Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 4182–4185, June 1985
Genetics

Evolutionary amplification of a pseudogene

(mouse major urinary protein/*Mup* genes/nonsense mutation/gene family)

P. GHAZAL, A. JOHN CLARK, AND JOHN O. BISHOP

Department of Genetics, University of Edinburgh, Edinburgh EH9 3JN, Scotland

Communicated by Alan Robertson, February 20, 1985

ABSTRACT **The family of mouse major urinary protein (MUP) genes has about 35 members, clustered together on chromosome 4. Most of the genes belong to two major subfamilies (group 1 and group 2) each with 12–15 members. Recently we showed that most of the group 1 and group 2 genes are arranged in pairs, each containing a group 1 and a group 2 gene in divergent transcriptional orientation, with 15 kilobases of DNA between the two cap sites. Here we present the nucleotide sequence of the first exon of six group 1 genes and four group 2 genes. The data confirm the close relationship of the genes within each group and the considerable divergence of the two groups from each other. The four group 2 genes all carry the same nonsense mutation in codon 7 of the sequence that specifies the mature protein. Thus, not only do these genes have a common ancestor, but also it seems that their amplification followed the mutation of the ancestor to a pseudogene. Taking into account the 3′ flanking regions of the two genes, the overall size of each gene-pair is about 45 kilobases. The sequencing data supports our earlier suggestion that this 45 kilobase domain is the unit of *Mup* amplification.**

The mouse major urinary protein (MUP) is a family of closely related polypeptides that are synthesized and secreted by the liver and excreted in the urine (1, 2). MUP mRNA makes up about 5% by weight of male liver mRNA (3, 4). Smaller amounts of biologically active mRNA are found in the lachrymal, salivary, and mammary glands. *In vitro* translation of hybrid-selected MUP mRNA from the different tissues shows that each directs the synthesis of a different subset of MUP polypeptides (5). The level of MUP mRNA in the liver is influenced by insulin, growth hormone, thyroxine, and testosterone (6). *In vitro* translation of mRNA from livers taken from mice maintained under different hormonal regimes shows that different species of mRNA (directing the synthesis of different polypeptides) respond differently to the various hormones. Testosterone is known to increase the rate of synthesis of MUP mRNA (7). The mouse genome contains about 35 MUP genes, defined as sequences that hybridize with MUP-specific probes. Most of these can be assigned to two main groups, group 1 and group 2, by hybridization with two canonical group 1 and group 2 probes (8). Most of the group 1 and group 2 genes are arranged in head-to-head (divergently orientated) pairs (9). Each pair contains a group 1 and a group 2 gene, homologous 5′ flanking sequences (two of 5 kb), 3′ flanking sequences (two of 11 kb) that contain regions of homology interspersed with nonhomologous regions, and 6 kb of DNA (located between the homologous 5′ flanking sequences) that is not duplicated within the pair. The overall size of the head-to-head pair, from the far end of one 3′ flanking sequence to the far end of the other, is about 45 kb. We have argued that this is the principal unit of MUP gene organization and evolution (9). Here we show that four group 2 genes are pseudogenes, in the sense that they contain at least one stop codon in the MUP reading-frame (the reading-frame of the group 1 genes). All of these genes contain the same stop codon in exon 1, showing that they are derived from a common ancestral pseudogene.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

MATERIALS AND METHODS

The MUP genes studied here were isolated from genomic clones that have been described (8–10). Plasmid subclones and M13 mp8 and M13 mp9 subclones were isolated by standard methods and sequenced as described (11).

RESULTS AND DISCUSSION

Four Group 2 MUP Genes Are Pseudogenes. The structure of the 45-kb gene pair is shown in Fig. 1*A*. Fig. 1*B* shows the seven-exon structure (12) of the group 1 MUP genes. The nucleotide sequences of the first exon of nine different MUP genes are summarized in Fig. 2. All of these were isolated from nuclear DNA of inbred BALB/c mice and, therefore, are different members of the gene family rather than allelic variants. They are all known to be different genes either because their sequences differ or because the genes themselves or their flanking regions contain different restriction enzyme recognition sites or because of deletions or insertions in their flanking sequences (8–10). Three of the nine genes were taken from clones (BS102-2, BS109-1, and BS109-2) that contain the central portion of a 45-kb gene pair, including the 5′ end of a group 1 gene and its 5′ flanking sequence and the 5′ end of a group 2 gene and its 5′ flanking sequence (see Fig. 1). Thus, these genes are definitely known to be part of the predominant 45-kb gene-pair organization (9). The other six genes are presumed also to be derived from 45-kb gene-pair units on the basis of restriction site homologies in their 5′-flanking regions.

