The EBI SRS server—new features

Evgeny M. Zdobnov, Rodrigo Lopez*, Rolf Apweiler and Thure Etzold

EMBL Outstation - European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

ABSTRACT

Summary: Here we report on recent developments at the EBI SRS server (http://srs.ebi.ac.uk). SRS has become an integration system for both data retrieval and sequence analysis applications. The EBI SRS server is a primary gateway to major databases in the field of molecular biology produced and supported at EBI as well as European public access point to the MEDLINE database provided by US National Library of Medicine (NLM). It is a reference server for latest developments in data and application integration. The new additions include: concept of virtual databases, integration of XML databases like the Integrated Resource of Protein Domains and Functional Sites (InterPro), Gene Ontology (GO), MEDLINE, Metabolic pathways, etc., user friendly data representation in '*Nice views*', SRSQuickSearch bookmarklets.

Availability: SRS6 is a licensed product of LION Bioscience AG freely available for academics. The EBI SRS server (http://srs.ebi.ac.uk) is a free central resource for molecular biology data as well as a reference server for the latest developments in data integration.

Contact: EBISRS@ebi.ac.uk

INTRODUCTION

SRS (Etzold *et al.*, 1996) is a powerful unified interface to over 400 different databases, which gained impressive popularity in bioinformatics community. It has been developed into an integration system for both data retrieval and sequence analysis applications.

The EBI SRS server (http://srs.ebi.ac.uk) provides flexible and up-to-date access to many major databases in the field of molecular biology, produced and maintained at the EBI and other institutions. The databases are grouped in specialized sections, which include nucleic acid and protein sequences, mapping data, macromolecular structures, sequence variations, protein domains and metabolic pathways (Figure 1). The EBI SRS server contains today more than 150 biological databases and integrates more than 20 applications. SRS is a constantly evolving system. New databases are being added and the interfaces to the old ones are always being enhanced. A robust hardware solution for running SRS has been identified at the EBI that guarantees optimal uptime and load tolerance in a wide variety of usage conditions. The EBI SRS server plays an important role as the reference server for the latest developments in data and application integration (Zdobnov *et al.*, 2001), especially as EBI is a major producer of these data.

WHAT'S NEW

The US National Library of Medicine (NLM) has granted a special licence to EBI to provide public access to the MEDLINE database containing over 11 million bibliographic citations and abstracts from more than 4000 biomedical journals. SRS interface to the primer reference database provides not only fast and well-grained querying access over a number of data fields and use of Unicode characters but also enables efficient use of the data in complex queries to other databases via indexed links between them. For example, retreaving all proteins with annotated similarity to receptors and having particular description in its reference abstract.

Another recent addition is the notion of *Virtual databases*. The normal data distribution cycle of a big database consists of major releases and incremental updates. Hence, users have to query various data sets to get full coverage of the data. The implementation of *virtual databases* hides this layer of complexity from the user and takes care of removing duplicate entries. At the EBI SRS server virtual databases are already defined for the EMBL Nucleotide Sequence Database, MEDLINE and Radiation Hybrid database (RHdb).

