THE ROYAL SOCIETY

doi 10.1098/rstb.2001.0903

The Ligand Gated Ion Channel database: an example of a sequence database in neuroscience

Nicolas Le Novère[1]* and Jean-Pierre Changeux[2]

[1]*Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK*
[2]*Neurobiologie Moléculaire, CNRS URA D1284, Institut Pasteur, 75724 Paris, France*

Multiple comparisons of receptor sequences, or receptor subunit sequences, has proved to be an invaluable tool in modern pharmacological investigations. Although of outstanding importance, general sequence databases suffer from several imperfections due to their size and their non-specificity. Room therefore exists for expert-maintained databases of restricted focus, where knowledge of the research field helps to filter the huge amount of data generated. Accordingly, neuroscientists have designed databases covering several types of proteins, in particular receptors for neurotransmitters.

Ligand-gated ion channels are oligomeric transmembrane proteins involved in the fast response to neurotransmitters. All these receptors are formed by the assembly of homologous subunits, and an unexpected wealth of genes coding for these subunits has been revealed during the last two decades. The Ligand Gated Ion Channel database (LGICdb) has been developed to handle this growing body of information. The database aims to provide only one entry for each gene, containing annotated nucleic acid and protein sequences.

Keywords: ion channel; sequence database; neurotransmitter receptor; sequence comparisons

1. INTRODUCTION

The comparison of the multiple sequences revealed by molecular cloning has driven a complete revolution of modern biological research. Gilbert (1991) referred to a 'paradigm shift in biology', in which experiments would be designed by the analysis of existing data using bioinformatic and biocomputational methods. This paradigm shift partly arose from the analyses of sequences and their comparison. In pharmacology—the study of drug receptors, their interactions with their ligands, and the consequences of these interactions for living matter—the revolution has been deep. It has affected both the theoretical and the methodological frameworks driving the investigations performed in the field.

Pharmacological diversity is indeed crucial to an understanding of the physiology of the organism and, in particular, of the nervous system. Many receptors for neurotransmitters are made of homologous subunits, coded by genes derived from a common ancestor. In the ionotropic receptors, for instance, the sequence variability between homologous subunits may generate differences in the permeability properties and in the affinities of the various structural states for the ligands. In turn, those variations not only affect binding events at a given concentration of ligand, but also affect the kinetics of the allosteric transitions between the different structural states. Consequences range from variation in mean opening time to modification of the desensitization properties (Galzi *et al.* 1996).

2. THEORETICAL INTEREST OF MULTIPLE SEQUENCE COMPARISONS

The discovery, for almost every kind of receptor, of numerous homologous genes, both orthologous (genes appearing by speciation) and paralogous (genes appearing by duplication within a species), challenged what we could call the 'typological thought'. It was a tacit rule to define, for each kind of receptor, what the palaeontologists call a 'holotype', that is a particular receptor considered as the model for the whole group—the other members being *de facto* defined with reference to this one (see the example of the nicotinic acetylcholine receptor of the *Torpedo* electric organ, for instance). The wealth of homologous sequences revealed a much more nuanced picture, with a continuum of variation invalidating the typological thought, and providing a basis for the wide and puzzling pharmacological spectra observed.

As a consequence, another conceptual outcome of the multiple sequence comparisons has been to modify the classification of receptors. Receptor classifications are heuristic and our understanding of a receptor family, namely the function of its different members and their relationships, is biased by its subdivisions and the notions hidden behind the name of each member. Indeed, a useful classification of any group member (or element) is that, within one group, the elements have more relationships between themselves than with any element of any other group. Such a classification can often lead to the elimination of numerous wrong directions for research.

The classical pharmacological classifications were based on ligand specificities because they were the first available data. Due to the typological thought, the classifications were often based on one particular ligand considered to be 'informative' or 'relevant'. This notion of relevance is largely subjective and driven by the type of study as well as the scientific background of the investigator. As a consequence, the emphasis on different characteristics has led to different classifications, the choice between them often being rather arbitrary. Moreover, these classifications have had to be constantly re-examined in parallel with the increases in knowledge.

The International Union of Pharmacology (IUPHAR) Committee on Receptor Nomenclature and Drug Classification recommended the integration of structural, operational (or pharmacological) and transductional data (Kenakin *et al*. 1992). However, because the functions of proteins depend on their three-dimensional structure, and because this structure is encoded in their amino-acid sequence, the operational and transductional characteristics strongly rely (although sometimes in a complex way) on the structural characteristics. Moreover, from a more theoretical point of view, the effect of a signal is determined by the receptor itself and not by the ligand. This is true at the cellular level; that is, the final effect upon the cell function of acetylcholine, for instance, depends on the receptor structure: a G-protein-coupled device for the muscarinic receptors, or an ionic channel for the nicotinic receptors. This remains valid even if, at the level of a tissue or an organism, we can sometimes speak about the 'cholinergic effects' or the 'retinoic effects'. Finally, the operational criteria are sometimes misleading. For instance, it is now well documented that the pharmacologies of orthologous receptors, i.e. the 'same' receptors in different species, can differ dramatically. Moreover, in the case of polymeric receptors the ligand binding sites are sometimes located at the interface between subunits. If heteromeric receptors exist, there is no actual one-to-one correspondence between the specific subunits and the particular pharmacological spectra.

The relationship between structure and function of a protein implies that the probability of similar properties for proteins will increase with the similarity of sequences. Even if a given characteristic is in contradiction with this 'law', the analysis of many characteristics will reinforce it. The general rule is that the phenotypic similarity will converge to the phylogenetic similarity with the number of characteristics taken into consideration. A good knowledge of the evolutionary relationship of genes, based on careful multiple sequence comparisons, is therefore a most favourable basis for establishing informative and robust classifications of receptors, or receptor subunits.

3. PRACTICAL ASPECTS OF MULTIPLE SEQUENCE COMPARISONS

Obvious uses of multiple sequence comparisons in pharmacology are the sequence analyses *per se*, performed, for instance, to unravel the evolutionary history of the receptors (e.g. Le Novère & Changeux 1995) or to predict certain structural features (e.g. Le Novère *et al*. 1999). But the information derived from multiple sequence comparisons can also be of outstanding interest for other kinds of experiments.

