SHORT COMMUNICATION

Olfactory Receptor Database (ORDB): A Resource for Sharing and Analyzing Published and Unpublished Data

Matthew D. Healy[1], Jason E. Smith[2], Michael S. Singer[1,2], Prakash M. Nadkarni[1], Emmanouil Skoufos[1,2], Perry L. Miller[1] and Gordon M. Shepherd[2]

[1]Center for Medical Informatics, Yale University School of Medicine, 333 Cedar Street, PO Box 208009, New Haven, CT 06520-208009 and [2]Section of Neurobiology, Yale University School of Medicine, 236 FMB, 333 Cedar Street, New Haven, CT 06510, USA

Correspondence to be sent to: Matthew Healy, Center for Medical Informatics, Yale University School of Medicine, 333 Cedar Street, PO Box 208009, New Haven, CT 06520-208009, USA

Abstract

An olfactory receptor database (ORDB) is being developed to facilitate analysis of this large gene family. ORDB currently contains over 400 olfactory receptor sequences and related information, and is available via the World Wide Web. We plan to incorporate functional data, structural models, spatial localization and other categories of information, toward an integrated model of olfactory receptor function. **Chem. Senses 22: 321–326, 1997.**

Introduction

Olfactory receptor (OR) molecules form one of the largest gene families yet identified. A mammalian species may have as many as 1000 different OR subtypes. They are believed to function as odor receptors, through differential affinities for various odor molecules (Buck and Axel, 1991; Ben-Arie *et al.*, 1993; Shepherd *et al.*, 1996), although experimental evidence for this function is still limited (cf. Raming *et al.*, 1993). ORs are members of the G protein-coupled receptor (GPCR) superfamily, which includes receptors for many neurotransmitters, hormones and other signal molecules (Schwartz, 1994).

This large number of OR genes poses severe problems in sequence determination and analysis. To date sequences have accumulated by the independent efforts of many laboratories worldwide. Many sequences remain unpublished for considerable amounts of time, resulting in barriers to access and duplication of research effort. There is also difficulty in locating related information on a given receptor, such as functional assays, molecular models and anatomical distribution studies. The well-established international databases such as GenBank and SwissProt have limited value in addressing these problems because they

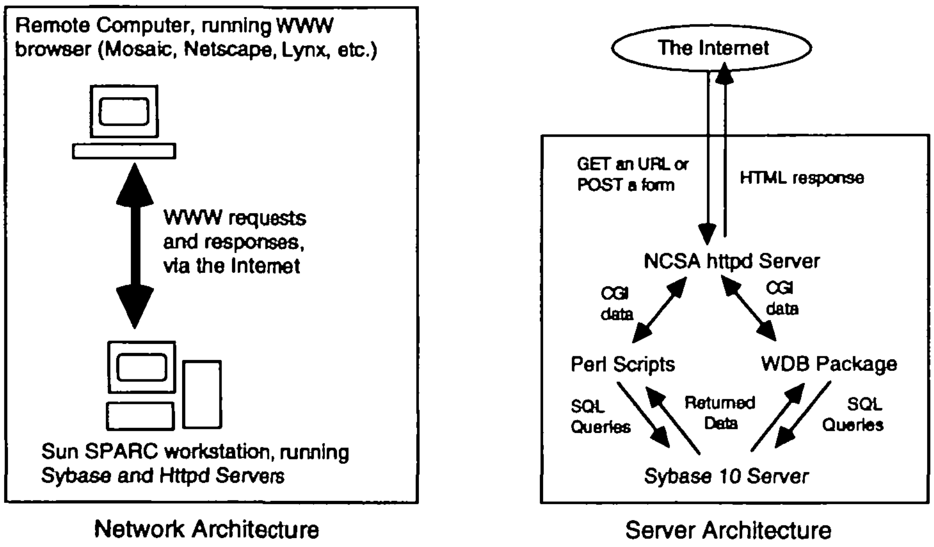

Remote Computer, running WWW browser (Mosaic, Netscape, Lynx, etc.)
WWW requests and responses, via the Internet
Sun SPARC workstation, running Sybase and Httpd Servers
Network Architecture
The Internet
GET an URL or POST a form
HTML response
NCSA httpd Server
CGI data
CGI data
Perl Scripts
WDB Package
SQL Queries
Returned Data
SQL Queries
Sybase 10 Server
Server Architecture

Figure 1 Overall system architecture.

do not yet offer services for unpublished data. Also, because of their broad nature, they do not include special fields and features desirable for an integrated OR database (Kolakowski, 1994).

