D26–D31 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 21 October 2008*
doi:10.1093/nar/gkn723

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, USA

Received September 16, 2008; Accepted September 30, 2008

ABSTRACT

GenBank® is a comprehensive database that contains publicly available nucleotide sequences for more than 300 000 organisms named at the genus level or lower, obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing 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 Molecular Biology Laboratory Nucleotide Sequence Database in Europe and the DNA Data Bank of Japan ensures worldwide coverage. GenBank is accessible through the National Center for Biotechnology Information (NCBI) Entrez retrieval system, which 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. (If you use the GenBank database in your published research, we ask that this article be cited.)

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. The US Office of Patents and Trademarks also contributes sequences from issued patents. GenBank participates with the European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL) (2) and the DNA Databank 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. 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).

ORGANIZATION OF THE DATABASE

From its inception, GenBank has grown exponentially, and continues to do so with a current doubling time of ~30 months. The traditional GenBank divisions contain over 95 billion nucleotide bases from more than 92 million individual sequences, with 16 million new sequences added in the past year. Contributions from Whole Genome Shotgun (WGS) projects supplement the data in the traditional divisions to bring the total to 213 billion bases. Complete genomes (www.ncbi.nlm.nih.gov/Genomes/index.html) continue to represent a rapidly growing segment of the database, with some 170 of the more than 740 complete microbial genomes in GenBank deposited over the past year. The number of eukaryote genomes with significant coverage and assembly continues to increase as well, with over 270 assemblies now available, including that of the reference human genome.

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

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

the valuable assistance of external advisers and curators. More than 300 000 species named at the genus level or lower are represented in GenBank, and new taxa are being added at the rate of over 2200 per month. About 12% of the sequences in GenBank are of human origin and 9% of all sequences are human ESTs. The top species in GenBank in terms of number of bases are *Homo sapiens* (13.1 billion bases), *Mus musculus* (8.4 billion), *Rattus norvegicus* (6.0 billion), *Bos taurus* (5.2 billion), *Zea mays* (4.6 billion), *Sus scrofa* (3.1 billion), *Danio rerio* (2.9 billion), *Oryza sativa* (1.5 billion), *Strongylocentrotus purpuratus* (1.4 billion), *Nicotania tabacum* (1.1 billion) and *Xenopus tropicalis* (1.0 billion).

GenBank records and divisions

Each GenBank entry includes a concise description of the sequence, the scientific name and taxonomy of the source organism, bibliographic references and a table of features (www.ncbi.nlm.nih.gov/collab/FT/index.html) listing areas of biological significance, such as coding regions and their protein translations, transcription units, repeat regions and sites of mutations or modifications. The files in the GenBank distribution have traditionally been partitioned into 'divisions' that roughly correspond to taxonomic groups, such as bacteria (BCT), viruses (VRL), primates (PRI) and rodents (ROD). In recent years, divisions have been added to support specific sequencing strategies. These include divisions for EST, GSS, high-throughput genomic (HTG), high-throughput cDNA (HTC) and environmental sample (ENV) sequences, making a total of 20 divisions. The newest division, Transcriptome Shotgun Assemblies (TSAs), was added in the past year and is described below. For convenience in file transfer, the GenBank data are partitioned into multiple files, currently more than 1600, for the bimonthly GenBank releases on the NCBI FTP site.

Expressed sequence tags. ESTs continue to be a major source of new sequence records and gene sequences, comprising over 30 billion nucleotide bases in GenBank release 167. Over the past year, the number of ESTs has increased by over 20% to a total of 54.8 million sequences representing more than 1640 different organisms. The top organisms represented in the EST division are *H. sapiens* (8.1 million records), *M. musculus* (4.9 million), *S. scrofa* (2.2 million), *Arabidopsis thaliana* (1.5 million), *B. taurus* (1.5 million), *Z. mays* (1.5 million) and *D. rerio* (1.4 million). As part of its daily processing of GenBank EST data, NCBI identifies through BLAST searches all homologies for new EST sequences and incorporates that information into the companion database, dbEST (www.ncbi.nlm.nih.gov/dbEST/index.html) (5). The data in dbEST are processed further to produce the UniGene database (www.ncbi.nlm.nih.gov/sites/entrez?db=unigene) of more than 3.4 million gene-oriented sequence clusters representing over 85 organisms (4).

Sequence-tagged sites, GSSs and ENV. The sequence-tagged site (STS) division of GenBank (www.ncbi.nlm.nih.gov/dbSTS/index.html) contains over 937 000 sequences, including anonymous STSs based on genomic sequence as well as gene-based STSs derived from the 3′-ends of genes and ESTs. These STS records usually include mapping information.

