# KEGG and DBGET/LinkDB: Integration of Biological Relationships in Divergent Molecular Biology Data

**Wataru Fujibuchi  Kazushige Sato**[†]
wataru@kuicr.kyoto-u.ac.jp  kazs@scl.kyoto-u.ac.jp
**Hiroyuki Ogata  Susumu Goto  Minoru Kanehisa**
ogata@kuicr.kyoto-u.ac.jp  goto@scl.kyoto-u.ac.jp  kanehisa@kuicr.kyoto-u.ac.jp
Institute for Chemical Research, Kyoto University
Uji, Kyoto 611-0011, Japan

### Abstract

A simple formulation to integrate various biological data is presented based on the concept of links, which are classified into three types: factual, similarity, and biological. Factual links are cross-reference information of entries among molecular biology databases. Similarity links are neighbor information of sequence entries computed by sequence similarity search programs. Biological links are the novel and powerful relations that are being organized in KEGG, including molecular interactions on the metabolic and regulatory pathways, physical closeness of genes on the genome, and the rest of binary relations in which divergent biological phenomena can be accommodated. DBGET/LinkDB was originally designed to handle factual and similarity links in the existing molecular biology databases, but it is now being extended to include biological links as well. The process of logical reasoning, for example, for functional assignment of newly sequenced genes can often be decomposed into a sequence of combining different types of links; thus, we expect it can be automated by the extension of the DBGET/LinkDB system and the database efforts of KEGG.

## Introduction

The power of making links is evident not only for navigation in the World Wide Web, but also for integration of knowledge in molecular biology that is based on various types of biological data. A tightly-coupled database, such as a relational database containing all types of data, is reliable in its framework and retrieval ability. However, because of its rigidity, it cannot follow the rapid changes of database formats and contents. On the contrary, a loosely-coupled approach that combines any database at the level of its entry, has been successful in a number of practical systems, such as our DBGET/LinkDB(Fujibuchi *et al.* 1997) as well as Entrez(Schuler *et al.* 1996) and SRS(Etzold & Argos 1993).

The links provided by the original databases as cross-references are useful for retrieving related entries in other databases, especially in the Web-based systems. In addition, the sequence neighbors computed by sequence similarity search programs such as BLAST and FASTA are extremely useful for finding functionally related sequences. However, we consider the knowledge of biological relationships, such as the information of biological partners in molecular pathways and molecular assemblies, is under-represented in the existing molecular biology databases. Such information may occasionally exist in text as comments, but it is not readily utilizable for computation.

Under the project named KEGG, Kyoto Encyclopedia of Genes and Genomes(Kanehisa 1997a), we are computerizing biological relationships of molecular interactions and genetic interactions in living cells. Our aim is to integrate such biological knowledge together with the cross-references and similarity neighbors based on the common concept of links between two items, or what we call "binary relations". Here we report the current status of the DBGET/LinkDB system, the KEGG databases, and the KEGG computational tools as well as our plan for future extensions.

## Integration of Public Databases

### Data Representation

In DBGET a database is simply considered a collection of entries stored in a flat file or multiple flat files. Our definition of flat files includes text files and other multimedia files such as GIF files for the pathway diagrams. Because an entry can be uniquely identified by an entry name or an accession number in each database, any entry of a flat file database can be retrieved uniformly in DBGET by the combination of the "database name" and the entry "identifier". The cross-reference data among a number of molecular biology databases can be represented as:

**database1:identifier1 → database2:identifier2**

This binary relation is the basis of our LinkDB system and its extension:

**organism:gene1 → organism:gene2**

[†] Permanent Address: Nihon Silicon Graphics Cray K.K., Umeda, Kita-ku, Osaka 530-0001, Japan

is the basis of representing biological relationships in KEGG.

