Trilateral Project DR2
Biotechnology

Trilateral Search Guidebook in Biotechnology

[Ver.1 – Publication]

Part I

26 April 2007

United States Patent and trademark Office
European Patent Office
Japan Patent Office

Trilateral Search Guidebook in Biotechnology

I. Introduction

- What is “Trilateral Search Guidebook in Biotechnology ”? -

The Trilateral Search Guidebook in Biotechnology is a non-binding reference manual comprised of two parts: the first part describing the databases/search tools most commonly used in Trilateral Office, and the second part providing some representative search strategies used by each office for common types of claim sets presented in biotechnology-related applications.

This Version 1 for publication of The Trilateral Search Guidebook in Biotechnology focuses on common databases and sequence search algorithms in its first part.

- Disclaimer -

A patent examiner is responsible for reviewing prior patent documents, both domestic and foreign, and other printed literature related to an application's subject matter during the examination process. This review, called the search, is performed by consulting the appropriate patent classification schemes, and classified and text searches of other patent document databases, and/or any other printed media (also known as "non-patent literature" or NPL), which might disclose the invention disclosed/claimed in a pending application for a patent. This search may include the use of various search tools or methodologies in the process. Once the search is performed, the examiner determines the patentability of a claimed invention in light of the prior art uncovered by this search.

When determining the appropriate field of search for an invention, the examiner must consider three sources of information: domestic patent documents, foreign patent documents, and NPL. Patent examiners are not required to search every listed resource in the examination of every application. However, an examiner should not eliminate any resource from consideration unless the examiner can justify to a reasonable certainty that no more pertinent references will be found in a further search. The examiner will determine the most appropriate resources to search for each application by relying upon their professional judgment and assessment of the disclosed and claimed subject matter in the application under consideration. What the search guidebook provides is a compilation of the resources found to most frequently discover the most relevant prior art disclosures for the particular subject matter. Searching is highly dependent upon the experience and job knowledge of the individual performing and/or reviewing the search. Prior art to be considered in a search can slightly vary depending on the national patent law.

This aspect can also influence the sources used by examiners to perform their search.

II. Main Subject

A. Search Database / Search Tools

1. Patent literature, Bibliography & Reference databases and Search tools

[USPTO]

a. Patent literature and Patent related databases

Main USPTO Patent Data Resources	
Resource Name	**Resource Description**
U.S .Patents Full-Text (USPAT)	○ Contains full text of most patents issued from 1971-present. ○ Updated weekly as patents are issued. ○Also includes minimal information (patent number, publication date, and current classification) on all other US patents dating back to 1790.
U.S. Optical Character Recognition Full-Text (USOCR)	○ Full text search of 2.5 million U.S. Patent Grants issued from 1920-1970. ○ Approximately 100,000 patents issued during 1971-1979 that were not electronically captured when the USPAT database was first created.
U.S. Pre-Grant Publication Full-Text (US-PGPUBS)	○ Contains full text of published U.S. patent applications (2001 to the present). ○ Updated weekly as Pre-Grant Publications are published.
Patent Abstracts of Japan (PAJ)	○ Contains English language abstracts of published unexamined Japanese patent applications, issued from 1976 to the present.
Derwent World Patents Index Abstracts (Derwent)	○ Contains English language abstracts of published patent documents from more than 40 patent-issuing authorities, often with family information. ○ Years of coverage vary, normally 1970-present.
	○ Contains published abstracts from the EPO, from Selected EPO member countries and from the World

European Patent Office Abstracts (EPO)	Intellectual Property Organization (WIPO). ○ Provides PDF images for the full text of patents from a number of countries. ○ Has text search capabilities in the title and abstract fields as well as retrieval by IPC and ECLA classification numbers.
JPO Industrial Property Digital Library http://www.ipdl.jpo.go.jp/homepge.ipdl	○ Database contains Japanese patent information since 1976. ○ Has text search capabilities in the title and abstract fields as well as retrieval by F and FI terms.

b. Search tools for Patent –Related Databases

EAST (Examiner's Automated Search Tool), a Windows ™-based client and WEST (Web Examiner's Search Tool), a browser-based interface with a Web-like look and feel are search tools used by USPTO personnel to search for U.S. and Foreign Patents and Pre-Grant Publications. EAST/WEST may be used to perform general patent database searches or to perform searches bound to specific document sections.

