The Involucrin Genes of Pig and Dog: Comparison of Their Segments of Repeats with Those of Prosimians and Higher Primates

Hung Tseng and Howard Green
Department of Cellular and Molecular Physiology, Harvard Medical School

The involucrin genes of the dog and the pig have been cloned and sequenced. Like the corresponding genes of the prosimians, each contains a homologous segment of short tandem repeats at the same position in the coding region. However, the codon sequence of the repeats in the prosimians differs significantly from that of the nonprimate mammals. This evolution has been brought about by a combination of genetic modifications (selective deletions, mutations, and gene conversions). In the anthropoids, this segment of repeats was replaced by a modern one differing in location, sequence, and repeat length. In several of its properties the modern segment has continued the prosimian trend away from the nonprimates. The overall direction of the evolution of this segment has therefore been maintained even though there have been sudden changes in the evolutionary processes acting on the gene.

Introduction

Involucrin is a tissue-specific protein produced in terminally differentiating keratinocytes. So far, the coding region of every primate involucrin gene sequenced has been found to contain a segment of short tandem repeats. Two types of segment of repeats have been identified: the modern type has been found in all six anthropoid species so far examined (Eckert and Green 1986; Djian and Green 1989a, 1989b, 1990; Teumer and Green 1989; Tseng and Green 1989), and the prosimian type has been found in Lemur catta (Tseng and Green 1988) and Galago crassicaudatus (Phillips et al. 1990). These two types of repeat segment differ in sequence as well as in location within the coding region. We have suggested that the modern segment, which is absent from the prosimian genes, developed after the anthropoid lineage diverged from the prosimian lineage (Tseng and Green 1988).

Here we describe the involucrin genes of the pig and the dog. While their segments of repeats are different, they are homologous both to each other and to those of the prosimians. It is therefore clear that the segments of repeats of the lemur and galago were not an invention of the prosimians but resulted from modification of a segment of repeats already present in nonprimate mammals.

Materials and Methods

Keratinocytes of an outbred Yorkshire pig (Sus scrofa) were derived from a skin biopsy provided by Dr. G. Gallico III. Keratinocytes of a purebred beagle (Canis familiaris) were derived from an oral biopsy provided by Dr. Ray Williams. The...
keratinocytes were grown by the method of Rheinwald and Green (1975), with subsequent modifications summarized by Allen-Hoffman and Rheinwald (1984) and Simon and Green (1985). Restriction enzymes were purchased from New England Biolabs and Bethesda Research Laboratories and were used as recommended by the manufacturers. DNA and RNA were extracted from cultured cells (Tseng and Green 1988), and polyadenylated RNA was selected by oligo-[dT] cellulose (Pharmacia) chromatography according to the method of Aviv and Leder (1972). Complementary DNA libraries of keratinocyte RNA were prepared using kits purchased from Amersham. The sequence alignment shown in figure 1 is based on GENALIGN (version 5.2), a computer program licensed to Intelligenetics, Inc. The alignment was then improved by eye.

Results
Cloning and Sequencing

The pig gene was cloned by the method described by Tseng and Green (1988, 1989). Using a genomic clone of lemur involucrin as probe (Tseng and Green 1988), we prepared a restriction map of the pig locus by analysis of genomic DNA (Southern 1975). No apparent polymorphism was observed in one individual when single and double digestions with at least 10 restriction enzymes were used. A BamHI-HindIII fragment of ~5 kb containing the entire coding sequence was cloned into pGEM-3. The pig gene was sequenced from both strands by double-strand dideoxy sequencing (Sanger et al. 1977).

The dog gene proved to be more difficult to clone, as genomic DNA failed to provide an unambiguous signal when probed with the pig gene under conditions of low-stringency hybridization. Since involucrin mRNA is abundant in terminally differentiating keratinocytes, a λgt10 cDNA library was constructed using RNA from cultured dog keratinocytes that were allowed to differentiate. When probed with a mixture containing both pig and lemur involucrin genes, ~10% of the plaques in this library were positive. DNA was prepared from 10 of these, and the two largest inserts (from plaques 411 and 1713) were subcloned into pGEM-3 at the EcoRI site. Subsequent DNA sequencing showed that both were full-length cDNAs derived from involucrin mRNA. Clone 1713 differed from clone 411 by a single nucleotide change and by a 6-bp deletion 3' of site M, indicating that the two clones were from different alleles. A restriction map of the dog locus was constructed using clone 411 as probe. Using this probe, we cloned the dog gene, contained in a 3.5-kb genomic BamHI fragment, by the same procedure as was used for the pig gene. The genomic clone was sequenced from the strand complementary to the mRNA strand and was found to be identical to cDNA clone 1713. Comparison of the sequences from the dog genomic and cDNA clones confirmed the predicted 3' splicing site of the gene (Eckert and Green 1986).