## Field Indexing

One of the advantages of DBGET system is to use the original database files as they are, requiring less disk space for storage and less computation time for daily updates. 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 table by entry and accession keys for the position offset and the size of each entry |
| db.acc | flat | primary and secondary accessions of each entry |
| db.tit | flat | title or definition field of each entry |
| db.tit.pag | dbm | hash table by entry and accession keys for db.tit |
| db.ref | flat | references in each entry |
| db.aut | flat | authors in each entry |
| db.lnk+.pag | dbm | hash table by entry keys for original links |
| db.lnk–.pag | dbm | hash table by entry keys for reverse links |

Some of them are hashed by the GNU database manager library. Extracting specific keys and field data from various databases requires lexicographical analysis routines dedicated for individual databases. We have developed a C++ class library for parsing a wide range of database formats by utilizing the advantages of C++ language, such as the inheritance of base class members.

## Network-Distributed Database

In DBGET all databases do not necessarily exist on one server but may reside on several servers in the network. Thus, DBGET can be used as a network-distributed database system as shown in Figure 1.

The network configuration of multiple databases is defined in the *dbtab* table file. The example shown in Table2 describes that genbank is retrieved from the server at dbget.genome.ad.jp, while embl is taken from the server of IP address 133.103.97.7 with port number 3266. The user can incorporate his/her own databases into DBGET as shown on the last line of the Table, which specifies that the database named mydb in the genbank format is taken from the user's directory '/usr/local/db'.

## Computation of Links

LinkDB is a repository of all cross-reference information among a number of databases. It contains original links extracted from each database, and also reverse links and indirect links that are computed from original links. Depending on the characteristics of individual databases, the routes of computing indirect links are defined in the *linktab* file. For instance, OMIM does not have original cross-references to external databases, though it has internal references. In contrast, SWISS-PROT has rich cross-references to other databases, including EMBL, PROSITE, PIR, PDB, Medline, and OMIM. Therefore, OMIM can go out to almost all other databases by defining in the *linktab* table the route of first using the reverse link to SWISS-PROT.

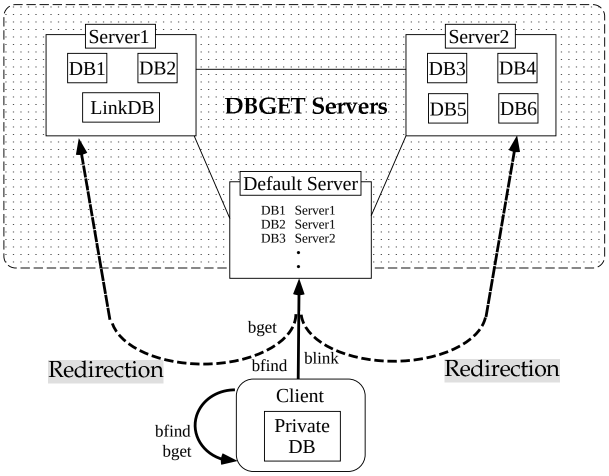


Figure 1: DBGET can be configured as a network-distributed database system.

Table 2: An example of network configuration defined in *dbtab*.

| database | dbtype | dbsite |
|---|---|---|
| genbank | genbank | @dbget.genome.ad.jp |
| embl | embl | @133.103.97.7:3266 |
| mydb | genbank | /usr/local/db/mydbdir |

Conceptually, there are three types of links that are to be utilized in LinkDB:

- **Factual Links:**
  links provided by the original databases, e.g., cross-reference of Medline ID and GenBank accession.
- **Similarity Links:**
  links produced by similarity search, e.g., the result of BLAST, FASTA and MOTIF programs.
- **Biological Links:**
  links by biological relationships, e.g., molecular or genetic interactions in the KEGG pathways.

LinkDB incorporates all three types of links to eventually perform integrated retrieval for biological reasoning, such as functional assignment of newly sequenced genes.

## DBGET/LinkDB on GenomeNet

Under the Japanese GenomeNet database service(Kanehisa 1997b) DBGET/LinkDB currently supports 18 databases and contains links to Medline. Table 3 shows a list of those databases maintained in DBGET/LinkDB, where all the major databases are daily or weekly updated. PATHWAY, LIGAND, GENES, and BRITE are the products of the KEGG project, in which biological knowledge of molecular interactions and pathways is being organized.

