

STN[®]

All about WIPO

A comparative review of sources of
WIPO sequence data

Robert Austin – FIZ Karlsruhe

Agenda

- Public sources of WIPO sequence data
- An introduction to PCTGEN
- The importance of value-added databases
- Comparisons and conclusions

See also *Biosequence searching in the patent literature* (detailed database content comparison and search example overview):

http://www.stn-international.com/training_center/bioseq/bioseqpat.pdf

Note: all facts & figures quoted in this presentation are current as of February 5th, 2007.

Public sources of WIPO sequence data

3

- International Nucleotide Sequence Database Collaboration (INSDC)
 - www.insdc.org
- EMBL-EBI EPO Protein Database
 - www.ebi.ac.uk
- WIPO/PCT Published Application Sequence Listings download
 - www.wipo.int/pct/en/sequences/listing.htm

International Nucleotide Sequence Database Collaboration (INSDC)

4

- NCBI/EMBL/DDBJ collaboration (Genbank)
- Direct submissions (>77% with no references)
- Information as given by the submission author
- Patent division data from USPTO, EPO and JPO
- 66.8 million nucleotide sequence records
 - Including 3.66 million in the patent division, of which 1.46 million come from WIPO/PCT publications
- Updated daily
- 1982- present

The INSDC: <http://www.insdc.org/>



International Nucleotide Sequence Database Collaboration

- ◆ The International Nucleotide Sequence Databases (INSD) have been developed and maintained collaboratively between [DDBJ](#) , [EMBL](#) , and [GenBank](#) for over 18 years.
- ◆ The INSDC advisory board, the [International Advisory Committee](#) , is made up of members of each of the databases' advisory bodies. At their most recent meeting, members of this committee unanimously endorsed and reaffirmed the existing data-sharing policy of the three databases that make up the INSDC, which is stated below.
- ◆ Individuals submitting data to the international sequence databases should be aware of the [INSDC policy](#) .

How to submit data





- ◆ For full details of how to submit data to the databases, please select a collaboration partner.
- ◆ [DDBJ](#) , [EMBL](#) , [GenBank](#)
- ◆ The INSDC Feature Table Definition Document is available [here](#) .

The INSDC collaboration and mutual exchange of data, is for **nucleotide sequences only**.

WIPO nucleotide patent sequence data is contributed by USPTO, EPO and JPO – but not directly from WIPO.



EMBL-EBI additionally provides a collection of WIPO peptide data

	WIPO Nucleotide	WIPO Peptide	Home Page
EMBL			www.ebi.ac.uk
NCBI			www.ncbi.nlm.nih.gov
DDBJ			www.ddbj.nig.ac.jp

WIPO sequence data at EMBL-EBI comes from a very specific source

- WIPO nucleotide sequence data in the EMBL Nucleotide Database is provided by the PCT International Search Authorities (ISAs)
 - Submitted by PCT applicants for search and/or preliminary examination under PCT Rule 13*ter**
 - Does **not** form part of the formal PCT application
- Similarly, some ISAs send their WIPO peptide sequences to the EPO, who incorporate it into the protein database they provide to EMBL-EBI
- Although more comprehensive than NCBI or DDBJ, the EBI collection still does not represent an authoritative WIPO sequence database

* See: <http://www.wipo.int/pct/en/texts/rules/r13ter.htm>

EMBL-EBI SRS: <http://srs.ebi.ac.uk/>

8

WIPO Nucleotide and Peptide sequences are included in the EMBL Nucleotide Database, and the Patent Proteins collection respectively.

The EPO Patent DNA sequence database, and the EPO, JPO and USPTO protein databases may also be searched separately.

Select Databanks to Search - Mozilla Firefox

File Edit View History Bookmarks Yahoo! Tools Help

http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz

STN International CRS STN/CAS Home STN on the Web CRS Summary Sheets CRS Workshop Schedule DWPI Reference INPADOC Home USPTO search Esp@cenet

EMBL-EBI EB-eye Search All Databases Enter Text Here Reset Advanced Search

Databases Tools Groups Training Industry About Us Help Site Index

Quick Search Library Page Query Form Tools Results Projects Views Databanks HELP Job Status

Reset

Search Options

1. Select the databanks you want to search
2. Enter your search terms in the Quick Search box, or choose a query form from below

Standard Query Form
Extended Query Form

You can browse through all the entries in any databanks. First, select the databanks you want to browse, then click:

