

LGICdb: the ligand-gated ion channel database

Nicolas Le Novère* and Jean-Pierre Changeux¹

Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK and ¹Récepteurs et Cognition, CNRS URA 2182, Institut Pasteur, 75724 Paris cedex 15, France

Received September 11, 2000; Accepted September 13, 2000

ABSTRACT

Ligand-Gated Ion Channels (LGIC) are polymeric transmembrane proteins involved in the fast response to numerous neurotransmitters. All these receptors are formed by homologous subunits and the last two decades revealed an unexpected wealth of genes coding for these subunits. The Ligand-Gated Ion Channel database (LGICdb) has been developed to handle this increasing amount of data. The database aims to provide only one entry for each gene, containing annotated nucleic acid and protein sequences. The repository is carefully structured and the entries can be retrieved by various criteria. In addition to the sequences, the LGICdb provides multiple sequence alignments, phylogenetic analyses and atomic coordinates when available. The database is accessible via the World Wide Web (<http://www.pasteur.fr/recherche/banques/LGIC/LGIC.html>), where it is continuously updated. The version 16 (September 2000) available for download contained 333 entries covering 34 species.

INTRODUCTION

Ligand-Gated Ion Channels (LGIC) are polymeric transmembrane proteins. Their physiological effect is carried out by the opening of an ionic channel upon binding of a particular ligand. We will not deal here with the intracellularly activated ion channels such as receptors for the inositol phosphate and the cyclic nucleotides, but only with the extracellularly activated LGIC, mainly responsible for the fast response to neurotransmitters. The Ligand-Gated Ion Channel database (LGICdb) scope is therefore currently limited to the superfamilies 1.1, 1.2 and 1.5 of Barnard (1).

The superfamily of nicotinicoid receptors (nicotinic receptors, GABA_A and GABA_C receptors, glycine receptors, 5-HT₃ receptors and some glutamate activated anionic channels) are made of five homologous subunits (2,3). The ATP gated channels (ATP P2x receptors) are made of three homologous subunits (4). Finally, the cationic channels activated by excitatory amino acid (NMDA receptors, AMPA receptors, kainate receptors etc., often referred to as cationic glutamate receptors) are made of four homologous subunits (5). The members of the three superfamilies are not homologous, i.e. the genes coding for the subunits do not descend from a common ancestor gene. Accordingly, these subunits do not display the same tri-dimensional

structure and present, for instance, different transmembrane organisations.

The last two decades have revealed an unexpected wealth of genes coding for the LGIC subunits. The LGICdb has been developed to handle this growing knowledge, initially during a molecular phylogenetic survey of the nicotinic receptor subunits (6). It was made available on the World Wide Web in 1995 and was initially presented in the literature in 1999 (7). The LGICdb has since evolved. Not only has its size increased, but its World Wide Web address has changed and its internal structure has been profoundly modified.

The release 16 of the LGICdb (September 2000) contained 333 subunit entries belonging to 34 different species (Table 1). The LGICdb is accessible via the World Wide Web (new address: <http://www.pasteur.fr/recherche/banques/LGIC/LGIC.html>), where it is continuously updated.

Table 1. Content of the LGICdb, in release 0.1, presented in the database issue of 1998 and release 16 (September 2000)

	0.1	16
Superfamily of trimeric ATP receptor subunits	7	19
Superfamily of tetrameric excitatory amino-acid receptor subunits	23	53
NMDA receptor subunits		18
δ subunits		4
Kainate and AMPA receptor subunits		31
Superfamily of pentameric receptor subunits		
Anionic	84	100
GABA receptor subunits		81
Glycine receptor subunits		13
Glutamate receptor subunits		6
Cationic	98	161
Serotonin receptor subunits		6
Acetylcholine receptor subunits		155

STRUCTURE OF THE DATABASE

The database is intended to provide one unique entry for each gene and all the information is now contained in an internal flat file. The exact format of this file is detailed on the database Web site. Briefly, it contains a unique identifier for the entry, the date of creation of the LGICdb entry, the Linnean name of the species, a one line definition of the entry, the references of

*To whom correspondence should be addressed. Tel: +44 1223 336623; Fax: +44 1223 336676; Email: lenov@pasteur.fr

the original publications and submissions to the general purpose databases, the accession numbers of the sequences related to the entry in various databases, a section containing any notes related to the entry or any particular sequence merged into the entry, and lastly the protein, transcripts and genomic sequences.

The current format uses a markup style resembling that of the typesetting formats L_AT_EX or RTF, where a specifier is active until the next one. To ease the planned increase in complexity of the flat files as well as the multiplication of the user-driven treatments, the current format will soon be converted into an XML compatible grammar, able to handle nested levels of specification.

