The Annotation-enriched non-redundant patent sequence databases

Weizhong Li¹, Bartosz Kondratowicz², Hamish McWilliam¹, Stephane Nauche² and Rodrigo Lopez¹,∗

¹European Bioinformatics Institute, EMBL Outstation, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CD10 1SD, UK and ²European Patent Office, Patentlaan 3-9, 2288 EE Rijswijk, The Netherlands

*Corresponding author: Tel: +44 1223 494423; Fax: +44 1223 494468; Email: rls@ebi.ac.uk

Citation details: Li W., Kondratowicz B., McWilliam H., et al. The Annotation-enriched non-redundant patent sequence databases. Database (2013) Vol. 2013: article ID bat005; doi: 10.1093/database/bat005

Submitted 9 November 2012; Revised 19 December 2012; Accepted 22 January 2013

The EMBL-European Bioinformatics Institute (EMBL-EBI) offers public access to patent sequence data, providing a valuable service to the intellectual property and scientific communities. The non-redundant (NR) patent sequence databases comprise two-level nucleotide and protein sequence clusters (NRNL1, NRNL2, NRPL1 and NRPL2) based on sequence identity (level-1) and patent family (level-2). Annotation from the source entries in these databases is merged and enhanced with additional information from the patent literature and biological context. Corrections in patent publication numbers, kind-codes and patent equivalents significantly improve the data quality. Data are available through various user interfaces including web browser, downloads via FTP, SRS, Dbfetch and EBI-Search. Sequence similarity/homology searches against the databases are available using BLAST, FASTA and PSI-Search. In this article, we describe the data collection and annotation and also outline major changes and improvements introduced since 2009. Apart from data growth, these changes include additional annotation for singleton clusters, the identifier versioning for tracking entry change and the entry mappings between the two-level databases.

Database URL: http://www.ebi.ac.uk/patentdata/nr/

Introduction

The patent data are a valuable resource, not only for the intellectual property world but also for the scientific community (1,2). During the past 15 years, the number of biological sequences appearing in patent documents has been increasing constantly (3). Today, >30 million nucleotide and protein sequences extracted from patent documents are available in the public domain (shown by the black lines in Figure 1). Searching this large amount of patent sequence data has become one of the key approaches in patent-related studies (4,5). Proprietary data also exist from the commercial sector providing alternative annotations of patent sequence data, such as GENESEQ™ (Thomson Reuters, http://thomsonreuters.com/products_services/science/science_products/a-z/geneseq/), GQ-PAT (GenomeQuest, http://wiki.genomequest.com/index.php/GQ_Pat), USGENE (SequenceBase, http://www.sequencebase.com/usgene-sequences-database), but these require commercial licenses, which impose usage restrictions on the data.

The EMBL-European Bioinformatics Institute (EMBL-EBI) offers free and unrestricted access to patent sequence resources, providing a valuable service to the intellectual property and bioscience communities (6). The two-level non-redundant (NR) patent sequence databases, based on sequence identity and patent family clusters, are comprehensive repositories for patent information on nucleotide and protein sequences provided by the European Patent Office (EPO), the US Patent and Trademark Office (USPTO), the Japanese Patent Office (JPO) and the Korean Intellectual Property Office (KIPO) and include the World Intellectual Property Organization (WIPO) patents from...
these offices. The NR patent sequences are enriched with biological annotation and additional data from patent documents. These databases, serving as a repository of scientific innovation and inspiration, are an important resource for patent-related searches, especially for determining potential commercial use of biological sequences and their patentability. In this article, we describe the sequence collection and annotation of the NR patent sequence databases, and introduce improvements and development of the databases over the past 3 years.

**Sequence Collection and Annotation**

The NR patent sequence data sources cover nucleotide and protein sequences in patent applications from the EPO, the JPO, the KIPO and the USPTO. The patent sequence data deposited to ENA (7), GenBank (8) or the DDBJ (9) are exchanged between these databases through the International Nucleotide Sequence Database Collaboration (Figure 2a). Sequence submissions by the inventors can be made as part of the patent application process using tools, such as BISSAP (http://www.epo.org/bissap/), an application developed to facilitate the creation of sequence listings (WIPO ST.25 and proposed XML format) for patent applications containing biological sequences by the EPO in collaboration with national patent offices and the EMBL-EBI.