Browse Entries

Available Databases

Expand all Collapse all

- Literature, Bibliography and Reference
- Gene Dictionaries and Ontologies
- Nucleotide sequence databases
 - EMBL
 - IPD-KIR
 - EMBL (Coding Sequences)
 - EMBL ID/Accession Mapping
 - Patent DNA
 - EMBL (Contig)
 - Genome Reviews
 - EMBL MGA
 - IMGT/LIGM-DB
 - EMBL (Contigs expanded)
 - RefSeq Genome
 - IMGT/HLA
 - EMBL (Annotated Cons)
 - LiveLists
- Nucleotide sequence databases - subsections
 - EMBL (Updates)
 - EMBL (Release)
 - EMBL (Whole Genome Shotgun release)
 - EMBL (Whole Genome Shotgun updates)
 - EMBL (Contig updates)
 - EMBL (Contigs expanded release)
 - EMBL (Annotated Cons release)
 - RefSeq Genome (Updates)
 - EMBL (Whole Genome Shotgun)
 - EMBL (Contig release)
 - EMBL (Contigs expanded updates)
- Nucleotide related databases
- UniProt Universal Protein Resource
- Other protein sequence databases
 - Active protein sequence databases
 - Patent Proteins
 - IPI
 - Refseq Proteome (Release)
 - RefSeq Proteome (Updates)
 - EPO Proteins
 - IPI History
 - SWISSCHANGE
 - JPO Proteins
 - MHCBN
 - RefSeq Proteome
 - USPTO Proteins
 - BCIEP
 - Deprecated Protein Databases
 - Swall(SPTR)
 - PIR
 - RemTrEMBL
- Protein function, structure and interaction databases
- Enzymes, reactions and metabolic pathway databases
- Mutation and SNP databases

Tips

- bookmark this link to return to your project
- Linking to SRS?
- Please read our Linking to SRS guide for important information regarding linking to our SRS server.

Entry Information

Entry from: [EPO](#)
[Proteins](#)

Entry Options

Launch analysis tool:

NCBI BLASTP

Launch

Link to related
information:

Link

Save entry:

Save

View:

Printer Friendly

Go to: [General](#) [Description](#) [References](#) [Sequence](#)

General Information

Accession #	A59671
SRS Entry ID	EPO_PRT:A59671
Molecule Type	linear protein
Sequence Length	361
Entry Division	UNC
Entry Data Class	PAT
Sequence Version	A59671.1
Creation Date	06-MAR-1998
Modification Date	31-MAY-2006
UniParc	UPI000002E33C

Description

Description	Sequence 2 from Patent WO9704106.
Keywords	.
Organism	unidentified
Organism Classification	unclassified sequences.

References

1. Scarcez, T.; Van, B. A.;
NEW SACCHAROMYCES CEREVISIAE PECTINASE SEQUENCES AND HETEROLOGOUS EXPRESSION SYSTEMS
DERIVED THEREFROM
Patent number [WO9704106-A/2](#), 06-FEB-1997. INNOGENETICS NV (BE).

Features

Key	Location	Qualifier	Value
source	1..361		
		organism	unidentified
		mol_type	protein
		db_xref	taxon:32644

Sequence

Characteristics **Length:** 361 AA

Sequence

```
>epo_prt|A59671|A59671 Sequence 2 from Patent WO9704106.
MISANSLLIISTLCAFAIATPLSKRDSCTLTGSSLSLSSLSVVKKCSSIVIKDLTVPAGQTLT
LTGLSSGTTVTFEGTTFQYKEWSGPLISISGSKI SVVGASGHTIDGQGAKEWWDGLGDSG
KVKPKFVKLALTGT SKVTGLNIKNAPHQVF SINKCSDLTISDITIDIRDGDSAGGHNTDG
FDVGSSSNVLIQGCTVYNQDDCIAVNSGSTIKFMNNYCYNGHGI SVGSVGGRS DNTVNGF
IAPNNHVTNSDNCLEIKVYECACGCVMMVNRISNKTSCIEKSCVYVRCVYINSKPMCTAT
```

EMBL WIPO peptide sequence records, like this one, are **not** available at the NCBI or DDBJ.

EMBL patent sequence records have minimal searchable patent bibliographic and text information.

Reset Previous Entry Entry 5 of 6 from Query 1 Next Entry

Entry Information

Entry from: [EMBL](#)

Entry Options

Launch analysis tool:

NCBI BLASTN

Launch

Link to related information:

Link

Save entry:

Save

View:

Printer Friendly

Go to: [General](#) [Description](#) [References](#) [Sequence](#)

General Information

Primary Accession #	A59674
Accession #	A59674
SRS Entry ID	EMBL:A59674
Molecule Type	linear unassigned DNA
Sequence Length	27
Entry Division	UNC (<i>Unclassified</i>)
Entry Data Class	PAT (<i>Patent</i>)
Sequence Version	A59674.1
Creation Date	06-MAR-1998
Modification Date	06-MAR-1998

Description

Description	Sequence 5 from Patent WO9704106.
Keywords	;
Organism	unidentified
Organism Classification	unclassified sequences.

References

1. Scarcez,T.; Van,B.A.;
NEW SACCHAROMYCES CEREVISIAE PECTINASE SEQUENCES AND HETEROLOGOUS EXPRESSION SYSTEMS DERIVED THEREFROM
Patent number [WO9704106](#)-A/5, 06-FEB-1997. INNOGENETICS NV (BE).

