BIOINFORMATICS

Vol. 18 no. 2 2002
Pages 368–373

The EBI SRS server—recent developments

Evgeni M. Zdobnov, Rodrigo Lopez, Rolf Apweiler and Thure Etzold

EMBL Outstation, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

Received on March 2, 2001; revised on March 15, 2001; accepted on March 22, 2001

ABSTRACT

Motivation: The current data explosion is intractable without advanced data management systems. The numerous data sets become really useful when they are interconnected under a uniform interface—representing the domain knowledge. The SRS has become an integration system for both data retrieval and applications for data analysis. It provides capabilities to search multiple databases by shared attributes and to query across databases fast and efficiently.

Results: Here we present recent developments at the EBI SRS server (http://srs.ebi.ac.uk). The EBI SRS server contains today more than 130 biological databases and integrates more than 10 applications. It is a central resource for molecular biology data as well as a reference server for the latest developments in data integration. One of the latest additions to the EBI SRS server is the InterPro database—Integrated Resource of Protein Domains and Functional Sites. Distributed in XML format it became a turning point in low level XML–SRS integration. We present InterProScan as an example of data analysis applications, describe some advanced features of SRS6, and introduce the SRSQuickSearch JavaScript interfaces to SRS.

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: Rodrigo.Lopez@EBI.ac.uk; Evgueni.Zdobnov@EBI.ac.uk

INTRODUCTION

Advances in data management systems provide us with tools to cope with an otherwise intractable data explosion in molecular biology. This opens new challenges in large-scale data analysis using, e.g. statistical approaches. The integration of heterogeneous databases is critically important. The numerous data sets become really useful when they are interconnected under a uniform interface representing the domain knowledge. Sequence Retrieval System (SRS) was originally aimed at facilitating access to biological sequence databases (Etzold and Argos, 1993; Etzold *et al.*, 1996). Today it provides a powerful unified interface to over 400 different scientific databases. It provides capabilities to search multiple databases by shared attributes and to query across databases fast and efficiently. Nowadays SRS has become an integration system for both data retrieval and applications for data analysis. Originally SRS was developed at the EMBL and then later at the EBI. In 1999 it was acquired by LION Bioscience AG. Since then SRS has undergone a major internal reconstruction and SRS6 was released as a licensed product freely available for academics. The EBI SRS server (http://srs.ebi.ac.uk) is a central resource for molecular biology data as well as a reference server for the latest developments in data integration.

SRS CONCEPTS

SRS is designed to retrieve data directly from text files. From the beginning of the computer era text files in ASCII format have been widely accepted as a format to exchange information. This makes them portable to any computer system. While the paradigm of computers expanded from just performing calculations to complex data management, it became obvious that plain text format is not efficient enough for these purposes. But text files are still widely used to exchange and distribute information. In fact, formatted text files are the *de-facto* standard for biological databases like EMBL (Stoesser *et al.*, 1999) and SWISS-PROT (Bairoch and Apweiler, 2000). Self-descriptive XML format has advanced features but it is still just text.

The key feature of SRS is its unique object oriented design. It uses meta-data to define a class for a database entry object and rules for text-parsing methods, coupled with the entry attributes. For object definition and parsing rules SRS uses its own scripting language, Icarus, for which a debugger has been recently implemented. While RDBMS are highly advanced for data management, SRS has advantages as a retrieval system: first, it is much faster (10–100 times) than retrieving whole records from large databases

with complex data schemas (like EMBL). Second, since it retrieves data directly from flat files it is less demanding in terms of storage space requirements than RDBMS tables. The average difference of 2–5 times is significant in the case of large databases as EMBL, which is about 28 Gb in flat file format at present. Third, it is reasonably easy to integrate new data with basic retrieval capabilities and extend it further to a more sophisticated data schema. The integrating power of SRS benefits from sharing the definitions of conceptually equal attributes amongst different data sets. This enforces uniformity and allows multiple-database queries. Searchable links between databases and customizable data representation are original features of SRS.

