WebDBGET: an integrated database retrieval system which provides hyper–links among related database entries

Yutaka Akiyama, Susumu Goto, Ikuo Uchiyama, and Minoru Kanehisa
Institute for Chemical Research, Kyoto University, Japan

1. Introduction

As useful tools for molecular biologists and biochemists, on–line services of various biological databases have been made freely available on the Internet. Now researchers can, from his desk–top, easily access a vast collection of databases constructed by public institutes or private groups in the world. Those databases include amino acid sequences, nucleic acid sequences, 3–D structures, enzyme reaction, genetic map, genetic diseases, literature information, and so on.

On the other hand, several attempts have been presented to integrate such databases into an integrated system in order to provide more powerful and convenient environment for retrieving and interpreting biological information.

For database integration, various approaches or levels can been considered. If the database is constructed upon relational model, so–called schema integration is a well–defined procedure to join two different databases. If original databases are in flat text format, joining is to design a new flat format to store fused information. This approach, we call **"tightly–coupling" approach**, is suitable to build strong retrieval system. User can perform complecated data searches specified by any logical query. But this approach requires careful designing of fused data format, which is not so suitable to cope with rapid format changes among a large number of databases. Usually this approach also requires physically local gathering of all databases.

On the other hand, more **"loosely–coupling" approach** has been also presented. In this approach, the system does not provide a fused view for databases, but provides many cross–links connecting database elements. In this approach, we have developed a HTTP–based (WWW–based) database retrieval system on which related entries among collected databases are connected by utilization of embedded "anchors" in HTML texts.

Once an entry is retrieved, the user can retrieve related entries by simply clicking on the marked anchors in the cross–references field. We think this "loosely–coupling" approach for database integration may have great advantages in terms of simplicity in maintenance of each independent database and ability in rapid correspondence to novel data formats. In our HTTP–based system called "**WebDBGET**", we have integrated 16 databases with mutual links, those are dynamically embedded in the original database texts by dedicated filter libraries. The HTTP–based approach even allows us to put an entry–by– entry cross–links between databases located at half the globe away.

2. Methods

DBGET commands as basic retrieval engine:

The main engine of our database retrieval system is composed of a few simple commands (bget, bfind, etc.) that we had developed for efficient manipulation of flat database files based on several indices. The command system we call "**DBGET**" has been used for e−mail based retrieval service (dbget@genome.ad.jp) at Kyoto University since 1991. (See text for detail)

Recently DBGET commands are improved using the dbm technique, and also a server−client version (NetDBGET) has been developed by Ogiwara (not published).

New WWW version of DBGET retrieval system (WebDBGET) was released for inter− national users in July 1994 and it also uses the DBGET commands as its basic retrieval mechanism.

For example an HTTP call:

"http://www.genome.ad.jp/htbin/bget_genbank?HUMHPRTB"

will directly invoke a DBGET command "bget genbank HUMHPRTB" (followed by an appropriate anchor filtering described below).

Embedding cross−reference anchors:

The anchor syntax in the HTML language provides us a way to invoke another HTTP call (which can realize cross−reference, annotation, drawing figures, etc.) by simply clicking the marked word. In our system, each retrieval result is passed to an appropriate filtering program, and cross−reference anchors are automatically embedded into the original output text.

We have developed over 20 filter programs (written in ANSI C) dedicated to each database. For example, a SwissProt entry may have links either to EMBL, Prosite, PIR, PDB, OMIM, MEDLINE, or LIGAND so that the "make_anchor_swiss" filter should correctly parse a SwissProt entry text and embed anchor links to some of those seven databases where it is appropriate.

For more concrete example, when a GenBank entry "HUMHPRTB" must be sent out to a WWW client, "make_anchor_genbank" is automatically invoked and embed the following anchors at the appropriate field of the GenBank text, EMBL:M26434, MEDLINE:83213350 and so on.

In order to assure the modularity of filter programs, common libraries were designed and used. The URL addresses for various database services are maintained in an external table (path−table). Lexicographical rules for the entry ID string of each database type are also stored in a separated external table (rule−table) for easy maintenance purpose.

3. Result

We have put the WebDBGET integrated database retrieval system on our WWW server at Kyoto University (http://www.genome.ad.jp) and have released it for world−wide Internet users.

Currently the WebDBGET system provides 16 databases mutually connected: literature (MEDLINE, LITDB), genome maps (GDB), nucleotide sequences (GenBank−today, EMBL), amino acid sequences (PIR, SwissProt, PRF), 3−D structures (PDB, PDBSTR), protein sequence motifs (Prosite), Promotor(EPD), enzyme reaction (LIGAND), amino acid mutations (PMD), amino acid indices (AA−Index), and genetic diseases (OMIM). (Transcription databases TRANSFAC, TFD are planned but not yet released.)

Figure 1 shows the schematic diagram of mutual links among the databases. From this diagram a user can invoke WebDBGET system just by clicking a label of database name in the figure.

