Nucleic Acids Research, 2006, Vol. 34, Database issue ***D267–D269***
doi:10.1093/nar/gkj104

LGICdb: a manually curated sequence database after the genomes

Marco Donizelli, Marie-Ange Djite and Nicolas Le Novère*

European Bioinformatics Institute, EMBL, Wellcome-Trust Genome Campus, Hinxton CB10 1SD, UK

Received September 14, 2005; Revised and Accepted October 17, 2005

ABSTRACT

Ligand-gated ion channels form transmembrane ionic pores controlled by the binding of chemicals. The LGICdb aims to be a non-redundant, manually curated resource offering access to the large number of subunits composing extracellularly activated ligand-gated ion channels, such as nicotinic, ATP, GABA and glutamate ionotropic receptors. Composed of more than 500 human curated entries, the XML native database has been relocated in 2004 to the EBI. Its facilities have been enhanced with a new search system, customized multiple sequence alignments and manipulation of protein structures (http://www.ebi.ac.uk/compneur-srv/LGICdb/). Despite the vast improvement of general sequence resources, the LGICdb still provide sequences unavailable elsewhere.

INTRODUCTION

Ligand-gated ion channels are transmembrane proteins that can exist under different conformations, at least one forming a pore through the membrane connecting the two neighbour compartments. The equilibrium between the various conformations is affected by the binding of ligands on the channels. Phenomenologically, the ligands 'open' or 'close' the channel (1,2). There are three different superfamilies of extracellularly activated ligand-gated ion channel subunits.

The receptors of the 'cys-loop' superfamily (named after a conserved 13 residue loop closed by a disulfide bridge) are made up of five homologous subunits (3). Each subunit contains an extracellular N-terminal domain, followed by four transmembrane segments. The loop located between TM3 and TM4 composes the intracellular domain, of variable length. The subunits of the cys-loop superfamily are distributed in two clear monophyletic groups, one containing the subunits forming anionic channels (GABAA and GABAC, glycine, GLUCl, histamine and 5HTmod1 receptors) and one containing the subunits forming cationic channels (5-HT3 and nicotinic receptors) (4,5). Although site-directed mutagenesis in the channel part succeeded to invert selectivity (6,7), few examples transgressing this frontier have been discovered so far in nature (some subunits from Lymnea stagnalis may be exceptions).

The ATP-gated channels (ATP2x receptors) are made of three homologous subunits (8,9). Each subunit displays two transmembrane segments separated by an extracellular domain.

Finally, the glutamate-activated cationic channels are made of four homologous subunits (10,11). Each subunit contains an extracellular N-terminal domain similar to the bacterial leucine, isoleucine, valine binding protein (LIVBP), followed by half of the agonist binding core, two transmembrane domains separated by a 'P-loop', the second half of the agonist-binding core and a third transmembrane segment. The agonist binding core is similar to the bacterial lysine, arginine, ornithine binding protein (LAOBP). The cytoplasmic tail has a variable length.

Many of the subunits from the three superfamilies possess multiple isoforms, generated by alternative splicing or editing. The Ligand-Gated Ion Channel database (LGICdb) was created in the mid-1990s, as a repository offering a unique entry per gene (12). All the data are manually curated, in order to reduce redundancy and correct the errors coming from sequencing or introduced by automated methods of data mining (such as gene prediction).

CONSTRUCTION AND CONTENT

The LGICdb evolved significantly since the latest report in the literature. The resource created at the Pasteur Institute of Paris is now hosted by the European Bioinformatics Institute. Seminal was the transformation of the native format from a dedicated markup to an XML format. This move permitted a better syntax checking, the design of a validating editor, easier generation of export formats and the further treatment of the data by a native XML database engine. The number of entries increased much, thanks to the numerous genome projects. The systematic use of Ensembl (13) and UniProt (14) to

*To whom correspondence should be addressed. Tel: +44 1223 494521; Fax: +44 1223 494468; Email: lenov@ebi.ac.uk

retrieve possible entries permitted a more comprehensive population of the database.

The detailed procedure we used to build LGICdb entries is described elsewhere (15). Briefly, for each gene, the various transcripts and proteins are identified based on the published experimental sequences, but also on the predicted gene structure. Predicted isoforms are taken into account only if they are backed by experimental reports. If several variants exist for the 'same' sequence, all the possibilities are taken into account and a decision is taken based on the frequency of description of each variant, the comparisons with close orthologous sequences, etc. The resulting sequence is sometimes a chimaera built from several primary data. In the infrequent case where no consensus can be achieved, the variants are all presented with adequate annotation. When a predicted gene structure is incomplete, a tentative reconstruction is proposed, based on the genomic sequence and homologous subunits.

The release 57 (September 9, 2005) of the LGICdb contained 516 entries, totalizing 7 million nucleotides, 400 000 residues and 30 3D structures.

