D52–D56 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 1 November 2009*
doi:10.1093/nar/gkp960

Non-redundant patent sequence databases with value-added annotations at two levels

Weizhong Li[1], Hamish McWilliam[1], Ana Richart de la Torre[2], Adam Grodowski[2], Irina Benediktovich[2], Mickael Goujon[1], Stephane Nauche[2] and Rodrigo Lopez[1,*]

[1]European Bioinformatics Institute, EMBL Outstation, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK and [2]European Patent Office, IQ Life Sciences, Patentlaan 3-9, 2288 EE Rijswijk, The Netherlands

Received August 25, 2009; Revised September 22, 2009; Accepted October 13, 2009

ABSTRACT

The European Bioinformatics Institute (EMBL-EBI) provides public access to patent data, including abstracts, chemical compounds and sequences. Sequences can appear multiple times due to the filing of the same invention with multiple patent offices, or the use of the same sequence by different inventors in different contexts. Information relating to the source invention may be incomplete, and biological information available in patent documents elsewhere may not be reflected in the annotation of the sequence. Search and analysis of these data have become increasingly challenging for both the scientific and intellectual-property communities. Here, we report a collection of non-redundant patent sequence databases, which cover the EMBL-Bank nucleotides patent class and the patent protein databases and contain value-added annotations from patent documents. The databases were created at two levels by the use of sequence MD5 checksums. Sequences within a level-1 cluster are 100% identical over their whole length. Level-2 clusters were defined by sub-grouping level-1 clusters based on patent family information. Value-added annotations, such as publication number corrections, earliest publication dates and feature collations, significantly enhance the quality of the data, allowing for better tracking and cross-referencing. The databases are available format: http://www.ebi.ac.uk/patentdata/nr/.

INTRODUCTION

Patents in the biotechnology domain cover a wide range of areas, including health (e.g. vaccines, antibodies and diagnostics), industrial microbiology (e.g. genetically modified microbes) and agriculture (e.g. GMO and cultivars). Thus, the patent data are a valuable resource, not only for the intellectual-property world but also for the scientific community. Information in patent data can be more detailed (1), appears earlier or is not available in the scientific literature (2). The European Bioinformatics Institute (EMBL-EBI) provides public access to patent data resources, including abstracts, chemical compounds and sequences (http://www.ebi.ac.uk/patentdata/). Patent abstracts contains abstracts of biology-related patent applications derived from data products of the European Patent Office (EPO). Chemical compounds appearing in patents are available in ChEBI (3), a dictionary of molecular entities focused on small chemical compounds.

The sequences appearing in patent applications are an important resource for patent-related searches. During the past decade, the number of biological sequences appearing in patent applications has been increasing dramatically (Figure 1). Today, millions of nucleotide and protein sequences extracted from the patent documents are available from both the commercial sector and the public domain. Proprietary efforts include GENESEQ™ (Thomson Reuters) http://thomsonreuters.com/products_services/science/science_products/life_sciences/biology/geneseq), GQ-PAT (GenomeQuest, Inc.; http://www.genomequest.com), CAS REGISTRY (Chemical Abstracts Service; http://www.cas.org), PCTGEN (FIZ Karlsruhe; http://www.fiz-karlsruhe.de/sci_tech_patent_information.html) and USGENE (SequenceBase Corporation; http://www.sequencebase.com/) (4) and the major public databases represented by the International Nucleotide Sequence Database Collaboration (INSDC; http://www.insdc.org/) member databases: EMBL-Bank (5), GenBank (6) and DDBJ (7). These include data provided by the EPO, Japan Patent Office (JPO), Korean Intellectual Property Office (KIPO) and United States Patent and Trademark Office (USPTO). EMBL-Bank has a specific data class (PAT) for nucleotide sequences obtained from patents. The EMBL-EBI also collates protein sequences provided by the EPO, JPO,

*To whom correspondence should be addressed. Tel: +44 1223 494423; Fax: +44 1223 494468; Email: rls@ebi.ac.uk.

