Nucleic Acids Research, 2005, Vol. 33, Database issue ***D141–D146***
doi:10.1093/nar/gki021

UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs

Flavio Mignone[1], Giorgio Grillo[2], Flavio Licciulli[2], Michele Iacono[1], Sabino Liuni[2], Paul J. Kersey[4], Jorge Duarte[4], Cecilia Saccone[2,3] and Graziano Pesole[1,2,*]

[1]Dipartimento di Scienze Biomolecolari e Biotecnologie, Università di Milano, via Celoria 26, 20133 Milano, Italy, [2]Sezione di Bioinformatica e Genomica, Istituto Tecnologie Biomediche del Consiglio Nazionale delle Ricerche (CNR), via Amendola 165/A, 70126 Bari, Italy, [3]Dipartimento di Biochimica e Biologia Molecolare, Università di Bari, via Orabona 4, 70126 Bari, Italy and [4]EMBL Outstation, The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK

Received September 15, 2004; Accepted September 18, 2004

ABSTRACT

The 5′ and 3′ untranslated regions of eukaryotic mRNAs play crucial roles in the post-transcriptional regulation of gene expression through the modulation of nucleo-cytoplasmic mRNA transport, translation efficiency, subcellular localization and message stability. UTRdb is a curated database of 5′ and 3′ untranslated sequences of eukaryotic mRNAs, derived from several sources of primary data. Experimentally validated functional motifs are annotated (and also collated as the UTRsite database) and cross-links to genomic and protein data are provided. The integration of UTRdb with genomic and protein data has allowed the implementation of a powerful retrieval resource for the selection and extraction of UTR subsets based on their genomic coordinates and/or features of the protein encoded by the relevant mRNA (e.g. GO term, PFAM domain, etc.). All internet resources implemented for retrieval and functional analysis of 5′ and 3′ untranslated regions of eukaryotic mRNAs are accessible at http://www.ba.itb.cnr.it/UTR/.

INTRODUCTION

One of the main challenges of the post-genomic era is the understanding of the mechanisms that control the spatio-temporal regulation of gene expression. The fate of newly synthesized mRNA with respect to its nucleo-cytoplasmic transport, stability, translation efficiency and subcellular localization is determined at the post-transcriptional level. Such regulation is mostly mediated by cis-acting elements located in the 5′ and 3′ untranslated regions of mRNAs (5′ UTR and 3′ UTR) (1).

In several cases, specific functional sequence elements have been identified and characterized. These usually correspond to short oligonucleotide tracts whose biological activity relies on a combination of their primary sequence and specific secondary structure. These motifs act either as target sites for RNA-binding factors or interact directly with the translation machinery.

The availability of a large collection of functionally related sequences—such as UTRs—is invaluable for the inference of structural and compositional features and for the identification of conserved candidate regulatory motifs. For this reason, we have developed UTRdb, a collection of 5′ and 3′ UTR sequences derived from eukaryotic mRNAs. Sequences collated in UTRdb were generated by custom software. UTRdb is a non-redundant database and annotation includes information not available in the primary databases such as genome localization and structure and presence of known regulatory elements.

We have also created UTRsite, a collection of regulatory elements located in 5′ and 3′ UTRs whose function and structure have been experimentally determined and published. The UTRsite collection may prove useful in automatic annotation projects of unknown sequences as well as for finding previously undetected signals in known sequences.

