# LIGAND database for enzymes, compounds and reactions

**Susumu Goto*, Takaaki Nishioka[1] and Minoru Kanehisa**

Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan and [1]Graduate School of Agricultural Sciences, Kyoto University, Sakyo-ku, Kyoto 606-8502, Japan

Received September 8, 1998; Revised October 1, 1998; Accepted October 14, 1998

## ABSTRACT

**LIGAND is a composite database consisting of three sections and containing the information of chemical substances, chemical reactions and enzymes that catalyze reactions. The COMPOUND section is a collection of metabolic compounds, as well as macromolecules, chemical elements and other chemical substances in a living cell. The ENZYME section is a collection of all known enzymatic reactions, together with the information of enzyme molecules, classified according to the EC (Enzyme Commission) numbers. The REACTION section is a new addition to the database containing metabolic reactions that appear in the pathway diagrams of the KEGG/PATHWAY database and/or in the ENZYME section. The LIGAND database can be accessed through the WWW (http://www.genome.ad.jp/dbget/ligand.html ) or may be downloaded by anonymous FTP (ftp://kegg.genome.ad.jp/molecules/ligand/ ).**

## INTRODUCTION

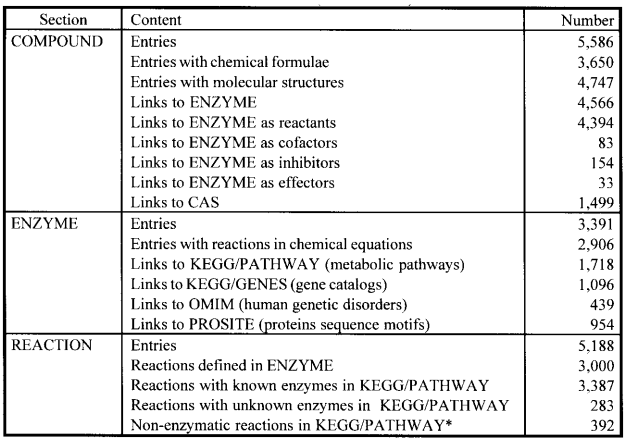
Life is a manifestation of both genetic and chemical information. While the genetic information is accumulated in the existing molecular sequence databases, the chemical information has not been well organized for the purposes of understanding cellular functions. LIGAND (1) is a chemical database to fill in the gap, especially for the metabolism in living cells. LIGAND is tightly integrated with the KEGG/PATHWAY and KEGG/GENES databases (2) and cross-referenced to many other molecular biology databases.

The original LIGAND database was designed for the sequence analysis of enzymes (3), and it was later expanded as the ENZYME section of LIGAND. The information in the ENZYME section is a mixture of enzyme molecules and enzymatic reactions. Since the initiation of the KEGG project (4), we started computerizing the information on chemical substances in the COMPOUND section of LIGAND in order to make a catalog of chemical building blocks of living cells (1). Recently we added a new section, REACTION, that is intended to represent the actual reaction pathways for computational purposes. We report here the current status of the LIGAND database and the addition of the REACTION section.

## ORGANIZATION OF THE DATABASE

LIGAND is constructed as a flat-file database consisting of COMPOUND, ENZYME and REACTION sections. The REACTION section is a new addition containing the information on chemical reactions that appear in the KEGG/PATHWAY database (2) and/or in the ENZYME section. At the moment, LIGAND is defined as a composite database of the ENZYME and COMPOUND sections, i.e., ligand = enzyme + compound, under the DBGET/LinkDB system (5) of the Japanese GenomeNet (6). The third REACTION section is available only through anonymous FTP. The number of entries in the current release is summarized in Table 1.

**Table 1.** The number of entries in release 18.0 (October 1998) of the LIGAND database



| Section | Content | Number |
|---|---|---|
| COMPOUND | Entries | 5,586 |
| | Entries with chemical formulae | 3,650 |
| | Entries with molecular structures | 4,747 |
| | Links to ENZYME | 4,566 |
| | Links to ENZYME as reactants | 4,394 |
| | Links to ENZYME as cofactors | 83 |
| | Links to ENZYME as inhibitors | 154 |
| | Links to ENZYME as effectors | 33 |
| | Links to CAS | 1,499 |
| ENZYME | Entries | 3,391 |
| | Entries with reactions in chemical equations | 2,906 |
| | Links to KEGG/PATHWAY (metabolic pathways) | 1,718 |
| | Links to KEGG/GENES (gene catalogs) | 1,096 |
| | Links to OMIM (human genetic disorders) | 439 |
| | Links to PROSITE (proteins sequence motifs) | 954 |
| REACTION | Entries | 5,188 |
| | Reactions defined in ENZYME | 3,000 |
| | Reactions with known enzymes in KEGG/PATHWAY | 3,387 |
| | Reactions with unknown enzymes in KEGG/PATHWAY | 283 |
| | Non-enzymatic reactions in KEGG/PATHWAY* | 392 |

*Non-enzymatic reactions include reactions that are not known whether enzymes are involved in catalysis.