The study of the molecular evolution of gene families is necessary to fully understand the current relationships between their members. In addition to the light shed on known proteins, the molecular phylogeny sometimes permits the discovery of new proteins. For instance, the existence of the dopaminergic receptor Dlc was predicted from molecular phylogenies, and it was thereafter cloned (Cardinaud *et al*. 1997). Another example is the cloning *in silico* of the prokaryotic glutamate receptor (Chen *et al*. 1999). The gene was already present in the general-purpose sequence databases, but was not properly annotated. Only a careful screen based on an iterative multiple alignment process (with the program PSI-BLAST; Altschul *et al*. 1997) permitted its discovery.

Careful analyses and comparisons of known nucleic acid sequences have led, and still lead, to the discovery of new homologous genes using cloning 'by similitude'. To perform the cloning we need adequate probes (or adequate primers for the PCR approaches), i.e. we need to know the closest sequences in the family. If we are able to infer, only by its phenotypical properties, the probable molecular position of a given receptor in the family, we should be able to improve the probability of success of the research. The design of primers has to incorporate enough diversity to maintain a good specificity, while nevertheless retaining sufficient conservation to grant good hybridization with the template. The best way to tackle this kind of problem is typically to perform a careful multiple sequence analysis to determine the optimal primers, and not simply to choose primers in the most conserved or the most variable regions as is often the case.

The study of the structural organization of receptors obviously relies on careful analysis of their sequences, for instance setting up experiments of site-directed mutagenesis (e.g. Galzi *et al*. 1992) or constructing chimeric proteins (e.g. Corringer *et al*. 1998). There are already numerous possibilities for mutations of a single residue. However, understanding the functional structure of receptors often requires taking into account local interactions between several residues. The combinatorial possibilities therefore explode, impeding any exhaustive investigation. The parallel comparisons of existing receptor sequences and functional characteristics restrain the space of interesting modifications within bench possibilities.

Consideration of multiple sequence alignments in the 1980s also boosted the accuracy of structure prediction (Rost & Sander 1996). Feeding the programs with multiple alignments resulted in distinction of the informative variability from the evolutionary noise. The quality of the multiple alignments therefore became an important factor driving the accuracy of the predictions.

Finally, although more indirectly, almost every design of an experiment in molecular biology involves a step of sequence comparisons. Not only is the preparation of the experiment based on sequence analysis (e.g. design of PCR primers, *in situ* hybridization probes, or choice of epitopes to produce specific antibodies), but its interpretation is also directed by the knowledge of the level of sequence resemblance (for instance the amount of cross-hybridization of probes or cross-recognition of antibodies).

All the experiments described above require fast access to a complete and accurate set of the sequences already known. To be efficiently used, those sequences have to be trustworthy, i.e. they have to contain a minimal number of errors (ideally none, of course), and they have to be available in a format that is immediately usable.

4. DEFECTS OF THE GENERAL SEQUENCE DATABASES

Most of the biological sequences uncovered so far are stored primarily in enormous public databases (Benson *et al.* 2000; Stoesser *et al.* 2001; Tateno *et al.* 2000). Those repositories are growing at a fast rate, which recently switched from geometric to exponential with the generalization of high-throughput sequencing of complete genomes (they contained *ca.* 9.6 billion nucleotides at the beginning of September 2000).

Although of outstanding importance, the general sequence databases suffer from several defects due to their size and their broad purpose. First of all, unwanted errors are sometimes made during the submission process; those errors are not always corrected. In addition, each gene is often represented by multiple entries. This multiplicity is generated from intrinsic causes, such as alternative splicing or editing, from methodology, for instance cDNA versus genomic cloning, but also from competition between laboratories, each submitting its own clone. For instance, the cDNA coding for the human nicotinic acetylcholine receptor subunit α7 is present in at least 11 different EMBL entries.

If a certain level of curation is sometimes achieved for the protein sequence repositories such as SWISS-PROT (Bairoch & Apweiler 2000) and PIR (Barker *et al.* 2000), for practical reasons this is almost never the case for the nucleic acid sequence repositories (the treatment of the 6400 new submissions entered daily in EMBL at the end of 1999 would require a prohibitively large number of highly skilled curators).

Finally, because of the lack of firm rules to select the name and the definition of the entries, the retrieval of a particular sequence within a general database can be cumbersome. For instance, the following three definitions are extracted from public database entries:

∇υμαν μϱ φωρ μυσψλε αψετθλψηωλινε ρεψεπτωρ αλπηα συβυνιτ•
ωυσε μϱ φωρ μυσψλε νιψωτινιψ αψετθλψηωλινε ρεψεπτωρ αλπηα•
∇•σαπιενσ νιψωτινιψ ρεψεπτωρ αλπηα <συβυνιτ μϱ ∵ ψωμπλετε ψδσ•

Searches with the keywords 'acetylcholine' or 'nicotine' would here both result in a partial retrieval.

A more reliable approach is often the screening of the database with a known sequence using tools such as BLAST (Altschul *et al.* 1997) or FASTA (Pearson & Lipman 1988). However, this process itself requires some preexisting knowledge of the sequences (such as the variability of the sequences across species and subtypes, as well as the distribution of this variability along the sequence) to use those tools efficiently. Moreover, the resulting output is generally not usable, in a raw state.

There is therefore room for expert-maintained databases, of restricted focus but higher quality, where the knowledge of the research field would help to filter the huge amount of data generated.

The computing tools needed to establish the databases have been in existence since the end of the 1970s, when the main sequence and structure databases were established. It was already possible to access the databases remotely—the 'Internet' was established back in the 1960s. However, those activities were still restricted to specialists. Two concomitant events drastically changed the handling of public data. The development of the World Wide Web provided a convenient tool to access a wide range of objects transparently. In parallel, a computer appeared on almost every scientist's desk, and was quickly connected to the global network.

Almost every researcher in the life sciences had (and probably still has) their own data collection, sometimes in notebooks, sometimes in a personal computer. That was especially true for the molecular biologists who carefully recorded the sequences. With the spreading of the network, they were henceforth able to make this knowledge widely available to the scientific community.

5. SEQUENCE DATABASES IN THE NEUROSCIENCES

The increasing size of the multigene families coding for proteins of neurobiological interest, and particularly for pharmacological receptors, gave rise to specialized sequence repositories. Those databases contain nucleic acid and protein sequences as well as atomic structures, when they are available. In addition, they often provide a first generation of sequence analysis, i.e. multiple sequence alignments and phylogenetic analysis (table 1). Note that the frequency of update is highly variable, and some of these databases could currently be moribund.

Although most of the databases deal with the fields of G-protein coupled receptors and extracellularly activated ligand-gated ion channels, some other collections exist, for instance ESTHER (Cousin *et al.* 1998). This server, created in 1994, is dedicated to the analysis of protein and nucleic acid sequences belonging to the superfamily of alpha/beta hydrolases, homologous to cholinesterases. The database was still maintained in January 2001.

