The Peculiar Evolution of Apolipoprotein(a) in Human and Rhesus Macaque

Graziano Pesole,* Angela Gerardi,* Fernando di Jesoti and Cecilia Saccone*

*Dipartimento di Biochimica e Biologia Molecolare, Università di Bari and CSMME-CNR, Bari, Italy and *Dipartimento di Biochimica, I Cattedra di Chimica Biologica, Università di Pavia, and Centro di Biochimica Cardiovascolare, Pavia, Italy

Manuscript received June 8, 1993
Accepted for publication September 27, 1993

ABSTRACT

Apo(a) is a low density lipoprotein homologous to plasminogen and has been shown to be involved in coronary atherosclerosis. In the present paper we will try to analyze the interesting evolutionary pattern of Apo(a). The plasminogen gene contains 5 cysteine-rich sequences, called kringles, followed by a protease domain. Apo(a), probably arisen by duplication of an ancestral plasminogen gene, contains many tandemly repeated copies of a sequence domain similar to the fourth kringle of plasminogen, 37 in human and at least 10 in the partially sequenced gene of rhesus, and the protease domain. We have found that the upstream kringles of apo(a) undergo Molecular Drive-like processes that produce high intraspecies similarity, whereas the downstream kringles evolve in a molecular clock-like manner and show an high interspecies sequence similarity. The latter regions are obviously suitable for dating the duplication event by which Apo(a) arose from plasminogen, but only if they evolve at the same rate in the two genes. Thus, we propose a "Molecular Clock Test" for assessing whether the comparison of two paralogous genes (or gene regions) can give reliable information on the dating of their origin by duplication. Applying this test to the kringle-4 domain of apo(a) and plasminogen gene, we demonstrate that the separation between the two genes by duplication dates back at about 90 Mya immediately before the radiation of mammals.

Lp(a) is a low density lipoprotein (LDL) that carries one copy of a protein called Apo(a) joined to apo B-100 by a disulphide linkage. High levels of Lp(a), i.e., above 30 mg/dl, double the risk of coronary atherosclerosis (Berg, Dalhen and Frick 1974; Utermann 1989). When LDL and Lp(a) both have a high concentration, the relative risk rises about five-fold (cf. Brown and Goldstein 1987).

McLean et al. (1987) showed that Apo(a) is a functionally altered relative of plasminogen, the precursor of plasmin, which dissolves fibrin clots. Plasminogen, a protein of 791 amino acids, contains 5 cysteine-rich sequences of 80–114 amino acids each, called kringles, followed by a serine protease domain. Each kringle contains three internal disulphide bridges, producing a structure resembling the Danish cake from which it derives its name. Kringles are also found in other proteases of the coagulation system, including tissue plasminogen activator (TPA) and prothrombin (Pattthy 1985). In plasminogen, kringles promote the binding to the substrate fibrin (cf. Brown and Goldstein 1987).

Apo(a) is one of the most polymorphic expressed sequences in the human genome as to date, at least 34 alleles, differing in the number of tandemly repeated sequences homologous to kringle 4 of plasminogen, have been characterized (Lackner et al. 1991; Lackner, Cohen and Hobbs 1993).

The Apo(a) gene has been completely sequenced in human and partially in rhesus monkey. The human gene contains a hydrophobic signal sequence followed by: 37 copies of a sequence domain similar to the fourth kringle of plasminogen, one copy of a sequence homologous to plasminogen kringle 5, and the protease domain. The rhesus Apo(a) sequence, which is still undetermined at the N-terminus, contains at least 10 copies of kringle 4-like sequence domain and the protease domain (Tomlison, McLean and Lawn 1989).

The protease domain contained in Apo(a) is likely to have lost the proteolytic activity because the arginine in the site cleaved by TPA in plasminogen is changed into serine both in human and rhesus macaque. In addition, the three amino acids of the catalytic triad are intact in human but two of them have been mutated in rhesus macaque.