Features

Key	Location	Qualifier	Value
source	1..27	organism	unidentified
		mol_type	unassigned DNA
		db_xref	taxon:32644

Sequence

Characteristics **Length:** 27 BP, **A Count:**7, **C Count:**8, **G Count:**4, **T Count:**8, **Others Count:**0

Sequence
 >[embl](#)|A59674|A59674 Sequence 5 from Patent WO9704106.
 tccttctagattaacagcttgcaccag

EMBL WIPO nucleotide sequence records, like this one, are typically also available at NCBI and DDBJ.

The EMBL-EBI SRS interface provides a searchable Patent Number field, and full-text links to Espacenet.



Differences for A59674 11-MAR-1998 / 19-JUN-2006

Lines unchanged Lines removed Lines inserted

```
ID A59674; SV 1; linear; unassigned DNA; PAT; UNC; 27 BP.
ID A59674 standard; DNA; UNC; 27 BP.
XX
AC A59674;
XX
NI e1260534
XX
DT 06-MAR-1998 (Rel. 54, Created)
DT 06-MAR-1998 (Rel. 54, Last updated, Version 1)
DT 06-MAR-1998 (Rel. 54, Last updated, Version 0)
XX
DE Sequence 5 from Patent WO9704106.
XX
KW .
XX
OS unidentified
OC unclassified sequences.
OC unclassified.
XX
RN [1]
RA Scarcez T., Van B.A.;
RT "NEW SACCHAROMYCES CEREVISIAE PECTINASE SEQUENCES AND HETEROLOGOUS
RT EXPRESSION SYSTEMS DERIVED THEREFROM";
RL Patent number WO9704106-A/5, 06-FEB-1997.
RL INNOGENETICS NV (BE).
XX
FH Key Location/Qualifiers
FH
FT source 1..27
FT /organism="unidentified"
FT /mol_type="unassigned DNA"
FT /db_xref="taxon:32644"
XX
SQ Sequence 27 BP; 7 A; 8 C; 4 G; 8 T; 0 other;
tccttctaga ttaacagctt gcaccag
//
```

EMBL provides historical version information for WIPO nucleotide sequence records. Often a record is not the same as the one which originally entered the database.

It can take months for WIPO data to enter EMBL databases. This typical EMBL WIPO record took 13 months from PCT publication to database entry.

The green shaded lines were inserted in June 2006.

EMBL-EBI provides a disclaimer from the European Patent Office (EPO)

“The purpose of the [EMBL] patent sequence database is to allow scientists and practitioners easy access to and search in published patent sequence information. **No guarantee is given as to the completeness and accuracy of the database**, in particular the conformity of the sequences in the database with the original publication where the sequence was first disclosed to the public or with the original disclosure in the respective patent application. All implied warranties including but not limited to the implied warranties of merchantability and fitness for a particular purpose are hereby disclaimed. In the case of discrepancies between information contained in the sequence database, on the one part, and the original publication or, as the case may be, the original disclosure in the respective patent application, on the other part, the latter prevails. For complete information on the bibliographic data, status and subject-matter of the relevant patent applications, users are referred to the official registers, publications, databases and other sources of information of the competent Patent Offices.”

See: http://www.ebi.ac.uk/embl/Documentation/Release_notes/relnotes54/54.html

WIPO/PCT Published Application Sequence Listings download

- Since 2001 this service provides access to sequence listings published in electronic form as a formal part of a PCT application
 - Under the *PCT Administrative Instructions Section 801(a)**, introduced Jan 11th, 2001
- There are 2,000+ PCT listings available, representing over 4.7 million sequences
- Critically, over 80% of these listings are unavailable to search at EMBL-EBI

* See: http://www.wipo.int/pct/en/texts/ai.html#_801

WIPO sequence listing download:
<http://www.wipo.int/pct/en/sequences/listing.htm>



Only WIPO sequence listings submitted and published in electronic form under the "PCT Administrative Instructions, Section 801(a)" are available for download.

Home > IP Services > PatentScope

Published Nucleotide and/or Amino Acid Sequence Listings Contained in Published PCT Applications

WinZIP 8.0 Compressed Sequence Listings (.txt files)

This data is also available for bulk download via anonymous ftp from ftp://ftp.wipo.int/pub/published_pct_sequences/.

Publication Date : February 01, 2007

WO Number	Size (Raw/Compressed)	Action	Applicant
WO07/012614	10.2 KB/2.56 KB	[Download]	PIERRE FABRE MEDICAMENT
WO07/013190	483 B/420 B	[Download]	NATIONAL INSTITUTE OF ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY
WO07/013709	1.24 KB/645 B	[Download]	
WO07/012119	1.14 MB/312 KB	[Download]	
WO07/013264	4.30 KB/1.28 KB	[Download]	
WO07/013265	5.60 KB/1.74 KB	[Download]	
WO07/013343	1.83 KB/726 B	[Download]	
WO07/013352	1.87 KB/691 B	[Download]	
WO07/012576	183 KB/31.5 KB	[Download]	BASF PLANT SCIENCE
WO07/012605	1.43 KB/597 B	[Download]	DRO BIOSYSTEMS, S.L.
WO07/012671	38.1 KB/8.33 KB	[Download]	CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE (CNRS)

It is cheaper for PCT applicants to choose this electronic publication route, if their sequence listings are greater than 400 pages in length.