The NR patent sequence databases have been created at two levels to remove sequence redundancy by using sequence MD5 (Message-Digest algorithm 5, http://www.faqs.org/rfcs/rfc1321.html) checksums and patent family information, comprising NR patent nucleotides level-1 and -2 and NR patent proteins level-1 and -2. Level-1 sequences are 100% identical over their entire lengths, arising from either the same or different patent families; level-2
sequences are 100% identical over their entire length and belong to the same patent family. Patent family information for source sequences is retrieved from the EPO Open Patent Services (OPS) (10). Level-1 databases include NR nucleotide patent sequence clusters level-1 (NRNL1, NRPL1, EPOPNR, JPOPNR, KPOPNR and USPOPNR) and level-2 (NRNL2 and NRPL2), the patent equivalent database and other relevant result files. (c) Data access through FTP, DbFetch, SRS, EBI-Search and SSS (Sequence Similarity/Homology Search).

In patents, a right of priority is a time-limited right triggered by the first filing of a patent application; a patent family refers to several patent applications or publications for an individual invention, claiming exactly the same priority or priorities; all of these family equivalents are related to each other by common priority numbers and associated priority dates (http://www.epo.org/searching/essentials/patent-families/about.html). The family information in the database covers patent family numbers, patent priority, master publications, patent equivalents, subsequent publication levels and patent classification. The database format is detailed in the user manual (http://www.ebi.ac.uk/patentdata/doc/Family_equivalents_database_v3.pdf).

The annotation of the NR sequences comprises cluster member annotation, patent family information and biological features. The cluster member annotation includes source sequence information, e.g. identifier (ID), molecular type, sequence length, source database, patent number and a general description. The patent family information consists of family number, master publication, patent priority, earliest publication date and the EPO and international classifications. The earliest publication date is determined to identify relevant prior art of the patent by comparing the patent publication dates of all the members of a NR sequence cluster. The biological features contain information on organisms, coding sequence regions, genes, variations, combined for both contig and singleton members. This combined annotation allows better exploration of the original patent applications for related intellectual property data. It also provides better cross-references to external data resources and improves the biological context at the sequence level. The annotation format is detailed in the user manual (http://www.ebi.ac.uk/patentdata/doc/Nonredundant_databases-user_manual_v3.pdf).

**Data Growth and Improvements**

The NR patent sequence databases are released every 3 or 4 months, but usually following EMBL-Bank's quarterly release cycles. The current release (Release 13, Oct 2012) contains 12 279 969 NRNL1, 14 920 929 NRNL2, 2 580 442 NRPL1 and 3 697 317 NRPL2, ~2.4-, 2.2-, 1.9- and 1.6-fold in size compared with the first release of NRNL1, NRNL2, NRPL1 and NRPL2, respectively, covering over 6 571 318 proteins and 24 364 832 nucleotides from 184 447 patents (130 538 unique patent families), which are provided by the patent offices of the EPO, the JPO, the KIPO and the USPTO (Table 1, Figure 1). The data coverage is slightly larger than the commercial patent sequence database GENESSEQ, which included ~27 million sequences from ~150 000 patents in Oct 2012 (http://thomsonreuters.com/products_services/science/science_products/a-z/geneseq/).

Patent publication numbers, sequence kind-codes and patent equivalents are corrected or updated in each release.
Database has two examples that illustrate how sequences from level-1 nucleotide and protein sequences are clustered into level-2 entries. These mappings offer a useful explanation of the relationship between identical sequences within or outside of a patent family.