Table 3: The DBGET databases on GenomeNet.

| Content | 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/regulatory pathways | *PATHWAY |
| regulatory relations | BRITE |
| 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.

### Generic Naming

Figure 2 is a WWW interface to DBGET/LinkDB and KEGG that are tightly integrated with each other. As indicated in the Figure DBGET allows generic naming for a composite database, for example, "DNA" database for GenBank and EMBL, and "PROTEIN" database for PRF, PIR, SWISS-PROT, and PDBSTR.

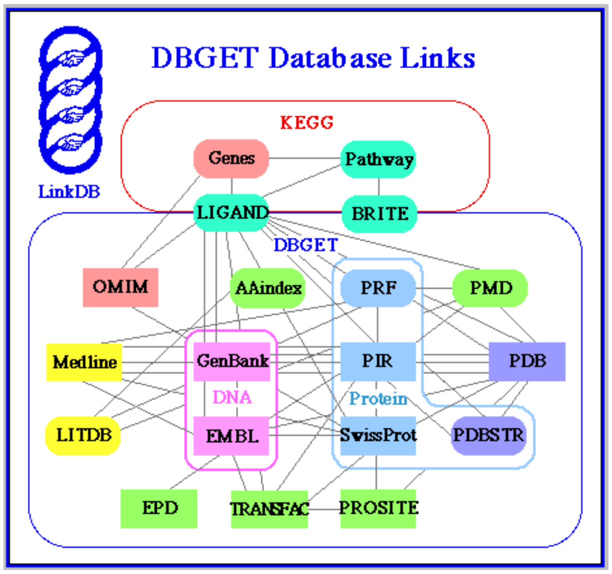


Figure 2: DBGET/LinkDB/KEGG database links diagram.

This generic naming is very useful in the daily updated databases, where the combination of the daily cumulative (e.g. genbank-upd) and the fixed release (e.g. genbank) is given the generic naming (e.g. genbank-today), which requires us to index only the cumulative portion every day in order to make the latest version of the database. It saves both the computation time and the disk space. Another useful feature of DBGET is aliasing for the databases, such as gb for genbank, and gbt for genbank-today.

### Integration with Other Tools

Because of the loose-coupling integration, DBGET/LinkDB system can easily be coupled with other Web-based search programs. In GenomeNet, for example, the result of similarity searches by BLAST and FASTA are directly linked to DBGET/LinkDB for retrieval of found entries and related entries, which help interpretation of biological meanings. Conversely, the entries retrieved by DBGET can be transferred to helper application programs, such as sequence similarity tools or 3D protein structure viewers.

## Integration of Biological Knowledge

### KEGG Databases

KEGG(Goto *et al.* 1996) is our attempt to describe, utilize, predict, and possibly design biological systems based on the genomic information. To accomplish this task, first, the current knowledge of biological functions in molecular and cellular biology is being computerized in terms of the information pathways that consist of interacting genes and molecules. This may be considered the process of making a functional catalog of living organisms. Second, gene catalogs being obtained by genome sequencing projects of different organisms are linked to individual components of the functional catalog. This may be considered the processing of mapping a gene catalog to the functional catalog.

KEGG thus provides the linkage between the catalog of molecular components and the network of molecular interactions in living cells and organisms. The latter is organized as the PATHWAY database, which is the primary product of the KEGG project, and the former is organized in the GENES database taken from the existing databases (genome databases, GenBank, and SWISS-PROT) as well as in the LIGAND database of chemical compounds(Goto, Nishioka, & Kanehisa 1998). The existing sequence databases are organized in such a way that the sequence is the core information with everything else as an attribute or an annotation to the sequence. The KEGG PATHWAY database is at a higher level of abstraction where the network of interacting genes or molecules is the core information. A gene or a molecule is the basic element that forms the network, just like a nucleotide or an amino acid is the basic element that forms the sequence. At the moment, most of the contents in the GENES database are automatically generated from a number of sources, but we plan to incorporate information on interacting partners that are manually identified in the KEGG project.