- PATENTSCOPE**
- Home
- About Patents
- Patent Search
- PCT Resources
- PCT Electronic Filing
- Patent & Technical Information
- Statistics
- Law of Patents
- Meetings
- Contact

- RELATED LINKS**
- International Patent Classification
- Natural Language IPC Search
- Standards & Documentation

- E-NEWSLETTERS**
- Subscribe to receive e-mails of news and updates on WIPO's activities regarding patents and the PCT

WIPO/PCT Published Application Sequence Listings download (cont.)

- WIPO sequences are not searchable at the site – they can only be downloaded
- The sequence listings are theoretically submitted by PCT applicants as plain-text WIPO ST.25 standard fielded data
- PCT applicants frequently do not supply sequences in acceptable ST.25 format
- For example, in 2006, over 42% of the available listings are incorrectly formatted

An example of a sequence entry in a WIPO ST.25 format sequence listing

```

<210> 1
<211> 314
<212> DNA
<213> Corynebacterium glutamicum
<220>
<221> CDS
<222> (1)..(291)
<400> 1
    cca ccg atc tac ttc tcc cac gac cgc gaa gtt t
    Pro Pro Ile Tyr Phe Ser His Asp Arg Glu Val P
        1             5             10
    atg tgg ctg acc gca ggc gag tgg ggt gga cca a
    Met Trp Leu Thr Ala Gly Glu Trp Gly Gly Pro Lys Lys Gly Glu Glu
                20             25             30
    atc gtc acc aag act gtc cgc tac cgc acc gtc ggc gat atg tcc tgc      144
    Ile Val Thr Lys Thr Val Arg Tyr Arg Thr Val Gly Asp Met Ser Cys
        35             40             45
    . .

```

<u>Information</u>	<u>ST.25</u>
SEQ ID NO	<210>
Length	<211>
Type	<212>
Organism	<213>
Feature	<220>
Name/key	<221>
Location	<222>
Sequence	<400>

For further details about WIPO ST.25 format, visit:

<http://www.wipo.int/scit/en/standards/pdf/03-25-01.pdf>

PCTGEN

- Produced by FIZ Karlsruhe and WIPO
- Sequences submitted & published *electronically* as a formal part of PCT patent applications
- Publication number and date, patent applicant name(s) and the original publication title are provided for each sequence
- Sequence length, SEQ ID, organism name and molecule type are included for each sequence
- Updated weekly – within **24 hours** of publication
- 4.7 million sequence records
- August 2001 – present

A large proportion of the non-standard WIPO data is processed for PCTGEN

- For example, in 2006, 1,000 electronic format listings were published by WIPO
- A total of 428 of these listings do not comply with WIPO ST.25 format rules
 - Including non-ASCII text, special characters, missing mandatory headings, incorrectly used headings, files in PDF and/or TIF format
- A total of 284 of these problematic listings were successfully converted for PCTGEN
 - Representing 2.3 million sequence records

Relationship between PCTFULL and PCTGEN databases

AN ... PCTFULL

TI

PA

PI WO A1

AB

DETD

CLM

AN Protein PCTGEN

PI WO A1

SEQ 1

AN DNA PCTGEN

PI WO A1

SEQ 2

AN Peptide PCTGEN

PI WO A1

SEQ 3

PCTFULL = WIPO/PCT patent applications full-text on STN®

PCTGEN = WIPO/PCT patent application biosequences on STN

A typical PCTGEN record

```
L1 ANSWER 1 OF 1 PCTGEN COPYRIGHT 2007 WIPO on STN
AN 2006069200.16112 PRT (1) PCTGEN
TI Group B Streptococcus
PA Tettelin, Herve (2)
   Massignani, Vega
PI WO 2006069200 20060629 (3)
RLI US 2004-638943P 20041222; US 2004-640438P 20041230
ED 20060630
DT Patent
ORGN Streptococcus agalactiae (4)
SQL 302 (5)
SEQ
```

```
1 mflmplasll gnltvwhhkl heikipfsr ldilihlrpt lmlflpqitm
51 qiylslnksm lgamdsvsva gyfdqsdki rilftivsai ggvlprlss (6)
. . . .
251 atlsgavlyy intqmsvslv nyviqslvav tiyvgivfit kapviql1XX
301 Xn
```

FEATURE TABLE: (7)

Key	Location
VARIANT	299, 300, 301 Xaa = Any Amino

Sequences are typically added to PCTGEN within 24 hours of publication by WIPO.

Note: this PCTGEN sequence record is an example of one which is not currently present in DGENE, REGISTRY or EMBL-EBI.

A typical PCTGEN record (cont.)

- 1) Accession Number (AN). This includes the sequence (SEQ ID) number. For example, AN 2006069200.16112 is SEQ ID 16112 from WO2006069200.
- 2) PCT publication title for the overall invention.
- 3) Patent bibliographic information: Patent Assignee (PA), Publication Number (PN) and, where given, Related Application Number (RLN) and/or Application Number (AP).

