Non-Redundant Patent Sequence Databases

Irina Benediktovich

The SLING project is funded by the European Commission within Research Infrastructures of the FP7 Capacities Specific Programme, grant agreement number 226073 (Integrating Activity)

Current Situation: Search process needs to be accelerated

EMBL-EBI
European Bioinformatics Institute

Get Nucleotide sequences for Go ? Site search Go

European Patent Office

Home | Search Tools | Database searches | Tools | EBI Site

EBI EPO SERVICES

Expert Fasta Summary Table

SUBMISSION PARAMETERS			
Title	SA304301_60_all	Database	uniprot gsp epop jpop uspop
Sequence length	282	Sequence type	p
Program	fasta	Version	3.4t23 April 23, 2004
Expectation upper value	100	Matrix	BL50
Sequence range	1-	Number of scores	500
Word size	2	Open gap penalty	-12
Gap extension penalty	-2	Histogram	false

Show Annotation | Fasta Result | MView | VisualFasta | SUBMIT ANOTHER JOB

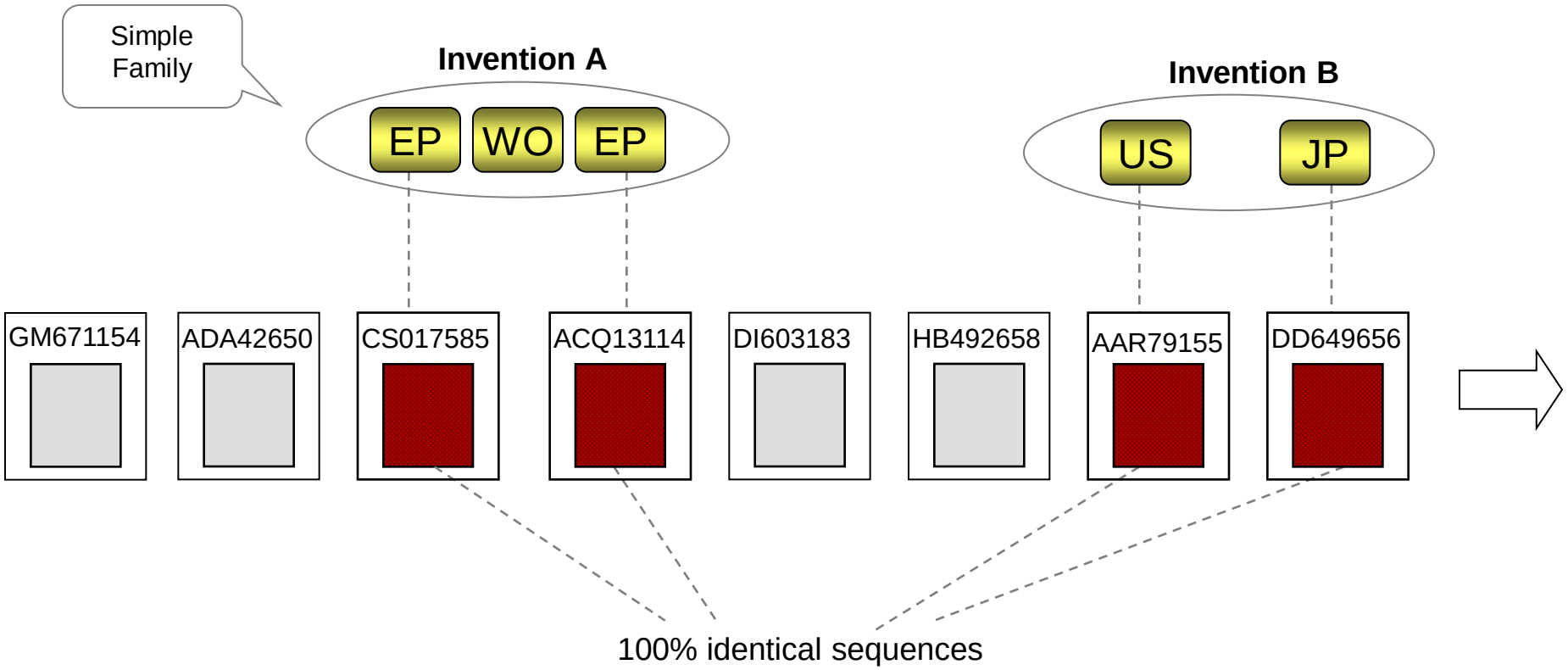
Clear all | Check all | Invert selection | Reset | Show Alignments | Download | fasta

Date of public availability: 9 Apr 2008

500 identical results. Too much to analyze!