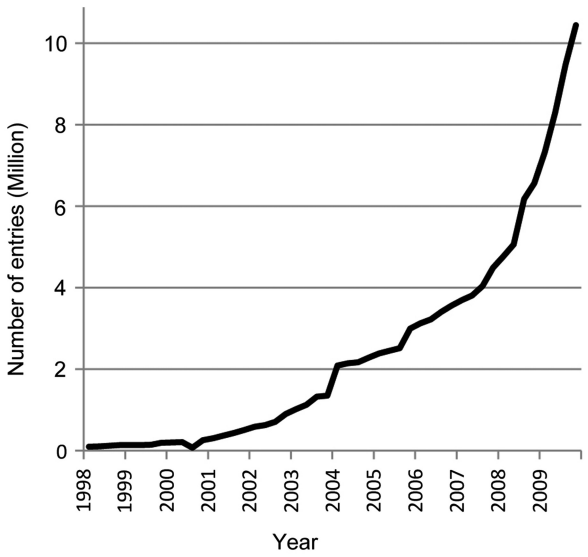

Figure 1. Data growth of EMBL-Bank patent class. The curve indicates the number of entries in the EMBL-Bank patent class has increased dramatically during the past decade.

KIPO and USPTO, into the Patent Proteins data set available from the EMBL-EBI ftp server (ftp://ftp.ebi.ac.uk/pub/databases/embl/patent/) and via the SRS server (http://srs.ebi.ac.uk/). FASTA format files for sequence searching are also available format: ftp://ftp.ebi.ac.uk/pub/databases/fastafiles/patent/.

Searching patent sequence databases can be used as inspiration for scientific innovation and discovery of existing inventions (e.g. industrial processes) with relevance to the work of the researcher. However, as mentioned earlier, sequences may appear multiple times due to the same invention being filed with multiple patent offices. Furthermore, the same sequence may be used by different inventors in different inventions. Information relating to the source patent may be incomplete, and biological information available in the patent document may not be reflected in annotation. Thus, search and analysis of these data have become increasingly challenging (8,9). Recent efforts have been made to create non-redundant patent sequence resources to improve access and direct analysis of the sequences. PatGen (10), a database containing non-redundant data from the public resources, allowed queries against patent bibliographic data and sequences. Unfortunately, the method of redundancy removal in PatGen has not been detailed to the public and the database is no longer available online. Patome (11) is a non-redundant patent sequence set also derived from the public resources, providing additional annotations with RefSeq (12), OMIM (13) and Gene ontology (GO) (14). Patome is useful for the identification of disease-related patent sequences. Duplicated sequences were removed in Patome according to the patent number (PN) and the sequence identifier in the sequence listing; however, identical sequences granted with different PNs by different patent offices are not classified in Patome. None of these studies attempts to establish publicly available non-redundant patent sequence databases based on the sequence level and the patent family level.

In this article, we describe a publicly available collection of non-redundant patent sequence databases, which have been created at two levels and cover the EMBL-Bank patent class nucleotides and the patent proteins from the EPO, JPO, KIPO and USPTO. The proprietary patent resources have been excluded due to the restrictions on their use. The non-redundant sequences are identified using MD5 (Message-Digest algorithm 5) (http://www.faqs.org/rfcs/rfc1321.html) checksums of the sequences. Members of a level-1 cluster are 100% identical over their whole length. Level-2 clusters are defined by sub-grouping level-1 clusters based on the patent equivalents which have been published by different patent offices for a single invention. The clusters contain value-added annotations, such as publication patent corrections and earliest publication dates. Level-2 clusters also offer merged biological features. The data collection significantly enhances the quality of patent sequence data and allows for better tracking and cross-referencing in patent search.