The *Receptor database* intends to cover the protein sequences of all types of receptors (Nakata *et al.* 1999). To achieve this ambitious goal, the entries are just unprocessed duplicates of the PIR and SWISS-PROT entries, although carefully ordered. In addition, the database provides secondary structure predictions for each protein presented (note that the predictions are performed with NNPREDICT, a program of very low accuracy according to current standards; see Lesk (1997)). It seems that the focus of the database is restricted to mammals. The database was still maintained in December 2000.

The G-protein-coupled receptors are monomeric transmembrane proteins that sometimes associate in dimers (Zoli *et al.* 1993). Signal transduction is mediated by the physical interaction with GTP-activated proteins, which in turn act on a wide range of effectors such as enzymes or ionic channels (Morris & Malbon 1999).

The *G protein Coupled Receptor Database* (GCRDB) was one of the first receptor databases to be established, back in 1989, and is still one of the most complete dealing with G-protein-coupled receptors (Kolakowski 1994). It contains a

Table 1. Sequence databases in neuroscience.

database	reference	content
G-protein-coupled receptors		
GCRDb	Kolakowski (1994)	sequences
GPCRDB	Horn *et al.* (1998)	sequences, alignments, phylogeny, structures
GRAP	Kristiansen *et al.* (1996)	list of mutants
ORDB	Skoufos *et al.* (1999)	sequences, alignments, phylogeny,
Receptor database	Nakata *et al.* (1999)	sequences, structure predictions
Ligand-gated ion channels		
GABAagent	Rachedi *et al.* (2000)	sequences, bibliography
LGICdb	Le Novère & Changeux (1999)	sequences, alignments, phylogeny, structures
Receptor database	See upper panel	
Other databases		
ESTHER	Cousin *et al.* (1998)	sequences, structures

huge amount of cleverly organized data. The last update of the GCRDB was made in August 1999. This database is now part of a network integrating the different databases dealing with G-protein-coupled receptors, organized around the *G Protein-Coupled Receptor Data Base* (GPCRDB) (Horn *et al.* 1998).

The *Olfactory Receptor DataBase* (ORDB) is focused on a particular group of G-protein-coupled receptors, the olfactory receptors (Skoufos *et al.* 1999). It was still maintained in December 2000.

Finally the GRAP lists all the mutants available for G-protein-coupled receptors (Kristiansen *et al.* 1996). It was still maintained in November 2000.

Ligand-gated ion channels (LGIC) are polymeric transmembrane proteins. Their physiological effect is mediated by the opening of an ionic channel upon the binding of a particular ligand. We will not deal here with the intracellularly-activated ion channels such as the receptors for inositol phosphate and cyclic nucleotides, but only with the extracellularly-activated LGIC, largely responsible for the fast response to neurotransmitters.

Two main databases collect the sequences of LGIC. Apart from the LGICdb presented below, the recently created GABAagent is focused on the receptor subunits homologous to the GABA ionotropic receptor subunits (Rachedi *et al.* 2000). It contains an unprocessed duplication of entries coming directly from the general purpose sequence databases. It is currently impossible to know how up-to-date the repository is.

6. THE LIGAND GATED ION CHANNEL DATABASE

The last two decades have revealed an unexpected wealth of genes coding for LGIC subunits. The Ligand Gated Ion Channel database (LGICdb) has been developed to handle this growing knowledge, initially during a molecular phylogenetic survey of the nicotinic receptor subunits (Le Novère & Changeux 1995). It was made available via the World Wide Web in 1995 and has since been described succinctly in the literature (Le Novère & Changeux 1999, 2001). The LGICdb's scope is currently limited to the extracellularly-activated transmitter-gated channels, i.e. the superfamilies 1.1, 1.2 and 1.5 of Barnard (1996).

The 'cys-loop' superfamily (nicotinic receptors, $GABA_A$ and $GABA_C$ receptors, glycine receptors, $5\text{-}HT_3$ receptors and some glutamate-activated anionic channels) contain receptors made up of five homologous subunits (Galzi & Changeux 1994; Ortells & Lunt 1995), each containing a characteristic loop of 13 residues flanked by cysteines; hence the name. The ATP-gated channels (ATP P2X receptors) are at this stage considered to be made up of three homologous subunits (Nicke *et al.* 1998). Finally, the cationic channels activated by excitatory amino acids (NMDA receptors, AMPA receptors, kainate receptors, etc., often referred to as cationic glutamate receptors) are made up of four homologous subunits (Dingledine *et al.* 1999). The members of the three superfamilies are not homologous, i.e. the gene coding for the subunits does not derive from a common ancestral gene. Accordingly, those subunits are not expected to display the same three-dimensional structure and have different transmembrane organizations; see figure 1.

The release 24 of the LGICdb (January 2001) contained 380 subunit entries belonging to 37 different species (table 2). The LGICdb is accessible via the World Wide Web (http://www.pasteur.fr/recherche/banques/LGIC/LGIC.html), where it is regularly updated.

(a) *Structure of the database entries*

The database intends to provide one unique entry for each gene, containing annotated nucleic acid and protein sequences, together with references to the cloning articles and the relevant accession numbers for other databases.

Contrary to numerous other types of data (see the other papers in this issue), the sequences are rather simple pieces of data. The sequence itself has just one dimension: a string of characters, and the surrounding information (references, various annotations) can be attached on the same level.

The structure of the LGICdb inner file containing the data (also called the 'flat file') is exemplified below, with the entry Ψ∇σαδδμ (only an informative subset of this entry's information is displayed; see figure 2 for the complete entry). The current format uses a mark-up style resembling those of the typesetting formats L_YX or RTF, that is, a section mark-up remains active until another section mark-up is reached.