The evolutionary origin of Apo(a) and its functional role is still obscure. To shed light on this issue, we report an accurate evolutionary analysis on Apo(a) and plasminogen genes for Human and Rhesus Macaque. We have found that the various regions of the two genes have a different evolutionary behavior. This implies that each protein may evolve under different evolutionary pressures. In particular, some domains undergo Molecular Drive-like processes that homogenize their nucleotide sequence composition,
and as a consequence these regions appear more similar within the species than between species. On the other hand, other domains evolve in a molecular clock-like manner. The latter regions are obviously suitable for dating the duplication event but only if they evolve at the same rate.

In this paper we propose a "Molecular Clock Test" for assessing whether the comparison of two paralogous genes (or gene regions) can give reliable information on the dating of their origin by duplication. Applying this test to the kringle-4 domains of Apo(a) and plasminogen genes, we demonstrate that the separation between the two genes by duplication is more ancient than 40 Mya, during primate evolution, as claimed by McLean et al. (1987); but it dates back at about 90 Mya immediately before the radiation of mammals. This implies Apo(a) should be present not only in primates but also in other mammals.

MATERIALS AND METHODS

Human and Rhesus Macaque Apo(a) and plasminogen genes were extracted from the EMBL database (accession numbers: X06290, J04665, X05199, J04697).

The multiple alignment of nucleotide sequences used in the evolutionary analysis has been constructed following that of protein sequences calculated by using the program PILEUP (GCG Package, 1993). The multiple alignments are available upon E-mail request from the authors at PE-SOLE@MVX36.CSATA.IT.

The nucleotide substitution rates have been calculated by using the Stationary Markov Process (SMC) program developed in our laboratory (Saccone et al. 1990) or the Li method (Li 1993). NEIGHBOR and DRAWGRAM programs of the PHYLIP package were used to draw phylogenetic trees (Felsenstein 1990) using relative times of divergence calculated by the SMC.

RESULTS

Evolutionary analysis of serine protease domain: It has been reported that bovine and human plasminogen genes, in both the protease domain and in the 3'-untranslated region, are more divergent than are the human plasminogen and human Apo(a) sequences, thus suggesting that Apo(a) comes from a recent duplication event (McLean et al. 1987). An alternative explanation, supported by the observation that Apo(a) and plasminogen are located very close to each other on chromosome 6 (Frank et al. 1988), is that the high sequence similarity between Apo(a) and plasminogen derives from exchange of DNA sequence as a result of Molecular Drive phenomena.

The evolution of the protease domain has already revealed interesting features. Otha and Basten (1992) have observed gene conversion with selection in proteases, which is made evident by a higher substitution rate of the reactive center with respect to the other regions.

The protease domains of Apo(a) and plasminogen genes are paralogous, i.e., they are the descendants of a duplicated ancestral gene (Creighton and Darby 1989). After gene duplication, they are likely to have evolved under different functional constraints. If nucleotide substitutions accumulate at different rates in the two paralogous genes, their comparative analysis cannot provide quantitative distance estimates. An exact estimate of the date of gene duplication can only be made if the two genes show to have evolved at comparable rates.

If the same Clock applies to Apo(a) and plasminogen, the same nucleotide substitution rates on synonymous and nonsynonymous positions should be observed for both genes in human-rhesus macaque comparisons. Thus, we calculated the nucleotide substitution rate of both Apo(a) and plasminogen between human and rhesus.

Table 1 shows the nucleotide substitution rates, calculated with both SMC and Li method on synonymous and nonsynonymous positions of Apo(a) and plasminogen, between human and rhesus monkey. In Apo(a) the replacement rate is significantly higher than in plasminogen and is comparable to the synonymous rate. This can be explained if the protease domain of Apo(a), deprived of its original proteolytic activity, began to accumulate neutral mutations in all codon positions, like a pseudogene. Alternatively, if the protease domain of Apo(a), after duplication, acquired a new functional role, the higher substitution rate observed on nonsynonymous positions, could be the result of positive selection (Creighton and Darby 1989). In both cases the Molecular Clock is not obeyed and quantitative estimates of gene distance are unreliable.