Linking

Data becomes more valuable in the context of other data. Besides enriching the original data by providing html linking, one of the original features of SRS is the ability to define indexed links between databases. These links reflect equal values of named entry attributes in two databases. It could be a link from an explicitly defined reference in DR (data reference) records in SWISS-PROT or an implicit link from SWISS-PROT to the ENZYME database by a corresponding Enzyme Commission (EC) number in the protein description. The links are bi-directional, operate on sets of entries, can be weighted and can be combined with logical operators (AND, OR and NOT). This is analogous to a table of relations in a relational database schema that allows querying of one table with conditions applied to others. The user can search not only the data contained in a particular database but also any conceptually related databases and then link to the desired data. Using the linking graph, SRS makes it possible to link databases that do not contain direct references to each other. Highly cross-linked data sets become a kind of domain knowledge base. This helps to perform queries like 'give me all proteins that share InterPro domains with my protein' by linking from SWISS-PROT to InterPro and back to SWISS-PROT, or 'give me all eukaryotic proteins for which the promoter is further characterized' by selecting only entries linked to the Eukaryotic Promoter Database (EPD) from the current set.

THE EBI SRS SERVER

The EBI SRS server plays an important role in EBI's mission to provide services in bioinformatics. It gives a flexible and up-to-date access to many major databases produced and maintained at the EBI and other institutions. The databases are grouped in specialized sections, including nucleic acid and protein sequences, mapping data, macromolecular structure, sequence variations, protein domains and metabolic pathways (Table 1). The EBI SRS server contains today more than 130 biological databases and integrates more than 10 applications. SRS is a constantly evolving system. New databases are being added and the interfaces to the old ones are always being enhanced. This server is in high demand by the bioinformatics community. Currently, requests and queries on the system total more than 3 genuine million hits (of all types) per month with a growth rate of more than 15% per month.

All SRS database parsers are available to external users and thus, the EBI SRS server plays an important role as a reference site for most other SRS servers. SRS has gained wide popularity and now there are more than 100 installations worldwide. To track the information available on publicly available databases on numerous SRS servers there is the 'Database of Data Banks'. It is based on a set of scripts that automatically gather information from SRS servers on the Internet and organizes these data into a searchable database (Kreil and Etzold, 1999). The SRS server at the EBI uses extensively the capability of the system to prepare indices off-line. This feature of SRS6.x solves the problem of a database not being available for querying during the updating process. Although there is a drawback in terms of storage the mere fact that the database is always on-line outweighs this disadvantage.

ADVANCED FEATURES & RECENT ADDITIONS

Multiple subentries

Data representation as a stream of entries in flat text implies restrictions to the underlying data schema. Since support for more advanced data schemas allows the resolution of more specific queries, SRS introduced subentries as logically independent concepts nested in the parent database entries (Etzold *et al.*, 1996). Probably the most commonly known examples of subentries are the elements of Feature Tables in sequence databases such as EMBL or SWISS-PROT. Other widely occurring cases are publication references. In SRS6, it is possible to define several subentries per database. In the case of SWISS-PROT there are now several definitions for subentries corresponding to elements of the feature table, publication references and comments. A special purpose subentry, called 'Counter', was introduced in order to make the number of links to other databanks and/or the number of certain features searchable. Using the 'Counter' subentry it is possible to query for all proteins with exactly seven transmembrane regions and with annotated similarities to receptors. The query can be easily constructed using the 'extended query form' in the SRS web interface.

Virtual data fields

SRS does not store any parsed data except indexes. The run time instance of an entry object gets its attribute values through defined text parsing methods. This works

Table 1. Some of the databases available through the EBI SRS server. Databases marked in bold are produced and maintained at the EBI