Individual laboratories have generated local archives to store OR sequences. While these archives can be custom-designed to meet the needs of olfaction researchers, they seldom include true database architecture and functionality, such as custom searches or systematic backup procedures. The archives also fail to facilitate sharing of data between laboratories. Most critically, they do not provide for a systematic analysis of all sequences, published or unpublished.

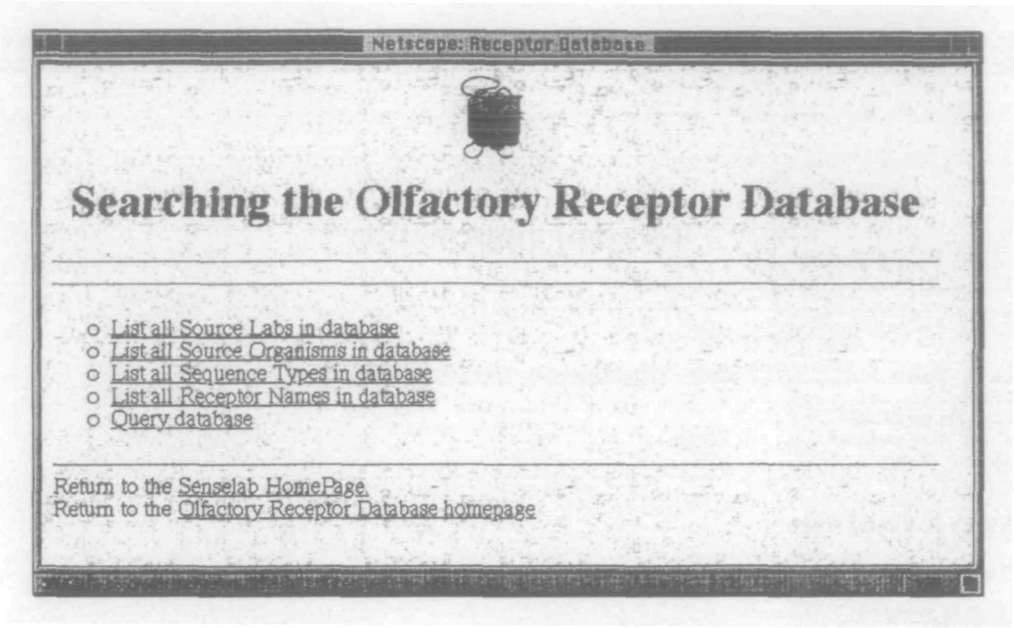
In order to begin to address these problems, several laboratories at the Association for Chemoreception Senses meeting in April 1994 discussed the need to develop a shared olfactory receptor database. They expressed particular interest in a mechanism for the controlled sharing of unpublished data. The availability of funding for this purpose under the new Human Brain Project provided an effective means for responding to this need. After two years of development, the first phase of the database, containing sequence information, is available on the World Wide Web (WWW) via browsers such as Netscape and Mosaic. It is named the Olfactory Receptor Database (ORDB), and serves as a common site for both nucleotide and amino acid sequences. The use of the WWW as a user interface makes ORDB available to olfactory researchers around the world, anywhere on the Internet, no matter what type of computer they use. It is a central, shared resource, like the other international sequence databases. However, it also has features specifically designed for the needs of the olfactory research community. ORDB is located at the URL http://senselab.med.yale.edu/ordb/.