Evolutionary analysis of the 3'-untranslated region: 3'-noncoding regions of Apo(a) and plasminogen mRNAs, each about 250 nucleotides long, show high sequence similarity, 87% in human and 88% in rhesus monkey.

McLean et al. (1987), assuming that differences accumulate in noncoding regions at a rate of about 0.3% per million years, as in noncoding sequences of the beta-globins of human and chimpanzee, dated the duplication about 40 million years ago, during primate evolution.
TABLE 2

<table>
<thead>
<tr>
<th></th>
<th>Nucleotide substitution rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plasminogen</td>
<td>1.46 ± 0.48</td>
</tr>
<tr>
<td>Apo(a)</td>
<td>3.28 ± 0.74</td>
</tr>
</tbody>
</table>

The rates are statistically different as proved by the t-test (t = 20.5, \( P = 0.00 \)).

mate evolution. If this is the case Apo(a) gene should be restricted to primates only.

We applied the same test shown before, i.e., the estimate of the nucleotide substitution rate, to the orthologous genes of human and rhesus monkey.

Table 2 shows the nucleotide substitution rate calculated between human and rhesus monkey 3'-non-coding regions in both Apo(a) and plasminogen. It is striking to note that the nucleotide substitution rate is more than double in Apo(a) with respect to the plasminogen. This demonstrates that the comparison of the untranslated regions of Apo(a) and plasminogen cannot provide reliable information about sequence divergence, i.e., they cannot be used as Molecular Clock. In the light of these data, the proposal of McLEAN et al. (1987) that the gene duplication occurred 40 Mya is unreliable.

**Evolutionary analysis of the kringle domain: Human Apo(a) gene contains 37 copies of a sequence resembling the plasminogen kringle 4.** For the rhesus macaque Apo(a) gene, only partially sequenced, 10 copies of the kringle 4 sequences are available. Indeed, it has been shown that the human alleles can contain a variable number of kringle 4 sequences, in the range 12–51 (LACKNER et al. 1991; LACKNER, COHEN and HOBBS 1993). A similar size variability has been detected in baboon (HIXSON et al. 1989).

Apo(a) kringle will be in the following denoted K41–K457, according to McLEAN et al. (1987). All cysteine residues, involved in the three disulfide bonds which generate the typical kringle structure are maintained in both human and rhesus monkey Apo(a). In K458 Apo(a) presents an extra unpaired cysteine, responsible for the disulphide linkage with apo B100. Human and Rhesus macaque K433 share a 8 aminoacid deletion.

The 37 kringle of humans can be classified into 11 different types: K42-23 and K425-26 are identical; K424 and K427-29, which differ from the previous kringle repeats by three nucleotides only; the remaining kringles, K41 and K430-37, are unique sequences. In rhesus monkey, all 10 kringles, K428-37, are unique sequences.

The presence of tandemly repeated kringle domains showing even 100% similarity suggest that the phenomenon that produces locus expansion or contraction is currently at work in Apo(a) gene, and thus individuals of the same species may present size polymorphic alleles with a variable number of kringle repeats. Indeed, considerable size heterogeneity is shown by human Apo(a), which exists as discrete glycoprotein isoform variants that range a molecular mass from approximately 400–800 kDa (GAUBATZ et al. 1990; KOSCHINSKY et al. 1990).

Northern blot analysis has demonstrated that transcript sizes are variable (8–12 kb) and in all cases closely correlated with protein masses as determined from immunoblots. Thus, it is very likely that Apo(a) isoform size variation is due to allelic differences in the number of tandemly repeated kringle sequences (KOSCHINSKY et al. 1990). LACKNER, COHEN and HOBBS (1993) estimated that at least 34 different alleles can be detected, and isoforms with as few as 12 and as many as 51 kringle repeats can be found in human plasma. The above authors have proposed that the high degree of length polymorphism is partly due to recombination between sister chromatids.

The SMC method has been applied to all kringle repeats of human and rhesus Apo(a) in order to try the reconstruction of their evolutionary origin. Figure 1 shows the calculated phylogenetic tree obtained by averaging the results on synonymous and non synonymous codon positions.