The GSS division of GenBank (www.ncbi.nlm.nih.gov/dbGSS/index.html) has grown over the past year by 13% to a total of 24.4 million records for over 800 organisms and contributes over 15.8 billion nucleotide bases. GSS sequences are the products of as many as 80 different experimental techniques, including metagenomic surveys of sequences arising from biological communities. However, more than one-quarter of all GSS records are single reads from bacterial artificial chromosomes ('BAC-ends') used in a variety of genome sequencing projects. The most highly represented species in the GSS division, including metagenomic surveys, are marine metagenome (2.6 million records), *M. musculus* (2.2 million), *Z. mays* (2.1 million) and *H. sapiens* (1.2 million). The human data have been used (www.ncbi.nlm.nih.gov/projects/genome/clone/) along with the STS records in tiling the BACs for the Human Genome Project (6).

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 various animal tissues, such as 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. As of GenBank release 167, the ENV division of GenBank contained over 860 000 sequences and 668 million base pairs.

HTG and 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 Phase 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. As of GenBank release 167, the HTG division contained 22.5 billion base pairs of sequence, an increase of more than 4 billion bases over the past year.

The HTC division of GenBank accommodates HTC sequences, which are of draft quality but may contain 5′-UTRs and 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. GenBank release 167 contained more than 520 000 HTC sequences totaling nearly 600 million bases. A project generating HTC data is described in Ref. (8).

WGS sequences. More than 118 billion bases of WGS sequence appear in GenBank as sets of WGS contigs, many of them bearing annotations originating from a single sequencing project. These sequences are issued accession numbers 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 'AAAA 01072744' is assigned to contig number '072744' of the

first version of the project 'AAAA'. WGS sequencing projects have contributed some 40 million contigs to GenBank, a 60% increase over last year's total. These primary sequences have been used to construct 5.6 million large-scale assemblies of scaffolds and chromosomes. WGS project contigs for *H. sapiens, Pan trodlodytes, Macacca mulatta, Equus caballus, Canis familiaris, Drosophila, Saccharomyces* and 800 other organisms and environmental samples are available. For a complete list of WGS projects with links to the data, see www.ncbi.nlm.nih.gov/projects/WGS/WGSprojectlist.cgi.

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 the number of standard inference types and 'text' is made up of structured text. These new qualifiers replace 'evidence = experimental' and 'evidence = non-experimental', respectively, which are no longer supported.

TSA sequences. In recent years a growing number of sequencing traces have been deposited in the NCBI Trace Archive (TA), which now contains almost 2 billion records (4). Given the advent of next-generation sequencing technologies, including those from Roche-454 Life Sciences, Illumina Solexa and Applied Biosystems SOLiD, NCBI deployed a Short Read Archive (SRA) in 2007. Neither of these archives is a part of GenBank, but beginning with release 166, GenBank added a new TSA division for TSA sequences, which are shotgun assemblies of sequences deposited in TA, SRA and the EST division of GenBank. 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.

Special record types

Third party annotation. Third party annotation (TPA) records support the reporting of published sequence annotations 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 two categories: *experimental*, in which case there is direct experimental evidence for the existence of the annotated molecule, and *inferential*, in which case the experimental evidence is indirect. 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:' at the beginning of each definition line and the keywords 'Third Party Annotation; TPA'. A TPA record also contains a primary block similar to that in a TSA record. Currently GenBank contains over 5900 TPA records, including 2170 for *Drosophila melanogaster*, 970 for *H. sapiens*, 640 for *O. sativa* and 300 for *M. musculus*. 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.

GenBank CON records for assemblies of smaller records. Small genomes, such as those from bacteria, can generally be conveniently represented, transferred between computers 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.

BUILDING THE DATABASE

The data in GenBank and the collaborating databases, EMBL and DDBJ, are submitted primarily 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/index.html), 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 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 almost 1600 per day. The accession number serves as 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 (www.ncbi.nlm.nih.gov/BankIt). Using BankIt, authors enter sequence information directly into a form and add biological annotation, such as coding regions or mRNA features. Free-form text boxes, list boxes and pull-down menus allow the submitter to describe the sequence further without having to learn formatting rules or restricted vocabularies. 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. BankIt is the tool of choice for simple submissions, especially when only one or a small number of records is being submitted (7). BankIt can also be used by submitters to update their existing GenBank records. NCBI plans to release a new version of BankIt in 2009 that streamlines the submission process and allows for easier submission of sets of related sequences.

Submission using Sequin and tbl2asn. NCBI also offers a standalone multi-platform submission program called Sequin (www.ncbi.nlm.nih.gov/Sequin/index.html) 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 for which BankIt and other web-based submission tools are not well suited. 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. Versions for Macintosh, PC and Unix computers are available 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* (described above) 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, usually a 648-bp fragment of the gene for cytochrome oxidase subunit I. NCBI, in collaboration with CBOL (www.barcoding.si.edu/index.htm) has created an online tool for the bulk submission of barcode sequences to GenBank (www.ncbi.nlm.nih.gov/BankIt/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 more than 15 000 barcode sequences in GenBank, over 11 000 of which were added in the last year.