To allow more specific querying, SRS uses a concept of *subentries*—logically independent entities nested in the *parent* database *entries*. Probably the most commonly known examples of *subentries* are the elements of a *Feature Table* in sequence databases such as EMBL or SWISS-PROT. Since SRS6, it is possible to define several *subentries* per database. In the case of SWISS-PROT there are now five types of *subentries* for the *Feature Table*, publication references, database cross-references, comments and a special purpose subentry called *Counter,* which was introduced in order to make the number of
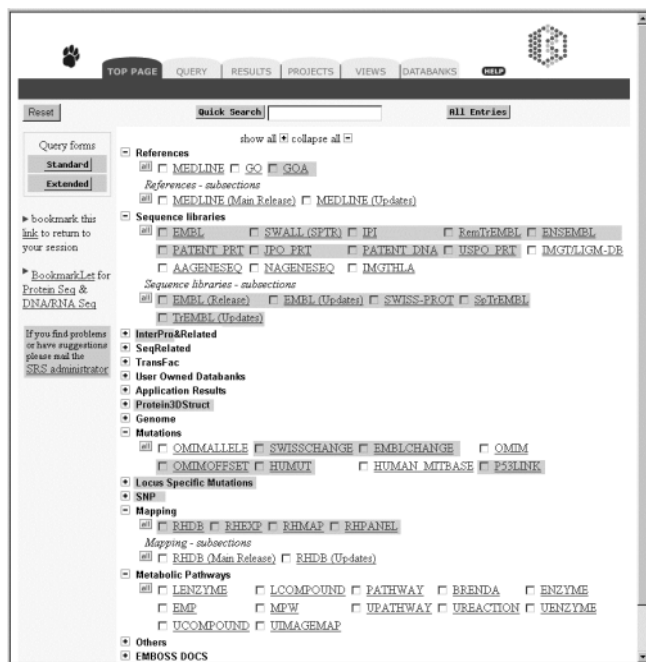
Fig. 1. Some of the databases available through the EBI SRS server. The marked databases are produced and maintained at the EBI.

certain features searchable, e.g. to allow querying for all proteins with exactly seven transmembrane regions. The queries can be easily constructed using the *extended query form* in the SRS web interface.

InterPro was the first XML formatted databank integrated in SRS using semi-generic *icarus* parsers, followed by MEDLINE, GO and RESID (Garavelli *et al.*, 2001), and it triggered the low-level implementation of XML support in SRS that allows the integration of XML formatted databanks without the need to write parsers.

SRS provides very flexible data representations. Enhancements to the object loader allow the creation of more user friendly—*Nicely* formatted html *views* of the data.

Advanced support for application launching and support of various UNIX queuing systems allows the EBI to provide public access to a choice of sequence analysis applications integrated under SRS. This integration benefits from the dynamic interpretation of an application's output

as a database and linking these results to other databases gives an enhanced overview of the results in the context of other data, e.g. filtering out Blast hits from viruses sequences. Using the linking graph, SRS makes it possible to perform link queries on databases that do not contain direct references to each other.

Being a keyword-based system, SRS's functionality is limited on free text descriptions. The technique of indexing all consecutive pairs of words makes the querying of concepts buried deep in free text much more powerful, e.g. the query of '*cytochrome c*' is different from that of '*cytochrome*' AND '*c*'.

Another unique feature of the EBI's SRS server are simple but very handy *SRSQuickSearch bookmarklets* (tiny JavaScript code) that can be customised for querying databases, using selected data fields, etc. and bookmarked as ordinary html links.

The SRS server at the EBI uses extensively the capability of the system to prepare indices off-line and thus data availability is never compromised. Data warehousing and server maintenance is simplified by the use of SRS PRISMA.

The EBI has implemented a very robust solution for running SRS in a hardware environment that guarantees optimal performance and updatedness. This is currently composed of a cluster of four Compaq ES40 sharing memory and storage over the EBI's SAN (Storage Area Network) where approximately 1 terabyte of data are on-line for SRS. More than 70% of all updates are done using PRISMA and LSF, a combination that permits efficient load sharing and balancing and very fast indexing. This particular solution permits planned or unplanned maintenance of individual components, something which is reflected in the fact this server is available 365 days per year.

REFERENCES

Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.

Garavelli,J., Hou,Z., Pattabiraman,N. and Stephens,R.M. (2001) The RESID Databases of protein structure modifications and the NRL-3D Sequence-Structure Database. *Nucleic Acid Res.*, **29**, 199–201.

Zdobnov,E.M., Lopez,R., Apweiler,R. and Etzold,T. (2001) The EBI SRS server-recent developments. *Bioinformatics* (accepted as GCB'2000 paper).