It is interesting to note that, starting from K431, each kringle in human is more closely related to the correspondent kringle in rhesus monkey than to the other kringles of the same gene. This means that the multiplication of kringle-4-like repeats, at least for the eight downstream kringles, occurred before the human-rhesus monkey split and not as recently as 2–3 Mya as claimed by IKEO, TAKAHASHI and GOJOBORI (1991). Indeed, if this was the case, considering that the split between human and rhesus monkey is dated at about 25 Mya (SACCONE et al. 1990) or 45 Mya as reported by IKEO, TAKAHASHI and GOJOBORI (1991), we should observe that all human kringle-4-like repeats clustered together as well as the rhesus monkey counterparts. This happens only for upstream kringles that appear instead more closely related between themselves than to the corresponding regions of the other species.

This supports the hypothesis that K451-37, the seven downstream kringles of Apo(a), unlike upstream kringles were not subject to unequal crossing-over or gene conversion events, two phenomena involved in the Molecular Drive, which generates sequence homogeneity between DNA tracts in the same species (DOVER 1982).

Looking at the phylogenetic tree shown in Figure 1, we can also argue that the ancestral Apo(a) gene, before the human-rhesus monkey split, contained at
least eight kringle-4-like repeats. The higher intraspecies similarity of Apo(a) upstream kringles suggests that they were affected by Molecular Drive mechanisms, which homogenized sequence divergence. In this context, only last seven kringles, unique sequences, are potentially useful for drawing quantitative evolutionary estimates.

To verify this hypothesis, we have carried out the Molecular Clock test shown before for protease domain and noncoding regions. We have calculated the synonymous and nonsynonymous rates for the kringles of Apo(a) and the K4 kringle of plasminogen between human and rhesus monkey. The data reported in Table 3 show that Apo(a) downstream kringles have an evolutionary rate comparable to that of kringle 4 of plasminogen, both at synonymous and nonsynonymous positions, in the human-rhesus monkey comparison. Therefore, this sequence domain in Apo(a) is under the same evolutionary pressure as the homologous domain in plasminogen and thus it can be used as a reliable Molecular clock.

Figure 2 reports the phylogenetic tree calculated by comparing Apo(a) K431-37 and a seven-fold repeat of kringle 4 of plasminogen between human and rhesus monkey. It is striking to note that the divergence time between human and rhesus, calculated on K431-37, is the same as that calculated on plasminogen K4. In addition, both nonsynonymous and synonymous positions give the same tree, thus supporting the

| Evolutionary rate (subs/site·10^-8) calculated according to the Stationary Markov Process (SMC) and Li methods between human and rhesus monkey kringle domain (downstream kringles 31–37 only) of Apo(a) and Plasminogen |
|---------------------------------|-------|
| Nonsynonymous rate              | Synonymous rate |
| Synonymous rate                 | Synonymous rate |
| SMC                             | Li     | SMC                  | Li                  |
| Apo(a)                          | 0.80 ± 0.27 | 0.82 ± 0.10 | 1.10 ± 0.38 | 1.40 ± 0.22 |
| Plasminogen                     | 0.70 ± 0.20 | 0.68 ± 0.08 | 1.15 ± 0.35 | 1.60 ± 0.24 |

Figure 3.—Schematic representation of human and rhesus monkey Apo(a) genes.
reliability of this sequence region for quantitative estimates. Fix the time of divergence between human and rhesus monkey at 25 Mya (SACCONE et al. 1990), then the gene duplication which originated Apo(a) can be dated about 93 Mya, much earlier the 40 Mya estimated by McLEAN et al. (1987) and in line with the 80 Mya estimate of IKEO, TAKAHASHI and GOJIBORI (1991). This means that Apo(a) is not restricted to primates and might be found in other mammals. These conclusion is in accord with the reliable human/rodent divergence time of 75-80 Mya. However, it is likely that not all mammals have a functional Lp(a) particle. Indeed, at least three distinct events must happen after duplication of plasminogen to create functional Lp(a) particles: (a) development of the crucial cysteine, which allows the disulphide bridge interaction with the apo-B100 in LDL; (b) large expansion of the kringle 4 copy number; (c) loss of the protease activity (probably by alteration of the arginine at the site cleaved in plasminogen when the zymogen is converted to the active plasmin).