## KEGG Tools

The functional catalog is represented in KEGG by graphical pathway maps describing molecular interaction networks and by hierarchically organized texts describing classifications of molecules. The gene catalog is represented by graphical genome maps and by hierarchical texts of gene classifications. KEGG provides tools to manipulate these different types of information. The tools allow users to browse, navigate, search, compute, and even modify the information.

KEGG is available in two versions, WWW and local (CD-ROM) versions. For the WWW version of KEGG, we first wrote CGI scripts for browsing pathway map graphics and hierarchical texts. However, it was not possible to implement genome map graphics by the server program alone, for they need be utilized not only for just browsing but also for local manipulations.

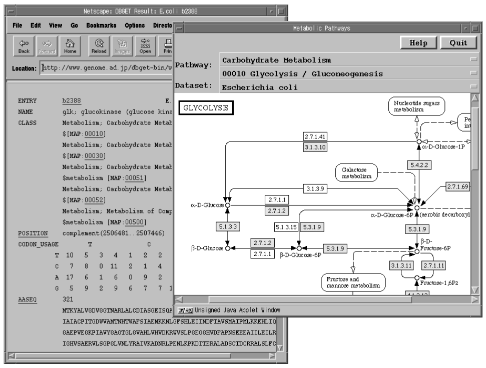
In the past, we developed genome map handling tools, such as Genomatica(Akiyama *et al.* 1994) and HyperGenome(Goto *et al.* 1994), under the client-server mechanism. From the user's point of view this mechanism requires downloading and installation of specialized client software, which often inhibits wide usage in the international genome research community. From the developer's point of view, the necessity of developing different client software for different platforms has also been a nightmare.

In order to solve these problems and to include more capabilities we have decided to use Java and developed a suite of Java applets for KEGG. Java has a potential of transforming the World Wide Web from a static information resource to be retrieved to a more dynamic resource to be computed. This is especially suited for KEGG which may be considered as a deductive database in the sense that additional information is logically deduced from the stored information. Thus, different types of computations are required in KEGG, and Java has enabled us to implement both computation and graphics handling capabilities to be performed locally on the user's machine.

However, server side CGI programs still have advantages for the WWW version of KEGG. Because CGI programs are executed on the server, less system resources, such as memory and CPU power, are required on the user side client machine. It means CGI version tools, although their functionality may be limited, can be used more conveniently by most users. In order to accommodate varying requirements, KEGG tools are written in both CGI scripts and Java programs.

## Java Map Browsers

KEGG contains two types of graphics data, pathway maps and genome maps, for which map browsers have been developed in Java. In the pathway map browser the user can easily switch to other organisms' maps or to other categories of maps by pop-up menus. Graphical objects on the map, such as a gene product or a compound, are linked to related information. These main features can also be implemented in CGI scripts and clickable map function of WWW, but Java allows more useful interactive features to be added.



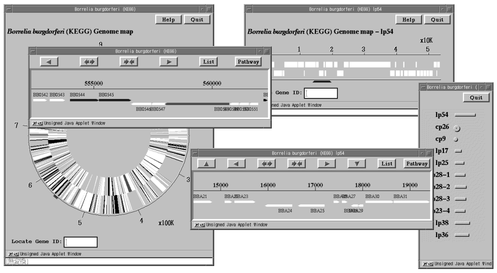
Location: http://www.genome.ad.jp/dbget-bin/w

```
ENTRY       b2388                        E.
NAME        glk; glucokinase (glucose kina
CLASS       Metabolism; Carbohydrate Metab
            $[MAP:00010]
            Metabolism; Carbohydrate Metab
            $[MAP:00030]
            Metabolism; Carbohydrate Metab
            $metabolism [MAP:00051]
            Metabolism; Carbohydrate Metab
            $[MAP:00052]
            Metabolism; Metabolism of Comp
            $metabolism [MAP:00500]
POSITION    complement(2506481..2507446)
CODON_USAGE        T            C
          T  10  5  3  4  1  2  2
          C   7  8  0 11  2  1  4
          A  17  6  1  6  0  9  2
          G   5  9  2  9  6  7  7  1
AASEQ       321
            MTKYALVGDVGGTNARLALCDIASGEISQA
            IAIACPITGDWVAMTNHTWAFSIAEMKKNLGFSHLEIINDFTAVSMAIPMLKKEHLIQ
            GAEPVEGKPIAVYGAGTGLGVAHLVHVDKRWVSLPGEGGHVDFAPNSEEEAIILEILR
            IGHVSAERVLSGPGLVNLYRAIVKADNRLPENLKPKDITERALADSCTDCRRALSLFC
```

Pathway: Carbohydrate Metabolism
00010 Glycolysis / Gluconeogenesis
Dataset: Escherichia coli

Figure 3: Pathway map browser.