DATA COLLECTION AND DATABASE CONTENTS

Redundancy removal

The source data used in this study were obtained from the EMBL-Bank PAT data class (see: http://www.ebi.ac.uk/patentdata/nucleotides/), these include the patent sequences submitted through the INSDC member databases; and the protein sequence data provided by EPO, JPO, KIPO and USPTO (see: http://www.ebi.ac.uk/patentdata/proteins/). The consistent use of data formats, especially the feature tables, simplifies downstream data processing.

Data redundancy was removed in two steps and two non-redundant databases were generated at different levels. The level-1 was defined by identifying sequences at 100% identity over the whole sequence length and grouped into clusters (Figure 2). A number of methods for identifying identical sequences were evaluated (Supplementary Data Section 2):

(i) Cyclic Redundancy Check (CRC) checksum (http://en.wikipedia.org/wiki/Cyclic_redundancy_check): the CRC32 (IEEE 802.3) and CRC64-ISO (ISO 3309) checksums commonly used in the sequence databases were found to be insufficiently robust to uniquely identify the sequences. The alternative CRC64-ECMA (ECMA-182) was also evaluated and while performing better than CRC64-ISO still found collisions, where the same checksum was generated by non-identical sequences.
(ii) nrdb: provided as part of the WU-BLAST package (15), proved not to be robust enough to deal with large-size data sets.
(iii) CD-HIT (16): clusters sequences efficiently using short word filters but discards short sequences and merges overlapping sequences of differing length.

(iv) SHA checksums (http://www.itl.nist.gov/fipspubs/fip180-1.htm): proved to be capable of distinguishing the currently available sequences; however, they are resource intensive.
(v) MD5 checksum: correctly distinguish the available sequences and are viewed by the database providers as a replacement for the current CRC checksums provided in the data.

On the basis of this evaluation MD5 checksums were chosen to identify identical sequences. Sequences with the same checksum are considered identical, conversely sequences with different checksums are considered non-identical. A level-1 cluster contains identical sequences originating from different patent documents, which may belong to different patent families, and thus may have biological annotations from different contexts (e.g. different organisms).

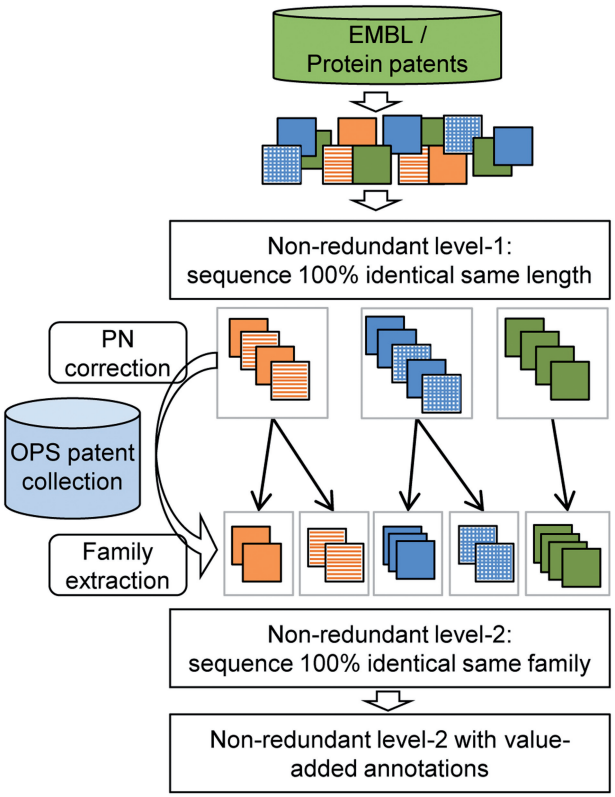

Figure 2. Steps to create non-redundant patent sequence databases at two levels. Squares of the same colour represent level-1 sequences, 100% identical over the whole length. Squares of the same colour and pattern represent level-2 sequences, which are identical and belong to the same invention (i.e. patent family).