PLUS (Patent Linguistics Utility System)

PLUS is a query by example search system for U.S. patents from 1971 to the present. PLUS produces a list of the most closely related patents to the application searched. In fall 2006, PLUS will offer search capabilities for the Pre-Grant Publication database. PLUS also provides a list of classes and subclasses where the suggested patents are classified.

Benefits:

PLUS searches can be requested and received before the Examiner reviews the application.

Examiner can gain a better understanding of the application this way.

Saves Examiner time -- PLUS searches are conducted by library staff.

Image File Wrapper cases -- Receive results the same day.

Paper cases -- Receive the results within 24 hours.

Emerging Technologies -- PLUS is especially useful in areas of emerging technologies, areas without art-specific terminology, and in areas needing reclassification.

[JPO]

a. Patent literature and Patent related databases

Main JPO Patent Data Resources	
Resource Name	**Resource Description**
JP Patents	○ Contains full text data of most patents issued from 1993- ○ Contains OCR full text data of most patents issued from 1986 to 1992. ○ Also includes minimal information (patent number, publication date, and current classification) on all other JP patents. ○ FI, F-term, keyword and full text (whole document) search is available
JPO Industrial Property Digital Library http://www.ipdl.jpo.go.jp/homepge.ipdl	○ Database contains Japanese patent information since 1885. ○ Has text search capabilities in the title, claim and abstract fields as well as retrieval by FI and F terms.
Foreign Patent database (Internal databases)	○ Contains CA, CH, DE, EP, FR, GB, US and WO documents ○ Contains gazette image/text (only English) data ○ ECLA, USC and full-text* search is available * full-text search (English): whole document of EP, US full-text search (Japanese): Japanese abstract of EP, US
WPI (World patent index) (Internal and commercial databases)	○ Contains English language abstracts of published patent documents from more than 40 patent-issuing authorities, often with family information. ○ Years of coverage vary, normally 1970-present.
Patent Electronic Business Center U.S. Patent Full-Text and Full-Page Image Databases http://www.uspto.gov/patft/index.html	○ Database contains U.S. patent information Issued Patents : full-text since 1976, full-page images since 1790 ○ Published Applications : published since 15 March 2001
European Patent Office Abstracts (EPO)	○ Contains published abstracts from the EPO, from Selected EPO member countries and from the World Intellectual Property Organization (WIPO). ○ Provides PDF images for the full text of patents from a number of countries. ○ Has text search capabilities in the title and abstract

	fields as well as retrieval by IPC and ECLA classification numbers.

b. Search tools for Patent –Related Databases

Cluster search (domestic patent DB/ foreign patent DB/ Non-patent-literature DB)

Cluster search is a search tool used by JPO patent examiner to search for JP, Foreign Patents and Non-patent-literature. It is used to perform general patent database searches or to perform searches bound to specific document sections.

Cluster search is done by single query (FI, F-term, Computer Software-term, ECLA, USC, Keyword, Full-text) for three databases (domestic patent DB/ foreign patent DB/ Non-patent-literature DB). And importing the document list from WPI to Cluster search window, we can see the original documents

[EPO]

a. Patent literature and Patent related databases

Search collections at the EPO are in most cases accessible via the EPOQUE integrated search environment.

Main EPO Patent Data Resources	
Resource Name	**Resource Description**
EPODOC (EPO documentation)	○ Contains references to patent documents which compose the systematically classified search documentation of the European Patent Office. ○ The documents consist of published applications, granted patents as well as classified and cited non-patent literature (XP documents). ○ The EPODOC database corresponds to the DocDB database which is the internal EPO master file used for the management of the search documentation. The bibliographic data (i.e. the publication, application and priority numbers and dates, the inventors and applicants data, the IPC classes, the EPO internal classification schemes ECLA and when available the US and JPO classification schemes, citation data, abstracts and title) are available for patent documents of most of the countries of the world. This database contains more than 55 million records that are usually linked to the image of the corresponding document. It covers more than

