Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods

Yoshihito Niimura* and Masatoshi Nei

Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University, 328 Mueller Laboratory, University Park, PA 16802

Contributed by Masatoshi Nei, March 9, 2005

Olfaction, which is an important physiological function for the survival of mammals, is controlled by a large multigene family of olfactory receptor (OR) genes. Fishes also have this gene family, but the number of genes is known to be substantially smaller than in mammals. To understand the evolutionary dynamics of OR genes, we conducted a phylogenetic analysis of all functional genes identified from the genome sequences of zebrafish, pufferfish, frogs, chickens, humans, and mice. The results suggested that the most recent common ancestor between fishes and tetrapods had at least nine ancestral OR genes, and all OR genes identified were classified into nine groups, each of which originated from one ancestral gene. Eight of the nine group genes are still observed in current fish species, whereas only two group genes were found from mammalian genomes, showing that the OR gene family in fishes is much more diverse than in mammals. In mammals, however, one group of genes, γ, expanded enormously, containing ~90% of the entire gene family. Interestingly, the gene groups observed in mammals or birds are nearly absent in fishes. The OR gene repertoire in frogs is as diverse as that in fishes, but the expansion of group γ genes also occurred, indicating that the frog OR gene family has both mammal- and fish-like characters. All of these observations can be explained by the environmental change that organisms have experienced from the time of the common ancestor of all vertebrates to the present.

Vertebrates can discriminate among thousands of different odor molecules in the environment. Odor molecules are detected by olfactory receptors (ORs) that are expressed in sensory neurons of olfactory epithelia in nasal cavities (see refs. 1–3 for review). ORs are encoded by a large multigene family, which consists of ~1,000 different genes in mammalian species. They are G protein-coupled receptors (GPCRs) that contain seven α-helical transmembrane regions. OR genes are ~310 codons long on average, do not have any introns in their coding regions, and form genomic clusters that are scattered on many chromosomes. The OR gene family in each species is considered to reflect the ability of olfaction of the species, because different OR genes are thought to bind to different sets of odor molecules.

Analysis of whole-genome sequences has shown that in humans, there are ~800 OR genes, but ~50% of them are pseudogenes (4, 5), whereas in mice, there are ~1,400 OR genes, and the fraction of pseudogenes is ~25% (6–9). OR genes were first identified from rats (10). In nonmammalian vertebrates, OR genes have been identified from lampreys (11), catfish (12), zebrafish (13, 14), medaka fish (16, 17), Japanese loaches (18), and chickens (19), and birds (20), although the numbers of genes identified are relatively small.

Currently, it is widely accepted that vertebrate OR genes can be classified into two different classes, class I and II genes (21). All of the fish genes are believed to belong to class I genes, and for this reason, they are referred to as “fish-like” genes. By contrast, class II genes are called “mammalian-like” genes, because the majority of mammalian OR genes belong to class II genes. X. laevis is known to have both types of genes. Class I genes are exclusively expressed in the water-filled lateral diverticulum of the nasal cavities and are similar to known fish genes, whereas class II genes are expressed in the air-filled medial diverticulum and are similar to known mammalian genes (19). From this observation and others (22), it has been proposed that class I genes are specialized for detecting water-soluble odorants, and class II genes are for recognizing airborne odorants (19). A small number of class I genes were found in humans, but they were considered to be evolutionary relics (23). Later, however, an extensive search of OR genes from the human and mouse genome sequences revealed that mammals have a substantial number of class I genes (4, 6). Therefore, the functional difference between class I and II genes is now unclear (6, 24).

The purpose of this paper is to study the evolutionary dynamics of this huge multigene family in vertebrates. For this purpose, we first identified OR genes from the draft genome sequences of zebrafish, pufferfish, Xenopus tropicalis, and chickens. We then conducted large-scale phylogenetic analyses by using these genes and compared OR gene families among various vertebrate species.