Level-1 cluster sequences were grouped into a level-2 cluster if they belonged to a same patent family; otherwise they were grouped into different level-2 clusters (Figure 2). A patent family is the collection of all the equivalent patents describing the same invention filed with different patent offices around the world. Please see Supplementary Data Section 1 for more patent-related terminology. Patent publication numbers associated with the sequences were corrected and then patent family information and international classifications (IPC classes) were extracted from the Open Patent Services (OPS) (17). Level-1 cluster sequences without patent family information were separated into individual level-2 clusters. For example, level-1 cluster NRP_AX000635 contains 15 members (Supplementary Data Section 4.2). Using the patent equivalent information (Supplementary Data Section 4.3), USPTO proteins AAN97218 and BE25759 belong to one patent family (family number 26846334); USPTO proteins AAO99687 and AAS33207 belong to another patent family (family number 27576013); the Korean protein DI578933 does not have family information; and the other members belong to 10 different patent families. Therefore, this cluster (NRP_AX000635) was split into 13 level-2 clusters (NRP00180079 through NRP00180085, Supplementary Data Section 4.4).

For the 8 300 915 patent nucleotide entries in EMBL-Bank release 99; 5 167 627 level-1 clusters (NRNL1) and 6 714 564 level-2 clusters (NRNL2) were obtained. For the 3 307 421 patent protein entries of 21 May 2009; 1 371 866 level-1 clusters (NRPL1) and 2 281 606 level-2 clusters (NRPL2) were obtained (Table 1). A nucleotide or a protein sequence appeared in ~1.30 or 1.66 patent families, respectively.

Annotations

The non-redundant patent sequence databases have added value with respect to the existing patent sequence repositories. Additional information that is legally important to the intellectual-property world was provided. Earliest publication dates are required to identify relevant prior art. The earliest publication date was determined for each cluster by comparing the patent publication dates amongst all the cluster members. Correct publication numbers are important to identify the legal status of the patent and to link to the patent full-text document.

Table 1. Summary of two-level non-redundant patent sequence databases (based on the EMBL Release 99 of March 2009)

NR databases	Abbreviation	Coverage	No. of entries	Redundancy before
NR patent nucleotides level-1	NRNL1	EMBL-Bank patents (8 300 915 entries)	5 167 627	1.61
NR patent nucleotides level-2	NRNL2	EMBL-Bank patents (8 300 915 entries)	6 714 564	1.24
NR patent proteins level-1	NRPL1	EPO, JPO, KIPO and USPTO patent proteins (3 307 421 entries)	1 371 866	2.41
NR patent proteins level-2	NRPL2	EPO, JPO, KIPO and USPTO patent proteins (3 307 421 entries)	2 281 606	1.45

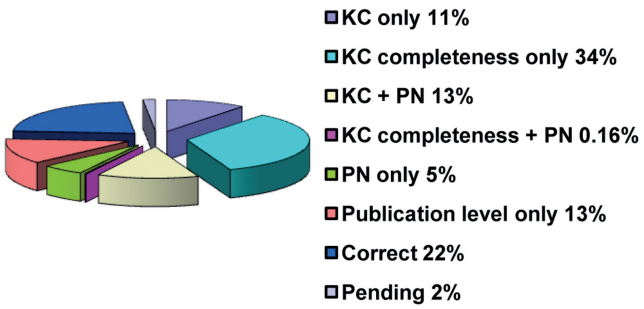

Figure 3. Publication number error types detected in the sequence data set (both nucleotide and protein). 'KC only' represents the errors of incorrect KC only; 'KC completeness only' represents the errors of incomplete KC only; 'KC + PN' represents the errors of wrong PN and wrong KC; 'KC completeness + PN' represents the errors of incomplete KC and wrong PN; 'PN only' represents the errors of wrong PN only; 'Publication Level only' represents errors of publication level only; and 'Pending' represents publication numbers which currently cannot be resolved and are pending for corrections.