For the most recent release of the UTRdb and UTRsite databases, we have focused on the improvement of data quality, increasing the degree of integration with other resources and the incorporation of genome-related facilities. Besides a new graphical interface, we have introduced new specific UTR collections: (i) UTRef from RefSeq database (2); (ii) UTRait from TRAIT database of muscle-specific transcripts (3); and (iii) UTRexp, a collection of UTR sequences whose functional activity has been experimentally investigated

*To whom correspondence should be addressed. Tel: +39 02 50314915; Fax: +39 02 50314912; Email: graziano.pesole@unimi.it

General Information

Entry Name	UTR:3HSA059740
Accession #	CC221872;
Molecule Type	mRNA
Sequence Length	922
Entry Division	HUM
Creation_Date	01-SEP-2003
Modification_Date	01-SEP-2003

Description

Description	3'UTR in Homo sapiens AK2B mRNA for adenylate kinase isozyme 2, complete cds.
Organism	Homo sapiens (human)
Organism Classification	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
Region	3'UTR
Status	Complete

Database Cross-references

EMBL	AB005622;
UTR	BB200376;

Genomic Features

Chromosome	1		
Start Position	32977996	End Position	32978902
Strand	-		
IPI	IPI00218988;		
ENSEMBL	ENSG00000004455;		
UTR-Genome	5;		

Features

Key	Location	Qualifier	Value
3'utr	1..922	source	EMBL::AB005622:742..1663
		gene	AK2B
		product	adenylate kinase isozyme 2
k-box	815..822	source	EMBL::AB005622:1556..1563
		evidence	Pattern Similarity
		standard_name	K-Box (KB)
		db_xref	UTRSITE:U0023
repeat_region	212..232	source	EMBL::AB005622:953..973
		evidence	Pattern Similarity
		repeat_type	AT_rich
		repeat_family	Low_complexity
repeat_region	886..922	source	EMBL::AB005622:1627..1663
		evidence	Pattern Similarity
		repeat_type	AT_rich
		repeat_family	Low_complexity

Sequence

Characteristics	**Length:** 922 BP, **A Count:**275, **C Count:**178, **G Count:**215, **T Count:**254, **Others Count:**0
Sequence	tatcagaagg ccaggcgaga ctgcaacact gctcatcacc ccgcggcgtg atccctgctc 60 ttaggtgctg ggcagagggg aagggtggtc agggtgagga tggtgaggga gggctggtga 120 ggggctcaga ggaatacttg gaacaatagc agtgttattg tagtgtggca gtttctttta 180

Figure 1. Sample entry of UTRdb database. The Genomic Features section includes information on genome mapping coordinates and links to the related transcript and protein sequences.

(see below for details). The UTRsite collection of functional motifs has also been significantly expanded. Moreover, we have mapped human UTRs on genome assemblies, facilitating the direct comparison and integration of several annotated genomic features available through batch queries of Ensembl databases.

The integration of UTRs and protein/genomic resources is potent in that it allows the retrieval of specific UTR subsets