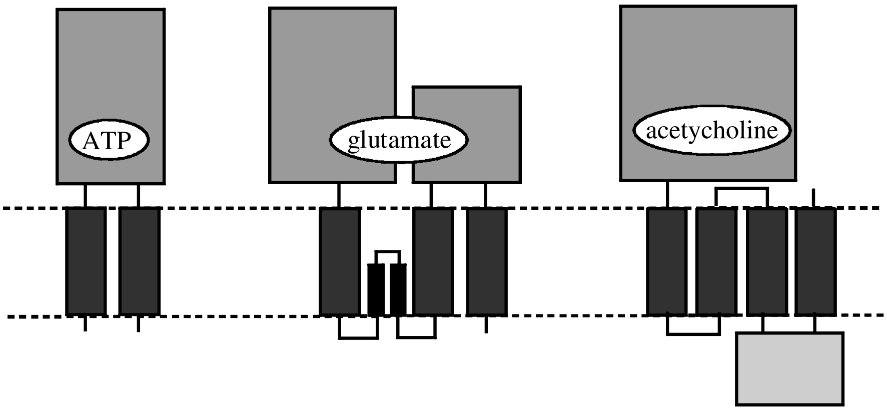


Figure 1. Transmembrane topology of the three superfamilies of neurotransmitter-gated ion channel: left, P2X receptor subunit; centre, excitatory amino-acid receptor subunit; right, nicotinic receptor subunit.

Table 2. Content of the LGICdb, release 24.

superfamily of trimeric ATP receptor subunits	21
superfamily of tetrameric excitatory amino-acid receptor subunits	
NMDA receptor subunits	18
δ subunits	4
kainate and AMPA receptor subunits	33
plant subunits	20
superfamily of pentameric receptor subunits	
anionic channels	
GABA receptor subunits	84
glycine receptor subunits	13
glutamate receptor subunits	7
cationic channels	
serotonin receptor subunits	6
acetylcholine receptor subunits	174

```
\LGICID ACHsaddm
```

Each entry possesses a unique, meaningful identifier, which permits the immediate classification of the entry within the database (see below for the naming system).

```
\CREATION 22/AUG/1999
\LASTMODIF 11/DEC/2000
```

The date of creation of the LGICdb entry is specified (which is unrelated to the date of appearance of the sequences related to the entry in the general databases). The date of last modification of the entry is also provided, which is an important feature used in some other sequence databases in neuroscience (e.g. the GCRDb), which allows, for instance, the automatic generation of customized updates.

```
\SPECIES
Drosophila melanogaster
```

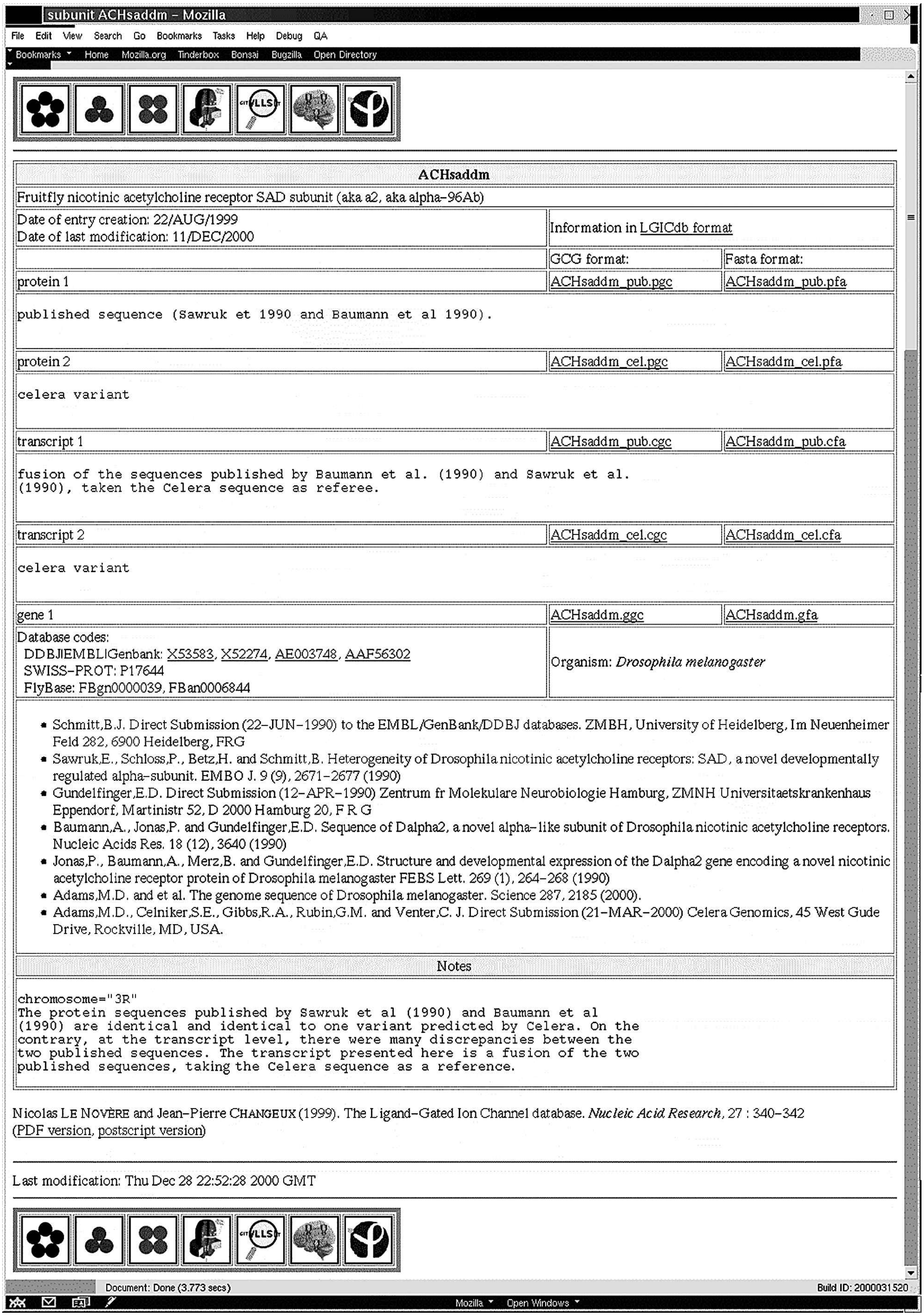
The two-component name of the species is notified. The complete classification, as present in EMBL or GenBank for example, has been omitted because it is of little direct interest. Tables, nevertheless, present the data ordered according to the species systematic (see below).

```
\DEFINITION
Fruitfly nicotinic acetylcholine recept\
or SAD subunit (aka alpha2, aka alpha-9\
6Ab)
```