Each entry of the database is composed of one XML file. A Perl script using BioPerl (16), XML::Simple (http://search.cpan.org/dist/XML-Simple/) and XML::Writer (http://search.cpan.org/~josephw/XML-Writer-0.600/) generates one HTML page per entry, and files in the FASTA, GenBank and EMBL formats for all sequences (Figure 1). In addition, browsing lists of entries, by entry accession and by organisms, are also generated.

User can use the central FASTA (17) search of the EBI to retrieve entries based on sequence similarity. String searching of the whole LGICdb content is implemented by using the API provided by the Apache Xindice native XML database (http://xml.apache.org/xindice/). This solution has proved to be adequate in terms of speed for the amount of data we currently have in the LGICdb. Custom multiple sequence alignments can now be generated on the result of a string search, using ClustalW (18). Other multiple sequence alignments methods should be implemented in the future. The atomic coordinates can be manipulated with the Jmol applet (http://www.jmol.org/). Users can currently download the whole database (~13 MB). Selective download following string and sequence similarity search are under development.

While the core of the database is available through an Apache HTTP (http://httpd.apache.org/) server, with HTML pages generated with PHP, the string search and the multiple alignments are provided by an Apache Tomcat server (http://jakarta.apache.org/tomcat/).

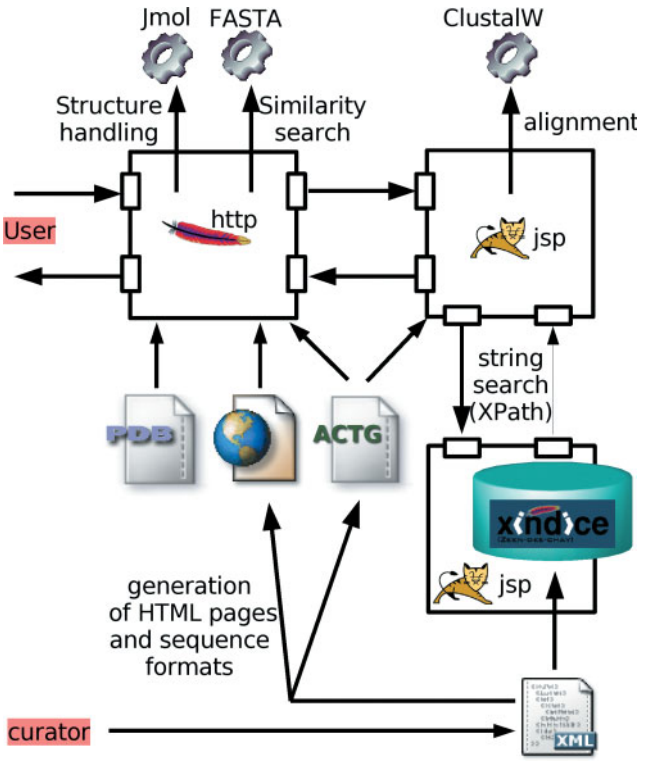

Figure 1. Schema describing the relationships between the various components of the LGICdb. Square boxes represent servers. http: Apache Hypertext Transfer Protocol server; jsp: Jakarta Tomcat Java Server Page. Gears represent external applications.

DISCUSSION

One could argue about the utility of manually curated databases of thematic focus, now that the community can benefit from large efforts such as Ensembl and UniProt. However, several problems are directly triggered by the large-scale aspect of those efforts.

The first issue, that triggered the creation of the LGICdb in the first place, is the redundancy. Although efforts have been undertaken to reduce this redundancy to a minimum, and to gather overlapping information together, there are still four entries for the human gene *CHRNA7* in UniProt. The situation is worse in non-curated resources. For instance, there are 10 proteins corresponding to *CHRNA7* in GenBank. While this is an unavoidable problem for a general resource, it can be easily solved when the resource is of limited focus. There is only one entry in the LGICdb for the human nicotinic subunit alpha7. In addition, while the LGICdb reports the various alternative splicing and editing, sequencing errors are corrected based on diverse criteria, such as the comparison with orthologous sequences.

Another problem is generated by the automatic annotation, such as the recognition of genes. For instance, the human GABA receptor rho3 subunit is splitted in two parts in Ensembl. The C-terminal exon is reported in UniProt, and therefore has been annotated as 'known transcript' in Ensembl, while part of the N-terminal portion has been predicted by Ensembl, and annotated as 'novel transcript'. As a consequence, the longest sequence for human rho3 is currently the one in the LGICdb, built by fusion.

The LGICdb actually belongs to a constellation of expert-maintained topical data resources. While their size is limited (by comparison with general purpose public resources), they serve data of high accuracy. In the field of transmembrane proteins, one could quote the GPCRDB (http://www.gpcr.org/7tm/) on G-protein coupled receptors (19), the VKCDB (http://vkcdb.biology.ualberta.ca/) on voltage-gated ion channels (20), the TCDB (http://www.tcdb.org/) on transporters and the protein kinase resource (http://www.kinasenet.org/) (21).

