The Involucrin Gene of the Orangutan: Generation of the Late Region as an Evolutionary Trend in the Hominoids

P. Djian and H. Green
Department of Cellular and Molecular Physiology, Harvard Medical School

In the evolutionary line leading to the higher primates, the coding region of the involucrin gene evolved a segment consisting of numerous repeats of a 10-codon sequence. Additions to this segment of repeats have been made successively, thus generating regions that can be defined as early, middle, and late. The involucrin gene of the orangutan (Pongo pygmaeus abelii) possesses a segment of repeats whose early region has the same repeat structure as that in other anthropoids. The middle region is not similar in repeat structure to that of all anthropoids but is similar to that of other hominoids. The late region is unique to the species; it does not correspond at all in its repeat structure to that of the human or gorilla and is much larger. The late region of the orangutan was generated by duplications of blocks of older repeats clearly belonging to the middle region. Continued duplications extending the late region are an evolutionary trend in the hominoids. The process of addition of repeats at a particular location is a more significant aspect of the evolution of involucrin than are random nucleotide substitutions; in addition, it has proceeded more rapidly.

Introduction

In several anthropoid primates, the gene for involucrin has been found to contain, at the same position within the coding region, a segment of repeats of a 10-codon sequence. Since this segment is unlike the segment of repeats found in other mammals and is confined to the anthropoid primates, it has been called the modern segment. From a comparison of the nucleotide sequence of the involucrin genes of the human (Eckert and Green 1986), the gorilla Gorilla gorilla (Teumer and Green 1989), and the owl monkey Aotus trivirgatus (Tseng and Green 1989), the modern segment of repeats has been divided into an early region present in all anthropoids, a middle region shared by the hominoids, and a late region different in the two hominoids so far examined. We report the nucleotide sequence of the orangutan involucrin gene and describe the expansion of its late region, a process that occurred after separation of the orangutan lineage from that of the other hominoids.

Material and Method

A vaginal biopsy of a Sumatran orangutan (Tupa) was performed at the Yerkes Primate Center (Atlanta) and provided to Dr. R. H. Rice (Harvard School of Public Health), who prepared fibroblast cultures. Genomic DNA was prepared from these fibroblasts according to the method of Maniatis et al. (1982, pp. 280–281). For constructing a restriction map (fig. 1), DNA was digested with various restriction enzymes.
FIG. 1.—Restriction map of the orangutan involucrin gene. The map shows restriction sites on a 25-kb fragment containing the gene. The coding region is drawn as a box with its divided ancestral segment (open) and its modern segment (cross-hatched). The asterisk (*) marks an EcoRI site present in the orangutan but absent from the human and gorilla.

Results
Restriction Map

Over a distance of 25 kb of genomic DNA containing the orangutan involucrin gene and its flanking DNA, the restriction map is very similar to that of the human and gorilla; the orangutan gene differs only by the presence of an EcoRI site in the 3'-untranslated region (fig. 1) and the absence of a HindIII site located <300 nucleotides downstream of the poly A addition site in the human and gorilla genes. Because of the numerous duplicated blocks in the coding region, it was necessary to sequence overlapping fragments including marker nucleotides to distinguish the different blocks. The clones sequenced are illustrated in figure 2.

The Coding Region Flanking the Modern Segment of Repeats

As in other higher primates, the coding region of the orangutan involucrin gene is divided into two parts by the segment of repeats. Both 5' and 3' of the segment of repeats, the orangutan coding region corresponds to that of the human, without any gaps (fig. 3).
Fig. 2.—Sequencing of the modern segment of the orangutan involucrin gene. The modern segment contains 64 repeats numbered 3' to 5'. The late region consists of repeats 29–59 and is built by duplicated blocks of repeats numbered 3' to 5' with Roman numerals. Duplicate blocks III and IV are boxed with continuous lines. Duplicate blocks IVb, V, and VI are boxed with dashed lines; block VI contains additional internal duplications of a repeat with the unusual GAT codon. Boxed codons contain marker nucleotides that coincide in more than one block of repeats. Encircled codons contain marker nucleotides that are confined to a single block and distinguish that block from others. In order not to miss entire repeat blocks, it is important that some of the very similar overlapping fragments sequenced (arrows) be distinguished by these encircled codons.
The Segment of Repeats: Early and Middle Regions

The segment of repeats in the orangutan consists of 64 repeats of a 10-codon sequence (fig. 3). This is a larger segment than that of the human, gorilla, and owl monkey (which have 39, 30–44, and 29–35 repeats, respectively). As in other anthropoids, the repeats are classified into two types, A and B, according to the nature of the first three codons (Teumer and Green 1989; Tseng and Green 1989): the A repeat contains AAG, CAC, and CTG, encoding KHL, and the B repeat contains GAG, CTC, and CCA, encoding ELP. A and B repeats are interspersed throughout the segment of repeats.

The early region (repeats 1–10) is present in all anthropoids so far examined (human, gorilla, and owl monkey). The early region of the orangutan has exactly the same alternation of A and B repeats as do the other anthropoids and shares with the early region of the human 31 coincident nonconsensus nucleotides (marker nucleotides) at corresponding positions (fig. 3). Coincident marker nucleotides indicate a common origin of repeats. As shown elsewhere (Tseng and Green 1989), repeats 3–6 of the early region are a duplicate of repeats 7–10.