For the genome map browser two versions were developed, one for completely sequenced genomes and the other for human chromosomes. The former consists of two viewer windows for circular genome maps and linear genome maps. For example, the *E. coli* chromosome is displayed in a circular genome map window, while the 16 *S. cerevisiae* chromosomes are displayed in a linear genome map window. Organisms that have both circular and linear plasmids, such as *B. burgdorferi*, are displayed in both circular and linear genome map windows.



Borrelia burgdorferi (KEGG)
Help
Quit
Borrelia burgdorferi (KEGG) Genome map
9
7
6
5
4
3
x100K
Locate Gene ID:
Unsigned Java Applet Window
Borrelia burgdorferi (KEGG)
List
Pathway
555000
560000
BB0542
BB0543
BB0544
BB0545
BB0546
BB0547
BB0551
Unsigned Java Applet Window
Borrelia burgdorferi (KEGG) lp54
Help
Quit
Borrelia burgdorferi (KEGG) Genome map – lp54
x10K
1
2
3
4
5
Gene ID:
Unsigned Java Applet Window
Borrelia burgdorferi (KEGG) lp54
List
Pathway
15000
16000
17000
18000
19000
BBA21
BBA23
BBA27
BBA30
BBA31
BBA24
BBA25
BBA29
Unsigned Java Applet Window
Borrelia burgdorferi (
Quit
lp54
cp26
cp9
lp17
lp25
lp28–1
lp28–2
lp28–3
lp23–4
lp38
lp36
Unsigned Java Applet Wind

Figure 4: Genome map browser (*B. burgdorferi*).

The genome map browser for human chromosomes is similar to the linear chromosome genome map viewer, but it has different functions to handle low resolution maps(Figure 5).

Each genome map viewer contains a zoom-up window and an overall view. The gene identifiers in the zoom-up window are clickable to retrieve the definition of the genes stored in the hierarchical gene catalog, and then additional information in the existing databases through the DBGET/LinkDB retrieval system. Conversely, each GENES database entry has a

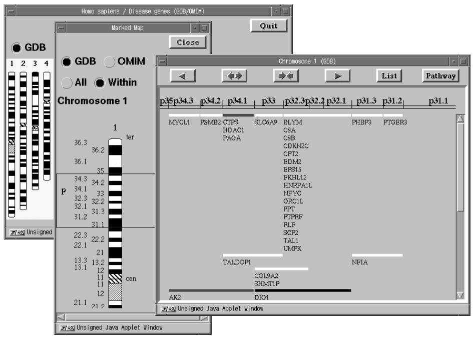

Homo sapiens / Disease genes (GDB/OMIM)
Marked Map
Close
Quit
Chromosome 1 (GDB)
● GDB
● GDB ○ OMIM
○ All ● Within
Chromosome 1
List
Pathway
p35 p34.3 p34.2 p34.1 p33 p32.3 p32.2 p32.1 p31.3 p31.2 p31.1
MYCL1 PSMB2 CTPS HDAC1 PAGA SLC6A9 BLYM C8A C8B CDKN2C CPT2 EDM2 EPB15 FKHL12 HNRPA1L NFYC ORC1L PPT PTPRF RLF SCP2 TAL1 UMPK PHBP3 PTGER3
TALDOP1 NFIA COL9A2 SHMT1P AK2 DIO1
Unsigned Java Applet Window

Figure 5: Genome map browser (*H. sapiens*).

link to the genome map view. Therefore, the user can easily navigate between genome maps and functional catalogs(Figure 6).