A definition of the entry is given in one line. This definition is intended to describe the entry, and not a particular sequence (the backslash, as in all further examples, means a continuation of the same line). This particular subunit has received three different names, from three different groups. The most rational, with respect to the whole group of homology, is retained. If the terminologies are logically equivalent, the one used most in the literature is retained.

```
\REFERENCES
Schmitt,B.J. Direct Submission (22-JUN-\
1990) to the EMBL/GenBank/DDBJ database\
s. ZMBH, University of Heidelberg, Im N\
euenheimer Feld 282, 6900 Heidelberg, F\
RG
Sawruk,E., Schloss,P., Betz,H. and Schm\
itt, B. Heterogeneity of Drosophila nic\
otinic acetylcholine receptors: SAD, a \
novel developmentally regulated alpha-s\
ubunit. EMBO J. 9 (9), 2671-2677 (1990)
Gundelfinger,E.D. Direct Submission (12\
-APR-1990) Zentrum fr Molekulare Neurob\
iologie Hamburg, ZMNH Universitaetskran\
kenhaus Eppendorf, Martinistr 52, D 200\
0 Hamburg 20, FRG
Baumann,A., Jonas,P. and Gundelfinger,E\
.D. Sequence of Dalpha2, a novel alpha-\
like subunit of Drosophila nicotinic ac\
etylcholine receptors. Nucleic Acids Re\
s. 18 (12), 3640 (1990)
Adams,M.D., Celniker,S.E., Gibbs,R.A., \
Rubin,G.M. and Venter,C. J. Direct Subm\
ission (21-MAR-2000) Celera Genomics, 4\
5 West Gude Drive, Rockville, MD, USA.
```

The references to the original publications and submissions to the general purpose databases are presented, one per line. The above example contains two publications by two different groups, and three direct database submissions, which were all used in the construction of this LGICdb entry. The references are not currently processed, as in the GCRDB where the authors are separated; it did not appear of primary importance to do so. In addition,


subunit ACHsaddm – Mozilla
File Edit View Search Go Bookmarks Tasks Help Debug QA
Bookmarks Home Mozilla.org Tinderbox Bonsai Bugzilla Open Directory
Document: Done (3.773 secs)
Build ID: 2000031520
Mozilla Open Windows

ACHsaddm		
Fruitfly nicotinic acetylcholine receptor SAD subunit (aka a2, aka alpha-96Ab)		
Date of entry creation: 22/AUG/1999 Date of last modification: 11/DEC/2000	Information in LGICdb format	
	GCG format:	Fasta format:
protein 1	ACHsaddm_pub.pgc	ACHsaddm_pub.pfa
`published sequence (Sawruk et 1990 and Baumann et al 1990).`		
protein 2	ACHsaddm_cel.pgc	ACHsaddm_cel.pfa
`celera variant`		
transcript 1	ACHsaddm_pub.cgc	ACHsaddm_pub.cfa
`fusion of the sequences published by Baumann et al. (1990) and Sawruk et al. (1990), taken the Celera sequence as referee.`		
transcript 2	ACHsaddm_cel.cgc	ACHsaddm_cel.cfa
`celera variant`		
gene 1	ACHsaddm.ggc	ACHsaddm.gfa
Database codes: DDBJIEMBLIGenbank: X53583, X52274, AE003748, AAF56302 SWISS-PROT: P17644 FlyBase: FBgn0000039, FBan0006844	Organism: *Drosophila melanogaster*	

- Schmitt,B.J. Direct Submission (22-JUN-1990) to the EMBL/GenBank/DDBJ databases. ZMBH, University of Heidelberg, Im Neuenheimer Feld 282, 6900 Heidelberg, FRG
- Sawruk,E., Schloss,P., Betz,H. and Schmitt,B. Heterogeneity of Drosophila nicotinic acetylcholine receptors: SAD, a novel developmentally regulated alpha-subunit. EMBO J. 9 (9), 2671-2677 (1990)
- Gundelfinger,E.D. Direct Submission (12-APR-1990) Zentrum fr Molekulare Neurobiologie Hamburg, ZMNH Universitaetskrankenhaus Eppendorf, Martinistr 52, D 2000 Hamburg 20, F R G
- Baumann,A., Jonas,P. and Gundelfinger,E.D. Sequence of Dalpha2, a novel alpha-like subunit of Drosophila nicotinic acetylcholine receptors. Nucleic Acids Res. 18 (12), 3640 (1990)
- Jonas,P., Baumann,A., Merz,B. and Gundelfinger,E.D. Structure and developmental expression of the Dalpha2 gene encoding a novel nicotinic acetylcholine receptor protein of Drosophila melanogaster FEBS Lett. 269 (1), 264-268 (1990)
- Adams,M.D. and et al. The genome sequence of Drosophila melanogaster. Science 287, 2185 (2000).
- Adams,M.D., Celniker,S.E., Gibbs,R.A., Rubin,G.M. and Venter,C. J. Direct Submission (21-MAR-2000) Celera Genomics, 45 West Gude Drive, Rockville, MD, USA.

Notes

```
chromosome="3R"
The protein sequences published by Sawruk et al (1990) and Baumann et al
(1990) are identical and identical to one variant predicted by Celera. On the
contrary, at the transcript level, there were many discrepancies between the
two published sequences. The transcript presented here is a fusion of the two
published sequences, taking the Celera sequence as a reference.
```

Nicolas LE NOVÈRE and Jean-Pierre CHANGEUX (1999). The Ligand-Gated Ion Channel database. *Nucleic Acid Research*, 27 : 340-342 (PDF version, postscript version)

Last modification: Thu Dec 28 22:52:28 2000 GMT

Figure 2. The main HTML page presenting one entry of the LGICdb. The chosen example is ACHsaddm which is a fairly complex entry. The underlined words are hyperlinks to sequence files, stored either locally (GCG or FASTA files) or remotely (GenBank entries).

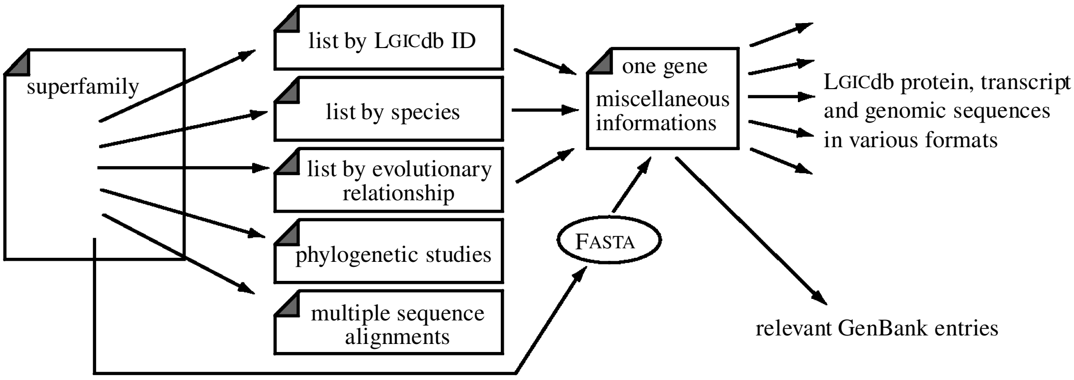

superfamily
list by LGICdb ID
list by species
list by evolutionary relationship
phylogenetic studies
multiple sequence alignments
one gene
miscellaneous informations
FASTA
LGICdb protein, transcript and genomic sequences in various formats
relevant GenBank entries