**DISCUSSION**

Two genes are said to be paralogous if they are derived from a duplication event, orthologous if they are derived from a speciation. Apo(a) and Plasminogen are thus paralogous genes and the time of gene duplication can be reliably estimated by measuring the nucleotide substitutions between the two genes if both gene evolved at the same rate. This can be verified through a Molecular Clock test that counts the nucleotide substitution rate for both Apo(a) and plasminogen genes or portions of them separately in orthologous comparisons (i.e., human vs. rhesus monkey). This test has been applied to the 3′-untranslated region, to the protease domain and to the K4 region of the two genes in the couple human-rhesus monkey. This test has shown that only the K4 region, in particular K4s1-s7, may be suitably used for dating gene duplication. This DNA segment indeed, both in Apo(a) and in plasminogen evolves at the same rate between human and rhesus monkey. Fixing the divergence time human/rhesus monkey at 25 Mya (SACCONE et al. 1990), the duplication event dates back about 90 Mya, immediately before the radiation of mammals that occurred about 75-80 Mya. The same divergence time was obtained when the K4 of mouse plasminogen was included in the analysis and the very reliable human/rodent divergence time of 75 Mya was used as calibration (DAYHOFF 1972). If the duplication event has been dated correctly, Apo(a) may not be restricted to primates and might be found in other mammals. These conclusion is in accord with the earlier finding of LAPLAUD et al. (1988) and RATH AND PAULING (1990), who detected Apo(a) in the hedgehog and guinea pig, nonprimate mammals.

One of the outstanding result of this work—the intriguing evolutionary pattern of Apo(a) kringle repeats—is schematically shown in Figure 3. These repeats can be clearly classified into two groups, the upstream kringles showing a higher intraspecies similarity that denotes the presence of Molecular Drive mechanisms; and the downstream kringles that, on the contrary, show higher interspecies similarity and a clocklike behavior. Obviously, due to the stochastical nature of concerted evolution processes, more sequence data might confirm more strictly the clear-cut division between the two kringle regions.

The expansion/contraction of Apo(a) gene locus is also cosupported by the data of HELMOLD et al. (1991), LACKNER et al. (1991) and LACKNER, COHEN and HOBBIS (1993), who investigated Apo(a) polymorphism in various ethnic groups. They found that the greatest ethnic variation is observed in plasma Lp(a) concentrations associated with the high molecular weight Apo(a) polymorphs. Our data suggest that contraction/expansion of gene locus occur at level of the K4 region, but only upstream kringles are involved.

The biological significance of polymorphic forms of Apo(a) is still unclear but a highly significant inverse correlation was found between the molecular weight of Apo(a) and the plasma Lp(a) concentration (BERG, DAHLN and FRICK 1974; UTERMANN 1989). This suggests that locus expansion should preserve from the risk of coronary atherosclerosis.

Another interesting outcome of our evolutionary analysis is that the serine protease domain of Apo(a) has should have not conserved its protease activity.

This work was financed by Progetto Finalizzato Biotecnologie e Bioinstrumentazioni (CNR, Italy), Progetto Finalizzato Ingegneria Genetica (CNR, Italy) and Murst, Italy.

**LITERATURE CITED**


DAYHOFF, M. O., 1972 Atlas of Protein Sequence and Structure, Vol. 5, National Biomedical Research Foundation, Silver Spring, Md.


LAPLAUD, P. M., L. BEAUBATTE, S. C. RALL, JR., G. LUC and M. SABOUREAU, 1988 Lipoprotein(a) is the major apolB-containing lipoprotein in the plasma of a hibernator, the hedgehog (Erinaceus europaeus). J. Lipid Res. 29: 1157–1170.


Communicating editor: W.-H. Li