PERSPECTIVES

While a long-recognized resource in the field of neurotransmitter receptor research, the LGICdb progressively attracts attention on a larger audience, as witnessed by the Science NetWatch on August 26, 2005. It becomes all the more important to complete the database in order to improve its comprehensivness. However, if the availability of complete genomes could have suggested a possible completion of the work, the results of the FANTOM consortium (22), describing an unforeseen number of transcripts, make perhaps such a prospect unrealistic for a project without dedicated resources.

All data contained in the LGICdb may be copied and redistributed freely, without any restriction. If one uses some of these data in a scientific publication, authors would welcome a citation of the resource in the list of references.

ACKNOWLEDGEMENTS

The development of the Ligand-Gated Ion Channel was initiated with the support of professor Jean-Pierre Changeux. The complete list of contributors is available at http://www.ebi.ac.uk/compneur-srv/LGICdb/. Funding to pay the Open Access publication charges for this article was provided by EMBL.

Conflict of interest statement. None declared.

REFERENCES

1. Changeux,J.-P. and Edelstein,S.J. (1998) Allosteric receptors after 30 years. *Neuron*, **21**, 959–980.
2. Colquhoun,D. and Sivilotti,L.G. (2004) Function and structure in glycine receptors and some of their relatives. *Trends Neurosci*., **27**, 337–344.
3. Galzi,J.L. and Changeux,J.-P. (1994) Neurotransmitter-gated ion channels as unconventional allosteric proteins. *Curr. Opin. Struct. Biol*., **4**, 554–565.
4. Le Novère,N. and Changeux,J.-P. (1995) Molecular evolution of the nicotinic acetylcholine receptor: an example of multigene family in excitable cells. *J. Mol. Evol*., **40**, 155–172.
5. Ortells,M.O. and Lunt,G.G. (1995) Evolutionary history of the ligand-gated ion-channel superfamily of receptors. *Trends Neurosci*., **18**, 121–127.
6. Galzi,J.L., Devillers-Thiéry,A., Hussy,N., Bertrand,S., Changeux,J.P. and Bertrand,D. (1992) Mutations in the channel domain of a neuronal nicotinic receptor convert ion selectivity from cationic to anionic. *Nature*, **359**, 500–505.
7. Keramidas,A., Moorhouse,A.J., French,C.R., Schofield,P.R. and Barry,P.H. (2000) M2 pore mutations convert the glycine receptor channel from being anion- to cation-selective. *Biophys. J*., **79**, 247–259.
8. Nicke,A., Bäumert,H.G., 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.
9. North,R.A. (2002) Molecular physiology of p2x receptors. *Physiol. Rev*., **82**, 1013–1067.
10. Hollman,M. (1999) Ionotropic glutamate receptors in the CNS.. *Handbook Exp. Pharmacol*., **141**, 1–98.
11. Mayer,M.L. and Armstrong,N. (2004) Structure and function of glutamate receptor ion channels. *Annu. Rev. Physiol*., **66**, 161–181.
12. Le Novère,N. and Changeux,J.P. (1999) The ligand gated ion channel database. *Nucleic Acids Res*., **27**, 340–342.
13. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al*. (2005) Ensembl 2005. *Nucleic Acids Res*., **34**, D447–D453.
14. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2005) The universal protein resource (UniProt). *Nucleic Acids Res*., **33**, D154–D159.
15. Le Novère,N. and Changeux,J.-P. (2001) The ligand gated ion channel database (LGICdb), an example of a sequence database in neuroscience. *Philos. Trans. R. Soc. Lond. B. Biol. Sci*., **356**, 1121–1130.
16. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al*. (2002) The bioperl toolkit: Perl modules for the life sciences. *Genome Res*., **12**, 1611–1618.
17. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
18. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.
19. Horn,F., Weare,J., Beukers,M.W., Hörsch,S., Bairoch,A., Chen,W., Edvardsen,O., Campagne,F. and Vriend,G. (1998) GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res*., **26**, 275–279.
20. Li,B. and Gallin,W.J. (2004) VKCDB: Voltage-gated potassium channel database. *BMC Bioinformatics*, **5**, 3.
21. Smith,C.M., Shindyalov,I.N., Veretnik,S., Gribskov,M., Taylor,S.S., Ten Eyck,L.F. and Bourne,P.E. (1997) The protein kinase resource. *Trends Biochem. Sci*., **22**, 444–446.
22. Carninci,P., Kasukawa,T., Katayama,S., Gough,J., Frith,M.C., Maeda,N., Oyama,R., Ravasi,T., Lenhard,B., Wells,C. *et al*. (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.