D32–D37 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 10 November 2010*
doi:10.1093/nar/gkq1079

GenBank

Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell and Eric W. Sayers*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received September 17, 2010; Revised and Accepted October 14, 2010

ABSTRACT

GenBank® is a comprehensive database that contains publicly available nucleotide sequences for more than 380 000 organisms named at the genus level or lower, obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects, including whole genome shotgun (WGS) and environmental sampling projects. Most submissions are made using the web-based BankIt or standalone Sequin programs, and accession numbers are assigned by GenBank staff upon receipt. Daily data exchange with the European Nucleotide Archive (ENA) and the DNA Data Bank of Japan (DDBJ) ensures worldwide coverage. GenBank is accessible through the NCBI Entrez retrieval system that integrates data from the major DNA and protein sequence databases along with taxonomy, genome, mapping, protein structure and domain information, and the biomedical journal literature via PubMed. BLAST provides sequence similarity searches of GenBank and other sequence databases. Complete bimonthly releases and daily updates of the GenBank database are available by FTP. To access GenBank and its related retrieval and analysis services, begin at the NCBI Homepage: www.ncbi.nlm.nih.gov.

INTRODUCTION

GenBank (1) is a comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation. GenBank is built and distributed by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM), located on the campus of the US National Institutes of Health (NIH) in Bethesda, MD, USA.

NCBI builds GenBank primarily from the submission of sequence data from authors and from the bulk submission of expressed sequence tag (EST), genome survey sequence (GSS) and other high-throughput data from sequencing centers. GenBank participates with the European Nucleotide Archive (ENA) (2) and the DNA Data Bank of Japan (DDBJ) (3) as a partner in the International Nucleotide Sequence Database Collaboration (INSDC), which exchanges data daily to ensure that a uniform and comprehensive collection of sequence information is available worldwide. The US Office of Patents and Trademarks also contributes sequences from issued patents. NCBI makes the GenBank data available at no cost over the Internet, through FTP and a wide range of Web-based retrieval and analysis services (4).

RECENT DEVELOPMENTS

Updated BankIt submission tool

In the past year NCBI released a major redesign of the popular BankIt submission tool (www.ncbi.nlm.nih.gov/WebSub/?tool=genbank). The new version offers several improvements: a depositor's contact information is stored and easily reused in future submissions; sets of sequences can be uploaded as one submission; feature table data can be uploaded from a file; and a submitter can leave a partially finished submission and return later to complete it. The revised BankIt consists of a series of forms labeled by tabs at the top of the display. Once a user enters the required data on a given form, the tool proceeds to the next tab. All previously completed tabs become links to allow users to go back and change previously entered data on those tabs without losing work on the current tab.

New tools for submitters of microbial genomes

As aids to submitters of annotated microbial genomes, NCBI now offers two new verification tools: the Discrepancy Report and the Genome Submission Check Tool. The Discrepancy Report (www.ncbi.nlm.nih.gov/genbank/asndisc.html) evaluates single or multiple ASN.1 files for common errors in genome submissions, such as missing protein IDs or gene features, inconsistent locus_tag prefixes and suspect product names. This tool is

*To whom correspondence should be addressed. Tel: +1 301 496 2475; Fax: +1 301 480 9241; Email: sayers@ncbi.nlm.nih.gov

Published by Oxford University Press 2010.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

available within Sequin (with a special configuration), as an argument for *tbl2asn* and as a separate standalone program. The Genome Submission Check Tool (www.ncbi.nlm.nih.gov/genomes/frameshifts/frameshifts.cgi) performs a series of self-consistency checks on a submission file. For example, it uses BLASTp to detect neighboring protein annotations that may be parts of a single gene, reports overlapping annotations and checks for rRNA or tRNA annotations made on the wrong strand.

New facility for retrieving coding sequences

Users are now able to download the CDS regions annotated on a set of retrieved nucleotide sequences in Entrez. After performing a search in Entrez Nucleotide and viewing one or more records, the 'Send' menu allows downloading either the full record or all annotated coding sequence regions as FASTA formatted sequences. This facility works with annotated CDS features on any nucleotide record from simple open reading frames on mRNA sequences to complex multi-exon genes on mammalian genomic regions. Each downloaded CDS has its own structured title that includes a unique identifier incorporating the parent sequence accessions, gene symbol, protein product, reading frame, protein identifier and location on the parent sequence.

