D1 Data Retrieval with Entrez and DBGET/LinkDB

Key Notes

Access to distributed data

Biological data is widely distributed over the WWW. As an alternative to standard search engines, dedicated data retrieval tools such as Entrez, DBGET and SRS can be used to search multiple biological databases and retrieve relevant information.

Entrez

Entrez is a WWW-based data retrieval tool developed by the NCBI, which can be used to search for information in 11 integrated NCBI databases, including GenBank and its subsidiaries, OMIM and the literature database MEDLINE, through PubMED.

Getting started with NCBI and Entrez

Entrez is accessed via the NCBI homepage and is a simple, user-friendly system. Text search terms (words or Boolean phrases) can be used to search individual databases, and sequences can be used as queries with utilities such as BLAST. Hits are listed in order of relevance or similarity, with hits on the target database known as neighbors and hits on other databases known as links.

DBGET/LinkDB

DBGET is a data retrieval tool maintained by Kyoto University and the University of Tokyo. It covers more than 20 databases and is closely associated with KEGG. A related system, LinkDB, finds relationships between entries in the various databases covered by DBGET and others. DBGET has a simpler and more limited search format than Entrez.

Related topics

- Bioinformatics and the Internet (A3)
- Annotated sequence databases (C2)
- Miscellaneous databases (C4)
- Data retrieval with SRS (sequence retrieval system) (D2)
- Database searches: FASTA and BLAST (E3)

Access to distributed data

A large amount of biological information is available over the World Wide Web (WWW; Topic A2), but the data are widely distributed and it is therefore necessary for scientists to have efficient mechanisms for **data retrieval**. One approach is to use standard **search engines** to find relevant web pages (Topic A3). However, it is sometimes difficult to find the desired information using this method, especially if the chosen search term has other connotations and pulls out many irrelevant sites. Alternatively, there are a number of dedicated **data retrieval tools** that can be used to access information for molecular biologists. The most widely used of these are **Entrez** and **DBGET** (discussed in this topic) and **SRS (sequence retrieval system;** discussed in Topic D2). Each of these tools allows text-based searching of a number of linked databases as well as sequence

searching with utilities such as BLAST (Topic E3). They differ in the databases they cover and how the retrieved information is accessed and presented.

Entrez

Entrez is a WWW-based data retrieval system, developed by the National Center for Biotechnology Information (NCBI), which integrates information held in all NCBI databases. These databases include nucleotide sequences (from GenBank and its subsidiaries), protein sequences, macromolecular structures and whole genomes. Other resources linked to the NCBI can also be searched using Entrez. These include OnLine Mendelian Inheritance in Man (OMIM; Topic C4) and the literature database MEDLINE, through PubMed. Entrez can be accessed via the NCBI web site at the following URL: http://www.ncbi.nlm.nih.gov/Entrez/. In total, Entrez links to 11 databases, which are listed in *Table 1*.

Getting started with NCBI and Entrez

Entrez is the common front-end to all the databases maintained by the NCBI and is an extremely easy system to use. The Entrez main page, as with all NCBI pages, is undemanding in its browser requirements and downloads quickly. Part of the front page is illustrated in *Fig. 1*. The databases available for searching can be accessed by hyperlinks at the top of the page, or by using the drop-down menu as shown. Once a database has been selected, a **search term** is then entered in the space provided. The search term may be a single word or a Boolean phrase. Clicking on 'GO' initiates the search. Hits in the selected database are displayed (these are known as **neighbors**) and matching records in other Entrez databases are also shown (these are known as **links**). Hits are ordered by similarity based on **precomputed analysis** of sequences/structures or the literature.