A typical PCTGEN record (cont.)

- 4) Organism name (ORGN) providing the name of the species from which the sequence derives.
- 5) Sequence Length (SQL). Searchable/sortable.
- 6) The sequence (SEQ) represented with one letter codes (following WIPO standard ST.25). Non-standard nucleotides are indicated with N. Uncommon amino acids are indicated with X.
- 7) Feature table (FEAT) describing modifications and features of the sequence, as given by the patent applicant.

PCTGEN Original Sequence (SEQO)

=> D SEQO

L1 ANSWER 1 OF 1 PCTGEN COPYRIGHT
SEQO

```
cgctcgcagt ctgtgggccc tccgggaggc ggcggaggtc accgcgggga gaggggcggg      60
cgcagc   atg gca gcc tcc tta cgg ctc ctc gga gct gcc tcc ggt ctc      108
          Met Ala Ala Ser Leu Arg Leu Leu Gly Ala Ala Ser Gly Leu
                1                5                10

cgg tac tgg agc cgg cgg ctg cgg ccg gca gcc ggc agc ttt gca gcg      156
Arg Tyr Trp Ser Arg Arg Leu Arg Pro Ala Ala Gly Ser Phe Ala Ala
    15                20                25                30

gtg tgt tct agg tca gtg gct tca aag act cca gtt gga ttc att gga      204
Val Cys Ser Arg Ser Val Ala Ser Lys Thr Pro Val Gly Phe Ile Gly
                35

ctg ggc aac atg ggg aat cca atg gc
Leu Gly Asn Met Gly Asn Pro Met AL
                50                55
```

The original input format of a PCTGEN sequence is available for display using the **SEQO** display field.

Often the original format includes the PCT patent applicant's alignment of the nucleotide sequence coding region with the corresponding protein sequence.

Two key value-added STN databases provide extensive WIPO sequence data

- Thomson Scientific GENESEQ (DGENE)
- Chemical Abstracts Service REGISTRY
- Both databases are unique sources of WIPO sequence data, filling in most of the gap in coverage of EMBL-EBI and WIPO
- The content and editorial indexing policies of both GENESEQ* and REGISTRY were presented at *PIUG 2007 Boston Biotech*

* See: http://www.stn-international.com/training_center/bioseq/geneseq.pdf

GENESEQ™ (DGENE)

- Produced by Thomson Scientific
- Sequences from the **basic** patents of the 40 authorities of the Derwent World Patents Index®
- Bibliography, enhanced title, abstract, indexing and patent location provided for each sequence
- Patent Family and Legal Status display
- 8.4 million patent sequence records
 - Including 5.3 million WIPO sequence records
- Updated every two weeks
- 1981 - present

Some editorial insights regarding WIPO sequences indexed in DGENE

- On average 120 WIPO/PCT basic patents have sequences indexed into DGENE each week
- Of these, about 15-20 may have electronic listings available – the rest are keyed manually
 - Sequences are independently double-keyed with a guaranteed accuracy of 99.995% (1 in 20,000)
- About 15% of PCTs with electronic listings have extra sequences indexed from the specification
- Typically 1 or 2 documents per week will also have intellectually derived sequences indexed, based upon the wording of the patent claims