The middle region of the human contains repeats 11–28. Most of the middle region of the orangutan can be aligned, repeat for repeat, with the middle region of the human. The human and orangutan (and the gorilla as well) contain a 3-codon deletion in the same position of repeat 13. There is coincidence of 32 marker nucleotides in the 15 corresponding repeats of the orangutan and human. Three repeats present in the human (15, 20, and 23) are absent from the orangutan, and three repeats present in the orangutan (23, 27, and 28) are absent from the human. Repeats 27 and 28 of the orangutan, although they have no human counterpart, are part of the middle region, since two corresponding repeats are present in two of the gorilla alleles sequenced (Teumer and Green 1989) and in the gene of *Macaca fascicularis* (P. Djian and H. Green, unpublished experiments); in the human, the two corresponding repeats must have been deleted.

The Late Region

This region of the involucrin gene is very different in the two hominoids previously examined. In the human, it consists largely of two blocks, each built of nearly identical repeats. One block consists of five typical B repeats, and the other block consists of three modified B repeats, designated as B\(^a\), in which the CCA codon has been changed.

---

**FIG. 3.**—Involucrin gene of the orangutan. The entire coding region, divided by the segment of repeats, is aligned with that of the human. Repeats of the modern segment are designated A or B according to their first three codons. The repeat pattern in the early region of the orangutan corresponds perfectly with that of the human. The middle region mostly corresponds with that of the human, but three repeats are lacking in the orangutan and three are lacking in the human. The vertical bar separating most of the late regions of the two species indicates that these regions do not correspond and are therefore species specific. Repeats 60–64 of the orangutan do not belong to the late region, for reasons described in the text. The two blocks of quasi-invariant repeats forming the late region of the human have been framed. The second of these blocks contains the B\(^a\) repeat, encoding serine in the third position. The late region of the orangutan gene has been built by five consecutive duplications of different blocks of repeats, numbered I–VI. Boxed marker nucleotides occupy identical positions in all six blocks. Encircled marker nucleotides occupy identical positions in more than one block but not in every block. A\(^D\) is a modified A repeat, encoding aspartic acid in the fourth position. Marker nucleotides present at the same position in both the orangutan and human are underlined in the human sequence. Nearly all of these are located in the early and middle regions.
Table 1

Nucleotide Comparison of Duplicated Blocks in the Late Region of the Orangutan

<table>
<thead>
<tr>
<th>Block Comparison</th>
<th>No. of Shared Marker Nucleotides</th>
<th>No. of Divergent Nucleotides (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block I vs. block II</td>
<td>6</td>
<td>0/90 (0)</td>
</tr>
<tr>
<td>Block III vs. block IV</td>
<td>20</td>
<td>1/360 (0.3)</td>
</tr>
<tr>
<td>Block IV(b) vs. block V</td>
<td>10</td>
<td>1/180 (0.6)</td>
</tr>
<tr>
<td>Block IV(b) vs. block VI (omitting A^D repeats)</td>
<td>9</td>
<td>2/180 (1.1)</td>
</tr>
<tr>
<td>Block of A^D repeats</td>
<td></td>
<td>0/150 (0)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>4/960 (0.4)</td>
</tr>
</tbody>
</table>

| A and B repeats of early and middle region vs.         | 76/831 (9.1)                      |
| their respective consensus                            |                                  |
| Early and middle region of orangutan vs.             | 26/741 (3.5)*                    |
| corresponding repeats of early and middle             |                                  |
| region of human                                       |                                  |

* Value is somewhat higher than that for the coding regions of the hemoglobin genes (Koop et al. 1986).

The late region of the orangutan gene is considerably larger than that of the human or gorilla, as it consists of 31 repeats, in comparison to 9–16 in the other two species. This accounts for the larger size of the modern segment and of the entire involucrin molecule (Parenteau et al. 1987). In contrast to that of the human, the late region of the orangutan contains no block of B repeats or any B' repeats. Instead, it has been created by a series of duplications of blocks of mixed A and B repeats, as in the gorilla, but the pattern of repeats is quite different. The orangutan late region also contains a unique block of identical modified A repeats, designated as A^D (fig. 3).

The late region of the orangutan is likely to have been generated as follows: block I, consisting of three repeats within the middle region, close to its 5' end, was duplicated to give block II, thus beginning the late region. Block III, consisting of blocks I and II together with six additional repeats near the end of the middle region, were then duplicated to give block IV, thus extending the late region. A part of this (block IVb) was then duplicated to produce block V, and block V was duplicated to produce block VI. Within block VI, one repeat containing GAT (an aspartic codon) in the fourth position instead of GAG (a glutamic codon) was duplicated four times to produce the invariant block of five A^D repeats, each containing the unusual GAT codon. Thus the entire late region was generated by successive duplications beginning with repeats 20–28 of the middle region.