For the newcomer, the following URL provides an overview of Entrez and a useful tutorial: http://www.ncbi.nlm.nih.gov/Database/index.html. This page

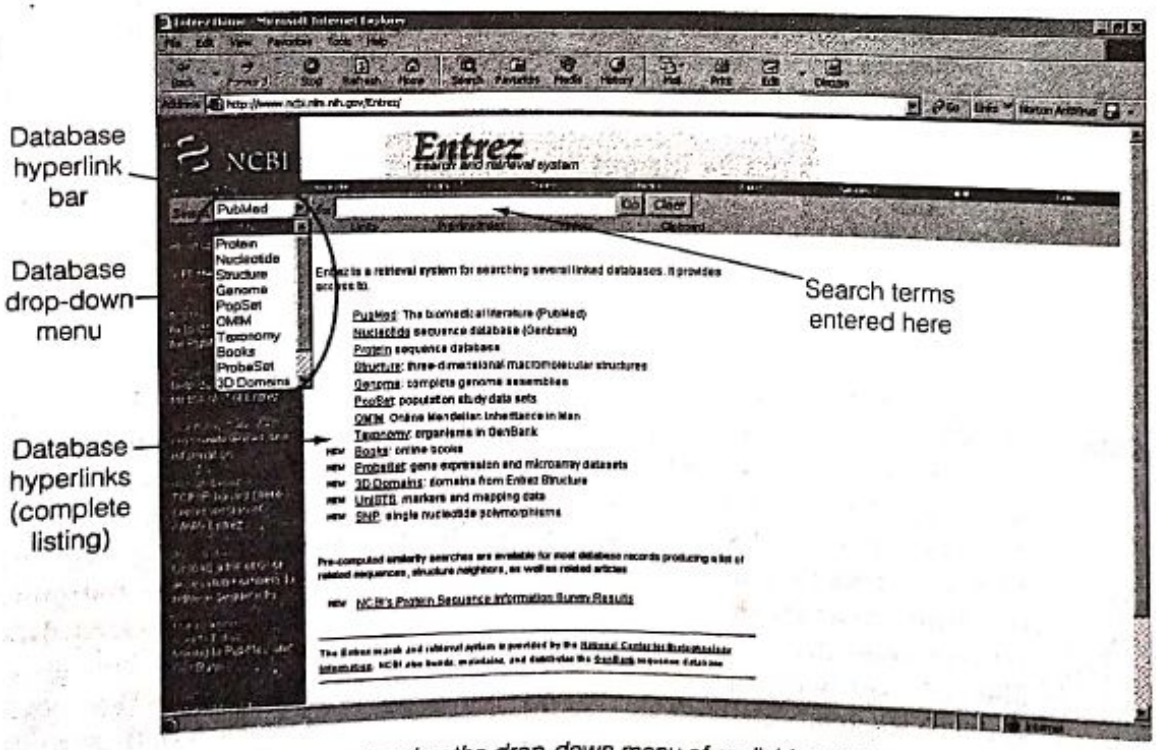

Fig. 1. The Entrez main page, showing the drop-down menu of available databases and the search field.

Table 1. The databases covered by Entrez, listed by category

Category	Database
Nucleic acid sequences	Entrez nucleotides: sequences obtained from GenBank, RefSeq and PDB
Protein sequences	Entrez protein: sequences obtained from SWISS-PROT, PIR, PRF, PDB, and translations from annotated coding regions in GenBank and RefSeq
3D structures	Entrez Molecular Modelling Database (MMDB)
Genomes	Complete genome assemblies from many sources
PopSet	From GenBank, set of DNA sequences that have been collected to analyse the evolutionary relatedness of a population
OMIM	OnLine Mendelian Inheritance in Man
Taxonomy	NCBI Taxonomy Database
Books	Bookshelf
ProbeSet	Gene Expression Omnibus (GEO)
3D domains	Domains from the Entrez Molecular Modelling Database (MMDB)
Literature	PubMED

is shown in *Fig.* 2, and includes a diagram showing the connectivity between eight of Entrez's databases.

DBGET/LinkDB

DBGET is an integrated data retrieval system developed and jointly maintained by the Institute for Chemical Research (Kyoto University) and the Human Genome Center (University of Tokyo). It is integrated with more than 20 data-