Source: Colin Williams, GENESEQ Editorial & Content Manager, Thomson Scientific

Relationship between DWPI patent family & DGENE sequence database

AN WPINDEX
 TI
 PA
 PI FR A1
 WO A1
 US A1
 US B2
 AB

AN Protein DGENE
 PI FR A1
 SEQ 1
 AB

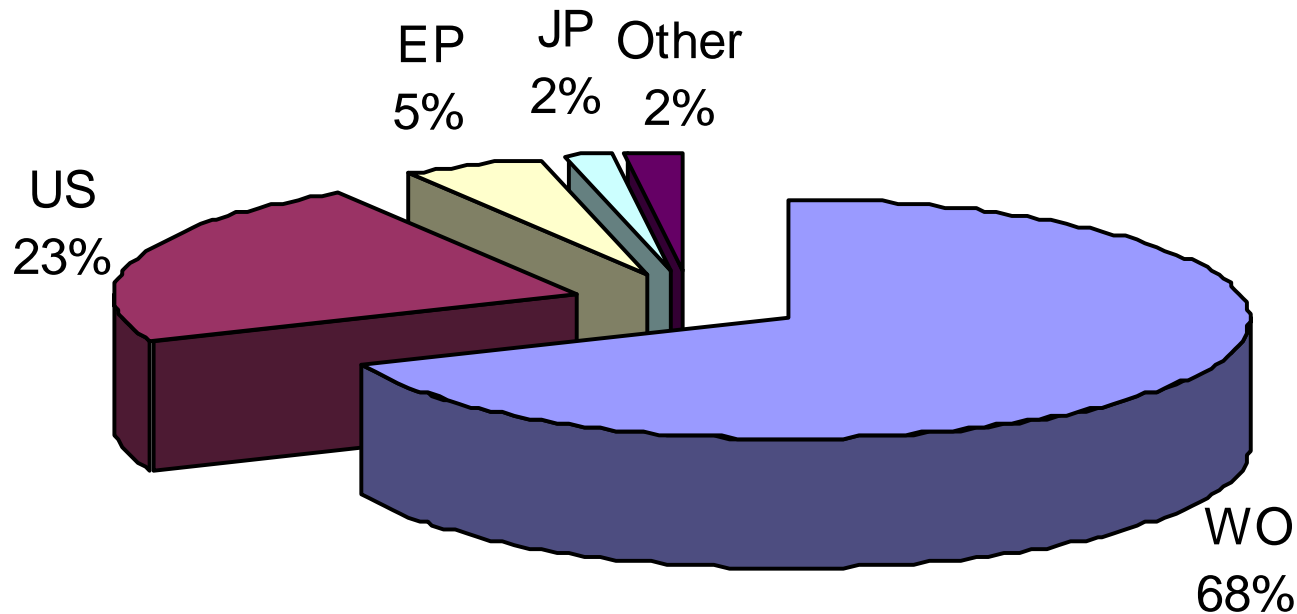
AN DNA DGENE
 PI FR A1
 SEQ 2
 AB

AN Peptide DGENE
 PI FR A1
 SEQ 3
 AB

WPINDEX = Derwent World Patents Index® on STN®

DGENE = GENESEQ™ on STN

DGENE sequences are indexed from DWPI basic patent publications



The majority of DGENE patent sequences are indexed from WIPO/PCT published patent applications (WO).

A typical DGENE sequence record for a WIPO patent sequence

```
L1 ANSWER 1 OF 1 DGENE COPYRIGHT 2007 THE THOMSON CORP ON STN
AN ABA00306 DNA DGENE
TI Use of peroxisome proliferator activated receptor delta modulator for
increasing e.g. atp-binding cassette protein activity and/or
expression
IN Willson T M
PA (SMIK) SMITHKLINE BEECHAM CORP.
PI WO 2002070011 A2 20020912
AI WO 2002-US3017 20020131
PRAI US 2001-266393P 20010202
PSL Example; Page 10
DED 09 DEC 2002 (first entry)
LA English
OS 2002-691739 [74]
DESC ABCA1 forward primer.
KW Primer; amplify; PCR; ATP-binding
transport; peroxisome proliferato
dyslipidaemia; diabetic dyslipida
syndrome; heart failure; hypercho
psoriasis; anorexia bulimia; anor
ORGN Homo sapiens
```

Each DGENE sequence record includes **full basic patent bibliography**, enhanced title, abstract, indexing and patent location, provided by Thomson.

Note: this DGENE sequence record is an example of one which is not present in REGISTRY, PCTGEN or EMBL-EBI (as of Feb 5th, 2007).

A typical DGENE sequence record for a WIPO patent sequence (cont.)

AB The sequences given in ABA00306-08 are primers and a probe used in the determination of ATP-binding cassette transporter 1 (ABCA1) activity and/or expression. ABCA1 activity and/or expression of ABCA1 transport is promoted by administration of a peroxisome proliferator activated receptor agonist for increasing ABCA1 activity and promoting cholesterol transport. Also for treatment of e.g. dyslipidaemia, diabetic dyslipidaemia, mixed dyslipidaemia, metabolic syndrome, heart failure, hypercholesterolaemia, atherosclerosis, arteriosclerosis, hypertriglyceridaemia, type II diabetes mellitus, type I diabetes, insulin resistance, hyperlipidaemia, inflammation, epithelial eczema, psoriasis, condition associated with lung, gut, regulation of appetite, obesity, anorexia bulimia and anorexia nervosa. The modulators of PPAR delta showed strong induction of ABCA1 expression, regulated cholesterol efflux from macrophages, and produced a 2-fold increase in cholesterol efflux to apolipoprotein (apo A1).

DGENE abstracts and indexing are prepared for each sequence record by Thomson Scientific experts.

SQL 26
SEQ
1 tgtccagtcc agtaatgggt ctgtgt

DGENE value-added indexing offers unique search refinement options

```
=> FILE DGENE
```

```
=> RUN BLAST L1 /SQP -F F
```

```
BLAST Version 2.2
```

```
. . . .
```

```
1492 ANSWERS FOUND BELOW EXPECTATION VALUE OF 10.0
```

```
. . . .
```

```
L2      RUN STATEMENT CREATED
```

```
L2      1492 MGGALLLALLLVSAAAAP . . . NMVPNFGGAVGRAIV/SQP.-F F
```

```
=> S L2 AND CLAIM/PSL AND PRY<2002
```

```
L3      874 L2 AND CLAIM/PSL AND PRY<2002
```

```
=> SOR L3 SCORE D
```

```
L4      874 SOR L3 SCORE D
```

```
=> D TRI OS ALIGN 1-
```

L1 is a sequence query previously uploaded via the STN Express 8.01 Sequence Query Upload Wizard.

Here a DGENE search has been limited to claimed sequences with a priority year before 2002 (L3).