Analysis of the PNs, extracted from the sequence records (both nucleotide and protein), which are associated with individual patents, discovered errors in 77.8% of the numbers. The errors detected include the publication number (e.g. WO0112792), the publication date, the Kind Code (KC; an indication of the legal status of the patent application and its progress through the patent process) and the publication level (first availability versus subsequent publication, of the patent in the prior art). The types and percentage of errors detected are summarized in Figure 3.

A level-2 cluster contains identical sequences originating from the same invention (same patent family) and has biological annotations from the same context. Therefore, the original annotations among the members of the same level-2 cluster are expected to be the same. However, differences have been discovered in some cases. To provide relevant biological information for level-2 clusters, the feature tables of the source entries have been merged and qualifiers added to indicate the origin of the annotation.

Result format

Identifiers were assigned to the clusters. A level-1 cluster identifier begins with 'NRN_' for a nucleotide cluster and 'NRP_' for a protein cluster, followed by the accession number of a source entry selected according to the priority order of EPO, USPTO, JPO and KIPO for members from different patent offices and alpha-numeric order for members from the same patent office, e.g. NRN_CS587094 and NRP_AX000635. A level-2 cluster identifier begins with 'NRN' for a nucleotide cluster and 'NRP' for a protein cluster, followed by eight hexadecimal digits, e.g. NRN001DEAFE and NRP00000018. The level-2 cluster identifiers are maintained across database releases and an identifier history is provided which includes information about deletions and insertions within the cluster. This is stored in a tab-delimited table, allowing users to track interesting clusters in different releases.

The sequence data are stored in FASTA sequence format, while the annotation data are stored in an EMBL-like format. The FASTA sequence format header begins with the cluster identifier, followed by a publication number, chosen from the set of publication numbers associated with the cluster, using the same rules, as were used to choose the accession number used in the identifier at level-1, and its KC. Examples of FASTA format sequences for the clusters can be found in Supplementary Data Section 3. Annotation files for both levels provide information about the earliest patent publication within the cluster (e.g. the earliest publication date, the corresponding publication number and its complete kind code in the ED line), data references of cluster members (DR block) and sequence information (e.g. the sequence length as well as the MD5 checksum in the SQ line). Corrected publication numbers are described in additional PN lines, followed by the type of correction (CC line). Level-2 annotation also indicates the master patent family number of the cluster (MF line), the earliest active priority number within the family (PR line) and the merged features and qualifiers (FT lines). Examples of annotation records for the clusters are shown in Supplementary Data Section 4.

DATA ACCESS

The FASTA format sequences and annotation files for the clusters are downloadable from the EMBL-EBI non-redundant patent data webpage (http://www.ebi.ac.uk/patentdata/nr/). Sequence similarity and homology searches [e.g. FASTA (18), NCBI BLAST (19) and WU-BLAST (15)] against the non-redundant databases are available through the web interface (http://www.ebi.ac.uk/Tools/sss/) and SOAP and REST style Web Services (20). The non-redundant patent sequence data and the patent equivalents are available for text searching through the EMBL-EBI's SRS server (http://srs.ebi.ac.uk/).

CONCLUSIONS

Sequence similarity and homology searches against non-redundant sequence libraries are faster and more sensitive than the equivalent search using redundant sequence databases since fewer entries have to be scanned, more reliable statistics are obtained due the database being smaller and the relationships between significant hits are easier to interpret. These databases are the first publicly available collection of non-redundant patent sequence databases, at both the sequence and patent-family levels. Searches against level-1 clusters result in identical or similar sequences from the patent literature (i.e. each sequence may correspond to multiple inventions); searches against level-2 clusters result in identical or similar sequences from the same invention and in effect this is searching for related patents and is not a pure sequence space search. In addition, the level-2 data

provide patent-family information, allowing the user to explore the original patent applications for related intellectual-property annotation. The corrected publication numbers enhance the data quality, enabling proper cross-referencing to patent full-text documents. Similarly, the earliest publication dates offer direct tracking of the patent-application history, enabling effective prior art searches. The collation of feature annotation from the members of the family provides better cross-referencing and improved biological context.