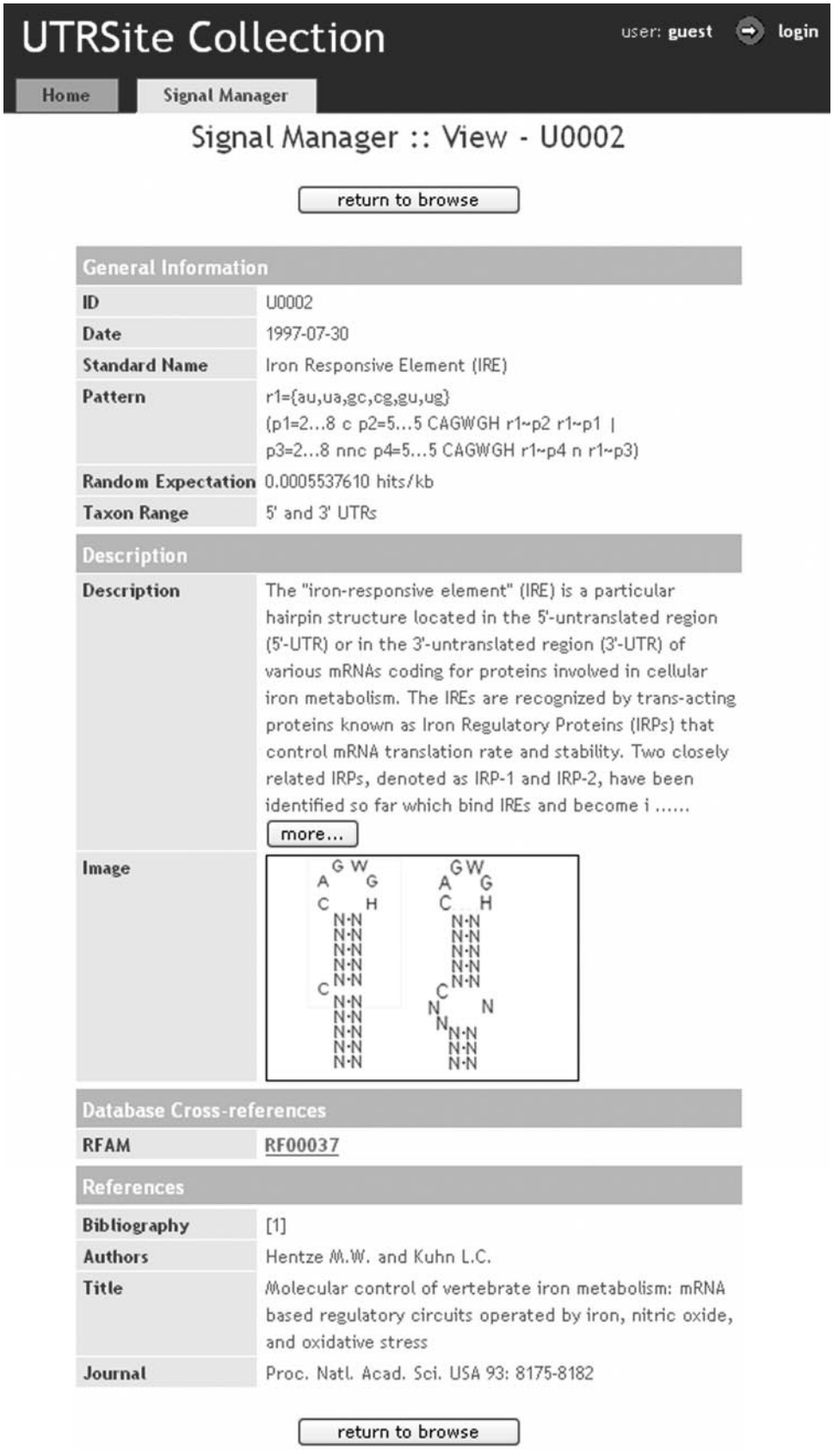

Figure 2. Sample UTRsite entry. The General Information section includes the pattern syntax of the regulatory motif in a format suitable for PatSearch analysis (9) and the number of hits/kb randomly expected in a sequence collection of the same nucleotide composition of UTRdb. The cross-link to the RFAM database (10) if available is also provided.

Figure 3. Home page of the Submission Tool for the management of UTRsite entries. The 'Guest' login only gives access to the 'View' option, whereas the 'Annotator' login allows to 'Edit/Create' UTRsite entries.

based on their genomic coordinates and/or features associated with the encoded proteins (e.g. GO terms, PFAM domains, etc.).

GENERATION OF UTRdb AND ITS INTEGRATION WITH OTHER DATABASES

UTRdb entries are automatically generated through the accurate parsing of the Feature Table of entries in primary databases (e.g. EMBL). Entry curation includes the detection of contaminating vector sequences, the removal of sequence redundancy and the annotation of repetitive elements and known regulatory motifs collected in the UTRsite database. Details of this process can be found in (4).

The current release of UTRdb contains three further specialized divisions: UTRef, UTRait and UTRexp. Sequences collected in UTRef and UTRait have been generated from the RefSeq (2) and TRAIT (3) databases, respectively. UTRexp contains UTRs that have been investigated experimentally and shown to contain functional motifs. Some of these sequences are not present in primary sequence databases and have been manually extracted from literature resources.