The database holds published sequences, which are visible to all users. In addition, a special feature is that unpublished sequences are available to 'private' users under an 'invisible data' scheme described below. Both 'public' and 'private' users can search the database by terms such as receptor name, species and source laboratory. Private users can also conduct BLAST similarity screens and submit data on-line. ORDB operates on a system of dynamic output and input interfaces (see below), which facilitate searching and navigation. The database includes a directory of source laboratories and e-mail links to encourage communication between users. For those interested in other members of the GPCR superfamily, ORDB maintains links with other members of the international consortium of GPCR databases. ORDB was developed in response to the specific needs of the olfactory research community, but it may also be a useful prototype for dedicated databases serving other research communities.

Methods, results and discussion

Database architecture

ORDB uses the client/client/server structure illustrated in Figure 1. The remote user's computer runs a standard Web browser (e.g. Netscape or Microsoft Internet Explorer),



Searching the Olfactory Receptor Database

- List all Source Labs in database
- List all Source Organisms in database
- List all Sequence Types in database
- List all Receptor Names in database
- Query database

Return to the Senselab HomePage.
Return to the Olfactory Receptor Database homepage

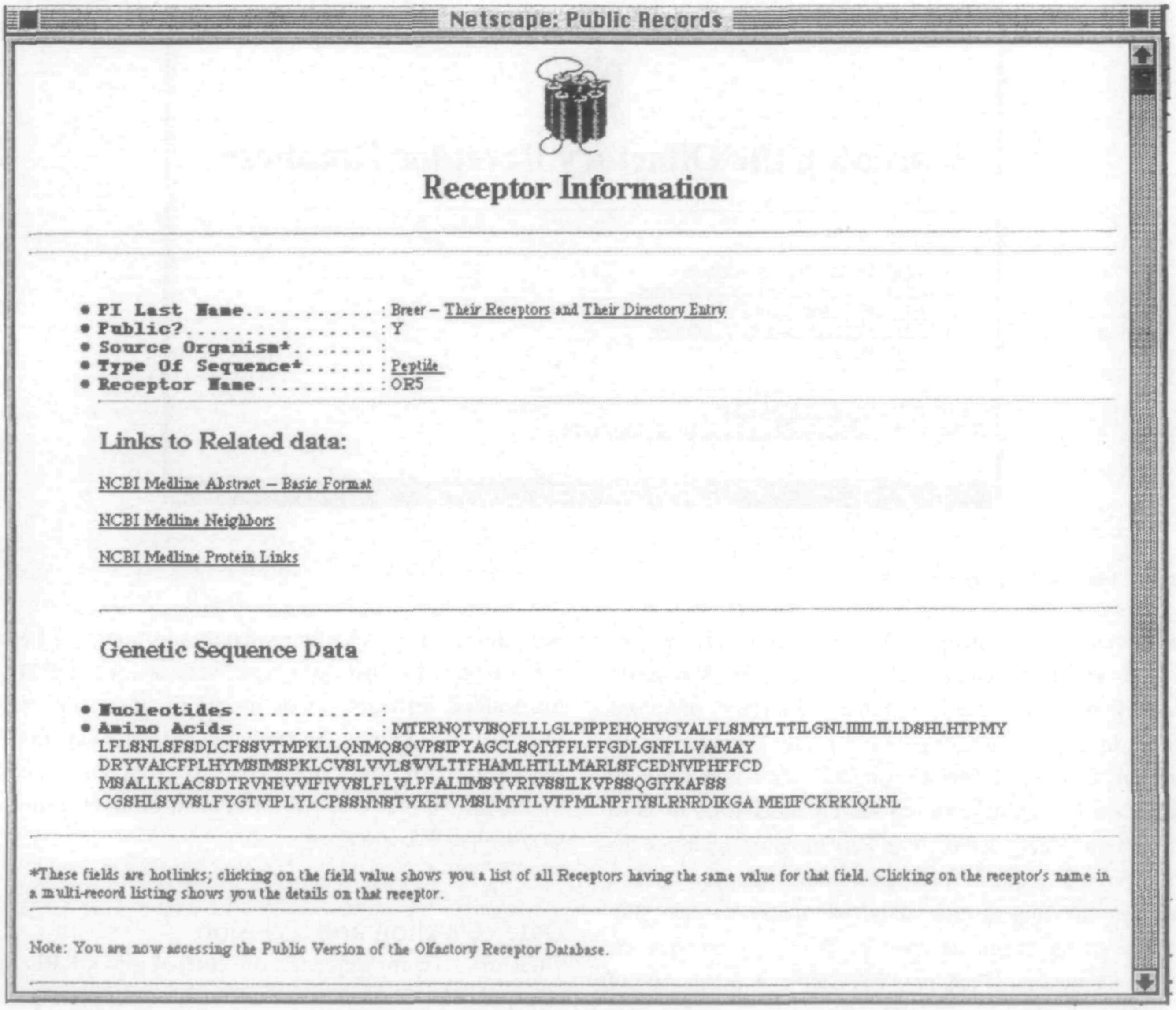
Figure 2 Initial search screen for ORDB.

