Olfactory Receptor Database: a database of the largest eukaryotic gene family

Emmanouil Skoufos1,2,*, Matthew D. Healy3, Michael S. Singer2, Prakash M. Nadkarni1, Perry L. Miller1 and Gordon M. Shepherd2

1Center for Medical Informatics and 2Section of Neurobiology, Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06511, USA and 3Research Division, Bristol-Myers Squibb, Wallingford, CT, USA

Received August 27, 1998; Revised and Accepted September 30, 1998

ABSTRACT

The Olfactory Receptor Database (ORDB) is a WWW-accessible database that stores data on Olfactory Receptor-like molecules (ORs) and has been open to the public since June 1996. It contains a public and a private area. The public area includes published DNA and protein sequence data for ORs, links to OR models and data on their expression, chromosomal localization and source organism, as well as (i) links to bibliography through PubMed and (ii) interactive WWW-based tools, such as BLAST homology searching. The private area functions as a service to laboratories that are actively cloning receptors. Source laboratories enter the sequences of the receptor clones they have characterized to the private database and can search for identical or near identical OR sequences in both public and private databases. If another laboratory has cloned and deposited an identical or closely matching sequence there are means for communication between the laboratories to help avoid duplication of work. ORDB is available via the WWW at http://crepe.med.yale.edu/ORDB/HTML

INTRODUCTION

Olfactory Receptors (ORs) are thought to be the largest eukaryotic gene family, including ~1000 different genes in the mouse (1). They belong to the superfamily of seven transmembrane domain, G protein-coupled receptors (GPCRs). ORs are thought to bind ~10 000 odor molecules, however, there is very little evidence regarding OR-ligand specificity. Furthermore, OR-like proteins are expressed in a variety of tissues including reproductive organs, olfactory epithelium and heart, suggesting that members of this family of proteins may have functions beyond odor recognition. To aid the cloning, sequencing and classification of the ORs, the Olfactory Receptor DataBase (ORDB) was created in 1995 as a response to a request from several laboratories in the field during the 1994 Association for Chemoreception Sciences meeting (2,3). The database contains sequence information about all the known ORs, supplemented with data about their properties. In its initial stages the database contained 147 OR sequences from eight species, information about source organism, sequencing laboratory and the nature (genomic DNA or cDNA) of the sequence. In addition, a private section was established where investigators that clone ORs may deposit their unpublished sequences and execute BLAST (4) searches of both public and private data to determine whether other laboratories have cloned ORs similar or identical to their own. ORDB is a member of the consortium of the GPCR databases that includes the GPCRDB (5), the GRAP mutant database (6), the mutation analysis of GPCRs database (7), and the GCRDb (8). As an example of particular use of this relationship, there are direct links at the record level from ORDB to GPCRDB receptor entries.

THE STATE OF ORDB, 1998

The latest version of ORDB contains 476 OR sequences from 13 species that represent the sequencing efforts of 52 laboratories worldwide. In addition to information, such as organism and nature of sequence, which has been included in the database since its debut, the newest version of ORDB contains information about the tissue from which an OR is cloned, the size (partial or full-length) of clones, chromosomal information (for 89 ORs), and direct links to the PubMed and/or GenBank records of the ORs. Furthermore, BLAST searches of the public database are now open to the public, and all users can search for similarities between any sequence of their interest and all sequences in the public database. A keyword search interface was also established, allowing users to access particular records based on keywords. To better serve the community of olfactory researchers using the database, an ORDB user group has been recently created. The user group is open to all olfactory researchers and it includes a discussion list that functions as a forum for the exchange of ideas and possible collaborations among the users. ORDB was originally UNIX-based, hosted on a Sybase server that communicated with an httpd WWW server using Perl. Recently ORDB has been upgraded to a Windows NT server, hosted on a Microsoft Access server communicating with Microsoft Internet Information Server using Active Sever Pages protocol. This upgrade increased the performance of the database and decreased access time, especially on tasks such as BLAST searches, 5- to 10-fold depending on the connection, for example, from 30 s to <5 s (data not shown).