The Coding Region

In figure 1 the entire coding regions of the involucrin genes of dog and pig are aligned with those of one prosimian and one higher primate. There are two sites at which a segment of repeats has been found in involucrin genes: the dog and pig genes, like the genes of the prosimians, contain a segment of repeats only at site P, whereas the anthropoids contain a segment of repeats only at site M.

Comparison of the coding regions of the dog and the pig genes with those of the galago and the owl monkey allows deletions and insertions to be identified. In the pig
gene, a sequence of 21 nucleotides was inserted at site M, where, in the anthropoids, the modern segment of repeats was generated. The deletion of a single nucleotide just upstream of site M resulted in both a frameshift and the introduction of a stop codon into the pig insertion; consequently, ~40 amino acids constituting the carboxyl end of involucrin in the other species have been eliminated from pig involucrin, leaving the protein with only 347 amino acids. This is consistent with the smaller size of pig involucrin detected electrophoretically (Simon and Green 1989). The prosimian gene also apparently underwent a shortening at the 3' end—but only by seven codons.

The Segment of Repeats

The segments of repeats of the dog and pig involucrin genes are not only located at the same site as in the two prosimian genes known, but the repeats of all four species are homologous. This is clearly shown in figure 2, where the consensus sequences of each pair of animals are compared; there is a definite divergence between the two consensus sequences at only two of 48 nucleotide positions.

The segment of repeats in the dog contains only six repeats, including one that is incomplete. The third repeat is an identical duplicate of the fourth, as indicated by their five shared marker nucleotides. The dog segment contains the smallest number of repeats found so far in an involucrin gene; as a result, the protein contains only 285 amino acids (two less for the protein corresponding to clone 1713), making it the smallest of the involucrins so far known (Simon and Green 1989). The pig gene, like the galago gene, has 13 repeats, one of which is incomplete, and the lemur, by means of a single duplication, has added six more repeats.

Although there are several deletions, the repeat length in the dog is 20 codons, whereas the repeat length in the pig is 16 codons. This difference could have arisen through either deletion or insertion, depending on which length is more primitive. In either case, the most common repeat length of the prosimians is shorter than that of the pig, as a result of the frequent deletion of as many as three codons at positions 10–12.

Discussion

By comparing the involucrin genes of the lemur and the human, we concluded earlier that the modern segment of repeats at site M in the human gene originated in the anthropoid lineage after its divergence from the prosimians. Since then, the involucrin genes of five other anthropoids have been sequenced, and all have been found to have exclusively the modern segment of repeats at site M.

The segment of repeats of the prosimians is not of modern origin but instead was derived from the segment of repeats of nonprimate ancestors. After divergence of the lineage of the nonprimates from that of the prosimians, some nucleotide positions underwent concerted evolution. As a result, an amino acid encoded uniformly at a given position in the segment of repeats of one species may be absent at the same position in the segment of repeats of another. For example, GTG, encoding valine, is the codon at position 7 of the pig segment in all of 12 repeats, but it is uniformly absent at this position in the segment of repeats of the other three species. At this position in the lemur segment, all of three synonymous codons encode leucine; ordinarily, this might imply sustained selective pressure for leucine, but this conclusion would be difficult to reconcile with the absence of leucine at the same position in the pig segment, unless selective pressure favored a different amino acid at the same position in different species. A similar discrepancy is seen for the aspartic acid codon GAT,
which is present at position 8 in 10 of 12 repeats of the pig but is uniformly absent at this position in the segment of repeats of the other three species. Similarly, AAG or AAA, encoding lysine, are the codons at the ninth position in the lemur segment of repeats but are uniformly absent from this position in the pig and dog. TCA, encoding serine, is either the dominant or the only codon at position 16 of the pig, dog, and galago but is only occasionally present at the same position in the lemur segment.

In some cases, it seems even more obvious that uniformity of nucleotide sequence of the repeats within a species is not the result of selective pressure acting on the encoded sequence. For example, CAT is almost uniformly the codon at position 6 in the pig; but the synonymous codon CAC is dominant at this position in the galago and the lemur. Similarly, CAA is almost uniformly the codon at position 13 in both pig and dog, but the synonymous codon CAG is almost uniformly the codon at that position in the galago and is more common than CAA in the lemur. These differences are more likely the result of a single silent substitution, followed by a homogenization process, probably gene conversion, as has been suggested elsewhere for the galago segment (Phillips et al. 1990).