Sequence				
	EMBL	**EMBLNEW**	**ENSEMBL**	
	SWISSPROT	**SPTREMBL**	**TREMBLNEW**	**REMTREMBL**
	SWALL	IMGT	IMGTHLA	
InterPro & Related				
	InterPro	**InterProMatches**	PROSITEDOC	BLOCKS
	PFAMA	PFAMB	PFAMHMM	PFAMSEED
	PRINTS	**NICEDOM**	PRODOM	PROSITE
SeqRelated				
	TAXONOMY	GENETICCODE	EPD	**HTG_QSCORE**
	UTR	UTRSITE	**EMESTLIB**	
TransFac				
	TFSITE	TFFACTOR	TFCELL	TFCLASS
	TFGENE	TFMATRIX		
Protein3DStruct				
	PDB	**DSSP**	**HSSP**	**FSSP**
Genome				
	HSAGENES	MOUSE2HUMAN	LOCUSLINK	
Mapping				
	RHDB	**RHEXP**	**RHMAP**	**RHPANEL**
	OMIMMAP			
Mutations				
	MUTRES	**MUTRESSTATUS**	OMIM	**OMIMALLELE**
	OMIMOFFSET	**SWISSCHANGE**	**EMBLCHANGE**	**HUMUT**
	HUMAN_MITBASE	**P53LINK**		
SNP				
	MITSNP	dbSNP_Contact	dbSNP_Method	
	dbSNP_Population	dbSNP_Publication	dbSNP_Assay	
	dbSNP_SNP	dbSNP_PopUse	dbSNP_IndUse	
	HGBASE	**HGBASE_SUBMITER**	**SNPLink**	
Metabolic Pathways				
	PATHWAY	LENZYME	LCOMPOUND	BRENDA
	EMP	MPW	UPATHWAY	UREACTION
	UCOMPOUND	UIMAGEMAP	ENZYME	UENZYME

only on demand, which is implemented by so called 'lazy parsing'. This is flexible enough to allow different data representations—views, constructed at the level of entry data fields. Nothing prevents the definition of an attribute coupled with a method generating 'on-the-fly' new data from the original data. These could be the graphical visualization of protein domains and/or functional sites, links to external data sources or precompiled SRS queries. As an example, the 'AllSeq' attribute of a PRODOM entry is the SRS query that leads to all SWISS-PROT proteins containing this PRODOM domain.

Composite views

SRS allows the definition of composite views that dynamically link entries from the main query database to other related databases. These views display external data as if they were original database attributes. An example is the visualization of the PRODOM (Corpet *et al.*, 2000) domains in SWISS-PROT protein sequences using a view called 'SW_NiceDom' available at the EBI SRS server. The PRODOM group compiled the graphical visualization of the protein domain structures as a set of images along with html maps, which are available as a supplement to the PRODOM database distribution. We used this set of images and maps to compile a derived database called NICEDOM that allows us to build a bridge between protein sequences and the graphical representation of the domains. A NICEDOM entry is composed of the protein accession number and image mapping information. The image data field is actually a generated link to the original images. The 'SW_NiceDom' view combines some data fields from a protein sequence database, which are dynamically fetched through NICEDOM graphics (Figure 1). Another useful