ORGANIZATION OF THE DATABASE

GenBank divisions

GenBank groups sequence records into various divisions based either on the source taxonomy or the sequencing strategy used to obtain the data. There are 11 taxonomic divisions (BCT, INV, MAM, PHG, PLN, PRI, ROD, SYN, UNA, VRL, VRT,) and seven high-throughput divisions (ENV, EST, GSS, HTC, HTG, STS, TSA). Finally, the PAT division contains records supplied by patent offices and the WGS division contains sequences from whole genome shotgun (WGS) projects. The size and growth of these divisions, and of GenBank as a whole, are shown in Table 1. Complete genomes (www.ncbi.nlm.nih.gov/Genomes/) continue to represent a rapidly growing segment of the database. GenBank now contains more than 1200 complete genomes from bacteria and archaea, and 20% of these were deposited during the past year. The number of eukaryote genomes with significant coverage and assembly continues to increase as well, with over 460 WGS assemblies now available.

Sequence-based taxonomy

Database sequences are classified and can be queried using a comprehensive sequence-based taxonomy (www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy) developed by NCBI in collaboration with EMBL and DDBJ and with the valuable assistance of external advisers and curators. More than 380 000 species named at the genus level or lower are represented in GenBank, and new taxa are being added at the rate of over 3800 per month. The top species in the non-WGS GenBank divisions are listed in Table 2.

Sequence identifiers and accession numbers

Each GenBank record, consisting of both a sequence and its annotations, is assigned a unique identifier called an accession number that is shared across the three collaborating databases (GenBank, DDBJ, EMBL). The accession number appears on the ACCESSION line of a GenBank record and remains constant over the lifetime of the record, even when there is a change to the sequence or annotation. Changes to the sequence data itself are tracked by an integer extension of the accession number, and this *Accession.version* identifier appears on the VERSION line of the GenBank flat file. The initial version of a sequence has the extension '.1'. In addition, each version of the DNA sequence is also assigned a unique NCBI identifier called a *GI* number that also appears on the VERSION line following the *Accession.version*:

ACCESSION AF000001
VERSION AF000001.1 GI : 987654321

When a change is made to a sequence in a GenBank record, a new GI number is issued to the updated sequence and the version extension of the *Accession.version* identifier is incremented. The accession number for the record as a whole remains unchanged, and will always retrieve the most recent version of the record; the older versions remain available under the old *Accession.version* identifiers and their original GI numbers.

A similar system tracks changes in the corresponding protein translations. These identifiers appear as qualifiers for CDS features in the FEATURES portion of a GenBank entry, e.g. /protein_id = 'AAA00001.1'. Protein sequence translations also receive their own unique GI number, which appears as a second qualifier on the CDS feature:

/db_xref = 'GI : 1233445'

BUILDING THE DATABASE

The data in GenBank and the collaborating databases, EMBL and DDBJ, are submitted either by individual authors to one of the three databases or by sequencing centers as batches of EST, STS, GSS, HTC, WGS or HTG sequences. Data are exchanged daily with DDBJ and EMBL so that the daily updates from NCBI servers incorporate the most recently available sequence data from all sources.

Direct electronic submission

Virtually all records enter GenBank as direct electronic submissions (www.ncbi.nlm.nih.gov/Genbank/), with the majority of authors using the BankIt or Sequin programs. Many journals require authors with sequence data to submit the data to a public sequence database as a condition of publication. GenBank staff can usually assign an accession number to a sequence submission within two working days of receipt, and do so at a rate of approximately 3500 per day. The accession number serves as

Table 1. Growth of GenBank Divisions (nucleotide base-pairs)