Materials and Methods

Database Search for Known Fish, Amphibian, and Avian OR Genes. To collect full-length sequences of functional OR genes in nonmammalian vertebrates, we conducted BLASTP searches (25) against the DADB database (all of the amino acid sequences) on the DNA Databank of Japan (DDBJ) web site (www.ddbj.nig.ac.jp) with the E value below 1e-05. A human class I OR gene (HsOR11.3.2, OR4F16) and a class II gene (HsOR11.4, OR52B4) were used as queries. For the names and sequences of human OR genes, see ref. 5. Here the nomenclature adopted by the HUGO Gene Nomenclature Committee is also presented (www.gene.ucl.ac.uk/ nomenclature).] From the hits obtained by homology search, mammalian genes and sequences <250 aa long were excluded. The remaining sequences were used for further analysis. We constructed a neighbor-joining (NJ) phylogenetic tree (26) by using the remaining sequences and 388 functional OR genes identified from the complete human genome sequences (5). OR genes formed a monophyletic clade supported with a high bootstrap value (>90%) and were therefore easily distinguishable from non-OR GPCR genes. We excluded these non-OR genes and obtained 106 non-redundant full-length functional OR genes.

Identification of OR Genes from Draft Genome Sequences. The procedure of the identification of functional and nonfunctional OR genes is drawn as a flow chart in Fig. 3, which is published as supporting information on the PNAS web site. The draft genome sequences of zebrafish (Danio rerio), pufferfish (Fugu rubripes), western clawed frogs (X. tropicalis), and chickens (Gallus gallus) were downloaded from the following web sites: www.ensembl.org (assembly Zv3, released in November 2003), genome.jgi-psf.org (assembly v3.0, released in August 2002; ref. 27), genome.jgi-psf.org (assembly v1.0, released in February 2003).

Abbreviations: OR, olfactory receptor; NJ, neighbor-joining; MRCA, most recent common ancestor; GPCR, G protein-coupled receptor; DDBJ, DNA Databank of Japan.

Downloaded from www.pnas.org/cgi/doi/10.1073/pnas.0501922102 on April 26, 2005
Fig. 1. Phylogenetic trees of vertebrate OR genes. (A) NJ tree for 1,026 functional OR genes. This tree contains 102 zebrafish, 44 pufferfish, 410 X. tropicalis, 82 chicken, and 388 human (5) OR genes. The number of amino acid sites used was 173 after deletion of all alignment gaps. A branch specific to each species is colored according to the color code at the bottom left. “Frog” indicates X. tropicalis. The scale bar indicates the estimated number of amino acid substitutions per site. The name of each OR gene group, defined in B, is shown. The bootstrap value obtained from 1,000 replications is shown for the clade determining each group. The arrow indicates the zebrafish group (Dr3ORS.4). Clade a and b and clade c indicate the rapid expansion of group γ genes in the X. tropicalis and chicken lineages, respectively. The GenBank accession numbers of the genes not assigned to the genome sequences are U42392, U42394, and AF283560–
We identified functional OR genes from the BLAST hits obtained above in the following way, which is modified from the method used in our previous study (5). We first excluded the sequences <250 aa (criterion 1 in Fig. 3). When a sequence contained interrupting stop codons or frameshifts, the longest sequence between a stop codon/frameshift and another stop codon/frameshift was considered, and if the longest sequence was <250 aa, it was excluded. Some of the remaining sequences (sequences of ≥250 aa) were also excluded by additional filtering processes (criteria 2 and 3). Because a BLAST hit may not cover the entire coding region of a gene, we extended the DNA sequence of each of the BLAST hits to both 3’ and 5’ directions along the chromosome and extracted the longest sequence that is clearly distinguishable from non-OR genes by using this procedure.

Phylogenetic Analysis. Phylogenetic trees were constructed in the following way. Translated amino acid sequences were aligned by the program FFT-NS-I (29). Poisson correction distances (31) were calculated after all alignment gaps were eliminated. A phylogenetic tree was constructed from these distances by using the NJ method given in the program LINTREE (32).

Classification of Functional and Nonfunctional OR Genes. The functional OR genes from zebradish, pufferfish, X. tropicalis, and chickens were classified into groups α–ω on the basis of the phylogenetic tree in Fig. 1. The classification of OR pseudogenes and partial sequences into the nine groups was done in the following way. In this analysis, we used 1,088 OR genes containing all of the functional OR genes identified from the four species and humans and all of the nonredundant functional OR genes in fishes, amphibians, and birds collected from the public database. We conducted a BLASTP search against the 1,088 functional OR genes by using the translated amino acid sequence of each of the OR pseudogenes and partial sequences as queries. The query sequence was assigned to the group to which the best hit of the query belonged.

