



GQ-PAT: GenomeQuest's Patent Sequence Database

Over the past two decades, the amount of data recorded in general biological sequence databases such as GenBank has grown exponentially. But recently the most critical sequence data, information about patented sequences, has been growing even faster. This document identifies the data sources behind GQ-PAT, GenomeQuest's unique patent sequence database, and describes its composition and structural features.

As GenomeQuest's non-redundant patent sequence database, GQ-PAT contains nucleotide and protein sequences listed in the patent literature.

Data Sources

With an exhaustive QC process, the unique GQ-PAT database is continuously built and released every two weeks. Data from various patent offices, publicly available sequence databases and web resources is processed in our proprietary automated pipeline to make the sequences and all of the annotations searchable and browsable in GQ-PAT. The major data sources for GQ-PAT include:

- **USPTO.** GQ-PAT has the complete USPTO patent archive (patents and published applications containing sequences) based on daily data feed of patents and published applications from USPTO. Documents with sequence listings are identified through a set of intricate rules.
- **INPADOC / EPO.** International Patent Documentation Center (INPADOC) is an international patent collection. The database is produced and maintained by the EPO. It contains patent families and legal status information, and is updated weekly. GQ-PAT fetches patent data from

the EPO weekly releases. This source collects bibliographic data from over 70 countries and legal status data from more than 40 patent authorities. Given the diverse content quality from our various sources, we regularly re-annotate all of our patents with data from INPADOC services to make sure key data fields are filled in. For example, Patent Title, Abstract, Publication Date, Priority Dates, Patent Assignee, etc.

- **WIPO / PCT.** Data fetched weekly contains sequence listings electronically filed with WIPO. An independent ongoing process extracts sequences from WO paper-only applications, converts images into a searchable format with Optical Character Recognition (OCR) software.
- **Public Databases.** Updates from patent divisions of Genbank, EMBL and DDBJ sequence databases are fetched weekly and reformatted so that critical data fields are extracted in a searchable format.

The Timeliness of the GQ-PAT Database Data:

Data Source	Source to GQ (Days)	Update Frequency	Earliest Year Coverage since
USPTO	7	2 weeks	1981
EPO	7	2 weeks	1979
WIPO	7	2 weeks	1980

GQ-PAT Statistics by major patent authorities (as of 2008.09.6): Visit our website:

http://www.genomequest.com/genomequest_sequence_patent_database.html for the most up to date statistics.

Country Code	Earliest Year of document	No. of documents	No. of sequences
US	1981	104,113	64,193,305
WO	1980	53,567	16,292,223
EP	1979	19,090	1,360,879
JP	1979	32,814	1,724,263

Data QC

There are many mistakes existing in the original data. The typical mistakes include sequence listing mal-format, typographical errors, incorrect numbering, miscounting, corrupted files, etc.

A set of intelligent rules is used in constructing GQ-PAT to detect these mistakes and fix them automatically whenever possible or manually when necessary without compromising GQ-PAT database content accuracy, and timeliness.

Note:

- To remove redundancy, a sequence search is only performed against the sequence listing of the representative patent from each equivalent set, thus the hit list of patent numbers will not contain the other members of an equivalent set, but these members are listed in the "Patent Family" field of the representative patent. Text search is performed against all members of an equivalent set.
- When picking the representative patent from an equivalent set, the US patent is preferred over other countries, a granted patent is preferred over an application, and an earlier patent is preferred over a later patent.

Very often, non-machine readable WIPO applications and their machine-readable US (or other national) applications are filed concurrently. In this case, our automated system uses Optical Character Recognition (OCR) software to acquire and retrieve sequence information, followed by human editing to correct remaining errors with guidance from other national machine-readable sources. This strict QC process delivers sequences of high quality.

Patent Family and Equivalent Construction

A patent family is re-constructed based on the INPADOC definition of a patent family: "All the documents directly or indirectly linked via a priority document belong to one patent family".