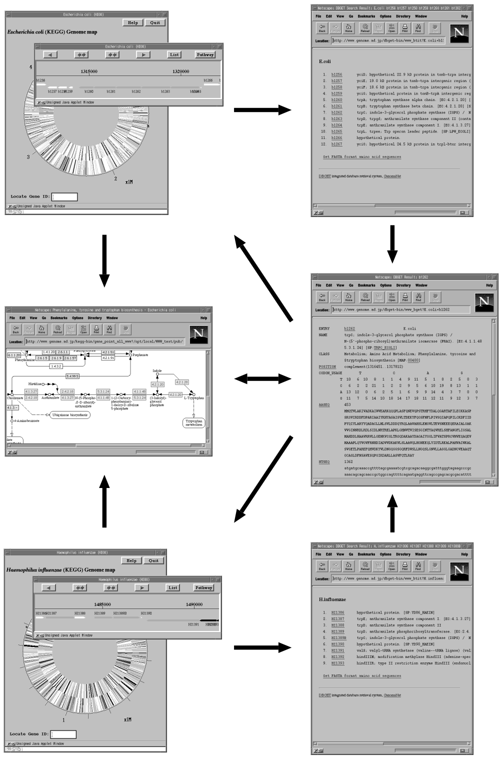

Escherichia coli (KEGG) Genome map
Help
Quit
Locate Gene ID:
Haemophilus influenzae (KEGG) Genome map
E.coli
H.influenzae
Unsigned Java Applet Window

Figure 6: Navigation through genome maps and functional catalogs.

## Links among KEGG

An example of local handling of the genome map is the following. Suppose that one wishes to see if tryptophan operon is conserved in *E. coli* and *H. influenzae*. Starting from the *E. coli* gene catalog of known operons, the genes are mapped on the functional catalog of the metabolic pathway, and then the corresponding *H. influenzae* genes are identified.

All these queries are done on the server. The last step of examining if the *H. influenzae* genes are localized in the genome to possibly form an operon is done locally by searching and marking the genes on the genome map.

In addition to the PATHWAY database, the GENES database is a key database in KEGG. Any gene name in the KEGG pathway or gene catalog has a link to this database and this database has links to the existing molecular biology databases. We also plan to add information of molecular interactions and genetic interactions in the GENES database.

While the KEGG databases provide numerous predefined links that can be used to browse a number of different aspects of genes and genomes, KEGG can also be used as a search tool for finding additional links. For example, genes can be searched on the known pathways in KEGG by EC numbers, by gene names, by compound accessions, or sequence similarity. Figure 7 shows an example of the search by gene names.

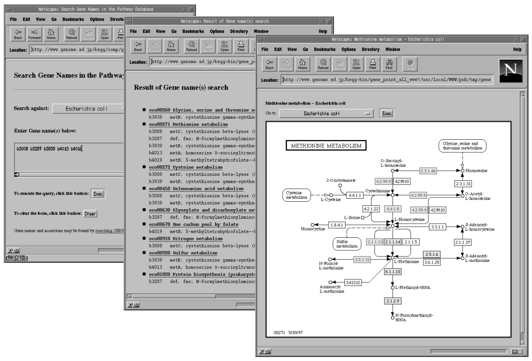

Search Gene Names in the Pathway Database
Result of Gene name(s) search
Methionine metabolism - Escherichia coli
Search against: Escherichia coli
Enter Gene name(s) below:
b3008 b3287 b3939 b4013 b4019
METHIONINE METABOLISM

Figure 7: Links made by pathway search.