Division	Description	Release 173 (8/2009)	Release 179 (8/2010)	Increase (%)
TSA	Transcriptome shotgun data	39 829 979	398 676 845	900.9
ENV	Environmental samples	1 091 072 890	1 723 286 428	57.9
PAT	Patented sequences	5 592 927 651	8 519 294 473	52.3
BCT	Bacteria	4 107 328 206	5 333 010 385	29.8
VRL	Viruses	779 481 462	970 125 245	24.5
PHG	Phages	36 100 172	43 456 808	20.4
MAM	Other mammals	576 977 646	679 274 390	17.7
INV	Invertebrates	1 734 996 371	2 036 240 836	17.4
WGS	WGS data	148 165 117 763	169 253 846 128	14.2
GSS	Genome survey sequences	16 738 219 857	18 442 479 673	10.2
PLN	Plants	3 695 552 256	4 038 424 961	9.3
SYN	Synthetic	131 361 806	142 548 355	8.5
VRT	Other vertebrates	2 366 300 257	2 533 789 261	7.1
EST	ESTs	34 522 977 161	36 803 930 321	6.6
HTC	High-throughput cDNA	636 472 189	659 355 057	3.6
PRI	Primates	5 751 413 009	5 943 029 356	3.3
ROD	Rodents	4 206 718 960	4 298 354 944	2.2
HTG	High-throughput genomic	23 895 733 886	24 276 862 305	1.6
UNA	Unannotated	119 348	120 289	0.8
STS	Sequence tagged sites	629 573 650	634 263 196	0.7
TOTAL	All GenBank sequences	254 698 274 519	286 730 369 256	12.6

Table 2. Top organisms in GenBank (Release 179)

Organism	Non-WGS base pairs
Homo sapiens	14 792 487 417
Mus musculus	8 859 010 528
Rattus norvegicus	6 443 768 086
Bos taurus	5 361 712 195
Zea mays	5 037 629 354
Sus scrofa	4 783 381 701
Danio rerio	3 137 945 523
Strongylocentrotus purpuratus	1 352 920 226
Oryza sativa Japonica Group	1 197 245 122
Nicotiana tabacum	1 187 388 273
Xenopus (Silurana) tropicalis	1 147 132 278
Drosophila melanogaster	1 047 707 620
Pan troglodytes	1 001 926 471
Arabidopsis thaliana	1 001 073 627
Canis lupus familiaris	943 043 649
Vitis vinifera	913 911 649
Gallus gallus	891 463 513
Glycine max	886 103 518
Macaca mulatta	821 393 285
Ciona intestinalis	748 350 657

confirmation that the sequence has been submitted and provides a means for readers of articles in which the sequence is cited to retrieve the data. Direct submissions receive a quality assurance review that includes checks for vector contamination, proper translation of coding regions, correct taxonomy and correct bibliographic citations. A draft of the GenBank record is passed back to the author for review before it enters the database.

Authors may ask that their sequences be kept confidential until the time of publication. Since GenBank policy requires that the deposited sequence data be made public when the sequence or accession number is published, authors are instructed to inform GenBank staff of the publication date of the article in which the sequence is cited in order to ensure a timely release of the data. Although only the submitter is permitted to modify sequence data or annotations, all users are encouraged to report lags in releasing data or possible errors or omissions to GenBank at update@ncbi.nlm.nih.gov.

NCBI works closely with sequencing centers to ensure timely incorporation of bulk data into GenBank for public release. GenBank offers special batch procedures for large-scale sequencing groups to facilitate data submission, including the program *tbl2asn*, described at www.ncbi.nlm.nih.gov/Sequin/table.html.

Submission using BankIt. About a third of author submissions are received through an NCBI Web-based data submission tool named BankIt (see 'Recent Developments' section). Using BankIt, authors enter sequence information directly into a form and add biological annotation such as coding regions or mRNA features. Text boxes and pull-down menus allow the submitter to describe the sequence further without having to learn formatting rules or controlled vocabularies. Additionally, BankIt now allows submitters to upload source and annotation using tab-delimited tables. Before creating a draft record in the GenBank flat file format for the submitter to review, BankIt validates the submissions by flagging many common errors and checking for vector contamination using a variant of BLAST called Vecscreen.