The evolution of the involucrin gene in nonprimates, prosimians, and anthropoids appears to have followed a consistent pattern in these respects:

1. Increase in number of repeats. The mean number of repeats in the two nonprimates is nine; in the two prosimians it is 16; and in the six anthropoid primates it is 41.
2. Decrease in repeat length. The repeat length has a mean value of 18 codons in the two nonprimates, a value close to 13 for the two prosimians, and 10 for the anthropoid primates. Since the increase in number of repeats is more extreme, the size of the protein has increased.
3. Change in the abundance of encoded amino acids. In figure 3 this is summarized for the seven different amino acids of the consensus of the anthropoid segment of repeats and for serine and aspartic acid, which are consensus amino acids in one or both of the nonprimate mammals but not in the anthropoid segment of repeats. It can be seen that there is a trend toward both increased abundance of at least three of the amino acids (fig. 3, L, G, and P) and decreased abundance of three (fig. 3, Q, S, and D). In each case, the amino acid composition of the anthropoid segment of repeats either is close to that of the prosimians or extends their trend away from the nonprimate mammals (also see Tseng and Green 1988).

This is remarkable, since the gene-altering processes in the prosimians and the anthropoids were different. In the prosimians, the changes from the codon abundances of the nonprimate mammals occurred in several ways. The reduction of glutamine content was achieved partly by selective removal of CAGs by deletion in the CAG-
<table>
<thead>
<tr>
<th>DOG</th>
<th>PIG</th>
<th>GALAGO</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAG GAG CAG GAA CAC CAG</td>
<td>CAG GAG CAG GAA CAC CAG</td>
<td>CAG GAG CAG GAA CAC CAG</td>
</tr>
<tr>
<td>CAG GAG CAG AAA CTG CAC CTG GAA CAG TGT CTG GAA CAG CAG</td>
<td>CAG GAG CAG AAA CTG CAC CTG GAA CAG TGT CTG GAA CAG CAG</td>
<td>CAG GAG CAG AAA CTG CAC CTG GAA CAG TGT CTG GAA CAG CAG</td>
</tr>
<tr>
<td>CAG GAG CAG AAA CTG TAC CGC GAG CAG TGT CTG GAA CAG CAG</td>
<td>CAG GAG CAG AAA CTG TAC CGC GAG CAG TGT CTG GAA CAG CAG</td>
<td>CAG GAG CAG AAA CTG TAC CGC GAG CAG TGT CTG GAA CAG CAG</td>
</tr>
<tr>
<td>CAG GAG CAG AAA CTG TAC CGC GAG CAG TGT CTG GAA CAG CAG</td>
<td>CAG GAG CAG AAA CTG TAC CGC GAG CAG TGT CTG GAA CAG CAG</td>
<td>CAG GAG CAG AAA CTG TAC CGC GAG CAG TGT CTG GAA CAG CAG</td>
</tr>
<tr>
<td>CAG CAC AGG GAA CAC CAT CTG GAA CAG TGT CTG GAA CAG CAG</td>
<td>CAG CAC AGG GAA CAC CAT CTG GAA CAG TGT CTG GAA CAG CAG</td>
<td>CAG CAC AGG GAA CAC CAT CTG GAA CAG TGT CTG GAA CAG CAG</td>
</tr>
<tr>
<td>CAG CAC AGG GAA CAC CAT CTG GAA CAG TGT CTG GAA CAG CAG</td>
<td>CAG CAC AGG GAA CAC CAT CTG GAA CAG TGT CTG GAA CAG CAG</td>
<td>CAG CAC AGG GAA CAC CAT CTG GAA CAG TGT CTG GAA CAG CAG</td>
</tr>
<tr>
<td>CAG GAG GAA CTG --- --- --- ---</td>
<td>CAG GAG GAA CTG --- --- --- ---</td>
<td>CAG GAG GAA CTG --- --- --- ---</td>
</tr>
</tbody>
</table>

**Image:** The image contains a table comparing nucleotide sequences from different species. The table is labeled as DOG, PIG, and GALAGO, with each species having a column of nucleotide sequences. The sequences are highlighted in a way that emphasizes certain patterns or comparisons across the species.
CONSENSUS

Do  CAG GAG CAG CAA CTG CAT GAC A AAG CACG AAG CAG TGT CTG GAA CAG CAG CAG CAA GAG TCA
Pi CAG GAG CAG GGA CTG CAT --- --- --- --- GTG GAT CAG CAG CAG CAG CAG CAA GAG TCA
Ga CAG GAG CAG CAA CTG CAT --- --- --- --- GTG GAT CAG CAG CAG CAG CAG CAA GAG TCA
Le  CAG CAC CAC CAA CTG CAT --- --- --- --- CTG GGA AAA CAG CAG CAG CAG CAG CAA GAG TCA