which connects as a client to ORDB's Web server over the Internet. To handle each ORDB request, the Web server launches one of several database interface programs (written in the Perl language) via a protocol known as the Common Gateway Interface, or CGI. This is a standard mechanism for adding custom search capabilities to most Web servers (Healy, 1996). The Perl program analyzes the remote user's request, performs some security checks, generates the appropriate database query in the SQL database query language used by Sybase, then uses the Sybperl extensions (Peppler, 1992–1996) to access one of our Sybase database servers. The script then must convert the returned data into HTML format for viewing by standard Web browsers. Some of these programs were custom-written for ORDB, and others are from the WDB database access package (Rasmussen, 1994–1996). For sequence similarity searches, a Perl script launches programs from the BLAST package (Altschul *et al.*, 1990) and reformats the output as HTML. This Perl script is the most complex component of ORDB because it must interact with both Sybase and BLAST to perform its task.

Each of the 'private' users has a user name and a password defined using the standard password security feature built-in to most Web clients and servers. Each of the programs comprising ORDB's Web interface has two names: a 'public' name, which has no password protection and is thus available to anyone with a Web browser, and a 'private' name, which is password protected. When a program is run, it checks its name and alters its behavior accordingly. In this way, two versions of ORDB are maintained with one set of programs. By using the CGI interface, we combine the relational database capabilities of Sybase and the powerful sequence comparison capabilities of BLAST with a cross-platform, location-independent and easy-to-use Web interface.

Data collection and inclusion

Olfactory receptor sequences are entered into ORBD in two ways. (i) The Genbank, SwissProt, Protein Identification Resource (PIR) and European Molecular Biology Laboratory (EMBL) sequence depositories are searched for olfactory receptor sequences for inclusion in ORDB. ORDB has also recently joined the consortium of the GPCR databases, which will be another potential source of sequence information. (ii) In addition to entering data found in other databases, receptor sequences are submitted directly from the cloning laboratories to the ORDB through a Web data entry form. Each sequence submitted, before entry to the ORDB, is screened for errors in the content as well as for the presence of the OR signature motifs. In determining the quality of the data submitted directly from sequencing laboratories, ORDB is in a better position than are the large sequence depositories such as Genbank, since it is a database of homologous molecules. BLAST searches are performed for each receptor submitted, thus establishing its position as a homologous member of the OR family



Receptor Information

- **PI Last Name**........... : Breer – Their Receptors and Their Directory Entry
- **Public?**................ : Y
- **Source Organism***....... :
- **Type Of Sequence***...... : Peptide
- **Receptor Name**.......... : OR5

Links to Related data:

NCBI Medline Abstract – Basic Format

NCBI Medline Neighbors

NCBI Medline Protein Links

Genetic Sequence Data

- **Nucleotides**............ :
- **Amino Acids**............ : MTERNQTVISQFLLLGLPIPPEHQHVGYALFLSMYLTTILGNLIIIILILLDSHLHTPMY
LFLSNLSFSDLCFSSVTMPKLLQNMQSQVPSIPYAGCLSQIYFFLFFGDLGNFLLVAMAY
DRYVAICFPLHYMSIMSPKLCVSLVVLSWVLTTFHAMLHTLLMARLSFCEDNVIPHFFCD
MSALLKLACSDTRVNEVVIFIVVSLFLVLPFALIIMSYVRIVSSILKVPSSQGIYKAFSS
CGSHLSVVSLFYGTVIPLYLCPSSNNSTVKETVMSLMYTLVTPMLNPFIYSLRNRDIKGA MEIIFCKRKIQLNL

*These fields are hotlinks; clicking on the field value shows you a list of all Receptors having the same value for that field. Clicking on the receptor's name in a multi-record listing shows you the details on that receptor.