In the current release, we have also determined the genomic coordinates of human UTR sequences using the program BLAT (5) with the human genome assembly (Release NCBI 34). Only those UTRs that unambiguously mapped to a single genomic location were considered. Exonic structure of mapped UTRs was then refined by applying the program Spidey (6) to compare the UTR and its corresponding genomic location.

Table 1. Number of unique 5′ and 3′ UTR entries in the different UTRdb sections and of annotated UTRsite motifs (release 19.0)

	5′ UTR	3′ UTR	Total
UTRdb	139.019	159.017	298.036
UTRef	83.326	87.969	171.295
UTRait	6.290	5570	11.860
UTRexp	18	34	52
UTRgenome	18.864	26.903	45.767
Neighbor proteins	25.362	25.648	51.010
UTRsite			52

We have tried to associate each mapped UTR to the specific protein encoded by the corresponding mRNA using the relevant Ensembl coordinates. A protein was defined a 'neighbor' of a 5′ UTR if its start site corresponds to the end of the 5′ UTR sequence (and the converse for 3′ UTRs) Once the neighbor protein of a given UTR entry had been defined, we were also able to identify the Ensembl transcripts cross-referenced to the neighbor protein. If, for a given UTR entry, no annotated protein matched our criteria, we associated any Ensembl gene overlapping the same genomic region with the UTR.

The cross-referencing of UTRs and Ensembl features (protein, transcript, gene) provides a valuable resource as UTRs automatically inherit the large body of functional features annotated with the Ensembl project (7).

We have also endeavored to cross-link the UTRdb human division with IPI (International Protein Index) (8), which contains a complete non-redundant data set representing the

