correct evolutionary tree, because this method does not take into consideration such probabilistic process of multiple and revertant changes. On the other hand, the matrix method has a difficulty in that it does not make full use of the information available in the original sequences. It has been also pointed out that simple clustering technique based on pairwise similarities can give inconsistent estimates of an evolutionary tree if rates of evolution are sufficiently unequal in different lineages.[9] A third approach invloves method which tries to avoid the above difficulties and to make explicit and efficient use of all of the sequence data by formulating a probabilistic or statistical model of evolution. As such an attempt, Felsenstein proposed an algorithm to evaluate the likelihood of an evolutionary tree.[8] The likelihood of the tree is defined by the product of the probabilities of change in each tree segment, times the prior probability of the ancestral state. The most probable evolutionary tree can be obtained by maximizing such a likelihood value. A computer program was written for the maximum likelihood method.

This method was then applied to our case of CaM genes. We took nucleotide sequence data of the four domains, as shown in Fig. 1. Each domain is composed of 96 nucleotides, which correspond to 32 amino acid residues. For each of the chicken and eel CaM genes, the result computed by the maximum likelihood method shows that the evolutionary tree given by Fig. 2 is the most probable. The maximum likelihood value, log likelihood, is calculated as −450.00 in the case of chicken, while it is −477.53 in the case of eel. Figure 2 shows that the domains 1 and 3 came from the common ancestor of B, and the domains 2 and 4, from the common ancestor of C. The common ancestors, B and C, had diverged from the common ancestor A. The evolutionary tree estimated by the cDNA sequence data coincides well with that described previously by the maximum parsimony method together with the amino acid sequence data.[3] Our previous two-step intragenic duplication model is reconfirmed by the present evolutionary tree (for the reason, refer to Ref. 3).

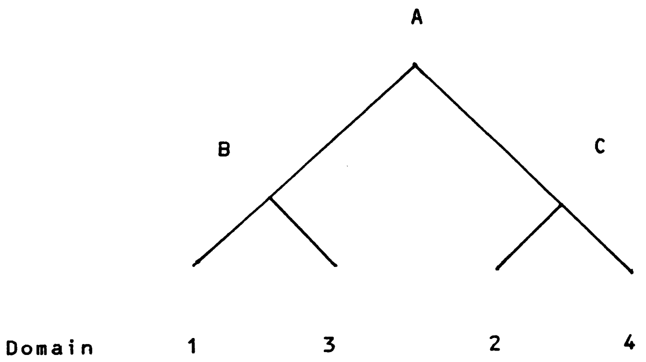

Fig. 2. The intramolecular evolutionary tree of the four domains of the chicken or eel calmodulin (CaM) gene estimated by the maximum likelihood method. This tree shows that the domains 1 and 3 diverged from the common ancestor B, while the domains 2 and 4, from the common ancestor C. B and C had diverged from the common ancestor A.

In Table 1, we show data on the optimal branch length, $\hat{p}$, in each tree segment of the evolutionary tree (Fig. 2), together with its 95% confidence limit value. In each of the biological species, the nucleotide sequence of the domain 1 has longer branch length from the

TABLE 1. DATA ON THE OPTIMAL BRANCH LENGTH, $\hat{p}$, IN EACH TREE SEGMENT OF THE EVOLUTIONARY TREE OF THE FOUR DOMAINS, TOGETHER WITH ITS 95% CONFIDENCE LIMIT VALUE

From	To	Chicken			Electric Eel		
		$\hat{p}$	95% Confidence Limits		$\hat{p}$	95% Confidence Limits	
Domain 1	B	0.4080	(0.2518,	0.5642)	0.4824	(0.3086,	0.6561)
B	Domain 3	0.2247	(0.0591,	0.3904)	0.3047	(0.1103,	0.4991)
B	C	0.3736	(0.2069,	0.5403)	0.3884	(0.1936,	0.5831)
Domain 2	C	0.3539	(0.2034,	0.5044)	0.4627	(0.2958,	0.6296)
C	Domain 4	0.2065	(0.0492,	0.3638)	0.2585	(0.0654,	0.4515)

For the nucleotide sequences of the four domains of chicken or eel calmodulin (CaM) gene, the evolutionary tree is given by Fig. 2.

common ancestor of B than does that of the domain 3. Similarly, the sequence of the domain 2 has longer branch from the common ancestor of C than does that of the domain 4. The distance between any two domains can be calculated by adding the $\hat{p}$ value of each tree path. It is found that distances between domains have appreciably large values of $\hat{p}$, that is, the sum of $\hat{p}$'s is always greater than 1/2. A reason for this is that nucleotides at the third positions of codons have changed considerably.[4] In spite of this, the distance between the domains 1 and 3 or between the domains 2 and 4 is nearest in both of the chicken and eel genes. The $\hat{p}$ data of Table 1 will be valuable for further discussion on the molecular evolution of CaM gene.

We wish to thank Dr. J. Felsenstein for showing us his computer programs.

References

1) Y. Iida, *Bull. Chem. Soc. Jpn.*, **55**, 2683 (1982).

2) Y. Iida, *J. Mol. Biology*, **159**, 167 (1982).

3) Y. Iida, *Bull. Chem. Soc. Jpn.*, **57**, 2665 (1984).

4) Y. Iida, *Bull. Chem. Soc. Jpn.*, **57**, 2667 (1984).

5) See, for example, "Calcium Binding Proteins and Calcium Function," ed by R. H. Wasserman *et al.*, North Holland, New York (1977).

6) L. Lagâce, T. Chandra, S. L. C. Woo and A. R. Means, *J. Biol. Chem.*, **258**, 1684 (1983).

7) J. A. Putkey, K. F. Ts'ui, T. Tanaka, L. Lagâce, J. P. Stein, E. C. Lai and A. R. Means, *J. Biol. Chem.*, **258**, 11864 (1983).

8) J. Felsenstein, *J. Mol. Evol.*, **17**, 368 (1981).

9) D. H. Colless, *Syst. Zool.*, **16**, 289 (1970).