Results

OR Genes in Zebrafish, Pufferfish, X. tropicalis, and Chickens. We identified OR genes from the draft genome sequences of zebradish, pufferfish, X. tropicalis, and chickens by conducting a homology search (see Materials and Methods). Table 1 shows the numbers of functional OR genes identified for these four species. Note that these numbers are the minimal estimates, because the genome sequencing of the four species has not been completed. Therefore, an OR gene may be split into two different contigs, or a part of a gene sequence may be missing from the draft genome sequence. In Table 1, therefore, the sequences identified as pseudogenes or partial sequences may contain a considerable number of functional genes.

As shown in Table 1, the numbers of OR genes found from the zebrafish and pufferfish genomes are close to the previous estimate from catfish (~100; ref. 12). We found that X. tropicalis has a larger number of functional OR genes than humans. We identified only 78 functional genes of >500 OR gene sequences in the chicken genome. However, the actual number of functional OR genes in the chicken genome appears to be much larger than 78, because the chicken draft genome sequence includes numerous short contigs that are a few kilobases long. (Approximately one-half of the contigs are shorter than 2 kb.) The lists and the nucleotide and amino acid sequences as queries. A BLASTP search was also conducted against the 626 functional OR genes from the four species and 388 functional OR genes from humans. When the best hit of a given query was an OR gene, we regarded the query sequence as an OR pseudogene or a partial sequence of an OR gene. Otherwise, we regarded the query sequence as a non-OR gene and excluded the sequence from our data set (criterion 5). OR pseudogenes were clearly distinguishable from non-OR genes by using this procedure.
a short period in the
6042
†The data for humans and mice were obtained from refs. 5 and 9, respectively.

Table 2. Numbers of functional and nonfunctional OR genes belonging to different groups in six species

<table>
<thead>
<tr>
<th>Group</th>
<th>Zebrafish</th>
<th>Pufferfish</th>
<th>X. tropicalis</th>
<th>Chicken</th>
<th>Mouse*</th>
<th>Human*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Funct Total</td>
<td>Funct Total</td>
<td>Funct Total</td>
<td>Funct Total</td>
<td>Funct Total</td>
<td>Funct Total</td>
</tr>
<tr>
<td>α (I)</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
<td>2 (0.5)</td>
<td>6 (0.7)</td>
<td>9 (11.0)</td>
</tr>
<tr>
<td>β (II)</td>
<td>1 (1.0)</td>
<td>2 (1.5)</td>
<td>1 (2.3)</td>
<td>1 (1.0)</td>
<td>5 (1.2)</td>
<td>19 (2.1)</td>
</tr>
<tr>
<td>γ (III)</td>
<td>1 (1.0)</td>
<td>1 (0.7)</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
<td>370 (90.2)</td>
<td>802 (90.3)</td>
</tr>
<tr>
<td>δ (IV)</td>
<td>44 (43.1)</td>
<td>55 (50.1)</td>
<td>28 (63.6)</td>
<td>61 (62.2)</td>
<td>22 (5.4)</td>
<td>36 (4.1)</td>
</tr>
<tr>
<td>ε (V)</td>
<td>11 (10.8)</td>
<td>14 (10.2)</td>
<td>2 (4.5)</td>
<td>2 (2.0)</td>
<td>6 (1.5)</td>
<td>17 (1.9)</td>
</tr>
<tr>
<td>ζ (VI)</td>
<td>27 (26.5)</td>
<td>40 (29.2)</td>
<td>6 (13.6)</td>
<td>8 (8.2)</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>η (VII)</td>
<td>16 (15.7)</td>
<td>23 (16.8)</td>
<td>5 (11.4)</td>
<td>24 (24.5)</td>
<td>3 (0.7)</td>
<td>6 (0.7)</td>
</tr>
<tr>
<td>θ (VIII)</td>
<td>1 (1.0)</td>
<td>1 (0.7)</td>
<td>1 (2.3)</td>
<td>1 (1.0)</td>
<td>1 (0.2)</td>
<td>1 (1.0)</td>
</tr>
<tr>
<td>χ (IX)</td>
<td>1 (1.0)</td>
<td>1 (0.7)</td>
<td>1 (2.3)</td>
<td>1 (1.0)</td>
<td>0 (0.0)</td>
<td>0 (0.0)</td>
</tr>
<tr>
<td>Total</td>
<td>102 (100.0)</td>
<td>137 (100.0)</td>
<td>44 (100.0)</td>
<td>98 (100.0)</td>
<td>410 (100.0)</td>
<td>888 (100.0)</td>
</tr>
</tbody>
</table>

The numbers in parentheses represent the percentage of genes for each group. Funct, functional genes; Total, total number of genes including functional genes, pseudogenes, and partial sequences.