Members of level-2 clusters in an old release can move to other clusters in a new release. This is due to changes in equivalents assignment in patent families. The ID versioning has been provided since Release 6 for direct tracking of entry history. This functionality is necessary for recovering information from old entries that have moved or have become obsolete in a new release.

| Table 1. Summary of the NR patent sequences and the patent families in Release 13 |
|---------------------------------|-------------------|
|                                 | Number of entries | Redundancy before |
| Patent nucleotides              | 24 364 832        |                   |
| NRNL1                           | 12 279 969        | 1.98              |
| NRNL2                           | 14 920 929        | 2.22              |
| Patent proteins                 | 6 571 318         |                   |
| NRPL1                           | 2 580 442         | 1.88              |
| NRPL2                           | 3 697 317         | 1.62              |
| Patents                         | 184 447           |                   |
| Unique patent families          | 130 538           | 1.41              |

Using the latest patent family data from the EPO’s OPS. Across all releases, 43 111 patent numbers and 14 330 sequence kind-codes have been corrected; 102 227 patent numbers have been involved in the patent family assignment. The corrected publication numbers link to the correct full-text patent documents; the corrected publication kind codes and the publication levels indicate the legal status and progress through the patent application process.

The ID mappings between level-1 and level-2 databases have been generated since Release 10 to clearly illustrate how identical sequences from level-1 databases are mapped to level-2 database entries according to their patent family information. Figure 3 has two examples that illustrate how sequences from level-1 nucleotide and protein sequences are clustered into level-2 entries. These mappings offer a useful explanation of the relationship between identical sequences within or outside of a patent family.

Members of level-2 clusters in an old release can move to other clusters in a new release. This is due to changes in equivalents assignment in patent families. The ID versioning has been provided since Release 6 for direct tracking of entry history. This functionality is necessary for recovering information from old entries that have moved or have become obsolete in a new release.

**Data Access and Usage**

Data access to the NR patent sequence resources has become more and more important to the user community as the volume of sequence data increases. The NR patent sequence databases can be accessed through four major routes (Figure 2c) at EMBL-EBI:

1. The flat files can be downloaded through the databases website (http://www.ebi.ac.uk/patentdata/nr/) and through the FTP site (ftp://ftp.ebi.ac.uk/pub/data bases/patentdata/).

2. The EMBL-like formatted annotation data can be retrieved on a per-accession through the Dbfetch/ WSDbfetch service (http://www.ebi.ac.uk/Tools/ dbfetch/dbfetchv) and also through the SRS server (http://srs.ebi.ac.uk/).

3. Sequence similarity/homology searches including FASTA (11), BLAST (12,13) and PSI-Search (14) against the databases are available through the web form submissions (http://www.ebi.ac.uk/Tools/sss/) and also through the corresponding EMBL-EBI SOAP/REST web services (15).

4. Keyword searches can be made using the EBI-Search engine (16) through both a web form (http://www.ebi.ac.uk/ebisearch) and the corresponding SOAP web services.

Approximately 10 000 sequence similarity/homology searches were performed using the databases during 2010. This grew to >36 000 searches in 2011, and it is estimated that ~37 500 searches will have taken place during 2012. The same trend is seen for data retrieval via Dbfetch/WSDbfetch, which have grown from 450 000 in 2011 to a projected 510 000 for 2012. FTP downloads of these sequence data have also grown from 394 downloads in 2011 to a projected 540 for 2012.

**Discussion and Future Implementation**

The NR patent sequence databases are the first publicly available collection of NR patent sequences, at both the sequence and patent-family levels. Other efforts in the public domain have been made to collate NR patent sequence data to improve access and use of these data, such as PatGen (17) and Patome (18). Unfortunately, PatGen is no longer available online; the sequence redundancy in Patome was defined according to the patent number and the sequence ID in the sequence listing. As a result, identical sequences granted with different patent numbers by different patent offices are not classified.