GENESEQ FASTAlert

- Produced by Thomson Scientific
- The rolling pre-production preview database for GENESEQ available for in-house subscription
- FASTA format sequence data compiled from the **basic** patents of DWPISM – records contain patent bibliography, patent title and sequence
- Like GENESEQ, FASTAlert sequences are indexed from the full patent specification
- Available within two weeks of publication date

A typical GENESEQ FASTAlert record for a WIPO patent sequence

```
>WO2006016172.1 /PT="New polypeptide, useful for treating, e.g.  
inflammation, cancer, colon cancer, inflammatory bowel disease, pancreatic  
disorder and/or interleukin-2 related disease" /QU="DPT" /PA="ARES TRADING  
SA" /PD="16-FEB-2006" /PR="11-AUG-2004" /ED="02-MAR-2006"  
atgaaatcattcagccggatcctcttctcgtcttctcctcctcgccggcctgaggtccaag  
gccgctccctcagcccctctgcctttgggctgtggctttccggacatggcccaccctct  
gagacttcccctctgaagggtgcttctgaaaattccaaacgagatcgccttaaccagaa  
tttctgggactccttaccctgagccttccaagctacctcatcagggtttccttgaaacc  
tcccacttgacttcaactgagcccctcaaccctgacctccgagaaaccccgacccagag  
tctcctgagacccccaaagctgactcactcacaacctcaatatcagaatccctggacatg  
cccaaaactaacctctccaaaatggcacacccagagtgcttctgagacccccacacctggc  
ccaactgaaatgccacaccaggatcccctgagacccccaaacctaacttctccaaaact  
tcacgcccagaatttctgagacccccaaactgaccttatgcaaactacacccaagaa  
tcccagagattctgcagcttaatgccactgaagtctcacaggcagaactccccgagacc  
tcaaactaacctaccaagaccctgacccccaaatccccagaaaagcatgacctcaac  
tccactgagacccccaaactctgaatttctccaagcttccatcctgaccttctaaaacc  
ccccaccagaatcccatgtgaccacaaatcccagccccaccgaaatttcccaaacagaa  
tccccacaacctactacaaaatgcaacagatgtaccaggacctccgacctcaaatc  
tccactagtctctaccagaaacacctgtgcccttcaaggatgacgccactgctctaaat  
Gagctgtccctgaatcccaaaccaggaacacctgcagccatccagccc . . . . .
```

REGISTRY

- Produced by the Chemical Abstracts Service
- Sequences from >3000 life science journals and the **basic** patents of the 50 patent issuing authorities of the CAplusSM file on STN
- Patent number, location and standardized nomenclature provided for each sequence
- 59.4 million sequence records
 - Including 12.1 million from patent publications
- Updated daily
- 1907 - present