Note: You are now accessing the Public Version of the Olfactory Receptor Database.

Figure 3 Example of screen with receptor data and links.

group. Furthermore, like the large depositories, ORDB allows revisions and updates to each entry.

Data access and security

ORDB contains two distinct categories of OR sequence records. The public category includes published sequences, as well as any unpublished sequences the source laboratory wishes to release. Public access is available to all users on the WWW. Private access is only available (as 'invisible data') to private users. Any laboratory that is willing to submit unpublished data for ORDB will receive a name and password for private access upon request via a WWW form found on the ORDB Web site. Submission is encouraged, particularly of unpublished sequences, so that the database will be as useful as possible.

Users can search or browse ORDB sequence records by receptor name, species or source laboratory (see Figure 2). ORDB makes extensive use of hotlinks to allow easy cross references, as shown in Figure 3. These hotlinks are generated in real-time from database lookups (Healy, 1996). For example, if one is viewing an OR sequence and would like to see a list of other ORs from the same source laboratory, one merely clicks on the name of the source laboratory. Then one can click on the name of any OR in that list to see its detailed sequence screen. Since the ORDB displays are generated on-the-fly from database contents, all internal cross-references are automatically updated whenever new OR sequence records are added to the ORDB.

Private users have all the functions available to public users, plus access to information about private sequences. However, private users cannot view sequences that have been marked 'private' by other source laboratories. Only such information as receptor name, species and source laboratory is visible. The sequence itself is hidden unless the owner of the data has instructed us to make it available. The database server and all programs implementing the WWW interface are maintained by the Yale Center for Medical Informatics; the Yale Neuroscience personnel involved in the ORDB project have the same access to unpublished data as do other registered users. This 'invisible data' mechanism, in conjunction with the BLAST searching facility described below, protects the confidentiality of unpublished data while helping laboratories to avoid duplication of effort.

Similarity searches using BLAST

BLAST (Altschul *et al.*, 1990) is a standard tool used for comparing genetic sequences for local similarity. With ORDB, the BLAST programs can be used to conduct similarity screens of a query sequence against all available sequences in the database. ORDB automatically launches *blastp*, *tblastn* or *tblastx* based on the nature of the matches sought. If the query sequence consists of amino acids, then *blastp* is used to compare the query sequence with every amino acid sequence in the database and *tblastn* is used to compare the query sequence with all six possible translations (three reading frames in both orientations) of every nucleotide sequence in the database. If the query sequence consists of nucleotides, then *blastx* is used to compare all six possible translations of the query sequence with every amino acid sequence in the database and *tblastx* is used to compare all six possible translations of the query sequence with all six possible translations of every nucleotide sequence in the database. The PAM120 scoring matrix is used because according to the BLAST documentation this is the usual general-purpose matrix used when comparing peptide sequences and there is no specific reason for selecting another scoring matrix. All other BLAST search parameters are left at their default values.

For each search, BLAST returns via the WWW interface the list of best matches, scores, reading frame (when applicable) and statistical significance (Altschul *et al.*, 1990). By clicking on a BLAST match, the user gains immediate access to species, source laboratory, address and e-mail address, and the sequence itself. For private users, BLAST also returns notice of any matches with the unpublished sequences, together with the source laboratory, address and e-mail address. The sequences themselves are suppressed in order to protect the confidentiality of the unpublished data. The user must contact the source laboratory to determine the status of work on the sequence, and to discuss who will carry the sequence forward to publication.