D/P  CAG GAG CAG GAA CTG CAC CAG CAG CAG CAG CAG CAG CAG CAG CAG CAG CAG GAG TCA
L/G  CAG GAG CAG GAA CTG CAC --- --- --- --- CTG GGA AAA CAG CAG CAG CAG CAG CAA GAG TCA

COMMON CONSENSUS

CAG GAG CAG CAA CTG CAT --- --- --- --- CTG GGA AAA CAG CAG CAG CAG CAG CAA GAG TCA

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

Fig. 2.—Nucleotide sequence of the segments of repeats at site P. The dog and pig repeats are aligned with each other and with the repeats of the two prosimians. Framed sequences in the lemur and dog segments indicate single duplications. Boxed nucleotides are those differing from the common consensus of the four species. A common consensus is also given for the two nonprimates and for the two prosimians, with asterisks indicating their divergent nucleotides. Codon positions are numbered at the top and bottom. Codons d1–d4 are present in dog only.
The image shows a graph titled "Abundance of amino acids encoded by the segments of repeats." The graph is divided into three categories: Non-primates, Prosimians, and Anthropoids. Each category has a line indicating the abundance of amino acids, with amino acids such as Q, L, E, P, G, K, H, S, D represented on the graph. The y-axis represents amino acid abundance in percentage (%), ranging from 0 to 50.

The text accompanying the graph reads:

"FIG. 3.—Abundance of amino acids encoded by the segments of repeats. The values have been averaged for the dog and pig (Non-primates), for the lemur and galago (Prosimians), and for the human and owl monkey (Anthropoids). The mean value for lysine of the Non-primates has been omitted because the values for the two species differ widely.

Rich region of the repeats. This is evident in both lemur and galago (fig. 2). Glutamine content was also reduced by nucleotide substitutions in codon position 9, thereby converting its consensus codon from CAG to AAG or AAA, which encode lysine. Aspartic acid and glutamic acid were eliminated as consensus amino acids at position 8 by substitutions in GAT and GAA codons, and glycine was increased by the resulting appearance of GGA and GGG codons. Serine was decreased by substitutions in TCA codons (position 16), and proline was increased by the resulting appearance of CCA codons. At least in the galago, nucleotide substitutions were evidently spread to adjacent repeats by a process of gene conversion (Phillips et al. 1990). The prosimian segments of repeats therefore acquired their changed codon abundance by a combination of deletion, mutation, and gene conversion.

The anthropoid lineage did not continue to use the same gene-altering processes. Instead, the anthropoids eliminated the entire old segment of repeats and constructed a new one; the choice of sequence duplicated (site M) produced an amino acid composition that either resembled that of the prosimians or continued their trend away from the nonprimate mammals. All seven consensus codons of the modern segment of repeats are the same as those present in the lemur and, with the exception of serine/proline, in the galago as well. In the prosimian coding region there is no other stretch..."
of 10 codons that resembles in its codon composition the prosimian segment of repeats as closely as does site M. This is true for the codons that are absent as well as for those that are present.

A further problem that seems to have been solved by the anthropoid lineage is how to accommodate all seven consensus amino acids—and the right abundance of each—into a shortened repeat length. This was accomplished by the introduction of A and B repeats, which differ in the first three codons (Tseng and Green 1989). Although a proline codon was present in the second position of the prosimian site M, this codon was not retained in the modern segment; instead, a proline codon appeared in the third position of the B repeat. Codons for lysine and histidine were introduced into the first and second positions of the A repeat. The fourth codon of both repeat types was mutationally altered from GGG, found in both galago and lemur, to GAG, leaving the last six codons virtually identical to those in the unduplicated site M of the prosimians. The introduction of A and B repeats by the anthropoid lineage must necessarily have followed the initial duplication at site M, since two or more repeats had to be present before one could be changed to the alternate type by mutations in the first three codons.

Acknowledgment

The careful reading and many valuable suggestions of Dr. Walter Fitch are gratefully acknowledged. This investigation was aided by a grant from the National Cancer Institute.

LITERATURE CITED


TEUMER, J., and H. GREEN. 1989. Divergent evolution of part of the involucrin gene in the


WALTER M. FITCH, reviewing editor

Received February 9, 1990; revision received March 23, 1990

Accepted March 28, 1990