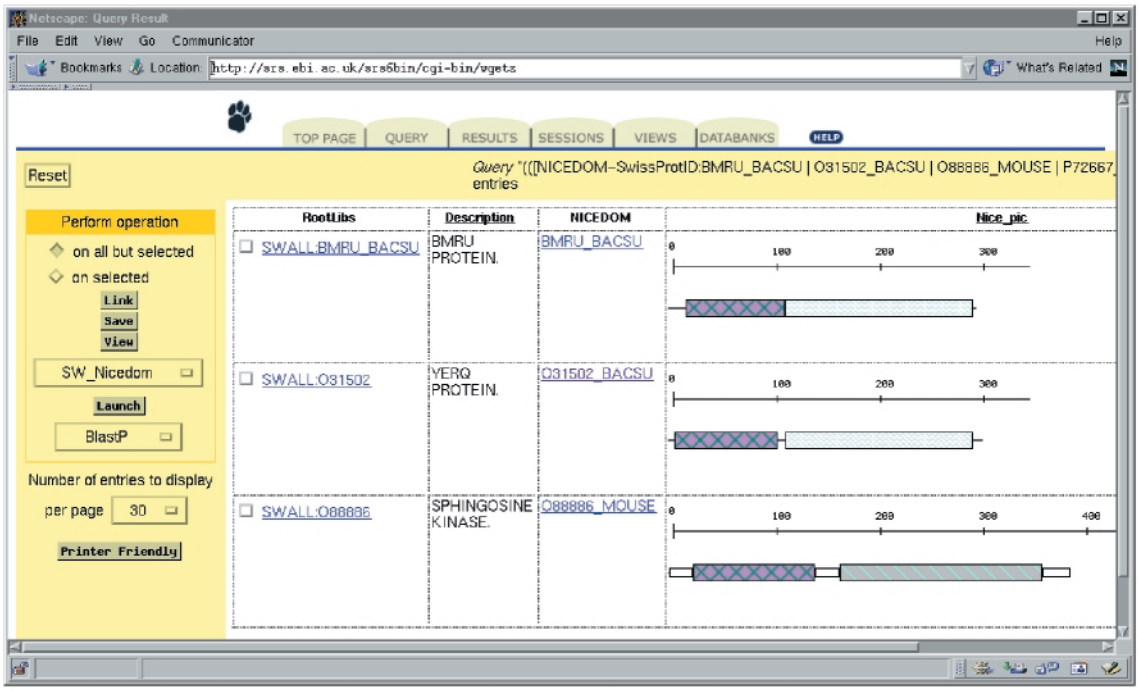

Fig. 1. Example showing the PRODOM domain structure of some protein sequences.

example is 'SW_InterProMatches' view that dynamically links protein sequences to the InterProMatches database.

InterPro & XML

One of the latest additions to the EBI SRS server is the InterPro database (The InterPro Consortium, 2000). InterPro (Integrated Resource of Protein Domains and Functional Sites) is an on-going initiative of several international groups co-ordinated at EBI. The main goal is to create and maintain a resource of known protein domains and functional sites with curated biological annotation. The first release (March 2000) is based on data from the PROSITE (Hofmann *et al.*, 1999), PRINTS (Attwood *et al.*, 2000) and PFAM (Bateman *et al.*, 2000) databases. PRODOM is now getting integrated into the system. It is essential to understand that InterPro is not going to replace the databases it integrates, but introduces checked cross-links between them and enriches the motif information by annotation. The databank is freely available through the EBI SRS server and downloadable in XML format (ftp://ftp.ebi.ac.uk/pub/databases/interpro/). The most up-to-date version is accessible through an ORACLE web interface (http://www.ebi.ac.uk/interpro/search.html). InterPro is the first XML formatted databank integrated in SRS and it represents a turning point in low level XML–SRS integration.

Data analysis applications

The introduction of the biosequence object in SRS allows the integration of various sequence analysis tools such as FASTA (Pearson and Lipman, 1988; Pearson, 1990) or CLUSTALW (Thompson *et al.*, 1994). This integration allows the treatment of text output of these applications like any other database. This enables linking to other databanks and user-defined data representations. Up to now about a dozen applications are already integrated into SRS and many others are in the pipeline. Expanding in this direction SRS becomes not only a data retrieval system but also a data analysis application server (Figure 2). Recent advances in application integration include different levels of user control over application parameters, support for different UNIX queuing systems (LSF, CODINE, DQS, NQS) and parallel threading. There is now also support for 'user-owned data' (the user's own sequences), which make SRS a more comprehensive research tool.

InterProScan

As an example of a data analysis application we present InterProScan, which was recently implemented at the EBI. InterProScan provides a single interface to a set of applications for scanning protein sequences against InterPro member databases. Currently it is based on:

(1) the FingerPRINTScan (Scordis *et al.*, 1999) application that searches the PRINTS database for protein signatures;

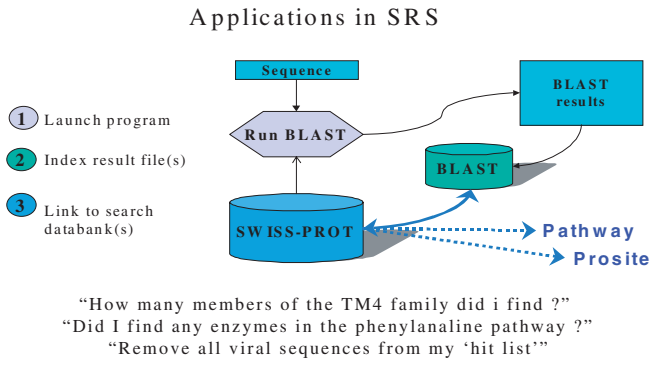

Fig. 2. The integration of applications in SRS has the advantage of treating the application output like any other database, which allows linking to other databanks and user-defined data representation.