Similar to the data formats of PIR (7) and GenBank (8) flat files, a fixed number of columns are assigned to specify each field of an entry. Tables 2 and 3 show the field identifiers and the corresponding data contents for the COMPOUND and ENZYME sections, respectively. The COMPOUND section contains the information on chemical compounds with links to the KEGG/PATHWAY database and the ENZYME section. An entry in this section is associated with the 2D molecular structure in a GIF image file and in an MDL-MOL file with stereochemistry, both

*To whom correspondence should be addressed. Tel: +81 774 38 3266; Fax: +81 774 38 3269; Email: goto@kuicr.kyoto-u.ac.jp

**Table 2.** The data content of the COMPOUND section

| Field | Data content | DBGET/LinkDB |
|---|---|---|
| ENTRY | Compound accession number | All related databases |
| NAME | Recommended and alternative names of the compound | |
| FORMULA | Chemical formula of the compound | |
| STRUCTURE | GIF image of the molecular structure (MDL-MOL format file is also available) | |
| PATHWAY | KEGG metabolic pathway diagrams in which the compound appears | PATHWAY |
| ENZYME | Enzymatic reactions (EC numbers) in which the compound appears | ENZYME |
| DBLINKS | CAS registry number | |

**Table 3.** The data content of the ENZYME section

| Field | Data content | DBGET/LinkDB |
|---|---|---|
| ENTRY | Entry identifier (EC number) | All related databases |
| NAME | Recommended and alternative names of the enzyme | |
| CLASS | Description of the EC numbering classification | |
| SYSNAME | Systematic name of the enzyme | |
| REACTION | Chemical reaction in the form of an equation or an English sentense | |
| SUBSTRATE<br>PRODUCT | Chemical compounds that appear, respectively, on the left and right sides of the reaction equation | COMPOUND |
| INHIBITOR<br>COFACTOR<br>EFFECTOR | Chemical compounds that act, respectively, as inhibitors, cofactors, and effectors of the reaction | COMPOUND |
| COMMENT | Text information commenting on the enzyme | |
| PATHWAY | KEGG metabolic pathway diagrams in which the enzyme appears | PATHWAY |
| GENES | Gene names that encode the enzyme for the organisms in the KEGG/GENES database | GENES |
| DISEASE | Human genetic disorders caused by lack or mutation of the enzyme in OMIM (12) | OMIM |
| MOTIF | Protein sequence motifs in PROSITE (13) | PROSITE |
| STRUCTURES | Protein 3D structures in the Protein Data Bank (14) | PDB |
| DBLINKS | Bairoch's ENZYME database (15)<br>WIT (16)<br>UM-BBD Biocatalysis/Biodegradation Database (18)<br>BRENDA (10)<br>SCOP structural classification of proteins (17) | |

of which can be retrieved in the WWW. The entries in the ENZYME section contain three types of information. First, the classification and nomenclature of enzymatic reactions are given according to the Enzyme Nomenclature by the IUBMB (International Union of Biochemistry and Molecular Biology) Nomenclature Committee (9). Second, the information of chemical substances that are relevant to the reaction is organized according to the Enzyme Handbook (10), textbooks and journal articles as well as the Enzyme Nomenclature. Third, the information of enzyme molecules is given as links to other databases. The link information shown in the third columns of Tables 2 and 3 is highly integrated in the DBGET/LinkDB system where, for example, links computed from multiple and/or reverse links can be retrieved from the entry identifier in the ENTRY field.

## REACTION section

The REACTION section is newly added in the latest release (Release 17.0, July 1998) of the LIGAND database. A reaction appearing as a chemical equation format in the ENZYME section and/or a reaction in the KEGG/PATHWAY database (2) is given a unique accession number and stored in this section. Because the KEGG/PATHWAY database consists of graphical diagrams representing known metabolic pathways and includes many reactions that are not described in the ENZYME section, it was not readily possible to extract the information of substrate-product relations in successive reactions. Table 4 shows an example of how the information is organized in the REACTION section.