Addition of data from commercial efforts (e.g. GENESEQ™) would improve data coverage; however, restrictions on the use of these data mean that the non-redundant sets could not be made publicly available. Sequences originating from other patent offices, more specifically other International Searching Authorities (ISAs), would improve the public domain coverage without imposing usage restrictions. The EPO and EMBL-EBI are collaborating in the development of computer applications that normalize patent sequence formats (WIPO ST.25 or future XML standards) and enable interoperability between these data and existing sequence analysis software. This application will enable us to include data from patents submitted to other patent offices in the public sequence databases. The broad coverage of the non-redundant sequence databases described here, and the enhanced accessibility to the sequence data through these databases, improves the utility of the existing patent sequence data for the life sciences and intellectual-property communities.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors wish to acknowledge all software developers, database administrators, data curators and users at the EMBL-EBI, NCBI, DDBJ and the European Patent Office, who have provided extremely valuable feedback and support throughout.

FUNDING

The European Commission under FELICS (contract number 021902 (RII3), within the Research Infrastructure Action of the FP6 'Structuring the European Research Area' Programme); the European Molecular Biology Laboratory (EMBL); and the European Patent Office. Funding for open access charge: EMBL.

Conflict of interest statement. None declared.

REFERENCES

1. Thangaraj,H. (2007) Information from patent office could aid replication. *Nature*, **447**, 638.
2. Seeber,F. (2007) Patent searches as a complement to literature searches in the life sciences—a 'how-to' tutorial. *Nat. Protoc.*, **2**, 2418–2428.
3. Degtyarenko,K., de Matos,P., Ennis,M., Hastings,J., Zbinden,M., McNaught,A., Alcántara,R., Darsow,M., Guedj,M. and Ashburner,M. (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, **36**, D344–D350.
4. Andree,P.J., Harper,M.F., Nauche,S., Poolman,R.A., Shaw,J., Swinkels,J.C. and Wycherley,S. (2008) A comparative study of patent sequence databases. *World Pat. Inform.*, **30**, 300–308.
5. 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.
6. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.
7. Sugawara,H., Ogasawara,O., Okubo,K., Gojobori,T. and Tateno,Y. (2008) DDBJ with new system and face. *Nucleic Acids Res.*, **36**, D22–D24.
8. Yoo,H., Ramanathan,C. and Barcelon-Yang,C. (2005) Intellectual property management of biosequence information from a patent searching perspective. *World Pat. Inform.*, **27**, 203–211.
9. Dufresne,G., Takács,L., Heus,H.C., Codani,J.J. and Duval,M. (2002) Patent searches for genetic sequences: how to retrieve relevant records from patented sequence databases. *Nat. Biotechnol.*, **20**, 1269–1271.
10. Rouse,R.J., Castagnetto,J. and Niedner,R.H. (2005) PatGen—a consolidated resource for searching genetic patent sequences. *Bioinformatics*, **21**, 1707–1708.
11. Lee,B., Kim,T., Kim,S.K., Lee,K.H. and Lee,D. (2007) Patome: a database server for biological sequence annotation and analysis in issued patents and published patent applications. *Nucleic Acids Res.*, **35**, D47–D50.
12. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI Reference Sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
13. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
14. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium. Nat. Genet.*, **25**, 25–29.
15. Lopez,R., Silventoinen,V., Robinson,S., Kibria,A. and Gish,W. (2003) WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res.*, **31**, 3795–3798.
16. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
17. Kallas,P. (2006) Open patent services. *World Pat. Inform.*, **28**, 296–304.
18. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
19. Altschul,S.F., Madden,T.L., Schäffer,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.
20. McWilliam,H., Valentin,F., Goujon,M., Li,W., Narayanasamy,M., Martin,J., Miyar,T. and Lopez,R. (2009) Web services at the European Bioinformatics Institute-2009. *Nucleic Acids Res.*, **37**, W6–W10.