DBGET/LinkDB: A Way of Solution to Integrate Diverged Biological Databases

Wataru Fujibuchi

To extract and utilize information from various forms of biological databases together with your own database are labor-intensive for resarchers. We present here an integrated database retrieval system DBGET/LinkDB, which is the backbone of the Japnese GenomeNet service, to manage those tasks. DBGET is used to search and extract entries from a wide range of molecular biology databases, while LinkDB is used to compute links between entries in different dtabases. It is designed to be a network distributed database system with an open architecture, which is suitable for incorporating local databases or establishing a server environment. The WWW version of DBGET/LinkDB is integrated with other search tools, such as BLAST, FASTA and MOTIF to conduct further retrievals instantly. Moreover, LinkDB can search biological links to get more abstract links of biological objects, and it is the first step toward computerization of logical reasoning process of biological data.

Keywords : Biological Database/ Database Integration/ GenomeNet Service/ DBGET/LinkDB/ Link Computation/ Biological Link

In order to effectively make use of the information in the network of molecular biology databases, it is essential to develop an integrated database retrieval system. The key interst in integrating different source of data is the types of data coupling. Loosely coupled approach has been successful where different databases are linked at the level of entries, rather than the fields that form an entry.

The DBGET/LinkDB has the following characteristics:

• Distributed database: DBGET is organized and accessible through a network configuration systm. Database can exist in different servers, but from the user's point of view they all exist in a single DBGET server.

• Simple architecture: DBGET/LinkDB emphasizes the manipulation of flat file databases at the level of entries. By keeping the search capabilities of individual fields at a minimal level, the updating of DBGET databases requires minimal indexing, which is suited for rapid daily updates of a number of databases.

• Open architecture: The user can set up his/her own DBGET world by integrating the local databases with the databases on the DBGET server. It is also possible to down load public databses from GenomeNet to the user's closed environment. Moreover, LinkDB contains links to other databases outside of DBGET.

RESEARCH FACILITY OF NUCLEIC ACIDS

Scope of research

The following is the current major activities of this facility.
With emphasis on regulatory mechanisms of gene expression in higher organisms, the research activity has been focused on analysis of signal structures at the regulatory regions of transcriptional initiation and of molecular mechanisms involved in post-transcriptional modification by the use of eukaryotic systems appropriate for analysis. As of December 1994, studies are concentrated on the molecular mechanism of RNA editing in mitochondria of kinetoplastids.

Assoc Prof
SUGISAKI, Hiroyuki
(D Sc)

Instr
FUJIBUCHI, Wataru

Techn
YASUDA, Keiko

• Different interfaces: The simplest way to access DBGET is to use the Web interface, but by installing the special client program NetDBget, the DBGET commands can be entered at the UNIX command level.

DBGET does not convert the original files but use them as they are. In order to accomplish rapid access and search of entries, a small number of auxiliary files are created by the indexing program *seqnew* during the update procedure as shown in Table 1.

Table 1: The list of auxiliary files created by *seqnew*.

filename	*file type*	*contents*
db.pag	dbm	hash by entry and accession
db.acc	flat	primary and secondary accession
db.tit	flat	title or definition field of entry
db.tit.pag	dbm	hash by db.tit
db.ref	flat	references in entry
db.auf	flat	authors in entry
db.lnk+.pag	dbm	hash by entry for original links
db.lnk-.pag	dbm	hash by entry for reverse links

LinkDB contains the original links provided by each database and the indirect and reverse links that are computed. They are defined in the *link table*, and the route of computing indirect links is predefined in the *route table*.

At present, 17 databases and links to Medline database are supported in DBGET/LinkDB. As shown in Table 2, all the major databases are daily or weekly updated. PATHWAY, LIGAND and GENES are the products of the KEGG project, where PATHWAY is the database of metabolic pathways and regulatory pathways, LIGNAD is a composite database of ENZYME and COMPOUND, and GENES is a collection of gene catalogs for a number of oraganisms.

Table 2: The DBGET database on GenomeNet service..

Group of Databases	*Database Names*
nucleic acid sequences	*GenBank, *EMBL
protein sequences	*SWISS-PROT, PIR, PRF, *PDBSTR
3D structures	*PDB
sequence motifs	PROSITE, EPD, TRANSFAC
enzyme reactions	*LIGAND
metabolic pathways	*PATHWAY
amino acid mutations	PMD
amino acid indices	AAindex
genetic diseases	*OMIM
literature	LITDB
genes and genomes	*GENES

Those marked by asterisks are daily or weekly updated.

Once an entry is retrieved in the WWW mode of DBGET, all links from this entry can be obtained by clicking on the entry name, which causes the search against LinkDB. Figure 1 shows one of the Web interfaces at GenomeNet to access DBGET/LinkDB, wihch illustrates the supported databases and the original links among them.

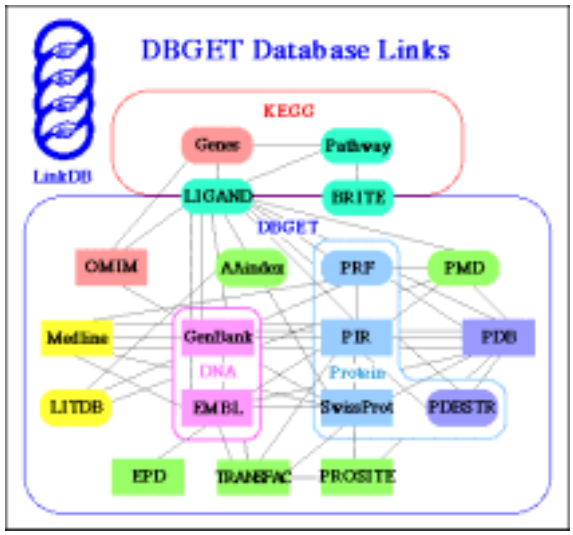

Figure 1. A framework diagram of LinkDB, a database of cross-links between molecular biology databases.

LinkDB started as a collecction of factual links only. Recently similarity links were added but they are not yet integrated with factual links for computing indirect and reverse links. Biological links are being identified by the KEGG project. They are stored as cross-references in the GENES databases that can be treated in a similar way as factual links.

Acknowledgment

This work was supported in part by the Grant-in-Aid for Scientific Research on the Priority Area 'Genome Informatics' 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. Fujibuchi W, Goto S, Migimatsu H, Uhicyama I, Ogiwara A, Akiyama Y and Kanehisa M, *Proc. Pacific Symposium on Biocomputing '98*, 683-694 (1997).
2. Kanehisa M, Fickett JW and Goad WB, *Nucl. Acids Res.*, **12**, 149-158 (1984).
3. Akiyama Y, Goto S, Uchiyama I and Kanehisa M, *MIMBD '95: Second Meeting on the Interconnection of Molecular Biology Databases* (1995).
4. Goto S, Akiyama Y and Kanehisa M, *MIMBD '95: Second Meeting on the Interconnection of Molecular Biology Databases* (1995).
5. Kanehisa M, *Trends Biochem. Sci.*, **22**, 442-444 (1997).