Four of the five group 1 genes (from clones BS1, BS5, BS6, and BL1) have identical exon 1 sequences. The fifth, from clone BS109-1, differs from the others in only two nucleotides, both in the leader sequence. Fig. 2 shows the sequence of the four identical group 1 genes and the deviations from that sequence found in the other genes. It is convenient to consider separately the leader sequence, the signal peptide region, and the remainder of exon 1, the region coding for the first 14 amino acids of the mature group 1 protein.

In the leader sequence, four of the five group 1 genes and three of the four group 2 genes are identical. These define group 1 and group 2 consensus sequences, which differ in 11 nucleotides (11/65 = 17%). In addition, the group 1 consensus sequence is 1 nucleotide longer than the group 2 consensus. One of the group 1 genes, from clone BS109-1, differs from the consensus in two positions. Similarly, one group 2 gene, from clone BL25, differs from the group 2 consensus in four positions and is the same length as the group 1 leader sequence, rather than 1 nucleotide shorter.

Abbreviations: MUP, mouse major urinary protein; kb, kilobases.

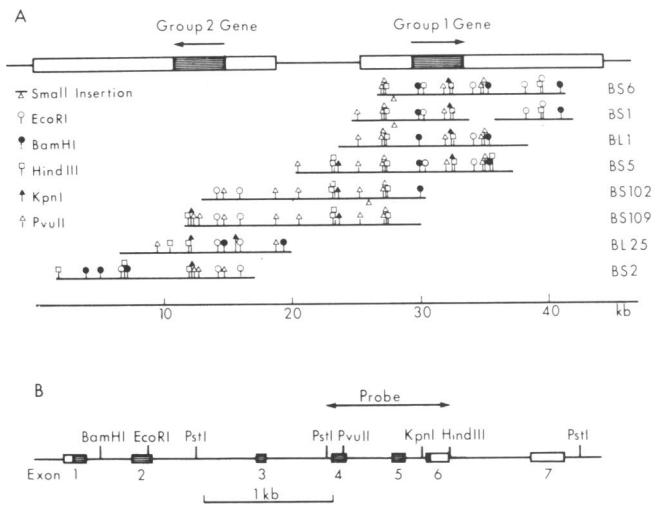

FIG. 1. Organization and structure of MUP genes. (*A*) The top line is a diagrammatic representation of the 45-kb unit. The group 1 and group 2 genes are shaded. Open rectangles are regions of homology between the flanking regions of group 1 and group 2 genes. The homology is not continuous over the 11-kb 3′ flanking region but rather is interrupted by regions of nonhomology. Restriction site maps of the genomic MUP clones used in this study are aligned with the diagram. Three small insertions and a 1.9-kb deletion are proposed in order to maximize the degree of restriction site homology between the clones. The isolation of these clones is described in refs. 8–10. (*B*) Structure of a group 1 MUP gene (12). Exons are shown as boxes, and introns, as lines. The coding region is shaded. The region of the homologous canonical group 1 and group 2 probes is shown.

The nucleotide sequence of the signal peptide region is identical in all five group 1 genes and specifies a signal peptide 18 amino acids long. In contrast, the signal peptide regions of the four group 2 genes (defined as the sequence from ATG to the NH_2 terminus of the mature group 1 protein) are all different. The signal peptides that they specify vary in length from 19 (clone BL25) to 25 (clone BS102-2) amino acids. Most of the additional codons are CTG (leucine) codons that may have arisen from adjacent CTG codons by polymerase "slippage" during replication or by unequal crossing-over. To either side of the additional codons, two of the four group 2 genes are identical and differ from the group 1 genes at two positions. BS109-2 and BL25 contain further nucleotide differences in the signal peptide region.