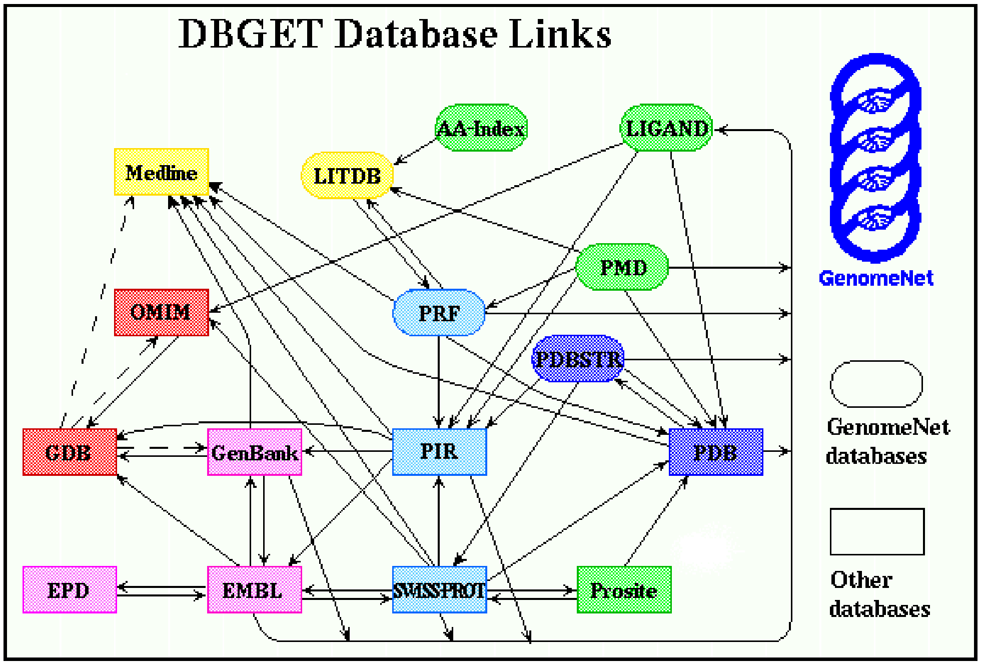

Figure 1. DBGET Database Links Diagram (clickable map)

Currently (as report of May.95) the server is accessed from 44 countries and we have over two thousands HTTP call a day. About 20% is from the outside of Japan. We are managing several databases daily–updated by cooperation with NCBI. On the WebDBGET, GenBank–today, the daily–update version of GenBank is already released. And other daily–updated databases including EMBL and PDB will be soon made released.

4. Discussion

LinkDB offers direct jump:

We have recently constructed a database of cross–link information among databases, called LinkDB (poster presentation at MIMBD'95). Now, when the WebDBGET user clicks on the identifier field of the entry, a list of all linked entries stored in the LinkDB database is given, which greatly facilitates identification of related entries. LinkDB contains not only the original links but also reverse links and indirect links derived by concatenating multiple links, and is daily–updated.

Related work:

SRSWWW(http://www.embl–heidelberg.de/srs/srsc) by EMBL, Heidelberg also provides mutual hyper–links among collected databases. As well as user interface designs, database collection is fairy different from our WebDBGET. WebDBGET provides Japanese originals: PRF, LITDB, PDBSTR, PMD, LIGAND and AAIndex.

Current limitations of the system:

By nature of the "loosely−coupled" approach, the system cannot effectively perform searches over multiple databases, though users may sometimes need to search a keyword across the whole database collection.

Another current limitation is on the accuracy of cross−link information. The links may be erroneous by any of the following reasons: a) database version mismatch, b) parsing confusion (ex. ECnumber vs. other dotted notation), c) combination of weak links resulted in a biologically inappropriate link.

Erroneous anchors caused by a) and b) can be partly detected and eliminated by testing accessability to the entry which the anchor describes. But examination of biological meaning for anchor linking is hard to be automated. Currently biological feasibility of our "link" information is not fully examined. One obvious improvement is to have an access test for the cross−referenced entry every time before embedding a suspected anchor. But it is currently not considered because of time.

Combination with sequence analysis programs:

On our WWW server, result of BLAST or MOTIF search is always processed by a dedicated filter for embedding links to the related database entries. For example homologous entries found by BLAST can be examined simply by clicking. User can even print related literature's abstracts by chaining links toward medline.

5. Conclusion

We have developed the WebDBGET integrated database retrieval system on which related entries among collected databases are connected by hyper−links in WWW. This approach has several advantages when used as a way to integrate a large number of rapidly emerging biological databases.

Acknowledgement:

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 and Culture in Japan.

References

1. Goto, S., Akiyama, Y., and Kanehisa, M.: LinkDB: A Database of Cross Links between Molecular Biology Databases, Poster Presentation, MIMBD−95 (1995).
2. Etzold, T. and Argos, P.: SRS − an indexing and retrieval tool for flat file data libraries, CABIOS, Vol.9, No.1, pp.49−57 (1993).