Figure 3. Structure of the LGICdb.

the references quoted in the general purpose sequence databases are entered by the submitters, and often contain abbreviations, such as *et al.*, which invalidate the recognition of individual authors. The retrieval of all the LGICdb entries containing a particular name would, nevertheless, be a trivial task for any pattern recognition program using regular expressions.

```
\OTHERIDS
DDBJ|EMBL|Genbank:X53583,X52274,AE00374\
8,AAF56302
FlyBase:FBgn0000039,FBan0006844
SWISS-PROT:P17644
```

The accession numbers of the sequences related to the entry in various databases are presented. For instance, accessions for DDBJ (Tateno *et al.* 2000), EMBL (Stoesser *et al.* 2001), GenBank (Benson *et al.* 2000), SWISS-PROT (Bairoch & Apweiler 2000) and FlyBase (The Flybase Consortium 1999) are presented here.

```
\NOTES
chromosome="3R"
The protein sequences published by Sawruk
et al (1990) and Baumann et al (1990) are
identical and identical to one variant
predicted by Celera. On the contrary, at
the transcript level, there were many
discrepancies between the two published
sequences. The transcript presented here
is a fusion of the two published sequences,
taking the Celera sequence as a reference.
```

A section can contain any notes related to the entry or any particular sequence merged into the entry. Those notes will be considered verbatim and thus can be formatted as desired by the submitter (for instance, carriage returns are conserved).

```
\PROTEINS
:pub
%published sequence (Sawruk et 1990 and
%Baumann et al 1990).
MAPGCCTTRPRPIALLAHIWRHCKPLCLLLVLLLLCETVQ
ANPDAKRLYDDLLSNYNRLIRPVSNNTDTVLVKLGLRLSQ
\\
:cel
%Celera variant
MAPGCCTTRPRPIALLAHIWRHCKPLCLLLVLLLLCETVQ
ANPDAKRLYDDLLSNYNRLIRPVSNNTDTVLVKLGLRLSQ
\TRANSCRIPTS
:pub
%fusion of the sequences published by
%Baumann et al. (1990) and Sawruk et al.
%(1990), taking the Celera sequence as
%reference.
GACAGCACGGAGCGGGGCCCAAAGGCTTGTTGAAATCAAG
TGAAAGTCCGCTTAAAACTGCACACAAAAATATTGAAAAA
\\
:cel
%celera variant
TCAGTTAACAAGTTTGAATATTTTTTAGAATTTTTTAAGC
ACGAAATTGAGTTGGTGAAAATTAAAAAGACTTTTTAAAT
\GENES
GCTTGTTGAAATCAAGTGAAAGTCCGCTTAAAACTGCACA
CAAAAATATTGAAAAATCAGTTAACAAGTTTGAATATTTT
```

Several protein sequences can be entered, separated by \\, together with mRNA and genomic sequences (the sequences in this example are truncated). Spaces, numbers and any letters are accepted (but dashes and question marks, for example, are not permitted). Each sequence may also possess an optional tag, which is used to construct the name of the files generated for this particular sequence. Individual annotations can also be attached, treated verbatim.

To accommodate both the planned increase of complexity of the flat files and the multiplication of the user-driven treatments, the current format will soon be converted into an XML-compatible grammar, which will be able to handle nested levels of specification.

(b) *Structure of the database*

There is a many-to-one relationship between the transcripts and their corresponding genes, and in most cases the relationship between the transcript and the protein is one-to-one. It is therefore not necessary to consider complex database designs (see, for instance, the extreme simplicity of GenBank and its equivalents).

The flat files are processed by a Perl script which converts the sequences into various usual formats. Currently, the FASTA format (Pearson & Lipman 1988) (one of the simplest) and the GCG format (Devereux *et al.* 1984) (one of the most frequently used) are provided. The script constructs one HTML page per entry (see figure 2) which contains all the information contained in the flat file and, in addition, hyperlinks to the files containing the various sequences and to the relevant entries of GenBank. Finally, the script constructs two tables which allow the database to be browsed by species and entry ID (figure 3). Browsable trees which allow the retrieval of entries by homology are also available; these are currently constructed by hand based on expert phylogenetic analyses.

If one knows a fragment of a sequence, the database can be quickly screened to retrieve entries presenting similar stretches of residues. This is achieved with the program FASTA (Pearson & Lipman 1988). The result of the search can be reformatted to give a multiple alignment by the program MVIEW (Brown *et al.* 1998). Communication between the different programs and the generation of the HTML interfaces is performed by the program PISE (Letondal 2000).

Snapshots of the database are available to be downloaded as compressed archive files. They contain only the flat files, but include the Perl script for reconstructing the whole database.

A unified system of cascading identification tags has been set up, which is congruent both with the sequence similarities of the subunits and the pharmacology of the receptors they form. A first tag, made of three capitalized digits, represents the main endogenous ligand (5HT, ACH, ATP, GAB, GLU, GLY). A second tag, of variable length, identifies the ortholog (e.g., a1 for α1, nmr2 for NMDAR2, etc.). Finally a two-digit tag identifies the species. The successive tags are cascading, that is, when read from left to right, their meaning is unambiguous within the framework defined by the previous one. For instance, the subunits ACHa1+ and ACHa2+ (where a '+' sign means one or more characters) belong to a natural group, both in term of pharmacology and of sequence; all form nicotinic receptors, and all the ACH+ form a monophyletic group. On the contrary, GABa1+ and ACHa1+ do not belong to a natural group. They do not participate to the same type of oligomeric receptors, and the +a1+ subunits do not form a monophyletic group. One can easily extend this explanation to the next level, taking as an example the pair ACHa1hs, ACHa1mm as a natural group, in contrast to the pair ACHa1hs, ACHa2hs.

(c) *Construction of the database*

The LGICdb is mainly fed from the general-purpose databases (DDBJ, EMBL, GenBank, SWISS-PROT), but also from the contributions of users (see Acknowledgements) and sometimes from published articles. Rather than automatically transforming the flat files of general purpose databases into a LGICdb flat file, as in the GCRDB, every bit of data is thoroughly scrutinized, and manually processed before inclusion in the LGICdb. This has been possible because of the small number of genes involved (until recently) and necessary because of the uneven quality of the original data.