In the third region of exon 1, which corresponds to the NH_2-terminal 14 amino acids of the mature group 1 proteins, the group 1 genes are again identical. The group 2 genes show a clear consensus, which differs by five nucleotides from the group 1 sequence (5/42 = 12%). Clones BS102-2 and BL25 each differ from the group 2 consensus in one position in this region.

One of the differences between the group 1 and group 2 consensus is between a glycine (GGA) in the group 1 sequences and a stop codon (TGA) in the group 2 sequences (Fig. 2, amino acid 7, position 160). Thus, in the context of the group 1 genes, all four group 2 genes are pseudogenes and contain an identical lesion. Other lesions are also present. BL25 contains a stop codon in place of amino acid 2 of the mature protein (Fig. 2, position 145), and BS2 contains a second stop codon and a frameshift mutation (unpublished data). However, the stop codon that is common to the group 2 genes is their most significant feature, implying as it does that it was present in an ancestral gene, which was therefore also a pseudogene and was ancestral to all four group 2 genes shown in Fig. 2.

We identify group 1 and group 2 genes on the basis of their hybridization with two homologous genomic probes (8) that contain exons 4, 5, and 6 (Fig. 1*B*). So far we have isolated only four group 2 genes that contain exon 1. Since all of these contain the common stop codon, it is likely that all of the approximately 12 group 2 genes in the BALB/c genome share this lesion and are descended from the same ancestral pseudogene.

Evolutionary Divergence of Group 1 and Group 2 Genes. The complete nucleotide sequences of a group 1 gene (clone BS6) and a group 2 gene (clone BS2) have been determined (unpublished data). The coding regions (excluding the signal peptide region) have been identified and compared (13) with each other and with a homologous rat α_{2u}-globulin gene (14–16). The replacement site divergence of the group 1 (clone BS6) and group 2 (clone BS2) mouse genes is ≈10% (BS6 × BS2 = 10.3%) while the divergence of each mouse gene from the rat α_{2u}-globulin (clone 207) gene is ≈20% (BS6 × 207 = 19.1%; BS2 × 207 = 22.4%).

The evolution of a multigene family is more complex than the evolution of a unique gene. In the latter case, the divergence time of two contemporary genes in different species can be taken to be the time since the divergence of the two phylogenetic lines from their common ancestor. In the case of a multigene family, genes that already have diverged from each other coexist within the same genome. The contemporary MUP genes show many examples of this, with divergences that vary from 1% (different group 1 genes) to 10% (group 1 genes compared with group 2 genes). Thus, extrapolating backwards in time, it is quite possible that genes ancestral to the group 1 genes, the group 2 genes, and the rat genes had already diverged from each other in the common ancestor of rats and mice.

The members of a multigene family do not necessarily diverge within a species at rates comparable to the divergence of single genes between species. Indeed, there is strong evidence to the contrary in the present case. A set of rat α_{2u}-globulin cDNA clones, which presumably represent the more abundantly transcribed genes of the rat multigene family, are all identical in sequence (14) and very similar to the corresponding regions of a gene (15). Similarly, the abundantly transcribed group 1 MUP genes are very closely related. The rat genes and the group 1 MUP genes must have arisen from a common ancestral gene, and yet they differ by about 20% in nucleotide sequence, while at the same time different rat genes differ from each other by only 1–2% and different group 1 MUP genes differ from each other to a similar extent. The group 2 MUP pseudogenes are a third reasonably homogeneous group of genes that differ from the rat genes by about 20% and from the group 1 MUP genes by about 10%.

The explanation of this phenomenon presumably relates to the clustering (17, 18) of both the group 1 and the group 2 MUP genes (8, 9) on mouse chromosome 4. One possibility is that the ancestor of rats and mice contained rather few urinary protein genes and that different members of that small set of genes were separately amplified, by tandem duplication, in the rat and mouse lines. According to this view, the group 1 and group 2 MUP genes would have been amplified together within the 45-kb unit of genomic organization. If the common group 2 nonsense mutation arose prior to or early in the course of this amplification, it could have been carried passively, so to speak, through the amplification process, in effect, as an inert DNA sequence within the 45-kb unit.