	80 patent offices and organizations worldwide and is widely considered as the world patent master database. A simplified access to this data is available via esp@cenet for the external public.
TXT (Full-text databases)	○ Represents the collection of the patent full-text databases available at the EPO. This collection contains more than 17 million records covering following countries and organizations: AU, AT, BE, CH, DE, EP,FR, GB, US and WO. An similar collection exists for major NPL collections.
WPI (World patent index)	○ Includes the bibliographic data and abstracts of patents from more than 40 countries as well as European and PCT patents.
Patent Abstracts of Japan (PAJ)	○ Contains English language abstracts of published unexamined Japanese patent applications, issued from 1976 to the present.

2. Non Patent literature, Bibliography and Reference databases

[USPTO , JPO, EPO]

The following resources are primarily used as bibliography and reference databases by USPTO, JPO and EPO biotech examiners. This list is non-exhaustive and does not include all other NPL sources used in other fields. Some Trilateral Offices may have additional sources at their disposal.

BIOSIS	○ Biological and biomedical Abstracts ○ Includes citation references to research articles and other source publications that include primary journals, books, reference texts, monographs, papers and abstracts from meetings of interest to the biological community and research reports. ○ The subjects covered by the BIOSIS database concerns the following technical fields: . Botany . Zoology . Microbiology . Agriculture . Pharmacology . Biochemistry . Biophysics . Ecology . Bioengineering . Experimental and clinical medicine
MEDLINE	○ Abstracts of the US National Library of Medicine ○ The NLM's premier bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, the health care system, and the preclinical sciences. ○ Contains bibliographic citations and author abstracts from more than 4,000 biomedical journals published in the United States and 70 other countries. ○ Contains over 12 million citations dating back to the mid-1960's. Coverage is worldwide, but most records are from English language sources or have English abstracts.
Science Direct Collection (Elsevier)	○ Contains over 25% of the world's science, technology and medicine full text and bibliographic information. ○ Apart from online reference works, handbooks and book series Science Direct offers a rich journal collection of over 2,000 titles. ○ In addition, the Back files program offers the ability to search a historical archive of over 6.75 million articles directly from your desktop, back to Volume 1, Issue 1. The collections contain 4 million articles prior to 1995, and 2.75 million articles from after 1994.

CA,CAPLUS (Chemical Abstracts)	○ Abstracts of the chemical literature back to 1907. ○ File CAplus covers the same material as File CA but contains early access to bibliographic information and patent family information for the documents in the process of being indexed. ○ Useful in text, inventor and registry number searching and a broad range of areas including chemical engineering, electrochemistry, biochemistry, and applied, macromolecular, organic, physical, inorganic, environmental and analytical chemistry.
MRCK (Merck Index Online)	○ 11th edition of The Merck Index, with descriptions of important chemicals, drugs, and biological, agricultural and natural products. ○ Coverage Late 19th century – present.
DIALOG	○ Dialog offers access to more than 15 Terabytes (Tb) of searchable content, including articles and reports from thousands of real-time news feeds, newspapers, broadcast transcripts and trade publications; In-depth repositories of scientific and technical data, patents, trademarks and other intellectual property data. ○ Additional content areas include government regulations, social sciences, food and agriculture, reference, energy and environment, chemicals, pharmaceuticals and medicine.
Registry (CAS Registry)	○ Chemical structure and dictionary database that contains unique substance records. ○ Protein and nucleic acid sequences are retrieved using codes for the amino acids and nucleotides. ○ Useful for searching sequences, registry numbers and chemical compounds/structures. ○ Coverage 1957 to date.

STN International	○ The online scientific and technical information service providing access to current and archival information from over 200 scientific, technical, business, and patent databases with links to the literature, patents, and chemical catalogs. ○ STN is operated by Chemical Abstracts Service (CAS), a division of the American Chemical Society, in North America; FIZ Karlsruhe (FIZ-K), in Europe; and Japan Science and Technology Agency, Information Center for Science and Technology (JST), in Japan. STN databases cover a wide range of scientific and technical topics. ○ STN allows access to the CAS Registry® database of chemistry drawings. ○ STN on the Web offers web-browser access to the complete content and functionality of STN for information professionals and advanced end-users. ○ STN on the Web combines the STN command functionality with web-browser technology. ○ STN express allows database searching via command-line interface.