Some mistakes have been detected in the sequences present in the general-purpose databases. Those detections were possible only because the readers of the sequences had a deep knowledge of the field, and thus identified improbable stretches of nucleotides or amino acids. Such error screening cannot be performed with the automated procedures available at present. When a clear mistake is identified, it is corrected either with the help of the paper describing the cloning or by comparison of the different entries in the general purpose databases.

As stated earlier, it is common to find redundant entries in the general-purpose databases. In such cases, we tend to keep the bigger clone, assuming that it might contain interesting regulatory sites. Sometimes the final sequence was obtained by the fusion of several clones to achieve the maximum length. The authors of every clone are, nevertheless, quoted in the final LGICdb entry.

The information presented in the general sequence databases is sometimes fragmented. In particular, genomic clones contain coding and non-coding sequences. When the description of the gene structure is present in the database entry (determined experimentally or automatically), the putative transcript sequence was reconstructed (although the whole gene remains present in the entry). In the case of genomic sequences, the gene is sometimes coded on the complementary strand, not on the one presented in the general databases. In such a case, we determined the 'reverse-complement' of the sequence to present the coding strand in the LGICdb.

When several alternative splicing products exist, all of them are included in one entry, with an explanation entered in the note section. The same rule will apply when the edited subunits of the glutamate receptors are treated. If several authors provide different sequences, without any obvious mistakes, all of the variants are presented, being considered as alleles.

All these processing steps imply some fading of the specific features of each clone. When several sequences, coming from different tissues, are merged, the origin becomes irrelevant, because the sequence of a gene coding for a LGIC subunit is supposedly the same in all the cells of an organism (the somatic mutations, considered as abnormalities, cannot be considered in the LGICdb) and therefore has to be suppressed. This loss of experimental precision is the price to pay for greater data accuracy and clarity. It is not, in fact, an actual loss, because the original sequence files are still present in the general-purpose databases. The simple duplication of all the relevant GenBank or EMBL entries, as is often the case in the specialized databases, does not bring much additional information to the user and obfuscates the issue. Moreover, because the DDBJ, EMBL or GenBank will always be more up-to-date, the outcome of such a compilation could actually be a more misleading framework than a potential improvement of efficiency for the experimentalist.

In addition to the gene entries, the LGICdb provides atomic coordinates when available. Those coordinates come from experimental determinations but also from modelling work. They are presented as standard PDB files.

Finally, multiple sequence alignments (Clustal format) and expert phylogenetic investigations are also available.

The database is continuously maintained, although with irregular updates. At the time of writing, approximately one upload is performed per month. New entries are, of course, regularly added, with previous entries updated according to the improvements in knowledge: sequences of proteins and transcripts are completed or corrected, and genomic information added when available.

7. CONCLUSION

Started back in 1993 as a personal limited collection of regular GenBank files, the Ligand Gated Ion Channel database has grown to a full scale public database, largely

Table 3. URLs of the databases quoted in the text

database	URL
DDBJ	http://www.ddbj.nig.ac.jp/
EMBL	http://www.ebi.ac.uk/embl/
ESTHER	http://www.ensam.inra.fr/cholinesterase/
FLYBASE	http://flybase.bio.indiana.edu/
GABAagent	http://gaba.ust.hk/Agent.html
GCRDb	http://www.gcrdb.uthscsa.edu/
GenBank	http://www.ncbi.nlm.nih.gov/
GPCRDB	http://www.gpcr.org/7tm/
GRAP	http://www-grap.fagmed.uit.no/GRAP/
LGICdb	http://www.pasteur.fr/recherche/banques/LGIC/
ORDB	http://ycmi.med.yale.edu/senselab/ordb/
PIR	http://pir.georgetown.edu/
Receptor database	http://impact.nihs.go.jp/RDB.html
SWISS-PROT	http://www.expasy.ch/sprot/

acknowledged in the field of neurotransmitter-activated channels. In addition to the outstanding knowledge of LGIC subunit sequences it brought to its curators, the LGICdb has been employed by several teams as a reliable source of sequences. Those sequences were subsequently used, for instance, to perform studies of molecular phylogeny or structure predictions, and also to generate multi-purpose sequence alignments. The structure of the database entries has been enriched, and the resulting dataset now incorporates an important body of manually added information beside the raw sequences picked from the general-purpose databases.

With the achievement of the various genome projects, access to the raw sequences is no longer a challenge. On the contrary, the emphasis is more on the ease of retrieval of a particular piece of data, and on the annotations which allow a faster treatment of this data. Moreover, there is evidence that the reliability of sequences found in the general-purpose databases has decreased with the high-throughput sequencing. The post-sequencing treatments, made by experts able to check the quality of the raw data, therefore become more important. The expert-maintained databases can be of outstanding interest for both experimentalists and theoreticians, who rely on the quality of the sequences they use.

The Service d'Informatique Scientifique of the Institut Pasteur provides the computing resources needed to maintain the database (http and ftp servers). Catherine Letondal constructed the interface to the program FASTA. Alain Bessis, Benoit Lacombe, Wladimir Saudek and Ralf Schoepfer helped us by providing sequences. Thanks to Howard Baylis, Jim Boulter, Alban de Kerchove D'Exaerde, Anne Devillers-Thiéry, Aymeric Duclert, Ronald Lukas, Yoav Paas and Hongjie Yang for their advice and corrections during the construction of the LGICdb. The authors are grateful to Matthew Levin for his thorough reading of the manuscript.

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.

Bairoch, A. & Apweiler, R. 2000 The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45–48.

Barker, W. C. (and 14 others) 2000 The protein information resource (PIR). *Nucl. Acids Res.*, **28**, 41–44.

Barnard, E. 1996 The transmitter-gated channels: a range of receptor types and structures. *Trends Pharmac. Sci.* **17**, 305–309.

Benson, D. A., Karsch-Mzrachi, I., Lipman, D. J., Ostell, J., Rapp, B. A. & Wheeler, D. L. 2000 GenBank. *Nucl. Acids Res.* **28**, 15–18.

Brown, N., Leroy, C. & Sander, C. 1998 Mview: A web compatible database search or multiple alignment viewer. *Bioinformatics* **14**, 380–381.