The flat files are processed by a Perl script which convert the sequences into various usual formats. Currently the FASTA format (8) (one of the simplest) and the GCG format (9) (one of the most used) are provided. The script then constructs one HTML page per entry, which contains all the information and the hyperlinks to the files containing the various sequences. The script also writes tables which permit browsing of the database by species and by entry Id. Browsable trees which allow the retrieval of entries by homology are also available, at present manually constructed based on expert phylogenetic analyses.

Finally, if one knows a piece of sequence, the database can be quickly screened to retrieve entries presenting similar stretches of residues. This is achieved with the program FASTA (8). The result of the search can be reformatted to give a multiple alignment by the program MVIEW (10). The communication between the different programs and the generation of the HTML interfaces is made by the program PISE (11). Snapshots of the database are made from time to time and are available for download as compressed archive files. They contain only the flat files, but include the perl script to reconstruct the whole database.

The LGICdb is mainly fed from the various general purpose databases but also from the contributions of users (see Acknowledgements) and sometimes from published articles. Rather than automatically transform the flat files of general purpose databases into a LGICdb flat file, every bit of data is thoroughly scrutinised, and manually processed before inclusion in the LGICdb. This has been possible because of the reasonable number of genes involved (until recently), and has been necessary because of the unequal quality of the original data.

Mistakes are sometimes detected in the sequences present in the general purpose databases. When a clear error is identified, it is corrected either with the help of the article presenting the cloning or by the comparison of the various entries present in the general purpose databases. In case of complete redundancy (without discrepancy), we try to keep the bigger clone, assuming that it could contain interesting regulatory sites. Sometimes the final sequence was obtained by the fusion of several clones in order to obtain the longest piece. 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, the 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 is reconstructed (the whole gene remaining present in the entry anyway). 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 determine the 'reverse-complement' of the sequence in order to present the coding strand in the LGICdb.

When several alternative splicing products exist, all of them are included in one entry, some explanation being entered in the note section. If several authors provide different sequences, without any obvious sequence mistake, all the variants are presented, considered as alleles. In the current flat file, the variants cannot be individually annotated. However, it is clear that this feature is required, and this will be implemented in the near future.

In addition to the gene entries, the LGICdb provides atomic coordinates when available. These coordinates come from experimental determination but also from modelling work. Finally, multiple sequence alignments and expert phylogenetic investigations are also available.

ACKNOWLEDGEMENTS

Catherine Letondal constructed the interface to the program FASTA. Alain Bessis, 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, Hongjie Yang for their advice and corrections during the construction of the LGICdb. The Service d'Informatique Scientifique of the Institut Pasteur provides the computing resources needed to maintain the database (<http> and <ftp> servers).

REFERENCES

- Barnard, E. (1996) The transmitter-gated channels: a range of receptor types and structures. *Trends Pharmacol. Sci.*, **17**, 305–309.
- Galzi, J.-L. and Changeux, J.-P. (1994) Neurotransmitter-gated ion channels as unconventional allosteric proteins. *Curr. Opin. Struct. Biol.*, **4**, 554–565.
- Ortells, M.O. and Lunt, G.G. (1995) Evolutionary history of the ligand-gated ion-channel superfamily of receptors. *Trends Neurosci.*, **18**, 121–126.
- Nicke, A., Bäumert, H., Rettinger, J., Eichele, A., Lambrecht, G., Mutschler, E. and Schmalzing, G. (1998) P2x1 and p2x3 receptors form stable trimers: a novel structural motif of ligand-gated ion channels. *EMBO J.*, **17**, 3016–3028.
- Dingledine, R., Borges, K., Bowie, D. and Traynelis, S. (1999) The glutamate receptor ion channels. *Pharmacol. Rev.*, **51**, 7–61.
- Le Novère, N. and 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. and Changeux, J.-P. (1999) The ligand-gated ion channel database. *Nucleic Acids Res.*, **27**, 340–342.
- Pearson, W.R. and Lipman, D.J. (1998) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
- Devereux, J., Haeblerli, P. and Smithies, O. (1984) A comprehensive set of sequence analysis programs for the vax. *Nucleic Acids Res.*, **12**, 387–395.
- Brown, N., Leroy, C. and Sander, C. (1998) Mview: a web compatible database search or multiple alignment viewer. *Bioinformatics*, **14**, 380–381.
- Letondal, C. (2000) A web interface generator for molecular biology programs in unix. *Bioinformatics*, in press.