More detailed information on this subject can be found at,

<http://gb.espacenet.com/espacenet/gb/en/help/161.htm>

The INPADOC family definition is sometimes very lenient and may include many remotely related patents and can cause particular problems with patents with sequences. In order to indicate closer relationship among patents (e.g. the same patent from different issuing authorities), GQ-PAT marks out as "equivalents", those patents sharing the same sequence listing. Specifically, 2 patents in GQ-PAT are defined as equivalent if, and only if:

1. They belong to the same family, and
2. They have the same number of sequences in the listing, and
3. The total sequence length is within 5% variation

Both family and equivalent document numbers are displayed in "Patent Family" field. The document numbers are sorted on the publication date with the earliest date first.

Description of Fields in GQ-PAT

Note:

Certain fields specific to each sequence within each patent are listed under “**Per Sequence Fields**” towards the end of the table.

Field Name	Description	Displayed / Searchable
Abstract	Patent abstract, summary, and related information section in US patents	Yes / Yes
Application Number	Application number	Yes / Yes
Application Pub Date	Application Publication date	Yes / Yes
Claimed SEQ ID NO	List of claimed SEQ ID numbers in this patent. Note: 1) This is currently only available to US patents; 2) This is generated automatically by our text mining algorithm, to recognize the SEQ ID NOs mentioned in the claim text.	Yes / No
Claims	Text of Claims	Yes / Yes
Database Name	Sequence database name, e.g. GQ-PAT	Yes / Yes
Date of Entry	Date of entry into the GQ-PAT database	Yes / Yes
Detailed Legal Status	Contains the description and the date of the latest kind code of a patent document	Yes / Yes
Earliest Priority Date	The earliest priority date	Yes / Yes
International Classification	International classifications, European classifications, and US classifications	Yes / Yes
Kind Code	Kind Code	Yes / No
Legal Status	Legal status of a patent document (Granted or Application)	Yes / Yes
Number of Claims	Number of claims in this patent	Yes / Yes
Number of SEQ	Number of sequences in this patent, it is the sum of nucleotide and protein sequences followed by the numbers of nucleotide and protein sequences in parenthesis. When performing filtering on this field, the total number of sequences is used	Yes / Yes
Patent Assignee	Patent assignee. If not available the correspondence address is used.	Yes / Yes
Patent Family	Contains two sections. 1) Equivalents according to GenomeQuest criteria; 2) INPADOC family information	Yes / Yes
Patent Family ID	The earliest priority number of a family of Patent Numbers (PNs). It is a hidden field and used to group PNs by family.	No / No*
Patent Filing Date	Patent filing date	Yes / Yes
Patent Inventors	Patent inventors	Yes / Yes
Patent Title	Patent title	Yes / Yes
PCT Pub Date	Date the patent is published by WIPO. Only available to the PCT patent that enters US national phase	Yes / Yes
PCT Pub Number	PCT publication number and PCT Number. Only available to the PCT patent that enters US national phase	Yes / No
PCT Related Dates	PCT filing date and 371 date and 102(e) date. Only available to the PCT patent that enters US national phase	Yes / No

Per Sequence Fields		
Field Name	Description	Displayed / Searchable
Equivalent Classification	Used to associate corresponding sequences in equivalent patent documents	No / No*
Features	Sequence features listed in the original sequence listing	Yes / No
Identifier	Sequence ID. usually composed of patent number, a dash, and seq id no. e.g. US20040181033-0147	Yes / Yes
Molecule Type	Sequence molecular type	Yes / No
Organism	The organism name (normalized) derived from what is mentioned in the document	Yes / Yes
Patent SEQ ID NO	Sequence's SEQ ID NO	Yes / No
Patent Sequence Location	Contains: "claim: x,y,..." if the Seq ID NO appears in Claims x,y...; "disclosure" if Seq ID NO does not appear in claims; "TBD" if GQ-PAT is currently missing claims text for this patent	Yes / Yes

* For future use



Corporate Headquarters

GenomeQuest, Inc.

1700 West Park Drive
Suite 260
Westborough, MA 01581
Phone: 508 616 0100
Fax: 508 616 0110

www.genomequest.com

contact-us@genomequest.com

France

GenomeQuest SA
147, Avenue Paul Doumer
92500 Rueil Malmaison France
Phone: +33 (0)1 41 96 80 30
Fax: +33 (0)1 41 96 80 31

European Sales

GenomeQuest
62 Main Road
Long Hanborough
Oxfordshire OX29 8BD
UK
Phone: +44 7900 69 89 28
Fax: +44 1993 88 19 33