*The data for humans and mice were obtained from refs. 5 and 9, respectively.

Table 2. Numbers of functional and nonfunctional OR genes belonging to different groups in six species
words, each group corresponds to at least one ancestral OR gene in the most recent common ancestor (MRCA) between fishes and tetrapods. For example, group e clade was supported with a 99% bootstrap value, and this clade contained a tetrapod- and a fish-specific clade, suggesting that the MRCA between fishes and tetrapods had at least one ancestral group e gene. Similarly, the clades determining groups β, γ, δ, η, θ, and κ were supported with high (>80%) bootstrap values, and each of them contained both fish and tetrapod genes. Groups α and ζ are tetrapod- and fish-specific, respectively. However, Fig. 1B indicates that both group α and ζ genes diverged from other group genes before the fish–tetrapod divergence, because all other groups (β, γ, δ, and e) in type 1 clade were well supported and contained both fish and tetrapod genes. Moreover, the monophyly of group α and ζ genes was not supported, because group α genes formed a monophyletic clade with group β genes in Fig. 1B, suggesting that group α and ζ genes originated from different ancestral genes in the MRCA of fishes and tetrapods. Group ζ was supported only with a 67% bootstrap value (see Fig. 4), and therefore this group is not represented as a monophyletic clade in Fig. 1B. However, for simplicity, we regarded that group ζ is a monophyletic group. In this way, we classified type 1 and 2 genes into six (α–ζ) and three groups (η, θ, and κ), respectively. Note that one group may correspond to two or more ancestral genes in the fish–tetrapod MRCA. Therefore, this classification gives a minimal estimate of the number of ancestral OR genes in the MRCA between fishes and tetrapods. All of the human class I genes belonged to group α. However, known fish OR genes found in the public database, which are also called class I genes, belonged to four different groups, β, δ, e, and ζ. Therefore, class I genes correspond to genes belonging to five groups (α, β, δ, e, and ζ) in our new classification, whereas class II genes correspond to group γ genes.

We also classified OR pseudogenes and partial sequences into the nine groups defined above. Table 2 shows the numbers of functional and nonfunctional OR genes belonging to different groups in six species. The distribution of OR genes in the nine groups is quite different among species. Zebrafish have genes belonging to all groups except group α. By contrast, humans and mice have only group α and γ genes. (We conducted a TBLASTN search against human and mouse genome sequences by using all of the functional OR genes identified in the study as queries. However, we did not find any OR functional genes or pseudogenes belonging to the groups other than α or γ.) Almost all chicken OR genes belong to either group α or γ, but one putatively functional gene belonging to group θ was also found. The OR genes in X. tropicalis belong to all of the different groups except group ζ. Most X. tropicalis genes are group γ genes, which are very rare in fishes but are abundant in mammals or birds. However, X. tropicalis also have a considerable number of group δ genes, which are common in fishes.

Discussion

Evolution of Vertebrate OR Genes. The results in Table 2 can be summarized as follows. (i) Although the total number of OR genes is much smaller in fishes than in mammals or birds, the diversity of an OR gene family is much larger in fishes. (ii) OR genes present in mammals or birds (groups α and γ) are absent or very rare in fishes and vice versa. (iii) The OR gene family in amphibians is as diverse as that in fishes, but the majority of OR genes in amphibians and mammals or birds belong to the same group (group γ). In other words, the OR gene repertoire in amphibians is similar to that in fishes in one aspect and to that in mammals or birds in another aspect. As we have mentioned, it is possible that type 2 genes are actually non-OR genes. Note, however, that observations i–iii do not essentially change even if we confine our argument only to type 1 OR genes. The presence or absence of OR gene groups in the organisms living in the aquatic environment (fishes and amphibians) and those in the terrestrial environment (amphibians, birds, and mammals) suggests that group α and γ OR genes are specialized for detecting airborne odorants, and other group genes are for detecting water-soluble odorants.

Observations i–iii can be well understood by considering the long-term evolution of vertebrate OR genes (Fig. 2). As we mentioned earlier, the divergence between type 1 and 2 genes appears to have occurred before the divergence between jawed and jawless vertebrates. Therefore, the MRCA between jawed and jawless vertebrates is thought to have had at least two OR genes. Here we should note that another family of OR genes in lampreys (L1OR) has been reported (1, 33). However, we did not use L1OR genes, because L1OR genes are more similar to non-OR GPCRs such as histamine receptors than to any OR genes used in this study (data not shown), and thus the evolutionary origin of L1OR genes is different from that of the OR genes.