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 and 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'

Ensuring stable access to sequence data

A convenient way to share data among a set of collaborators is to post the data to a locally maintained web site. However, if original data and updates are not simultaneously submitted to a central repository, significant problems can arise.

The access lifetime of the data may be reduced. The ephemeral nature of much of the content on the web is part of the common experience. In one attempt to quantify content lifetime, 360 randomly selected web pages were tracked for a period of 4 years, and a half-life of only 2 years was measured for the set (9). While a well-maintained web page can certainly persist for longer than 2 years, the relatively short half-life reported for this set of pages is worth noting.

The full biological context of the data may not be realized. Even during the accessible lifetime of locally

posted sequence data, the full biological context of a sequence may not be fully understood. This is particularly true if the sequence cannot be conveniently compared with others, perhaps derived from distantly related organisms that are beyond the scope of the host web page.

Existing data in heavily used, centralized databases will become outdated. If updates to sequences contained within centralized databases are made to a local page, but not made to corresponding records in a central database, the newer data will not reach the wider research community and much of its impact will be lost.

Submission of sequence data to a centralized repository solves these problems. Centralized databases, such as GenBank and the other members of the INSDC, ensure stable access to sequence data by providing versioned releases via FTP, uniform data sets via web interfaces and archival redundancy. Combining new data with that of other researchers worldwide within a central database provides a broad biological context that stimulates discovery. Moreover, the process of keeping each sequence up–to-date enhances the usefulness of all the sequences in the database.

RETRIEVING GENBANK DATA

The Entrez system

The sequence records in GenBank are accessible through the NCBI Entrez retrieval system (www.ncbi.nlm.nih.gov/sites/gquery) that covers 35 biological databases. 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. Other Entrez databases contain protein sequences derived from GenBank and other sources, genome maps, population, phylogenetic and environmental sequence sets, gene expression data, the NCBI taxonomy, protein domain information and protein structures from the Molecular Modeling Database, MMDB (10). Each database is linked to the scientific literature via PubMed and PubMed Central.

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 Project Database (www.ncbi.nlm.nih.gov/sites/entrez?db=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.

A new 'PROJECT' line appearing in GenBank flat files identifies the sequencing projects with which a GenBank sequence record is associated. The PROJECT line may contain multiple identifiers of the form 'type' and 'value', respectively, separated by a semicolon. As an example, the PROJECT line below associates a GenBank sequence record with Genome Project record '18787'.

PROJECT GenomeProject: 18787

Genome 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 Genome 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 the GenBank data. NCBI offers the BLAST family of programs (blast.ncbi.nlm.nih.gov/) to detect similarities between a query sequence and database sequences (11,12). BLAST searches may be performed on the NCBI web site (13) 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, are available by anonymous FTP from NCBI at ftp.ncbi.nih.gov/genbank as well as from a mirror site at the University of Indiana (ftp://bio-mirror.net/biomirror/genbank/). The full release in flat file format is available as a set of compressed files with a noncumulative set of updates at ftp.ncbi.nih.gov/daily-nc/. Uncompressed, a complete copy of release 167 occupies almost 360 GB. A script is provided in ftp.ncbi.nih.gov/tools/ to convert a set of daily updates into a cumulative update.

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 Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.
2. Kulikova,T., Akhtar,R., Aldebert,P., Althorpe,N., Andersson,M., Baldwin,A., Bates,K., Bhattacharyya,S., Bower,L., Browne,P. *et al.* (2007) EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res.*, **35**, D16–D20.
3. Sugawara,H., Ogasawara,O., Okubo,K., Gojobori,T. and Tateno,Y. (2008) DDBJ with new system and face. *Nucleic Acids Res.*, **36**, D22–D24.
4. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, in press.
5. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST–database for 'expressed sequence tags'. *Nat. Genet.*, **4**, 332–333.

6. Smith,M.W., Holmsen,A.L., Wei,Y.H., Peterson,M. and Evans,G.A. (1994) Genomic sequence sampling: a strategy for high resolution sequence-based physical mapping of complex genomes. *Nat. Genet.*, **7**, 40–47.
7. Kans,J.A. and Ouellette,B.F.F. (2001) Submitting DNA sequences to the databases. 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., New York, 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. Koehler,W. (2002) Web page change and persistence - a four-year longitudinal study. *J. Am. Soc. Inf. Sci. Technol.*, **53**, 162–171.
10. Wang,Y., Addess,K.J., Chen,J., Geer,L.Y., He,J., He,S., Lu,S., Madej,T., Marchler-Bauer,A., Thiessen,P.A. *et al.* (2007) MMDB: annotating protein sequences with Entrez's 3D-structure database. *Nucleic Acids Res.*, **35**, D298–D300.
11. 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.
12. 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.
13. Ye,J., McGinnis,S. and Madden,T.L. (2006) BLAST: improvements for better sequence analysis. *Nucleic Acids Res.*, **34**, W6–W9.