However, it is unlikely that separate amplification processes occurred independently in the rat and mouse lines. It seems more probable that the genes were already amplified in the common ancestor. If so, what we have to explain is an apparently concerted evolution of evolutionarily diverging arrays of genes in each of the two lines. One unavoidable implication of this model is that the ancestral gene array must have been lost or replaced in one or both of the descendant

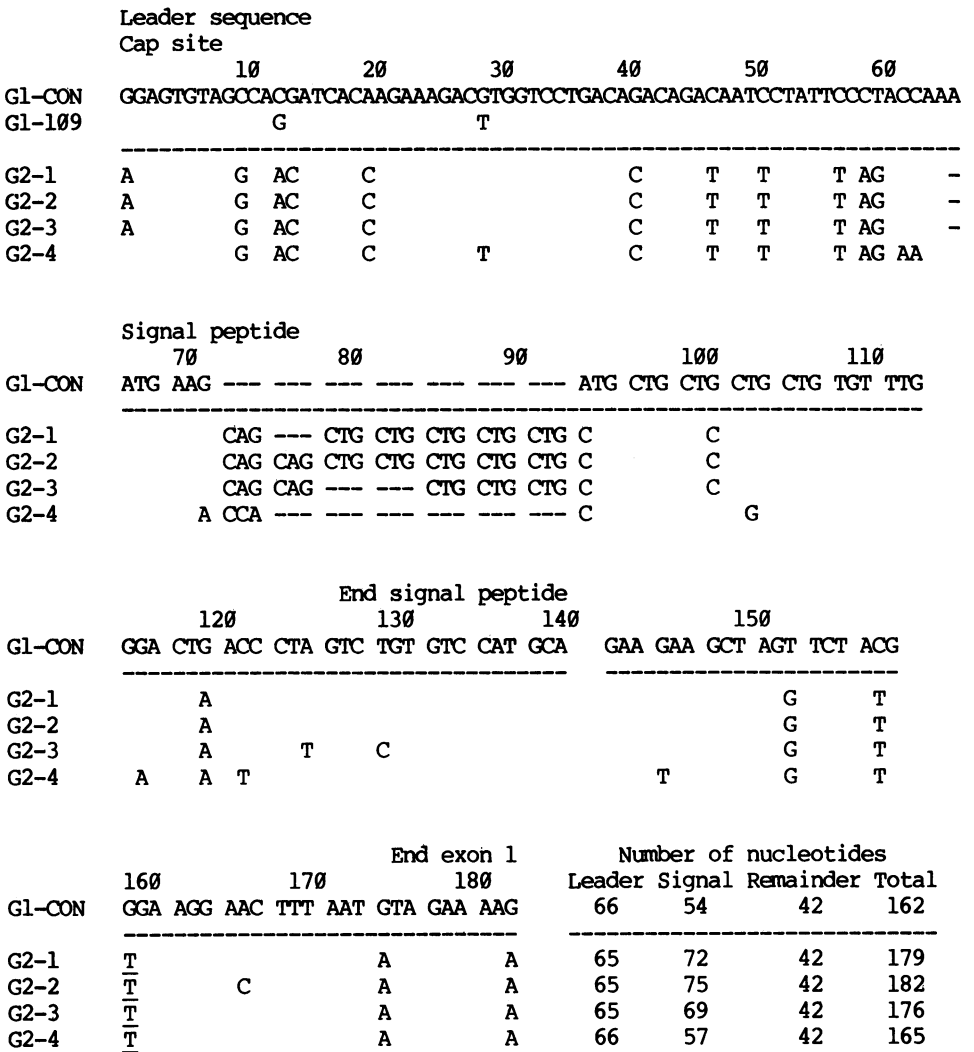

FIG. 2. MUP gene exon 1 sequences. The complete sequence of exon 1 of nine MUP genes is shown. Above the dashed line, five group 1 (G1) genes are shown: Four of these, clones BS1, BS5, BS6, and BL1, are identical in exon 1 and are shown as the sequence labeled G1-CON (for consensus). The fifth, G1-109, differs from the consensus in only two positions. Below the dashed line, the differences between four group 2 (G2) genes and the group 1 consensus are shown. G2-1, G2-2, G2-3, and G2-4 are, respectively, clones BS2, BS102-2, BS109-2, and BL25. The absence of a nucleotide relative to other sequences is signified by a dash. The G→T stop-codon mutation (nucleotide 160) is underlined. The cap site was defined by S1 nuclease mapping and primer extension (unpublished data).