Table 1 Projected databases within the ORDB group

Molecular models
Chromosome maps
Topographic maps
Phylogenetic relations
Expression vectors
Second messenger pathways
Odor ligands
Specific anosmias

Solutions and future directions

ORDB was built to address three main issues: cross-laboratory accessibility, duplication of effort and links to related information. Toward broad and convenient accessibility, the WWW was selected as the fastest, easiest and most universal medium. ORDB's Sun server processes searches in minimal time; searches from most sites within the USA are conducted in under 30 s. Access times from other locations will vary with network conditions. To minimize duplication of effort on the part of cloning laboratories, ORDB's 'invisible data' scheme allows one laboratory to learn that another laboratory has identified OR sequences very similar to its own, without revealing confidential data to either laboratory. The value of this feature should increase as more laboratories identify more of the estimated 300–1000 OR subtypes in each species. The 'invisible data' functions of ORDB have already received significant use. The database thus helps to address the desire of the National Center for Human Genome Research to encourage the swift dissemination of sequence data (Dickson, 1996). Finally, toward our goal of linking related sequence information, we have linked MEDLINE and the NCBI databases to records in ORDB.

The longer term goal of research on olfactory receptors is to understand the specificity of odor–receptor interactions, and the role that the receptors play in the development of the olfactory pathway and in odor perception and discrimination. This will require the integration of a range of related data, including molecular models, chromosome maps, phylogenetic trees, spatial distributions in the olfactory epithelium and olfactory bulb, and the results of functional assays. The sequence data are therefore only a first step toward making ORDB a database cluster to facilitate this integration. Table 1 lists several efforts, planned or in progress, which depend on the OR sequence data. Suggestions from users for ways we can improve ORDB and incorporate these new types of data are welcome.

ACKNOWLEDGEMENTS

We are grateful to Gerrit Vriend, Jason Mirsky and Gustavo Glusman for helpful discussions. This work is supported in part by NIH grant RO1 DC02307 from NIMH, NIDCD and NASA under the Human Brain Project, and by NIH grant G08 LM05583 from the National Library of Medicine.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Ben-Arie, N., Lancet, D., Taylor, C., Khen, M., Walker, N., Ledbetter, D.H., Carrozzo, R., Patel, K., Sheer, D. and Lehrach, H. (1993) Olfactory receptor gene cluster on human chromosome 17: possible duplication of an ancestral receptor repertoire. *Hum. Molec. Genet.*, **3**, 229–235.

Buck, L. and Axel, R. (1991) A novel multigene family may encode odorant receptors: a molecular basis for odorant recognition. *Cell*, **65**, 175–187.

Dickson, D. (1996) NIH seeks rapid sequence release. *Nature*, **380**, 279.

Healy M. (1996) Custom database query scripts. In Dwight, J. and Erwin, M. (eds), *Special Edition Using CGI*. QUE Corporation, Indianapolis, IN, pp. 323–362.

Kolakowski, L. (1994) GCRDb: a G protein-coupled receptor database. *Receptors Channels*, **2**, 1–7.

Peppler, M. (1992–1996) *Sybperl*. Available from the Internet sites: ftp://ftp.demon.co.uk/perl/perl4/sybperl and ftp://ftp.funet.fi/pub/languages/perl/CPAN plus many others.

Raming, K., Kreiger, J., Strotmann, J., Boekhoff, I., Kubick, S., Baumstark, C. and Breer, H. (1993). Cloning and expression of odorant receptors. *Nature*, 361, 353–356.

Rasmussen, B. (1994–1996) *WDB: A Web to Database Interface*. On-Line Document, available on the Internet at the URL: http://venus.dtv.dk/~bfr/wdb/

Schwartz, T.W. (1994) Locating ligand-binding sites in 7TM receptors by protein engineering. *Curr. Opin. Biotech.*, **5**, 434–444.

Shepherd, G.M., Singer, M.S. and Greer, C.A. (1996) Olfactory receptors: a large gene family with multiple functions. *The Neuroscientist*, in press.

Received on September 23, 1996; accepted on February 10, 1997