PatGen DB – a consolidated genetic patent database implementing standard data mining resources

R. Rouse*

*HTS < Resources/PatentInformatics
P.O. Box 948586, La Jolla, CA, USA, rjdrouse@htsresources.com

ABSTRACT

Compared to the wealth of online resources covering genomic, proteomic and derived data, the bioinformatics community is rather under served when it comes to genetic patent information. This paper describes how PatGen DB has been compiled. This is a case study demonstrating how integrated theme based patent databases can be developed without resorting to expensive commercial data providers.

Keywords: open patent services, XML, patent information, relational database, open source

1 THE STATE OF PATENT INFORMATION

Patent information is voluminous. According to the 2003 United States Patent and Trademark Office (USPTO) annual report, the office received 333,452 applications; this accounts for 913 applications a day (1). There is approximately 12,000 new patents issued every month by the USPTO. Given this scenario, effectively accessing patent information is important and yet a challenging task for a non patent professional.

Despite the presence of many commercial patent databases, public patent resources such as the European Patent Office Espacenet and INPADOC are excellent resources for accessing comprehensive patent information. Although these databases contain a wider collection of patent information than the commercial counterparts, effectively accessing this data has been difficult for a professional patent searcher.

This report demonstrates how freely available open source tools can be used to generate high quality patent information that can be reliably used when defining intellectual property space. As an example, a brief overview of how PatGen DB, our integrated open patent database has been compiled.

2 COMPOSITION OF PATGEN DB

Essentially PatGen DB is a consolidated database containing non-redundant data from public resources. Data from the EBI (2), DDBJ (3) and the NCBI (4) was collected from GenBank, EMBL and fasta formatted batch files vial ftp. Sequence data was also retrieved from the World Intellectual Property Organization (WIPO) (5). These sequences are derived from world patent applications in accordance with the Patent Cooperation Treaty (PCT). These files are in the required patent sequence-listing format (6).

As well as containing sequence data, PatGen DB also contains bibliographic data obtained from the European Patent Office (OPS). This is a web service where one can access current bibliographic, family and legal information in XML format (7). This can encompass multiple patent offices. This type of information enables the user to determine the legal status of both patent applications and issued patents. The data is accessed in real-time via a SOAP-based web service, delivering up-to-date information in a seamless and completely transparent manner. Full text patent information (i.e., background, specification and claims) can also be accessed.

The PatGen DB software architecture is entirely based on open source tools and deployed on a Suse 9.0 Linux server (8). Data acquisition and parsing of the flat files from the various sources is implemented using Perl (9) standard libraries. The consolidated dataset is stored and served from a MySQL relational database management system (10) using Bioperl-DB schema and parsers (11). The software interface that accesses the data is written in PHP (12). The OPS web service is being accessed using SOAP through PEAR – the PHP extension and application library (13). Full text is accessed using Perl libwww-perl library (14).

3 FEATURES OF PATGEN DB

PatGen DB is significantly more comprehensive than any public repositories. This database currently contains almost twice as many nucleic acids as well as significantly more amino acids. To remain current, PatGen DB is updated monthly. In order to keep PatGen DB non

redundant we add patent documents that are not already in the database.

In PatGen both issued patents and pending patent applications can be searched via fulltext queries against the bibliographic data to retrieve disclosed genetic sequences. The simple query form provides fields for searches based on title, abstract, inventors, applicants (i.e., the inventors' assignees) and date of publication. The interface also provides direct access to patent-related sequences via the patent publication code. Each search displays the retrieved sequences as a tabulated list with links to detailed sequence information such as sequence taxonomy, genetic code and a brief description. Alternatively the entire list of sequences can be accessed in fasta format for bioinformatic analysis. Sequence taxonomy searching is available as well as a sequence search feature using BLAST.

PatGen DB is a resource where one can perform both patent and bioinformatic analysis. Patent analysis is used to determine whether to enter into licensing agreements and an essential component in profiling the technology of a given industry and thus relevant in many business activities. Bioinformatics analysis creates opportunities to develop new types of patent strategies, in particular now that annotation of newly sequenced genomes and comparing sequences across organisms have become straightforward and commonplace in biological laboratories.

In establishing a consolidated patent genetic database using open source tools, our intent is to integrate genetic sequence data with patent information. As a result we have established a modular system that can be used to compile customized theme based patent information databases upon request.

REFERENCES

[1] USPTO 2003 Performance and Accountability Report, http://www.uspto.gov/web/offices/com/annual/2003/index.html

[2] European Bioinformatics Institute, ftp://ftp.ebi.ac.uk/pub/databases/embl/patent

[3] DNA Database of Japan, ftp://ftp.ddbj.nig.ac.jp/database/ddbj

[4] National Center for Biotechnology Information, for nucleic acids - ftp://ftp.ncbi.nlm.nih.gov/genbank for amino acids - ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA

[5] World Intellectual Property Organization, http://www.wipo.int/pct/en/sequences/listing.htm

[6] Information Concerning the Filing of International Applications Containing Large Nucleotide/Amino Acid Sequence Listings and/or Tables in the United States Receiving Office, http://www.uspto.gov/web/offices/pac/dapps/pct/aipart8.html

[7] European Patent Office Open Patent Service, http://ops.espacenet.com/

[8] Suse Linux, http://www.suse.com/us/index.html

[9] Comprehensive Perl Archive Network, http://www.cpan.org/

[10] MySQL, http://www.mysql.com/

[11] BioPerl, http://www.bioperl.org/

[12] PHP, http://www.php.net/

[13] PEAR::SOAP, http://pear.php.net/packages/SOAP