lines. The contemporary arrays would have been developing at the same time. As in the case of the simpler model, the principal unit of MUP gene evolution would be the 45-kb gene pair.

The urinary protein genes of rats and mice invite comparison with the rDNA of *Xenopus laevis* and *X. borealis*. The spacer sequences of the tandemly arranged rDNA genes have diverged widely between the two species, but within each species they are relatively homogeneous (19). This has been explained by a model incorporating two main features: unequal sister-strand crossing-over within the tandem arrays and selective constraints on their size (20). Under these circumstances it can be shown that the entire contemporary array in a given species may be directly descended from a single member of the array at some past time. Thus, genetic drift can go hand-in-hand with the preservation of homogeneity within the array. The degree of homogeneity preserved will depend on the mutation rate, frequency of unequal crossing-over, selection pressure, and so on (21). At least some of the 45-kb MUP gene pairs are arranged tandemly and in direct orientation (9). If this arrangement were general, the unequal crossing-over model would provide a sufficient explanation for the replacement of the ancestral array with a new one.

The phenomenon also can be explained on the basis of gene conversion (22, 23). Unequal crossing-over and gene conversion were discussed previously in relation to MUP gene evolution (9). Ohta (21) has shown that unequal crossing-over and gene conversion can provide formally equivalent explanations of the concerted evolution of a gene family.

We have suggested two models to explain the contemporary relationships of the rat genes and the group 1 and group 2 MUP genes: (*i*) separate *de novo* amplification of different genes in the rat and mouse lines and (*ii*) the replacement of a preexisting array by a new one in each line by unequal crossing-over, gene conversion, or both. The idea central to both models, that the unit of MUP gene amplification is the 45-kb gene pair, is suggested by the 10–15 copies of the gene pair that are present in the genome of the laboratory (BALB/c) mouse (9). The same idea also can explain the divergence of the group 1 and group 2 genes during a time period in which each group remained reasonably homogeneous. According to the models, the two main parts of the unit, the group 1 and group 2 genes and their respective flanking sequences, cannot replace each other and so would have been able to diverge. On the other hand, the 45-kb unit as a whole replaces other 45-kb units, and it is to this that we would attribute much of the uniformity of the 45-kb units and,

in particular, the uniformity of the group 1 and group 2 genes themselves. Thus, we would date the onset of divergence of the group 1 and group 2 genes to a time close to that at which the 45-kb unit originated (presumably by an inversion), whether in the mouse line or in a line ancestral to the divergence of mice and rats.

The group 1 genes are more homogeneous than the group 2 genes (ref. 8; Fig. 2). This can most easily be explained by supposing that selective constraints are superimposed on the amplification–replacement process. Selection acting against an unfavorable newly arisen group 1 gene would tend to lead to the elimination of the 45-kb unit within which it was located. Similarly, selection may have maintained the homogeneity of the group 2 genes up to the time at which the pseudogene mutation occurred but presumably did not operate on the pseudogene and its descendants. If so, mutational changes in the group 2 pseudogenes would have accumulated more rapidly. This raises the question as to why the group 2 pseudogene mutation was tolerated in the first instance. The most satisfactory explanation is that the inversion and the pseudogene mutation arose at about the same time. If this is the case, we can view (*i*) the homogeneity of the group 2 genes as a function of the concerted evolution of 45-kb units, driven by the group 1 genes, and (*ii*) the inhomogeneity of the group 2 genes as a function of the underlying mutation rate, acting against the homogenization process but unaffected by selection.