**Table 4.** An example of the REACTION entry

| | |
|---|---|
| Reaction ID | R01502 |
| Description | C00197 + C00002 <=> C00236 + C00008<br>(3-Phospho-D-glycerate + ATP <=> 3-Phospho-D-glyceroyl phosphate + ADP ) |
| Link to ENZYME | 2.7.2.3 (Phosphoglycerate kinase) |
| Link to KEGG/PATHWAY | map00010 (Glycolysis/Gluconeogenesis)<br>map00710 (Carbon fixation) |

While the reactions extracted from the ENZYME section always have fixed EC numbers, the reactions extracted from the KEGG pathway diagrams do not necessarily have EC numbers either because the EC numbers are not yet assigned or because the reaction is non-enzymatic. Thus, the reaction data in the REACTION section are classified into three groups according to the EC number and its extension. The first group is for the reactions that are catalyzed by specific enzymes described in the ENZYME section. A reaction in this group has a valid EC number with four numerals separated by periods. The second group is for the reactions that are catalyzed by enzymes not described in the ENZYME section. An EC number is not yet assigned to the enzyme by the Enzyme Commission and KEGG tentatively assigns an EC number with a minus sign in place of the fourth digit, such as '1.1.1.–'. The last group is for non-enzymatic reactions that are represented by the EC number '0.0.0.0'. The numbers of reactions classified in the second and the third groups are shown in the last two rows of Table 1.

The reaction data can be used for computing possible reaction pathways, which is especially useful in the pathway reconstruction process from the completely sequenced genomes (1). For this purpose, each reaction is converted into a set of substrate–product binary relations (11). For example, the reaction in Table 4 would produce four binary relations. In practice, however, it is desirable to exclude coenzymes, donors and acceptors of phosphates or electrons in order to obtain only the main routes of reaction pathways and to avoid the explosion of computation time and space. Thus, we prepare another set of reaction data, tentatively named reaction.main, to represent major compounds along the KEGG pathways. In the case of the reaction in Table 4, ATP and ADP are omitted in the reaction.main data.

Figure 1 shows the result of actually using the reaction.main data and computing possible reaction paths from 3-phospho-D-glycerate (C00236) to 3-phospho-D-glyceroyl phosphate (C00197) of Table 4. Figure 1a shows the shortest path, which is the reaction catalyzed by phosphoglycerate kinase (EC 2.7.2.3). Figure 1b shows an alternative path, which uses two enzymes. Figure 1c is a graphical view summarizing all alternative reaction pathways. This path computation tool is available in the KEGG system (http://www.genome.ad.jp/kegg-bin/mk_pathcomp_html ). We also note that this type of computation is also applicable to analyzing possible gene regulatory networks from gene–gene binary relations obtained from expression profile experiments.

## ACCESS TO THE DATABASE

The LIGAND database is accessible through the WWW at: http://www.genome.ad.jp/dbget/ligand.html . The user can then invoke the DBGET/LinkDB system to retrieve the COMPOUND and ENZYME sections or enter the KEGG system to view and