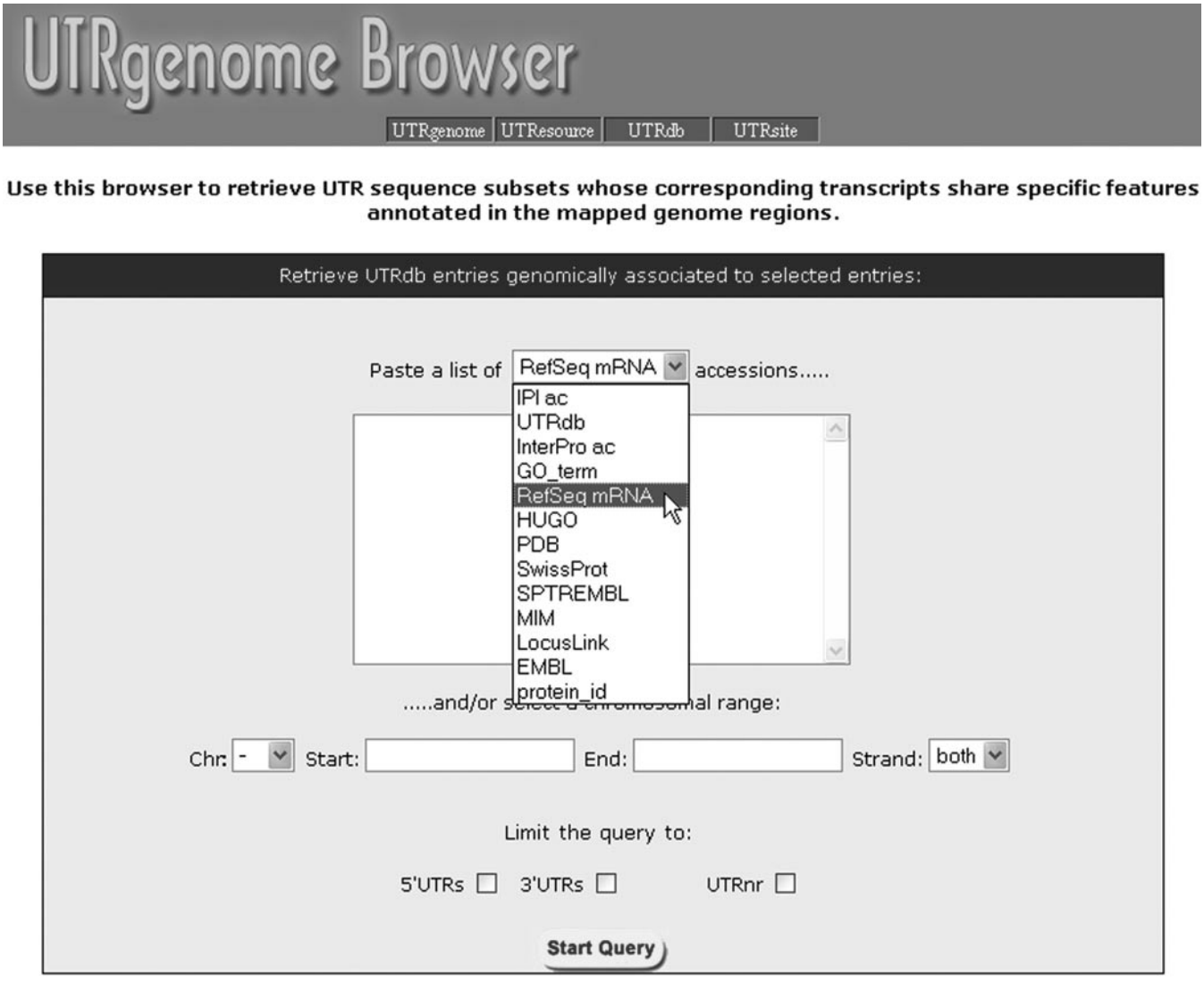

Figure 4. Home page of the UTRgenome browser for the retrieval of UTR subset based on their genomic coordinates and/or features of the protein encoded by the corresponding mRNAs.

human proteome, derived from different curated protein databases.

In future releases of UTRdb, we plan to extend the cross-referencing between UTRs and protein/genomic resources to all other organisms included in Ensembl.

UTRdb entries (see Figure 1 for an example) are annotated for the occurrence of regulatory motifs whose activity has been assessed by experimental investigation, located in the 5′ or 3′ UTR of eukaryotic mRNAs. All these motifs are collected in the UTRsite database. Each UTRsite entry (Figure 2) is prepared/reviewed/updated by expert scientists (in many cases, those who performed the experimental analysis). We have now developed a Submission Tool for the generation/management/update of UTRsite entries (Figure 3). This tool allows selected annotators, to annotate/update all the information in the entry in a user-friendly manner via a personal login.

The databases UTRdb, UTRsite and the new specific UTR collections (UTRef, UTRait, UTRexp) have been organized into MySQL relational database management system.

UTRdb CONTENT

The main section of UTRdb (Release 19) contains nine sequence collections, one for each of the eukaryotic divisions of the EMBL nucleotide database (Release 78), namely (i) human; (ii) mouse; (ii) rodent; (iv) other mammal; (v) other vertebrate; (vi) invertebrate; (vii) plant; (viii) fungi; and (ix) virus.

UTRef was generated from Reference Sequence collections (RefSeq Rel. 3). Table 1 reports a summary description of UTRdb which contains 298 036 entries and 128 286 081 nucleotides. UTRsite collects a total of 52 regulatory motifs, including upstream Open Reading Frames (uORFs) with known regulatory activity, whose occurrences have been annotated in 30 370 entries of UTRef collection.

AVAILABILITY OF UTRdb

UTRdb and UTRsite are accessible through an SRS retrieval system, which has been updated to include the new fields

added. In particular, all the information derived from genome mapping of UTRs, the IPI cross-link, and cross-referencing to 'neighbor proteins' and Ensembl genes/transcripts is available through a new field named 'Genomic Features' as reported in Figure 1. It is also possible to browse these fields by querying the SRS 'Extended query form' where relevant query fields have been added.