Sequence similarity/homology searching against the NR patent sequence databases has become a fundamental approach in patent-related studies. Searches against NR sequences are faster and more sensitive than the equivalent searches against redundant libraries, and the search results are easier to interpret. Searches against level-1 clusters can result in identical or similar patent sequences; searches against level-2 clusters can result in identical or similar sequences from the same invention. These searches can be used to find the published patents that cite a sequence and the patent families associated with a sequence, to discover the earliest priority data and the equivalents of a patent family, and to retrieve biological annotation extracted from patent documents.
The NR patent sequence databases are an important resource for patent-related searches, especially for determining potential commercial use of biological sequences. The earliest publication dates offer direct tracking of patent-application history, enabling effective searches on prior art. The corrections on the publication numbers and kind codes enhance the data quality, enabling proper cross-referencing to full-text patent documents. These databases are also a repository of scientific innovation and inspiration.

We will continue to make improvements and add new features in the future. For example, to broaden data coverage by including data from other national and regional patent offices, to shorten the release cycle to a monthly schedule and to integrate cross-references to claimed sequences and provide claimed status. Currently, users can download the ID history tables to track entry changes, such as status, and entry additions, deletions, merging and unmerging; in the future, an online searchable system will be implemented.

Acknowledgements

The authors want to acknowledge all database administrators, data curators and users at EMBL-EBI, DDBJ, NCBI and EPO, who have offered important support and valuable feedback throughout. Thanks to Andrew Cowley for additional manuscript input and language corrections.

Funding

This work has been supported by European Molecular Biology Laboratory (EMBL) and the European Patent Office (EPO). Funding for open access charge: EMBL.

Conflict of interest. None declared.

References

1. Seeber,F. (2007) Patent searches as a complement to literature searches in the life sciences—a ‘how-to’ tutorial. Nat. Protoc., 2, 2418–2428.
2. Thangaraj,H. (2007) Information from patent office could aid replication. Nature, 447, 638.
3. Dufresne,G. and Duval,M. (2004) Genetic sequences: how are they patented? Nat. Biotechnol., 22, 231–232.
4. Dufresne,G., Takács,L., Heus,H.C. et al. (2002) Patent searches for genetic sequences: how to retrieve relevant records from patented sequence databases. Nat. Biotechnol., 20, 1269–1271.
5. McDowall,J. (2011) Prioritizing patent sequence search results using annotation-rich data. World Pat. Inform., 33, 235–239.
6. Li,W., McWilliam,H., Richart de la Torre,A. et al. (2010) Non-redundant patent sequence databases with value-added annotations at two levels. Nucleic Acids Res., 38, D52–D56.
7. Leinonen,R., Akhtar,R., Binney,E. et al. (2011) The European nucleotide archive. Nucleic Acids Res., 39, D28–D31.
8. Benson, D.A., Karsch-Mizrachi, I., Clark, K. et al. (2012) GenBank. Nucleic Acids Res., 40, D48–D53.
9. Kaminuma, E., Kosuge, T., Kodama, Y. et al. (2011) DDBJ progress report. Nucleic Acids Res., 39, D22–D27.
10. Kallas, P. (2006) Open patent services. World Pat. Inform., 28, 296–304.
11. Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA, 85, 2444–2448.
12. Altschul, S.F., Madden, T.L., Schäffer, A.A. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389–3402.
13. Lopez, R., Silventoinen, V., Robinson, S. et al. (2003) WU-Blast2 server at the European Bioinformatics Institute. Nucleic Acids Res., 31, 3795–3798.
14. Li, W., McWilliam, H., Goujon, M. et al. (2012) PSI-Search: iterative HOE-reduced profile SSEARCH searching. Bioinformatics, 28, 1650–1651.
15. McWilliam, H., Valentim, F., Goujon, M. et al. (2009) Web services at the European Bioinformatics Institute-2009. Nucleic Acids Res., 37, W6–W10.
16. Valentim, F., Squizzato, S., Goujon, M. et al. (2010) Fast and efficient searching of biological data resources—using EB-eye. Brief. Bioinform., 1, 375–384.
17. Rouse, R.J., Castagnetto, J. and Niedner, R.H. (2005) PatGen—a consolidated resource for searching genetic patent sequences. Bioinformatics, 21, 1707–1708.
18. Lee, B., Kim, T., Kim, S.K. et al. (2007) Patome: a database server for biological sequence annotation and analysis in issued patents and published patent applications. Nucleic Acids Res., 35, D47–D50.