Submission using Sequin and tbl2asn. NCBI also offers a standalone multi-platform submission program called Sequin (www.ncbi.nlm.nih.gov/projects/Sequin/) that can be used interactively with other NCBI sequence retrieval and analysis tools. Sequin handles simple sequences, such as a single cDNA, as well as segmented entries, phylogenetic studies, population studies, mutation studies, environmental samples and alignments. Sequin has convenient editing and complex annotation capabilities and contains a number of built-in validation functions for quality assurance. In addition, Sequin is able to accommodate large

sequences, such as the 5.6 Mb *Escherichia coli* genome, and read in a full complement of annotations from simple tables. The most recent version, Sequin 10.0, was released in April 2010 and is available for Macintosh, PC and Unix computers via anonymous FTP at ftp.ncbi.nih.gov/sequin. Once a submission is completed, submitters can e-mail the Sequin file to gb-sub@ncbi.nlm.nih.gov. Submitters of large, heavily annotated genomes may find it convenient to use *tbl2asn* to convert a table of annotations generated from an annotation pipeline into an ASN.1 (Abstract Syntax Notation One) record suitable for submission to GenBank.

Submission of Barcode sequences. The Consortium for the Barcode of Life (CBOL) is an international initiative to develop DNA barcoding as a tool for characterizing species of organisms using a short DNA sequence. For animal species, a 648-bp fragment of the gene for cytochrome oxidase subunit I is used as the barcode. The plant and fungal communities are investigating other loci. NCBI, in collaboration with CBOL (www.barcoding.si.edu/) has created an online tool (BarSTool) for the bulk submission of barcode sequences to GenBank (www.ncbi.nlm.nih.gov/WebSub/?tool=barcode) that allows users to upload files containing a batch of sequences with associated source information. The Nucleotide query 'barcode[keyword]' retrieves the almost 200 000 barcode sequences in GenBank, over 160 000 of which were added in the last year.

Notes on particular divisions

Transcriptome Shotgun Assembly (TSA) sequences. The TSA division contains transcriptome shotgun assembly sequences that are assembled from sequences deposited in the NCBI Trace Archive, the Sequence Read Archive (SRA) and the EST division of GenBank. The TSA division has grown dramatically in the past year (Table 1) in response to the over 40 Terabasepairs of data deposited into SRA in the same period from next-generation sequencing technologies, including those from Roche-454 Life Sciences, Illumina Solexa and Applied Biosystems SOLiD. Neither the Trace Archive nor SRA is a part of GenBank and are described elsewhere (4,5). TSA records (e.g. EZ000001) have 'TSA' as their keyword and a Primary block that provides the base ranges and identifiers of the sequences used in the TSA assembly.

Environmental sample sequences (ENV). The ENV division of GenBank accommodates non-WGS sequences obtained via environmental sampling methods in which the source organism is unknown. Many ENV sequences arise from metagenome samples derived from microbiota in various animal tissues, such as within the gut or skin, or from particular environments, such as freshwater sediment, hot springs or areas of mine drainage. Records in the ENV division contain 'ENV' in the keyword field and use an '/environmental_sample' qualifier in the source feature.

Whole genome shotgun sequences. WGS sequences appear in GenBank as sets of WGS sequence overlap contigs, each of which is issued an accession number consisting of a four-letter project ID, followed by a two-digit version number and a six-digit contig ID. Hence, the WGS accession number 'AAAA01072744' is assigned to contig number '072744' of the first version of project 'AAAA'. WGS sequencing projects have contributed over 64 million contigs to GenBank, and these primary sequences have been used to construct 8 million large-scale assemblies of scaffolds and chromosomes. For a complete list of WGS projects with links to the data, see www.ncbi.nlm.nih.gov/Traces/wgs/.

Although WGS project sequences may be annotated, many low-coverage genome projects do not contain annotation. Because these sequence projects are ongoing and incomplete, these annotations may not be tracked from one assembly version to the next and should be considered preliminary. Submitters of genomic sequences, including WGS sequences, are urged to use evidence tags of the form '/experimental=*text*' and '/inference=*TYPE*:*text*', where *TYPE* is one of a number of standard inference types and *text* consists of structured text.

ESTs. ESTs continue to be a major source of data for gene expression and annotation studies, and at almost 37 billion base pairs, it remains the largest non-WGS division in GenBank. EST data are available for download from ftp.ncbi.nih.gov/repository/dbEST/ (6) as well as from the GenBank FTP site. The data in dbEST are clustered using the BLAST programs to produce the UniGene database (www.ncbi.nlm.nih.gov/unigene) of more than 4.3 million gene-oriented sequence clusters representing over 120 organisms (4).