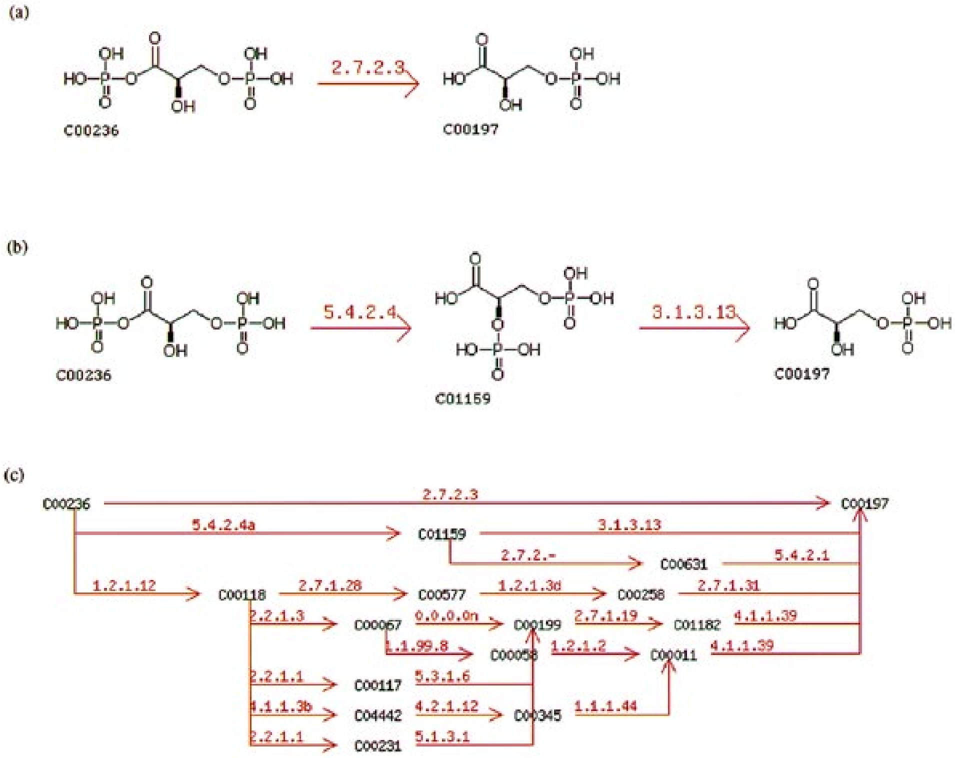


**Figure 1.** The result of pathway computation between two compounds of Table 4 by using the binary relation data: (**a**) the shortest path; (**b**) an alternative path; and (**c**) a graphical view of all possible reaction paths.

search different representations of the LIGAND database including the hierarchical classifications of enzymes and the periodic table for chemical elements.

The LIGAND database can be downloaded via anonymous FTP at: ftp://kegg.genome.ad.jp/molecules/ligand/ . This directory contains all sections, COMPOUND, ENZYME and REACTION, including the GIF image files and MDL-MOL files for compound structures. The same data set is mirrored at the NCBI repository: ftp://ncbi.nlm.nih.gov/repository/LIGAND/

The basic concept of the LIGAND database has been published elsewhere (1). The present article reflects the most up-to-date version of the database and should be cited accordingly.

## ACKNOWLEDGEMENTS

We thank Hiroko Ishida and Sanae Asanuma for extraction of reaction data from the KEGG pathway diagrams. This work was supported by the Grant-in-Aid for Scientific Research on the Priority Area ‘Genome Science’ from the Ministry of Education, Science, Sports and Culture of Japan. The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

## REFERENCES

1 Goto,S., Nishioka,T. and Kanehisa,M. (1998) *Bioinformatics*, **14**, 591–599.
2 Ogata,H., Goto,S., Fujibuchi,W., Sato,K., Bono,H. and Kanehisa,M. (1999) *Nucleic Acids Res*., **27**, 29–34.
3 Suyama,M., Ogiwara,A., Nishioka,T. and Oda,J. (1993) *Comput. Applic. Biosci*., **9**, 9–15.
4 Kanehisa,M. (1997) *Trends Genet*., **13**, 375–376.
5 Fujibuchi,W., Goto,S., Migimatsu,H., Uchiyama,I., Ogiwara,A., Akiyama,Y. and Kanehisa,M. (1998) *Pacific Symp. Biocomput*., 683–694.
6 Kanehisa,M. (1977) *Trends Biochem. Sci*., **22**, 442–444.
7 Barker,W.C., Garavelli J.S., Haft,D.H., Hunt,L.T., Marzec,C.R., Orcutt,B.C., Srinivasarao,G.Y., Yeh,L.S.L., Ledley,R.S., Mewes,H.W., Pfeiffer,F. and Tsugita,A. (1998) *Nucleic Acids Res*., **26**, 27–32.
8 Benson,D.A., Boguski,M.S., Lipman,D.J., Ostell,J. and Ouellette,B.F. (1998) *Nucleic Acids Res*., **26**, 1–7.
9 IUBMB (1992) *Enzyme Nomenclature: Recommendations (1992) of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology.* Academic Press, New York.
10 Schomburg,D. (ed.) (1990) *Enzyme Handbook* 1–16. Springer-Verlag, Heidelberg.
11 Goto,S., Bono,H., Ogata,H., Fujibuchi,W., Nishioka,T., Sato,K. and Kanehisa,M. (1997) *Pacific Symp. Biocomput*., 175–186.
12 Pearson,P., Francomano,C., Foster,P., Bocchini,C., Li,P. and McKusick,V. (1994) *Nucleic Acids Res*., **22**, 3470–3473.
13 Bairoch, A, Bucher,P. and Hofmann,K. (1997) *Nucleic Acids Res*., **25**, 217–221.
14 Bernstein,F.C., Koetzle,T.F., Williams,G.J., Meyer,E.F., Brice,M.B., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol*., **112**, 535–542.
15 Bairoch,A. (1996) *Nucleic Acids Res*., **24**, 221–222.
16 Overbeek,R., Larsen,N., Maltsev,N., Pusch,G.D. and Selkov,E. In Letovsky,S. (ed.) *Molecular Biology Databases*, Kluwer (in press).
17 Hubbard,T.J.P., Murzin,A.G., Brenner,S.E. and Chothia,C. (1997) *Nucleic Acids Res*., **25**, 236–239.
18 Ellis,L.B.M. and Wackett,L.P. (1995) *Soc. Industrial Microbiol. News*, **45**, 167–173.