Relationship between CAPLUS patent family and CAS Registry databases

AN CAPLUS

TI

PA

PI WO A1

FR A1

US A1

US B2

AB

IT RN

RN Protein REGISTRY

PI WO A1

SEQ 1

RN DNA REGISTRY

PI WO A1

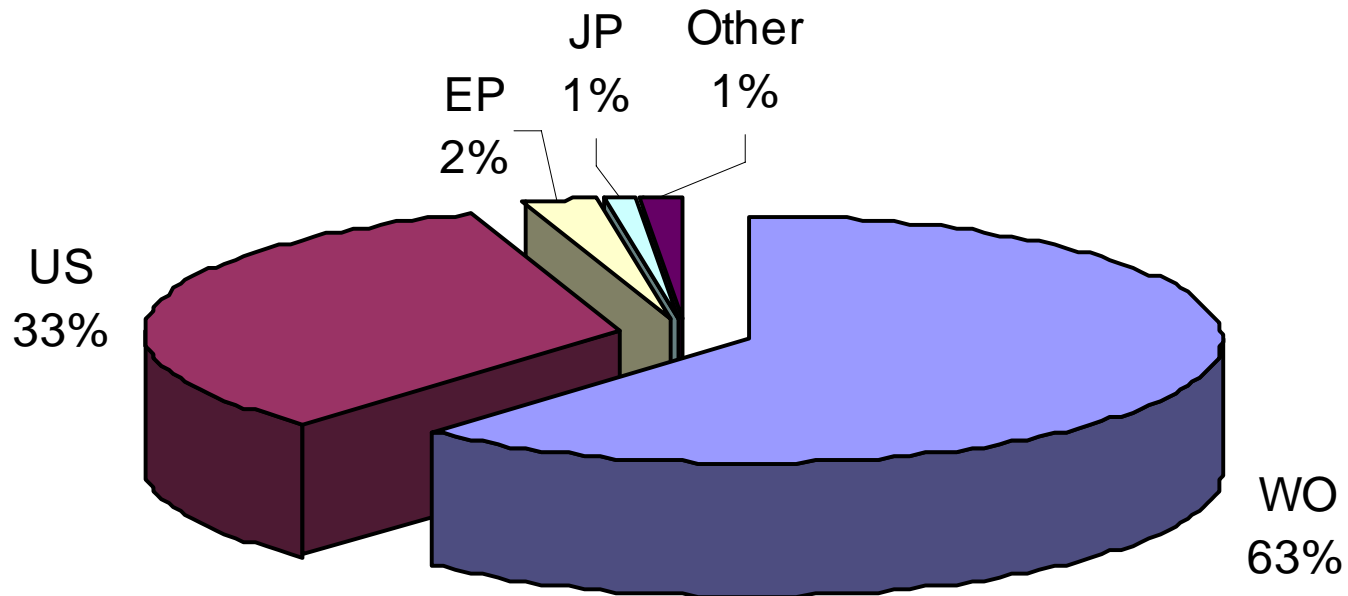
SEQ 2

RN Peptide REGISTRY

PI WO A1

SEQ 3

REGISTRY patent sequences are indexed from CAplus basic patents



The majority of REGISTRY patent sequences are indexed from WIPO/PCT published patent applications (WO), but the proportion is not exactly the same as DGENE (see slide 28).

A typical REGISTRY sequence record for a recent WIPO patent sequence

L1 ANSWER 1 OF 1 REGISTRY COPYRIGHT 2007 ACS on STN
 ED Entered STN: 16 Jan 2007
 RN 917531-59-0 REGISTRY
 CN 20: PN: WO2006137596 SEQID:
 FS PROTEIN SEQUENCE
 SQL 128

WIPO and other patent sequences typically enter the REGISTRY file within 27 days of publication – in this example only 20 days.

PATENT ANNOTATIONS (PNTE):

Sequence	Patent
Source	Reference
=====+=====	
Not Given	WO2006137596
	unclaimed
	SEQID 21

Since October 1999, CAS REGISTRY patent sequence records include the publication number, SEQ ID number and an intellectually assigned claimed/unclaimed notation.

SEQ 1 KCDLALDPDL ARIMAHSRDY DEQLHVWLAW RDAIGPQIRD KYIQYVQMAN
 51 HAARLNGFHD AGQQQREAYE DSDINSQLTE LWATLAPLYR ELHAYVRRHL
 101 VQRYGPERVR PDGPMPAHLL GNMWSRA

MF Unspecified

CI MAN

SR CA

LC STN Files: CA, CAPLUS

DT.CA Caplus document type: Patent

RL.P Roles from patents: PRP (Properties)

Note: this REGISTRY sequence record is an example of one which is not present in DGENE, PCTGEN or EMBL-EBI (as of Feb 5th, 2007).

The corresponding WIPO basic patent record is in the CAplus file on STN

L1 ANSWER 1 OF 1 CAPLUS COPYRIGHT 2007 ACS on STN
AN 2006:1357195 CAPLUS
DN 146:95066
TI Screening for effectors of insecticides and indexing for the WIPO basic patent.
IN Shimokawatoko, Yasutaka; Craen, Marc Van De; Nooren, Irene; Turconi, Sandra; Naudet, Yann; Nys, Guy; Debaveye, Jurgen
PA Sumitomo Chemical Company, Limited, Japan

PATENT NO.	KIND	DATE	APPLICATION NO.	DATE
-----	----	-----	-----	-----
PI WO 2006137596	A2	20061228	WO 2006-JP313039	20060623
JP 2007000060	A	20070111	JP 2005-183031	20050623
PRAI JP 2005-183031	A	20050623		

AB Methods of screening for effector of the dipeptidyl carboxypeptidase A (I) of insects for use as insecticides is described. The method. . . .
IC ICM A01N
CC 5-4 (Agrochemical Bioregulation) Section cross-reference(s)
. . . .



IT 917531-51-2 917531-53-4 917531-55-6 917531-57-8 917531-59-0
RL: PRP (Properties)
(unclaimed protein sequence; screening for effectors of insect dipeptidyl carboxypeptidase A as insecticides)

STN value-added databases provide a unique source of WIPO sequence data

<u>Database</u>	<u>Sequences</u>	<u>Documents</u>
REGISTRY*	7,110,000	38,800
DGENE	5,325,000	63,500
PCTGEN	4,740,000	2,000
EMBL-EBI	2,094,000	32,500

* Statistics from 10/1999 onwards.

Overview of timeliness of the various sources of WIPO sequence data

	Update Frequency	Typical Timeliness	Value added
PCTGEN	Weekly	24 hours	
FASTAlert	Biweekly	14 days	
REGISTRY	Daily	27 days	
DGENE	Biweekly	65 days	
EMBL-EBI	Daily	1-3 months	

Conclusions

- Web-based resources for searching WIPO sequence data are significantly incomplete
- STN offers three key databases for effective searching of WIPO patent sequence data
- DGENE is the “industry-standard” database and must be used in every patent sequence search
- PCTGEN and REGISTRY often provide more timely patent sequence data than DGENE
- REGISTRY offers complementary value-added patent sequence indexing; CAplus basic PCTs may also represent WIPO data not in DGENE

STN[®]

All about WIPO

A comparative review of sources of
WIPO sequence data

Robert Austin – FIZ Karlsruhe