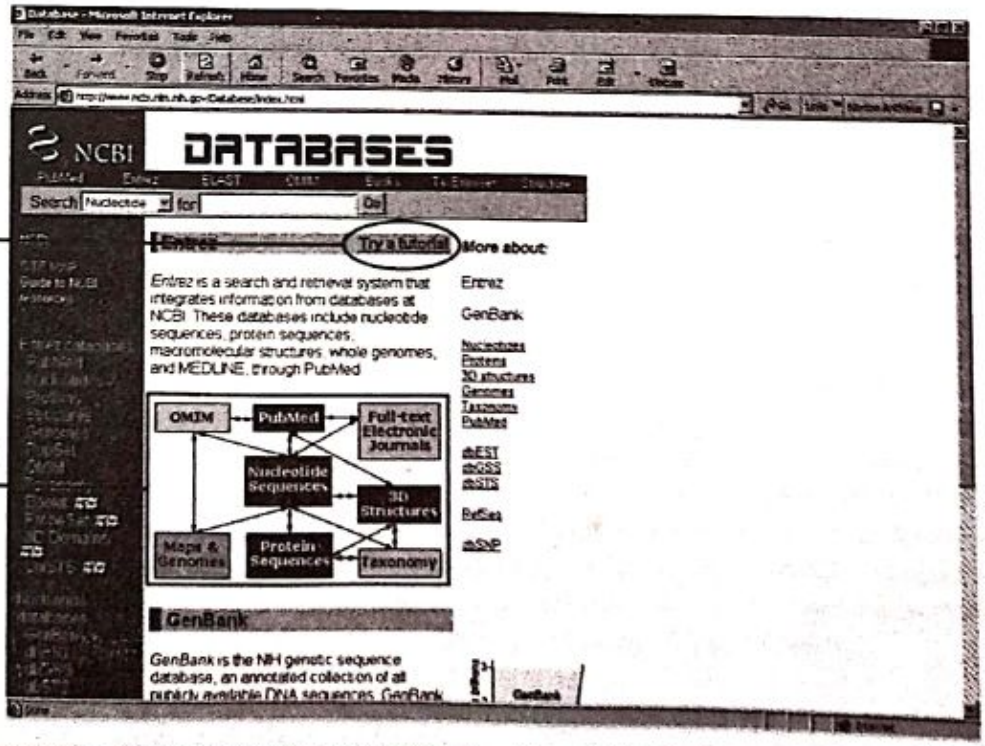

Fig. 2. The Entrez overview page, showing the tutorial link and the relationship between eight Entrez databases.

bases (*Table 2*), which can be searched one at a time or in combination using the commands **bfind** (for text searches) or **bget** (for searches based on accession number). Hits are presented as a list of results together with any available associated information. **LinkDB** is an associated database of links (**binary relationships**) between entries in the different databases available to DBGET and further organism-specific databases, such as the *C. elegans* Database (ACeDB), Flybase and the *Saccharomyces* Genome Database (SGD) (Topic C3). DBGET is closely associated with KEGG, the Kyoto Encyclopedia of Genes and Genomes, which is maintained by the same group (Topic C4).

Table 2. The databases covered by DBGET/LinkDB, listed by category

Category	Database
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 Medline
Organism-specific gene catalogs	*E. coli, H. influenzae, M. genitalium, M. pneumoniae, M. jannaschii, Synechocystis, S. cerevisiae*

D2 Data retrieval with SRS (sequence retrieval system)

Key Notes

SRS and Entrez/DBGET

SRS (sequence retrieval system) is a data retrieval tool that, like Entrez and DBGET, can be used over the WWW. However, unlike these other systems, SRS is open source software and can be installed and run on a local computer network.

Using SRS

SRS databases are grouped but using different principles to those used by Entrez and DBGET. For example, all sequences (nucleic acid and protein) are grouped together, while these are separated by Entrez. The use of SRS involves selecting one or more of these groupings and, within each selected group, selecting one or more of the available databases. Queries can be submitted using two styles of query form, Standard or Extended.

Installing SRS

The advantage of SRS is that it can be installed locally. This allows SRS to be tuned to local databases, which use novel data formats. The tuning of SRS to deal with local databases involves programming in SRS's own scripting language, Icarus.