(2) ProfileScanner (pfscan) from the Pftools package for searching protein sequences against a collection of generalized profiles in PROSITE (http://www.isrec.isb-sib.ch/software/PFSCAN_form.html);

(3) Ppsearch (Fuchs, 1994) for PROSITE pattern matching;

(4) HMMPfam from the HMMER package (http://hmmer.wustl.edu/) or HMMS implemented on a Decypher machine from TimeLogic that scans sequences against the Pfam collection of protein domain HMMs (hidden Markov models). InterProScan provides an efficient way to analyze protein sequences for known domains and functional sites by launching the applications in parallel, parsing their output and combining the results at the level of unified attributes into one representation with graphical visualization of the matches.

New services based on 'SRS Objects'

LION Bioscience has made available with SRS6 some APIs to popular programming languages, namely C++, JAVA, PERL and PYTHON (provided as a separate product). This permits the development of highly customized stand-alone interfaces, which use 'SRS Objects' for data retrieval, application launching and protected user sessions. While the SRS web interface can become at times complicated for beginners, it is possible to create programs with a specialized interface. To broaden the community using InterProScan we implemented it as both an integrated SRS sequence analysis tool and as a stand-alone web interface (http://www.ebi.ac.uk/interpro/interproscan/ipsearch.html) with the simplest possible interface using the Perl API. This client program automatically creates a user session and generates interfaces to all InterPro related applications within SRS. It effectively reuses SRS parsing of the results into memory objects and uses SRS to lookup related data. This approach trades the SRS integrity with the simplicity of the user interface, implementing 'one-click-away' results (Figure 3). The provision of these APIs represents a big step in the integration of common languages with SRS but it implies that the client program and the SRS server share the same file system (e.g. over NFS). Fortunately, SRS has a CORBA API as well (Coupaye, 1999), which allows development of truly distributed networked systems. For example, to enhance the searching capabilities of the simple interfaces for the CluSTr and InterPro databases stored in ORACLE, we use the SRS CORBA interface to extend the user query through an 'all-text search' in linked databases under SRS.

Free text indexing

Many biological databases contain free text descriptions. The simplest indexing of all individual words in free text lacks the ability to reflect the word's semantic meaning and does not represent underlying concepts specifically enough. A recently introduced technique of indexing all consecutive pairs of words makes the querying of concepts buried deep in free text descriptions much more powerful without significant compromise on index size or search speed. As an example: the result of the query of 'cytochrome c' is quite different from the query of 'cytochrome'AND 'c'.

Bookmarklets

It is worthwhile to mention the simple but very handy JavaScript interfaces to SRS that have also been developed recently (SRSQuickSearch). These have the advantage that they can be bookmarked as ordinary html links. In WWW parlance they are called BookmarkLets (http://www.bookmarklets.com/). Modern browsers such as Netscape or Internet Explorer allow the user to rearrange their bookmarks so that they appear as buttons on the browser window from where the BookmarkLets can be conveniently called at any time. The user can highlight one or more words on the current page and click on the SRSQuickSearch bookmarklet button to execute an SRS query. These scripts are especially useful when customized for particular needs. To make the users life easier we provide a set of the most popular pre-configured SRS bookmarklets as well as a tool to generate customized SRSQuickSearch bookmarklets. These scripts are used extensively by the curators at the EBI.

Simple search

To simplify the user interface to SRS we introduced a number of simple web forms based on JavaScript code. These are shortcuts for simple queries. All the required code is in the page source and users are encouraged to take it from the EBI web pages and use it for particular local needs.