T. Ohta writes (personal communication):

> The theory of concerted evolution is already available and is applicable to the present data. Referring to Ohta (21), let us assume the following parameter values: n (no. of amplification units) = 10, N (effective population size) = 10^4, v (mutation rate of nucleotides per generation) = 10^{-9}, β (interchromosomal recombination rate between units) = $10^{-4} \sim 10^{-6}$, and λ (rate by which a unit is replaced by another unit, or the rate of one cycle of unequal crossing-over or duplication–deletion) = 10^{-6}. Then the average divergence between the nonallelic genes belonging to the family becomes about 1%. By using the same set of parameter values except that the mutation rate (v) is five times as large (5×10^{-9}), one gets an average divergence of about 4.5%. The former is appropriate for group I, and the latter, for group II genes.
>
> The above application has several implications. (*i*) Even if tentative, the effective rate of unequal crossing-over is estimated. (*ii*) In the above model, nucleotide substitution is assumed to be selectively neutral. In view of available data, most nucleotide substitutions are neutral [Kimura (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, England], and the present case is not likely to be an exception. Group II genes are free to change and the rate is high. (*iii*) The time for spreading of a unit is estimated with the above set of parameters to be about 10^7 generations (see Ohta, *Genet. Res.* **41**, 47–55).

This work was supported by the Medical Research Council and the Cancer Research Campaign.

1. Rümke, Ph. & Thung, P. J. (1964) *Acta Endocrinol.* **47**, 156–164.
2. Finlayson, J. S., Asofsky, R., Potter, M. & Runner, C. C. (1965) *Science* **149**, 981–982.
3. Hastie, N. & Held, W. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 1217–1221.
4. Clissold, P. M. & Bishop, J. O. (1981) *Gene* **15**, 225–235.
5. Shaw, P. H., Held, W. & Hastie, N. D. (1983) *Cell* **32**, 755–761.
6. Knopf, J. L., Gallagher, J. R. & Held, W. A. (1983) *Mol. Cell. Biol.* **3**, 2232–2240.
7. Derman, E. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 5425–5429.
8. Bishop, J. O., Clark, A. J., Clissold, P. M., Hainey, S. & Francke, U. (1982) *EMBO J.* **1**, 615–620.
9. Clark, A. J., Hickman, J. & Bishop, J. O. (1984) *EMBO J.* **3**, 2055–2064.
10. Clark, A. J., Clissold, P. M. & Bishop, J. O. (1982) *Gene* **18**, 221–230.
11. Anderson, S., Gait, M., Mayol, L. & Young, I. G. (1980) *Nucleic Acids Res.* **8**, 1731–1745.
12. Clark, A. J., Clissold, P. M., Al Shawi, R., Beattie, P. & Bishop, J. O. (1984) *EMBO J.* **3**, 1045–1052.
13. Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dodgson, J. (1980) *Cell* **20**, 555–565.
14. Unterman, R. D., Lynch, K. R., Nakhasi, H. L., Dolan, K. P., Hamilton, J. W., Cohn, D. V. & Feigelson, P. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 3478–3482.
15. Dolan, K. P., Unterman, R., McLaughlin, M., Nakhasi, H. L., Lynch, K. R. & Feigelson, P. (1982) *J. Biol. Chem.* **257**, 13527–13543.
16. Laperche, Y., Lynch, K. R., Dolan, K. P. & Feigelson, P. (1983) *Cell* **32**, 453–460.
17. Bennett, K., Lalley, P., Barth, R. & Hastie, N. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 1220–1224.
18. Krauter, K., Leinwald, L., D'Eustachio, P., Ruddle, F. & Darnell, J. (1982) *J. Cell Biol.* **94**, 414–417.
19. Brown, D. D., Wensink, P. C. & Jordan, E. (1972) *J. Mol. Biol.* **63**, 57–73.
20. Smith, G. P. (1973) *Cold Spring Harbor Symp. Quant. Biol.* **38**, 507–513.
21. Ohta, T. (1983) *Theor. Pop. Biol.* **23**, 216–240.
22. Baltimore, D. (1981) *Cell* **24**, 592–594.
23. Dover, G. & Coen, E. S. (1981) *Nature (London)* **290**, 731–732.