Related topics

- Data retrieval with Entrez and DBGET/LinkDB (D1)
- Annotated sequence databases (C2)
- Genome and organism-specific databases (C3)
- Miscellaneous databases (C4)
- Protein families and pattern databases (F2)

SRS and Entrez/DBGET

SRS (sequence retrieval system) is a retrieval tool developed by the European Bioinformatics Institute (EBI) that integrates over 80 molecular biology databases. These are listed at http://srs6.ebi.ac.uk/srs6bin/cgi-bin/wgetz?-page+databanks+-newId, and are summarized in *Table 1*. Like Entrez and DBGET, SRS can be used over the WWW. However, the difference between SRS and Entrez/DBGET is that SRS is **open source software** that can be downloaded and installed locally. The result is that the databases and utilities that are available to the end user are not restricted by the activities of curators [at the National Center for Biotechnology Information (NCBI) in the case of Entrez, or GenomeNet in the case of DBGET]. Several large sites, including SWISS-PROT, use SRS as standard. The main ExPASy site (www.expasy.ch), one of the most useful bioinformatics gateways listed in Topic A3, is an example of SRS in action.

Using SRS

To start using SRS, access the appropriate site on your local network (if SRS has been installed locally) or alternatively try the SRS homepage at

Table 1. The databases covered by the SRS at http://srs6.ebi.ac.uk, listed by category

SRS description	Examples
Literature	MEDLINE, GO, GOA
Sequence	EMBL, EMBLNEW, SWISSPROT, SPTREMBL, REMTREMBL, TREMBLNEW, ENSEMBL, PATENT_PRT, USPO_PRT, IMGTLIGM, IMGTHLA
InterPro&Related	INTERPRO, IPRMATCHES, IPRMATCHES_ENSEMBL, PROSITE, PROSITEDOC, BLOCKS, PRINTS, PFAMA, PFAMB, PFAMHMM, PFAMSEED, PRODOM
SeqRelated	UTR, UTRSITE, TAXONOMY, GENETICCODE, EPD, HTG_QSCORE, CPGISLAND, EMBLALIGN, EMESTLIB
TransFac	TFSITE, TFFACTOR, TFCELL, TFCLASS, TFMATRIX, TFGENE
User Owned Databanks	USERDNA, USERPROTEIN
Application Results	FASTA, FASTX, FASTY, NFASTA, BLASTP, BLASTN, CLUSTALW, NCLUSTALW, PPSEARCH, RESTRICTIONMAP
Protein3DStruct	PDB, DSSP, HSSP, FSSP, PDBFINDER, RESID
Genome	MOUSE2HUMAN, LOCUSLINK, HGNC, HSAGENES
Mapping	RHDB, RHDBNEW, RHEXP, RHMAP, RHPANEL, OMIMMAP
Mutations	OMIMALLELE, MUTRES, SWISSCHANGE, EMBLCHANGE, MUTRESSTATUS, OMIM, OMIMOFFSET, HUMUT, HUMAN_MITBASE, P53LINK
Locus Specific Mutations	*41 entries omitted here*
SNP	MITSNP, dbSNPSubmitter, dbSNPAssay, dbSNPSNP, HGBASE, HGBASE_SUBMITTER
Metabolic Pathways	LENZYME, LCOMPOUND, PATHWAY, ENZYME, EMP, MPW, UPATHWAY, UREACTION, UENZYME, UCOMPOUND, UIMAGEMAP
Others	REBASE, SRSFAQ, BIOCATAL
System	PRISMASTATUS

http://srs6.ebi.ac.uk. On the WWW version, the top page lists 17 classifications By clicking on the adjacent '+' symbol, each classification can be expanded t reveal the associated databases, and by clicking on the '–' symbol, the classifica tions can be collapsed. It is also possible to expand collapse all classifications Note that SRS classifications, unlike those in Entrez, are grouped by the type o data not the type of molecule. Therefore, the *Sequence Libraries* classificatio covers all sequences (nucleotide and protein) whereas these data are separate by Entrez. Most of the other classifications are self-explanatory, but new user may find some unfamiliar. For example, *InterPro&Related* refers to the secondar databases of protein motifs (Topic F2), *SeqRelated* refers to specialized sequenc databases such as UniGene (Topic C4) and *TransFac* refers to transcription facto databases (Topic C4).