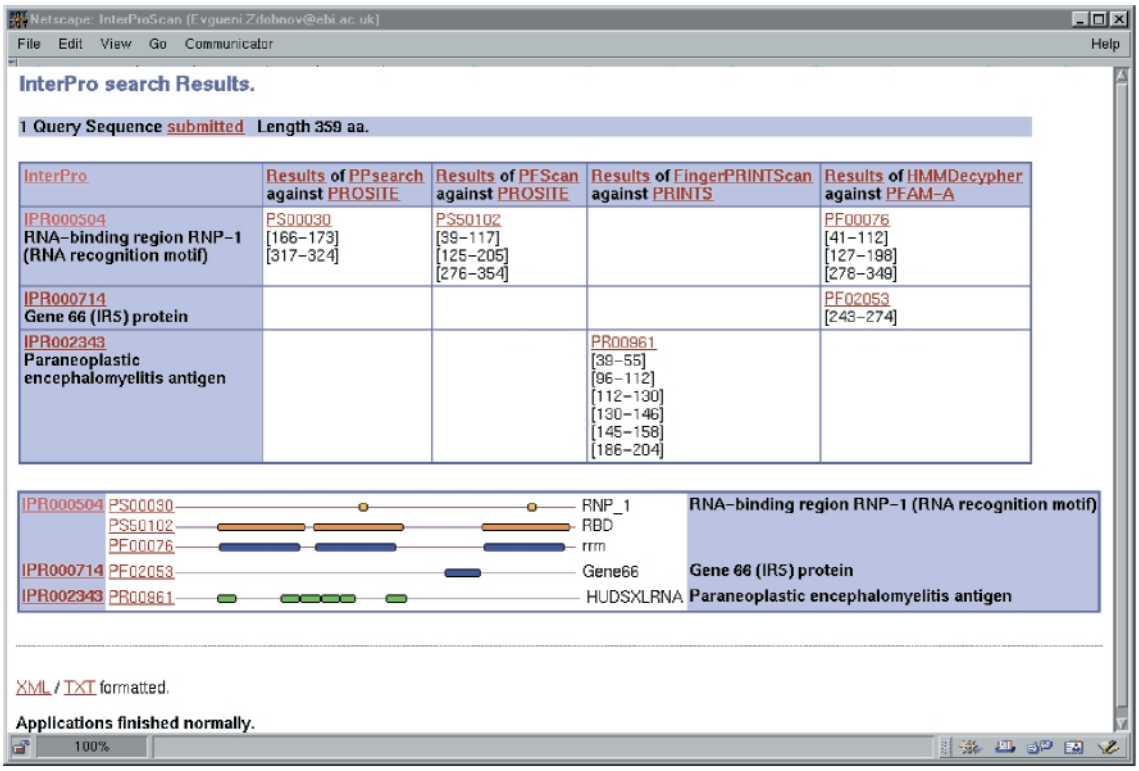

Fig. 3. Output of InterProScan, showing the InterPro domains in the query protein sequence.

REFERENCES

Attwood,T.K., Croning,M.D.R., Flower,D.R., Lewis,A.P., Mabey,J.E., Scordis,P., Selley,J.N. and Wright,W. (2000) PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.*, **28**, 225–227.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L.L. (2000) The pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.

Corpet,F., Servant,F., Gouzy,J. and Kahn,D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.*, **28**, 267–269.

Coupaye,T. (1999) Wrapping SRS with CORBA: from textual data to distributed objects. *Bioinformatics*, **15**.

Etzold,T. and Argos,P. (1993) Transforming a set of biological flat file libraries to a fast access network. *Comput. Appl. Biosci.*, **9**, 59–64.

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

Fuchs,R. (1994) Predicting protein function: a versatile tool for the Apple Macintosh. *Comput. Appl. Biosci.*, **10**, 171–178.

Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.

Kreil,D.P. and Etzold,T. (1999) DATABANKS—a catalogue database of molecular biology databases. *TIBS*, **24**, 155–157.

Lehvaslaiho,H., Ashburner,M. and Etzold,T. (1998) Unified access to mutation databases. *Trends Genet.*, **14**, 205–206.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence analysis. *Proc. Natl Acad. Sci.*, **85**, 2444–2448.

Pearson,W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Meth. Enzymol.*, **183**, 63–98.

Scordis,P., Flower,D.R. and Attwood,T.K. (1999) FingerPRINTScan: intelligent searching of the PRINTS motif database. *Bioinformatics*, **15**, 799–806.

Stoesser,G., Tuli,M.A., Lopez,R. and Sterk,P. (1999) The EMBL nucleotide sequence. *Nucleic Acids Res.*, **27**, 18–24.

The InterPro Consortium (*Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Copet,F., Croning,M.D.R., Durbin, Falquet,L.R., Fleischmann,W., Gouzy,J., Hermjakob,H., Hulo,N., Jonassen,I., Kahn,D., Kanapin,A., Karavidopoulou,Y., Lopez,R., Marx,B., Mulder,N.J., Oinn,T.M., Pagni,M., Servant,F., Sigrist,C.J.A. and Zdobnov,E.M.) (2000) InterPro—an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics*, **16**, 1145–1150.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.