High-throughput genomic (HTG) and high-throughput cDNA (HTC) sequences. The HTG division of GenBank (www.ncbi.nlm.nih.gov/HTGS/) contains unfinished large-scale genomic records, which are in transition to a finished state (7). These records are designated as belonging to Phases 0–3 depending on the quality of the data, with Phase 3 being the finished state. Upon reaching Phase 3, HTG records are moved into the appropriate organism division of GenBank.

The HTC division of GenBank contains high-throughput cDNA sequences that are of draft quality but may contain 5′ UTRs, 3′ UTRs, partial coding regions and introns. HTC sequences which are finished and of high quality are moved to the appropriate organism division of GenBank. A project generating HTC data is described in ref. (8).

Special record types

Third party annotation. Third Party Annotation (TPA) records are sequence annotations published by someone other than the original submitter of the primary sequence record in DDBJ/EMBL/GenBank (www.ncbi.nlm.nih.gov/Genbank/TPA.html). TPA records fall into one of three categories: *experimental*, in which case there is direct experimental evidence for the existence of the annotated molecule; *inferential*, in which case the

experimental evidence is indirect; and *reassembly*, where the focus is on providing a better assembly of the raw reads. TPA sequences may be created by assembling a number of primary sequences. The format of a TPA record (e.g. BK000016) is similar to that of a conventional GenBank record but includes the label 'TPA_exp:', 'TPA_inf:' or 'TPA_reasm:' at the beginning of each Definition Line as well as corresponding keywords. TPA experimental and inferential records also contain a Primary block similar to that in a TSA record. Currently GenBank contains over 5.6 million TPA records, >99% of which are derived from a recent submission of an individual human genome. TPA sequences are not released to the public until their accession numbers or sequence data and annotation appear in a peer-reviewed biological journal. TPA submissions to GenBank may be made using either BankIt or Sequin.

Contig (CON) records for assemblies of smaller records. Small genomes, such as those from bacteria, can generally be conveniently represented and analyzed as single sequences. For very long sequences, such as a eukaryotic chromosome, where the sequence is not complete but consists of several contig records with uncharacterized gaps between them, the entire chromosome is represented in GenBank as a CON record. Rather than listing the sequence itself, CON records contain assembly instructions involving the several component sequences. An example of such a CON record is DP000010 for rice chromosome 11.

RETRIEVING GENBANK DATA

The Entrez system

The sequence records in GenBank are accessible through the NCBI Entrez retrieval system (4). Records from the EST and GSS divisions of GenBank are stored in the Entrez EST and GSS databases, while all other GenBank records are stored in Entrez Nucleotide. GenBank sequences that are part of population or phylogenetic studies are collected together in Entrez PopSet, and conceptual translations of CDS sequences annotated on GenBank records are available in Entrez Protein. Each of these databases is linked to the scientific literature via PubMed and PubMed Central. Additional information about conducting Entrez searches is found in the NCBI Help Manual (www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helpsequences) and links to related tutorials are provided on the NCBI Education page (www.ncbi.nlm.nih.gov/Education/).

Associating sequence records with sequencing projects

The ability to identify all GenBank records submitted by a specific group or those with a particular focus, such as metagenomic surveys, is essential for the analysis of large volumes of sequence data. The use of organism or submitter names as a means to define such a set of sequences is unreliable. The Genome Projects database (www.ncbi.nlm.nih.gov/genomeprj), developed at NCBI and subsequently adopted across the INSDC, allows sequencing centers to register projects under a unique project identifier, enabling reliable linkage between sequencing projects and the data they produce. Genome Projects, soon to be renamed BioProjects, is currently expanding to include a wider variety of projects, such as metagenome, environmental sampling, comparative genomics and transcriptome projects as well as projects focused on a particular locus, such as 16S ribosomal RNA, or a viral disease, such as SARS. A 'DBLINK' line appearing in GenBank flat files identifies the sequencing projects with which a GenBank sequence record is associated and, as of GenBank release 172, replaces the earlier 'PROJECT' line. As an example, the DBLINK line below associates a GenBank sequence record with Project record 18787.

