**PatGen—a consolidated resource for searching genetic patent sequences**

Richard J. D. Rouse\(^1\),*, Jesus Castagnetto\(^2\) and Roland H. Niedner\(^2\)

\(^1\)PatentInformatics, PO Box 948586 La Jolla, CA 92037, USA and \(^2\)San Diego Supercomputer Center, La Jolla, CA 92037, USA

Received on September 8, 2004; revised on November 15, 2004; accepted on December 1, 2004

Advance Access publication February 22, 2005

**ABSTRACT**

Summary: Compared to the wealth of online resources covering genomic, proteomic and derived data the Bioinformatics community is rather underserved when it comes to patent information related to biological sequences. The current online resources are either incomplete or rather expensive. This paper describes, PatGen, an integrated database containing data from bioinformatic and patent resources. This effort addresses the inconsistency of publicly available genetic patent data coverage by providing access to a consolidated dataset.

Availability: PatGen can be searched at http://www.patgendb.com

Contact: jrdrouse@patentinformatics.com

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 (http://www.uspto.gov/web/offices/com/annual/2003/index.html). Currently, the average pendency period for an application is 26.7 months. The number of applications filed increases every year. With this growing backlog, the USPTO is faced with the problem of expediting the patent issuance process while maintaining the quality of issued patents.

Currently, three web resources offer sequence searching of patent data: the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI) and the DNA Database of Japan (DDBJ). These resources offer both sequence downloads and alignment searching; however, the data across these sources is variable. Thompson Derwent and the Chemical Abstract Service maintain commercial databases containing more comprehensive information. The commercial resources are expensive to use and are normally accessed by private commercial entities having patent information specialists on staff.


Besides containing sequence data, PatGen also contains bibliographic data obtained from the European Patent Office’s (EPO) Open Patent Services (OPS). This is a web service where one can access current bibliographic, family and legal information in XML format (http://ops.espacenet.com/). 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.

The PatGen software architecture is entirely based on open source tools and deployed on a Suse 9.0 Linux server (http://www.suse.com/us/index.html). Data acquisition and parsing of the flat files from the various sources is implemented using Perl (http://www.cpan.org/) and standards. The consolidated dataset is stored and served from a MySQL relational database management system (http://www.mysql.com/) using Bioperl-DB schema and parsers (http://www.bioperl.org/). The software interface that accesses the data is written in PHP (http://www.php.net). The OPS web service is being accessed using SOAP through PE4R—the PHP extension and application library (http://pear.php.net/packages/SOAP).

Our database is significantly more comprehensive than that of any of the public repositories. As of October 1, 2004, PatGen had almost twice as many nucleic acids than the public resources as well as significantly more amino acid sequences. To remain current, PatGen is updated monthly. In order to keep the PatGen 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 as well as a sequence search feature using BLAST is available.

PatGen is a database where one can access data to perform both patent and bioinformatic analysis. Patent analysis is used to determine commercial resources, patent classification, inventor and assignee information. 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.

*To whom correspondence should be addressed.*
whether to enter into licensing agreements and is an essential component in profiling the technology of a given industry. It is 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. With the advent of genomic and proteomic analysis, more biochemical and metabolic pathways will be mapped out with an increase in the richness of relationships to be discovered.

In establishing a consolidated patent genetic database using open source tools, our intent is to integrate genetic sequence data with patent information. Future efforts will focus on implementation of open standards for patent information.

REFERENCES

European Bioinformatics Institute (2004) Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.
National Center for Biotechnology Information, for nucleic acids (2004) National Center of Biotechnology Information, National Library of Medicine, Building 38A, Bethesda, MD 20894, USA.
PHP (2004).
Suse Linux (2004).
United States Patent and Trademark Office (USPTO) 2003 Performance and Accountability Report, (2003), mail stop USPTO Contact Centre, P.O. Box 1450, Alexandria, VA 22313–1450, USA.
World Intellectual Property Organization (2004) 34, Chemin des Colombettes, P.O. Box 18, CH-1211 Geneva 20, Switzerland.