Cardinaud, B., Sugamori, K., Coudouel, S., Vincent, J., Niznik, H. & Vernier, P. 1997 Early emergence of three dopamine d1 receptor subtypes in vertebrates. Molecular phylogenetic, pharmacological, and functional criteria defining d1a, d1b, and d1c receptors in European eel *Anguilla anguilla*. *J. Biol. Chem.* **272**, 2778–2787.

Chen, G., Cui, C., Mayer, M. & Gouaux, E. 1999 Functional characterization of a potassium-selective prokaryotic glutamate receptor. *Nature*, **402**, 817–821.

Corringer, P.-J., Bertrand, S., Bohler, S., Edelstein, S. J., Changeux, J.-P. & Bertrand, D. 1998 Critical elements determining diversity in agonist binding and desensitization of neuronal nicotinic acetylcholine receptors. *J. Neurosci.* **18**, 648–657.

Cousin, X., Hotelier, T., Giles, K., Toutant, J.-P. & Chatonnet, A. 1998 aCHEdb: the database system for ESTHER, the α/β fold family of proteins and the cholinesterase gene serve. *Nucl. Acids Res.* **26**, 226–228.

Devereux, J., Haeberli, P. & Smithies, O. 1984 A comprehensive set of sequence analysis programs for the VAX, *Nucl. Acids Res.*, **12**, 387–395.

Dingledine, R., Borges, K., Bowie, D. & Traynelis, S. 1999 The glutamate receptor ion channels. *Pharmac. Rev.*, **51**, 7–61.

Galzi, J.-L. & Changeux, J.-P. 1994 Neurotransmitter-gated ion channels as unconventional allosteric proteins. *Curr. Opin. Struct. Biol.* **4**, 554–565.

Galzi, J., Devillers-Thiéry, A., Hussy, N., Bertrand, S., Changeux, J. & Bertrand, D. 1992 Mutations in the channel domain of a neuronal nicotinic receptor convert ion selectivity from cationic to anionic. *Nature* **359**, 500–505.

Galzi, J.-L., Edelstein, S. J. & Changeux, J.-P. 1996 The multiple phenotypes of allosteric receptor mutants. *Proc. Natl Acad. Sci. USA* **93**, 1853–1858.

Gilbert, W. 1991 Towards a paradigm shift in biology. *Nature*, **349**, 99.

Horn, F., Weare, J., Beukers, M. W. Hörsch, S., Bairoch, A., Chen, W., Edvardsen, Y., Campagne, F. & Vriend, G. 1998 GPCRDB: an information system for G protein-coupled receptors. *Nucl. Acids Res.* **26**, 275–279.

Kenakin, T., Bond, R. & Bonner, T. 1992 Definition of pharmacological receptors. *Pharmac. Rev.* **44**, 351–362.

Kolakowski, L. 1994 GCRDb: A G protein-coupled receptor database. *Recept. Channels* **2**, 1–7.

Kristiansen, K., Dahl, S. G. & Edvardsen, O. 1996 A database of mutants and effects of site-directed mutagenesis experiments on G-protein coupled receptors. *Proteins: Struct. Funct. Genet.* **26**, 81–94.

Le Novère, N. & Changeux, J.-P. 1995 Molecular evolution of the nicotinic acetylcholine receptor subunit family: an example of multigene family in excitable cells. *J. Mol. Evol.* **40**, 155–172.

Le Novère, N. & Changeux, J.-P. 1999 The ligand gated ion channel database. *Nucl. Acids Res.* **27**, 340–342.

Le Novère, N. & Changeux, J.-P. 2001 LGICdb: the ligand-gated ion channel database. *Nucl. Acids Res.* **29**, 294–295.

Le Novère, N., Corringer, P.-J. & Changeux, J.-P. 1999 Improved secondary structure prediction of a nicotinic receptor subunit. Incorporation of solvent accessibility and experimental data into a 2D representation. *Biophys. J.* **76**, 2329–2345.

Lesk, A. M. 1997 CASP2: reports on ab initio prediction. *Proteins: Struct. Funct. Genet.* **29** (suppl 1), 151–166.

Letondal, C. 2000 A web interface generator for molecular biology programs in Unix. *Bioinformatics* **17**, 73–82.

Morris, A. J. & Malbon, C. C. 1999 Physiological regulation of G protein-linked signaling. *Physiol. Rev.* **79**, 1373–1430.

Nakata, K., Takai, T. & Kaminuma, T. 1999 Development of the receptor database (RDB): application to the endocrine disruptor problem. *Bioinformatics* **15**, 544–552.

Nicke, A., Bäumert, H., Rettinger, J., Eichele, A., Lambrecht, G., Mutschler, E. & Schmalzing, G. 1998 P2X1 and P2X3 receptors form stable trimers: a novel structural motif of ligand-gated ion channels. *EMBO J.* **17**, 3016–3028.

Ortells, M. O & Lunt, G. G. 1995 Evolutionary history of the ligand-gated ion-channel superfamily of receptors. *Trends Neurosci.*, **18**, 121–126.

Pearson, W. R & Lipman, D. J. 1988 Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* **85**, 2444–2448.

Rachedi, A., Rebhan, M. & Xue, H. 2000 GABAagent: a system for integrating data on GABA receptors. *Bioinformatics* **16**, 301–312.

Rost, B. & Sander, C. 1996 Bridging the protein sequence–structure gap by structure predictions. *A. Rev. Biophys. Biomol. Struct.* **25**, 113–136.

Skoufos, E., Marenco, L., Healy, M., Singer, M., Nadkarni, P., Miller, P. & Shepherd, G. 1999 Olfactory receptor database: a database of the largest eukaryotic gene family. *Nucl. Acids Res.* **27**, 343–345.

Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., Kulikova, T., Lombard, V., Lopez, R., Parkinson, H., Redaschi, N., Sterk, P., Stoehr, P. & Tuli, M. 2001 The EMBL nucleotide sequence database. *Nucl. Acids Res.* **29**, 17–21.

Tateno, Y., Miyazaki, S., Ota, M., Sugawara, H. & Gojobori, T. 2000 DNA Data Bank of Japan (DDBJ) in collaboration with mass sequencing teams. *Nucl. Acids Res.* **28**, 24–26.

The FlyBase Consortium 1999 The flyBase database of the drosophila genome projects and community literature. *Nucl. Acid Res.* **27**, 85–88.

Zoli, M., Agnati, L., Hedlund, P., Li, X., Ferre, S. & Fuxe, K. 1993 Receptor–receptor interactions as an integrative mechanism in nerve cells. *Mol. Neurobiol.* **1997**, 293–334.