In addition, to access all of the information indirectly linked to UTRs, we have developed a custom browsing system—the 'UTR genome browser' (Figure 4). Through this retrieval system, it is possible to select and extract specific UTR subsets defined by accession numbers derived from a variety of databases including IPI (8), Interpro (11), GO (12), GENEW (13), MIM (14), etc. as well as by genomic coordinates.

The user can choose to download selected entries in Fasta format or to display a summary of their relevant genomic features, including genomic coordinates and cross-references to a variety of genomic resources. A graphical representation displays selected UTRs in terms of genomic coordinates and shows other human UTRs, cDNAs, proteins and ESTs in the same genome location.

With this new tool, it is now possible to obtain specific UTR subsets from mRNAs coding for proteins of a selected protein family, containing a specific domain or belonging to a specific GO class. Further investigations of such homogeneous sets of UTRs may allow the identification of common features or conserved regions whose potential functional activity may then be experimentally characterized.

Further on-line utilities are UTRscan and UTRblast. The UTRscan feature allows the enquirer to search user submitted sequences for any of the motifs collected in UTRsite. The UTRblast utility allows database searches against any of the UTRdb sections.

UTRdb, UTRsite and other related resources are publicly available at http://www.ba.itb.cnr.it/UTR/.

ACKNOWLEDGEMENTS

We thank David Horner for helpful comments on the manuscript. This work was supported by Ministero dell'Istruzione e Ricerca, Italy (projects: MIUR Cluster C03/2000-CEGB, PON 2000-2006 Progetto BIG, FIRB project 'Bioinformatica per la Genomica e la Proteomica') and Telethon. F.M. was the recipient of a EU Marie Curie fellowship at the European Bioinformatic Institute, Hinxton, UK.

REFERENCES

1. Mignone,F., Gissi,C., Liuni,S. and Pesole,G. (2002) Untranslated regions of mRNAs. *Genome Biol.*, **3**, REVIEWS0004.
2. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2003) NCBI Reference Sequence project: update and current status. *Nucleic Acids Res.*, **31**, 34–37.
3. Toppo,S., Cannata,N., Fontana,P., Romualdi,C., Laveder,P., Bertocco,E., Lanfranchi,G. and Valle,G. (2003) TRAIT (TRAnscript Integrated Table): a knowledgebase of human skeletal muscle transcripts. *Bioinformatics*, **19**, 661–662.
4. Pesole,G., Liuni,S., Grillo,G., Licciulli,F., Mignone,F., Gissi,C. and Saccone,C. (2002) UTRdb and UTRsite: specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs. Update 2002. *Nucleic Acids Res.*, **30**, 335–340.
5. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
6. Wheelan,S.J., Church,D.M. and Ostell,J.M. (2001) Spidey: a tool for mRNA-to-genomic alignments. *Genome Res.*, **11**, 1952–1957.
7. Birney,E., Andrews,T.D., Bevan,P., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cuff,J., Curwen,V., Cutts,T. *et al.* (2004) An overview of Ensembl. *Genome Res.*, **14**, 925–928.
8. Kersey,P.J., Duarte,J., Williams,A., Karavidopoulou,Y., Birney,E. and Apweiler,R. (2004) The International Protein Index: an integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.
9. Grillo,G., Licciulli,F., Liuni,S., Sbisa,E. and Pesole,G. (2003) PatSearch: a program for the detection of patterns and structural motifs in nucleotide sequences. *Nucleic Acids Res.*, **31**, 3608–3612.
10. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
11. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
12. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32** (Database issue), D258–D261.
13. Wain,H.M., Lush,M.J., Ducluzeau,F., Khodiyar,V.K. and Povey,S. (2004) Genew: the Human Gene Nomenclature Database, 2004 updates. *Nucleic Acids Res.*, **32** (Database issue), D255–D257.
14. Hamosh,A., Scott,A.F., Amberger,J., Bocchini,C., Valle,D. and McKusick,V.A. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.