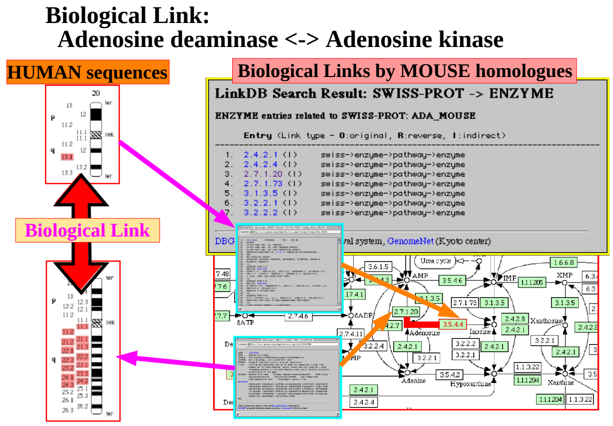
## Biological Reasoning by Computing Links

In LinkDB the link search is configured in the *route table*, for example, the combination of original/reverse links to derive indirect links is defined here. links are to be combined in This mechanism can be extended to the biological reasoning as shown in Table 4. In this example, the processing starts with a human sequence by similarity search against mouse sequences, finds interacting molecules in the KEGG pathways by biological links, and again by similarity search, finds human homologues of these interacting molecules.

Table 4: An example of defining integrated link search.

| database | route |
|---|---|
| h.sapiens | +s(m.musculus):+b(pathway):+s(h.sapiens) |

An illustration of this reasoning process is shown in Figure 8, where biological links in mouse can be utilized to identify human counterpart molecules. Here



```
LinkDB Search Result: SWISS-PROT -> ENZYME

ENZYME entries related to SWISS-PROT: ADA_MOUSE

     Entry (Link type - O:original, R:reverse, I:indirect)
---------------------------------------------------------------------------
 1.  2.4.2.1 (I)      swiss->enzyme->pathway->enzyme
 2.  2.4.2.4 (I)      swiss->enzyme->pathway->enzyme
 3.  2.7.1.20 (I)     swiss->enzyme->pathway->enzyme
 4.  2.7.1.73 (I)     swiss->enzyme->pathway->enzyme
 5.  3.1.3.5 (I)      swiss->enzyme->pathway->enzyme
 6.  3.2.2.1 (I)      swiss->enzyme->pathway->enzyme
 7.  3.2.2.2 (I)      swiss->enzyme->pathway->enzyme

DBG                  val system, GenomeNet (Kyoto center)
```

Figure 8: Biological link search in mouse may provide information about human counterpart molecules.

"adenosine-deaminase" is known to have biological relation to "adenosine-kinase" in the metabolic pathway. In this example, starting with the "adenosine-deaminase" gene on human chromosome 20, the suggested genes of "adenosine-kinase" are marked on the human chromosome 10 by following the combination of similarity and biological links.

## References

Akiyama, Y.; Yakoh, T.; Mori, H.; and Ogasawara, N. 1994. A server-client version of genomatica integrated genome information browser. In *Proc. Genome Informatics Workshop 1994*, 202–203.

Etzold, T., and Argos, P. 1993. Srs–an indexing and retrieval tool for flat file data libraries. *Comput Appl Biosci* 9:49–57.

Fujibuchi, W.; Migimatsu, H.; Uchiyama, I.; Ogiwara, A.; Akiyama, Y.; and Kanehisa, M. 1997. DBGET/LinkDB: an integrated database retrieval system. In *Pacific Symposium on Biocomputing '98*, 683–694.

Goto, S.; Kuhara, S.; Takagi, T.; and Kanehisa, M. 1994. Extension of the integrated database hypergenome for genome maps and sequence information. In *Proc. Genome Informatics Workshop 1994*, 204–205.

Goto, S.; Bono, H.; Ogata, H.; Fujibuchi, W.; Nishioka, T.; and Kanehisa, M. 1996. Organizing and computing metabolic pathway data in terms of binary relations. In *Pacific Symposium on Biocomputing '97*, 175–186.

Goto, S.; Nishioka, T.; and Kanehisa, M. 1998. Ligand: Chemical database for enzyme reactions. *Bioinformatics* accepted.

Kanehisa, M. 1997a. A database for post-genome analysis. *Trends Genet.* 13:375–376.

Kanehisa, M. 1997b. Linking databases and organisms - genomenet resources in japan. *Trends Biochem. Sci* 22:442–444.

Schuler, G.; Epstein, J.; Ohkawa, H.; and Knas, J. 1996. Entrez: molecular biology database and retrieval system. *Methods Enzymol* 266:141–162.