3. Nucleic acid and Protein sequence repositories

The USPTO sequence repositories reside on an internal system called the Automated Biotechnology Sequence Search System (ABSS System).

The JPO sequence repositories reside on an internal system called Genetic Information Analysis Services (GIAS).

The EPO sequence search is performed in the EPO-EBI Environment.

(1) Type of databases and their content

a. EMBL / GenBank / DDBJ

The EMBL, GenBank or DDBJ Nucleotide Sequence Databases constitute USA, Europe's or Japanese primary nucleotide sequence resources. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications.

The database is produced in an international collaboration with EMBL, GenBank and DDBJ. Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis.

b. Uniprot

UniProt (Universal Protein Resource) is the world's most comprehensive catalogue of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR. UniProt is comprised of three components, each optimized for different uses. The UniProt Knowledgebase (UniProt) is the central access point for extensive protein information, including function, classification, and cross-reference. The UniProt Non-redundant Reference (UniRef) databases combine closely related sequences into a single record to speed searches.

c. Geneseq

GENESEQ is the most comprehensive source of information on nucleic and amino acid sequences from Derwent WPI World patent index. It contains sequence information taken from patent applications and granted basic patents (all nucleic acids which are 10 or more bases in length, amino acids which are four or more residues in length plus all PCR primers and probes, of any length are included in the database). Sequences from patents from 1981 onwards are included in GENESEQ

d. PDB

The Protein Data Bank (PDB) is an archive of experimentally determined three-dimensional structures of biological macromolecules, serving a global community of researchers, educators, and students.

The archives contain atomic coordinates, bibliographic citations, primary and secondary structure information, as well as crystallographic structure factors and NMR experimental data. The PDB Newsletter and CD ROM are published quarterly.

(2) The search algorithms

a. The Smith-Waterman algorithm

The Smith and Waterman algorithm is a biological sequence comparison algorithm. It runs a search on cluster, using single and parallelized versions of the software. It allows an rigorous search in a reasonable computational time. It is an exhaustive algorithm, which is recognized as the most sensitive sequence comparison method available in Blast, Fasta and it. As a consequence, it is capable of identifying hits in cases where Blast and Fasta fail and also reports fewer false-positive hits.

b. FASTA

FASTA stands for FAST-All, reflecting the fact that it can be used for a fast protein comparison or a fast nucleotide comparison. This program achieves a high level of sensitivity for similarity searching at high speed. This is achieved by performing optimized searches for local alignments using a substitution matrix, in this case a DNA identity matrix.

The high speed of this program is achieved by using the observed pattern of word hits to identify potential matches before attempting the more time consuming optimized search. The trade-off between speed and sensitivity is controlled by the ktup parameter, which specifies the size of the word. Increasing the ktup decreases the number of background hits. Not every word hit is investigated but instead it initially looks for segment's containing several nearby hits. This program is much more sensitive than blast programs, which is reflected by the length of time required to produce results. Fasta produces optimal local alignment scores for the comparison of the query sequence to every sequence in the database. The majority of these scores involves unrelated sequences, and therefore can be used to estimate lambda and K values. These are statistical parameters estimated from the distribution of unrelated sequence similarity scores. This approach avoids the artificiality of a random sequence model by employing real sequences, with their natural correlations.

c. BLAST

BLAST is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences. It emphasizes speed over sensitivity. This emphasis on speed is vital to making the algorithm practical on the huge genome databases currently available.

d. GenePast / Fragment Search

GenePast does not produce local alignments. Instead it reports a best-fit alignment where the largest possible number of residues of the smallest sequence is aligned to the largest sequence with the minimal number of errors. GenePAST algorithm uses an approximate string-matching (or purely mathematical) approach.
This algorithm try to find an alignment between 2 sequences with edit cost at most k, i.e. can we transform part of text string to pattern string using at most k additions, deletions, and substitutions.