As indicated in Fig. 1B, the MRCA between fishes and tetrapods was estimated to have had at least six type 1 and three type 2 OR genes. Group α genes are specific to tetrapods, suggesting either that this group of genes have been lost in the fish lineage or that fish group α genes have not yet been found. Fish-specific group ζ genes can be explained in a similar manner. The MRCA between fishes and tetrapods is believed to have lived in the sea, and the OR genes in the MRCA therefore should have detected water-soluble odorants. Fishes retain OR genes belonging to almost all of the nine groups at present, probably because their environment has not changed substantially compared with that of the MRCA. In the tetrapod lineage, group α and γ genes apparently acquired the ability to detect airborne odorants at the time of terrestrial adaptation. After that, the OR genes specific to water-soluble odorants, i.e., genes other than group α or γ genes, appear to have been lost in the mammalian and avian lineages. Although the number of different OR gene groups substantially decreased in mammals and birds, the number of different OR genes, especially those belonging...
to group γ, have enormously increased. In the amphibian lineage, both the OR gene groups for detecting water-soluble odorants and those for airborne odorants are still present, reflecting that amphibians have adapted to both aquatic and terrestrial environments. Group γ gene expansion was observed in every tetrapod species examined, including amphibians. The reason for the expansion of group γ genes is unclear, but it is possible that more OR genes would have been required in the terrestrial than in the aquatic environment.

As we have seen above, the evolutionary dynamics of an OR gene family is characterized by processes in which some gene lineages have produced many descendants, whereas others have been completely lost. Therefore, the OR multigene family is an excellent example of birth-and-death evolution (5, 9, 34).

**Classification of Vertebrate OR Genes.** As we have mentioned, it is currently accepted that all vertebrate OR genes can be grouped into fish-like class I genes and mammalian-like class II genes. However, our phylogenetic analysis (Fig. 1B) suggested that vertebrate OR genes reported so far (i.e., type 1 OR genes) are classified into at least six different groups. This difference can be explained in the following way. The classification of OR genes into class I and II genes is primarily based on Freitag et al.’s study (19), which proposed that *X. laevis* have two different classes of OR genes, one (class I) being similar to known fish genes and the other (class II) being close to known mammalian genes. Our phylogenetic analysis supported Freitag et al.’s study, because *X. laevis* genes used in their study were assigned to group γ and ε in our classification, which contained mammalian and fish genes, respectively. (These *X. laevis* genes are not included in Fig. 1B, because they are partial, but we confirmed they are closest to group ε genes by homology search. Four *X. laevis* group ε genes in Fig. 1B are from ref. 22.) Moreover, many studies supported the view that mammalian OR genes are grouped into class I and II genes (4–6, 9). Our current analysis is also consistent with this view, because mammalian class I (group ε) and II genes (group γ) were clearly separated in Fig. 1B. Therefore, this portion of Fig. 1B is not inconsistent with previous studies. However, classification of all vertebrate OR genes into two classes would not be appropriate, because Fig. 1B indicated that class I genes have originated from at least five ancestral genes in the MRCA between fishes and tetrapods, whereas class II genes are likely to have originated from one ancestral gene. Moreover, as we have mentioned, the function of class I genes was unclear, because both fishes and mammals have a substantial number of class I genes. According to our new classification, however, OR genes belonging to each group are likely to be specific to either water-soluble or airborne odorants. Therefore, our classification of OR genes appears to be more reasonable than the previous one from both evolutionary and functional aspects. Our hypothesis of the functional specificity of each group of genes, however, should be tested experimentally. One way to test this hypothesis might be to examine the difference in gene expression pattern at the whole-genome level between larval and adult amphibians.

We thank Shozo Yokoyama, Alex Rooney, Jianzhi Zhang, Jan Klein, and Blair Hedges for helpful comments and discussion. This work was supported by National Institutes of Health Grant GM20293 (to M.N.). The sequence data for the zebrafish genome were produced by the Zebrafish Sequencing Group at the Sanger Institute and can be obtained from wwwensembl.org/Danio_rerio. The data for the pufferfish genome have been provided freely by the Fugu Genome Consortium for use in this publication only. The data for the *X. tropicalis* genome were provided by the Department of Energy’s Joint Genome Institute. The data for the chicken genome were provided by the Genome Sequencing Center, Washington University School of Medicine (St. Louis).