DBLINK Project:18787

Project record 18787 provides details of the progress made in the effort to sequence the green anole, *Anolis carolinensis* (www.broad.mit.edu/models/anole/). Within the Entrez system, such a sequence record is linked directly to the appropriate Genome Project record; these links are bidirectional, so that the Project records also link back to associated sequence records.

BLAST sequence-similarity searching

Sequence-similarity searches are the most fundamental and frequent type of analysis performed on GenBank data. NCBI offers the BLAST family of programs (blast.ncbi.nlm.nih.gov) to detect similarities between a query sequence and database sequences (9,10). BLAST searches may be performed on the NCBI Web site (11) or by using a set of standalone programs distributed by FTP (4).

Obtaining GenBank by FTP

NCBI distributes GenBank releases in the traditional flat file format as well as in the ASN.1 format used for internal maintenance. The full bimonthly GenBank release along with the daily updates, which incorporate sequence data from EMBL and DDBJ, is available by anonymous FTP from NCBI at ftp.ncbi.nih.gov/genbank. The full release in flat file format is available as a set of compressed files with a non-cumulative set of updates at ftp.ncbi.nih.gov/daily-nc/. For convenience in file transfer, the data are partitioned into multiple files; for release 179 there are 1443 files requiring 484 GB of uncompressed disk storage. A script is provided in ftp.ncbi.nih.gov/tools/ to convert a set of daily updates into a cumulative update.

MAILING ADDRESS

GenBank, National Center for Biotechnology Information, Building 45, Room 6AN12D-37, 45 Center Drive, Bethesda, MD 20892, USA. Tel: +1 301 496 2475; Fax: +1 301 480 9241.

ELECTRONIC ADDRESSES

www.ncbi.nlm.nih.gov - NCBI Home Page.

gb-sub@ncbi.nlm.nih.gov - Submission of sequence data to GenBank.

update@ncbi.nlm.nih.gov - Revisions to, or notification of release of, 'confidential' GenBank entries.

info@ncbi.nlm.nih.gov - General information about NCBI resources.

CITING GENBANK

If you use the GenBank database in your published research, we ask that this article be cited.

FUNDING

Funding for open access charge: Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2010) GenBank. *Nucleic Acids Res.*, **38**, D46–D51.
2. Leinonen,R., Akhtar,R., Birney,E., Bonfield,J., Bower,L., Corbett,M., Cheng,Y., Demiralp,F., Faruque,N., Goodgame,N. *et al.* (2010) Improvements to services at the European nucleotide archive. *Nucleic Acids Res.*, **38**, D39–D45.
3. Kaminuma,E., Mashima,J., Kodama,Y., Gojobori,T., Ogasawara,O., Okubo,K., Takagi,T. and Nakamura,Y. (2010) DDBJ launches a new archive database with analytical tools for next-generation sequence data. *Nucleic Acids Res.*, **38**, D33–D38.
4. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Federhen,S. *et al.* (2011) Database resources at the national center for biotechnology information. *Nucleic Acids Res.*, this issue.
5. Leinonen,R., Sugawara,H. and Shumway,M. (2011) The sequence read archive. *Nucleic Acids Res.* (in press).
6. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST–database for "expressed sequence tags". *Nat. Genet.*, **4**, 332–333.
7. Kans,J.A. and Ouellette,B.F.F. (2001) In Baxevanis,A.D. and Ouellette,B.F.F. (eds), *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*. John Wiley and Sons, Inc., NY, pp. 65–81.
8. Kawai,J., Shinagawa,A., Shibata,K., Yoshino,M., Itoh,M., Ishii,Y., Arakawa,T., Hara,A., Fukunishi,Y., Konno,H. *et al.* (2001) Functional annotation of a full-length mouse cDNA collection. *Nature*, **409**, 685–690.
9. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
10. Zhang,Z., Schaffer,A.A., Miller,W., Madden,T.L., Lipman,D.J., Koonin,E.V. and Altschul,S.F. (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.*, **26**, 3986–3990.
11. Johnson,M., Zaretskaya,I., Raytselis,Y., Merezhuk,Y., McGinnis,S. and Madden,T.L. (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res.*, **36**, W5–W9.