To search with SRS, expand the relevant classifications and check boxe corresponding to the required databases as shown in *Fig. 1*. Clicking on th 'standard' or 'expanded' buttons then brings up the query form ('standard' i recommended for newcomers). Search terms can be entered in a number o fields by selecting from the drop-down menu (e.g. Accession numbe Description, Keywords, Organism) or alternatively it is possible to search a fields simultaneously. Up to four different search terms can be used, linked b Boolean operators across multiple fields, so quite specific searches can be carrie out. After clicking the 'submit query' button, the query results are displayed a

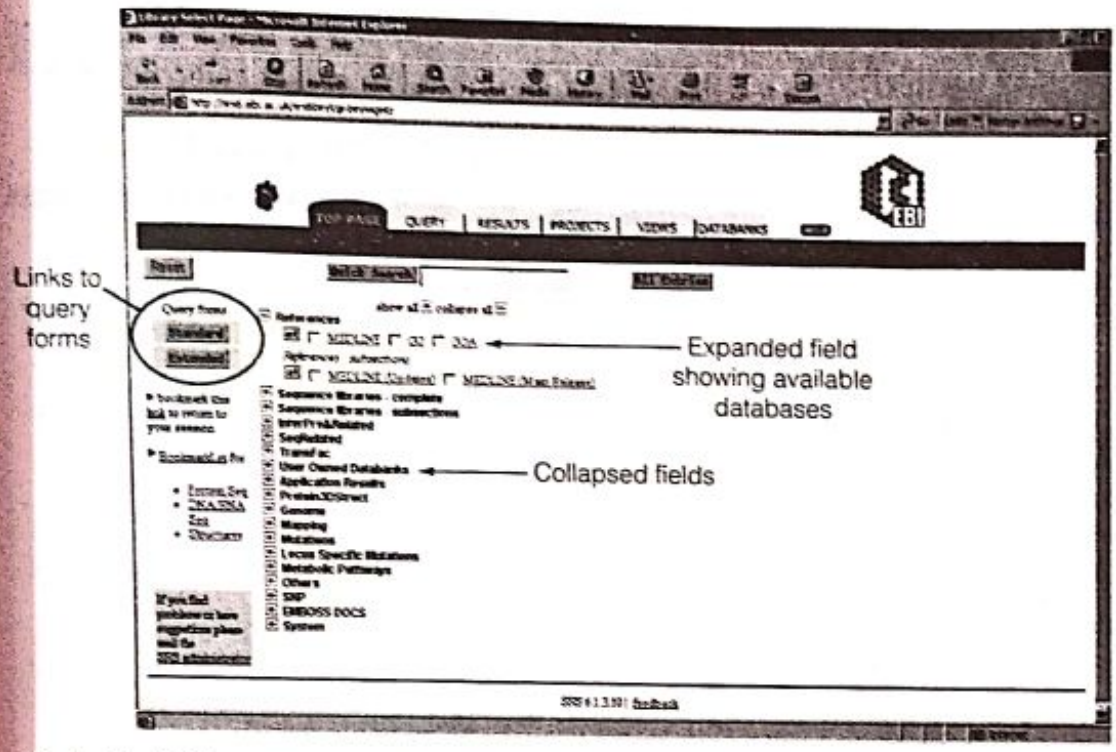

Fig. 1. The SRS top page at http://srs6.ebi.ac.uk. The user has selected the top page and has expanded the Literature classification. One or more of the databases (MEDLINE, GO and GOA) can now be selected and searched.

a table, with hyperlinks to files in the appropriate databases. A helpful tutorial for SRS users is available at http://srs6.ebi.ac.uk. From here, follow the "Information" link and look at the "User Guide".

Installing SRS

SRS is available to be installed locally. Essentially, it provides both the graphical interface and files to allow databases (such as those in *Fig. 1*) to be accessed. SRS works by constructing indexing files and should be updated regularly. The installation set comes with several configuration files and module files for the most popular bioinformatic databases ('databanks' in SRS). SRS configuration files establish which databases are available to SRS. Some of these files are used to add databases and others specify a grouping mechanism for databases or create new built-in views for data results. Module files correspond to the databases to be indexed.

Thus, an SRS system can be installed and operated with the minimum of effort. However, there are several points to consider when planning (or asking for help with) an SRS installation. Although the databases may all be remote, the indexing files are very large and, on a PC or other small system, the indexing process can take a lot of processor time (days). Nevertheless, a modest SRS system using example configuration module files is easily maintained and updated. The strength of SRS lies, in part, with providing the installer the capacity to use databases of his or her own even if they use a novel format. In order to do this, SRS has an associated scripting language called **Icarus**.