*To whom correspondence should be addressed. Tel: +1 203 785 3730; Fax: +1 203 785 6990; Email: skoufos@spine.med.yale.edu
ORDB INTERFACE AND NAVIGATION

Frame technology is implemented, in order to make it easier for the users to navigate ORDB (Fig. 1). Users have four navigation choices directly from the home page. They can log in to the private database, search the database using keywords, search the OR sequences in the database using BLAST, or browse the database records. From the home page, and at any time using the constant navigating frame, users can visit an array of olfaction links, apply for access to the private database, and send their feedback to the database administrators. The browsing of the database is directly linked to keyword searching, thus providing the users with a variety of conceptual start points of the database: users can access the ORDB records based on sequencing laboratory, data source, type of sequence, organism, tissue, size of OR clone, GenBank accession number, external database reported or chromosomal localization (Fig. 1). The receptor entry screen contains detailed information about receptor properties, nucleotide and amino acid sequences of the clones, linked to the BLAST interface of ORDB and external links to PubMed and GenBank information as well as links to their entry in GPCRDB, if included (Table 1). The implementation of direct BLAST links within each receptor entry allows investigators to perform searches and see the relations of a particular receptor with the rest of the ORs in the database without leaving the screen that contains the other properties of the receptor.

Table 1. Lists of specific information for each receptor in ORDB

<table>
<thead>
<tr>
<th>Data source (where did we find the receptor)</th>
<th>Sequencing laboratory</th>
<th>Other database information (e.g. GenBank entry)</th>
<th>Organism</th>
<th>Tissue</th>
<th>Size of clone (full-length or partial)</th>
<th>Type of clone (cDNA, genomic)</th>
<th>Chromosome localization</th>
<th>Name</th>
<th>ORDB UID</th>
<th>Nucleotide sequence</th>
<th>Amino acid sequence</th>
<th>Links to PubMed, Genbank</th>
<th>Links to 2D receptor models</th>
</tr>
</thead>
</table>

Figure 1. ORDB entry point. The frame interface provides ease of navigation for the users. The users can log in to the private database, browse and search the database using keywords or BLAST and communicate with the database administrators. Once users select to browse the database they have a variety of options to retrieve receptor entries. This screen is automatically generated based on the fields of ORDB included in the keyword-searching engine of the database. In this way, the browsing of the database can start with a variety of different interest points.
DATA SOURCES AND CONSIDERATIONS

When ORDB started in 1995, the major data sources were direct submissions of sequences by the sequencing laboratories. Since the beginning of 1997 systematic searching of GenBank sequences using both keyword searching and BLAST similarity searching has been the main source of data in ORDB. Recently, the ORDB investigators discovered 25 new ORs based on sequence homology, produced in the human genome sequencing project that were added to the public version of ORDB. In this way, the ORDB administrating group, in addition to performing a hosting service, with the curation of the database, has become a data source. Furthermore, high quality sequences from the EST projects are routinely scanned for ORs, which are included in the database. In this way, ORDB users have a more complete idea about the expression patterns of OR genes. Now about one-third of the ORDB sequences are from genome and EST project data and their percentage is expected to increase as data from these projects accumulate faster than data from individual laboratories.

USES OF ORDB

ORDB has been used as a central source of Olfactory Receptor information. The use of the OR sequence information in ORDB has led to several studies aiding the understanding of the structure of the active site of the ORs (9,10), possible OR–odor molecule interactions (11), and identifying common domains among the receptors (E.Skoufos and G.M.Shepherd, manuscript submitted). In addition, ORDB has aided in new OR gene discovery by the ORDB group. Furthermore, ORDB has been of importance in developing the structure for a comprehensive nomenclature of the ORs (12). In order for the ORDB administrators to keep track of the uses of ORDB as well as ensure future funding for the database, it would be appreciated if any study that uses ORDB and leads to publication, includes this manuscript as a reference.

FUTURE CONSIDERATIONS

As additional types of information about ORs become available, they will be incorporated into ORDB. The structure of the database and its WWW interface are flexible enough to facilitate such changes to the database. In the near future, alignments and phylogenetic trees of the ORs will be included in the database, to complement the public BLAST searches and to give investigators a more complete idea about the relationships among the receptors. An additional concern of ORDB is that of OR nomenclature. ORs have been named by the sequencing laboratories in a variety of ways. ORDB has the potential to provide a central utility that laboratories can implement to name their receptors in a mutually agreed manner. It can function as a resource where laboratories will input their sequences to be included in the database and receive a name for each receptor assigned automatically by the database.

ACKNOWLEDGEMENTS

This work has been supported in part by NIH grant RO1 DC02307 (Human Brain Project) and NIH grants GO8 LM05583 and T15 LM07056 from the National Library of Medicine.

REFERENCES