This postulated sequence of block duplications is supported by the coincidences of marker nucleotides between the different blocks (table 1). Duplicated blocks also show very little nucleotide divergence. For all five duplications, there are a total of four (0.4%) divergent nucleotides in a total of 960. This value is much lower than the divergence of the A and B repeats of the early and middle regions from their respective consensus (table 1). This indicates that the multiple block duplications of the orangutan late region were very recent events relative to the creation of the early and middle regions.

It is worth noting that repeat 23, which is identical to the multiply duplicated.
repeat of the late region (fig. 3), has no counterpart in the human or gorilla (Teumer and Green 1989). In the human, there is a single repeat (i.e., 25) containing GAT, but it differs, in four other nucleotides, from the duplicated repeat of the orangutan.

Repeats 60–64 of the orangutan gene correspond to five repeats present in the modern segment of *M. fascicularis* (P. Djian and H. Green, unpublished experiments). As these repeats must have been generated in an ancestor of the catarrhines, we now exclude them from the late region. Three of these five repeats (corresponding to repeats 60–62) have apparently been deleted from the human, and all five have been deleted from the gorilla.

As specified by the consensus sequence, a GAG codon encoding glutamic acid predominates at the seventh position of B repeats in the modern segment of the owl monkey (Tseng and Green 1989), gorilla (Teumer and Green 1989), and human (fig. 3). In the middle region of the hominoids, three B repeats containing a GTG codon are present; as a consequence of successive duplications of two of these B repeats to produce the late region of the orangutan, 11 of 19 B repeats present in the modern segment of the orangutan contain T instead of A as the second nucleotide at this codon position.

In the early regions, B repeats are not as numerous as A repeats (B:A = 0.7). During the development of the late regions of the gorilla and human, B repeats tended to be overduplicated, so that the ratio B:A increased several fold. This has not occurred in the orangutan, whose B:A ratio in the late region is 0.4.

**Discussion**

In the human, gorilla, and orangutan, the nucleotide divergence between the parts of the coding region flanking the segment of repeats is <2.2%. The repeat pattern and the many coincident marker nucleotides in the modern segment make the latter more valuable, for tracing the relatedness of different higher primates, than either the rest of the coding region or other genes. Because of the repeat structure, the presence of a marker nucleotide at any position indicates that a mutation occurred in the lineage of that species. This contrasts with nucleotide divergences in more ordinary genes, where two species with divergent nucleotides must be compared with an outgroup in order to decide in which lineage the mutation occurred.

That the gorilla and the human are more closely related to each other than to the orangutan is clear from the nearly identical repeat structure of their middle regions, including three repeats that have no counterpart in the orangutan. In addition, there are four marker nucleotides shared by the early and middle regions of the human and gorilla but not by the orangutan; also, the seventh codon of the B repeats is predominantly GTG in the orangutan and GAG in the human and the gorilla. These correspondences are consistent with the most prevalent view of a more distant relatedness of the orangutan to the other hominoids (Ciochon 1985; Koop et al. 1986; Miyamoto et al. 1988) and not consistent with the view that the human and the orangutan are closely related (Schwartz 1984).

The parts of the coding region flanking the modern segment as well as the early region of the modern segment are very similar in gorilla, human, and orangutan: there have been no duplications or deletions that distinguish one species from the others. This is consistent with the idea that the coding region flanking the modern segment was generated prior to the evolution of the primates (Tseng and Green 1988; Simon and Green 1989) and that the early region of the modern segment was created in a common ancestor of all higher primates (Tseng and Green 1989). Since the middle
regions of the orangutan, human, and gorilla are similar to each other but different from any group of repeats in the owl monkey, it was concluded that the middle region of the hominoid gene was generated by repeat addition after separation of the hominoid lineage from that of the new-world monkeys (Tseng and Green 1989); after separation of the different hominoid lineages from each other, their middle regions underwent relatively little further modification.

It is now clear that rapid expansion of the late region of the involucrin gene is a characteristic feature of the evolution of the hominoids. This process can be described as an evolutionary trend. Not specified in the trend was which repeats were to be duplicated or the size of the blocks to be duplicated. Consequently, each of the hominoid species has a unique late region.

During the period following separation of the orangutan lineage from that of the human, the late region of the orangutan added 31 repeats or 310 codons, thereby nearly doubling the size of the involucrin molecule. This was accomplished by duplications of blocks of repeats, a single block consisting of as many as 12 repeats. For this reason, the expansion of the late region took place very rapidly compared with the process of nucleotide divergence (table 1). Involucrin of the orangutan is larger than that of the other hominoids (98 kD, compared with 68 kD in the human), and nearly all of this difference is due to expansion of its late region.

Expansion of the involucrin gene has continued through multiple speciations. That the trend may still be current is suggested by the polymorphism of involucrin in many higher primates (Parenteau et al. 1987).

Sequence Availability

These sequences have been deposited in GenBank under accession number M25312.

Acknowledgment

This investigation was aided by a grant from the National Cancer Institute.

LITERATURE CITED


Walter M. Fitch, reviewing editor

Received January 24, 1989; final revision received May 16, 1989

Accepted May 16, 1989