Alignment	DB:ID	Source	Length	Identity%	Ungapped%	Overlap	Date of entry	E()
1	EPOP:CS255624	Sequence 60 from Patent EP1621	282	100.000	100.000	282	28-FEB-2006	8.6e-118
2	EPOP:CS112270	Sequence 8 from Patent EP14453	282	100.000	100.000	282	22-JUN-2005	8.6e-118
3	EPOP:CS110541	Sequence 208 from Patent WO200	282	100.000	100.000	282	22-JUN-2005	8.6e-118
499	GSP:ABM07995	Human secreted polypeptide PRO12	282	100.000	100.000	282	20-SEP-2003	8.6e-118
500	GSP:ABM08300	Human secreted polypeptide PRO12	282	100.000	100.000	282	20-SEP-2003	8.6e-118

Why we can have 500 identical hits?

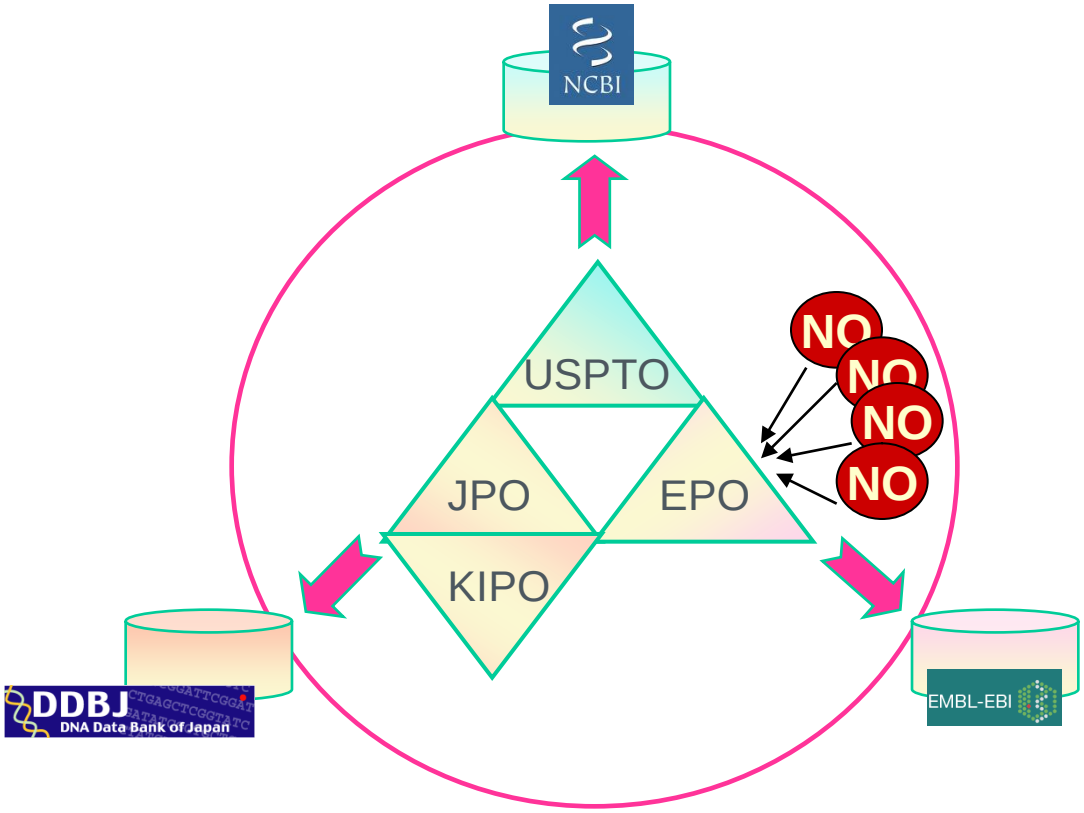


The same sequence can appear multiple times in the database due to:

1) The same invention is filed multiple times in different offices
2) Different Inventors use the same sequence in different contexts

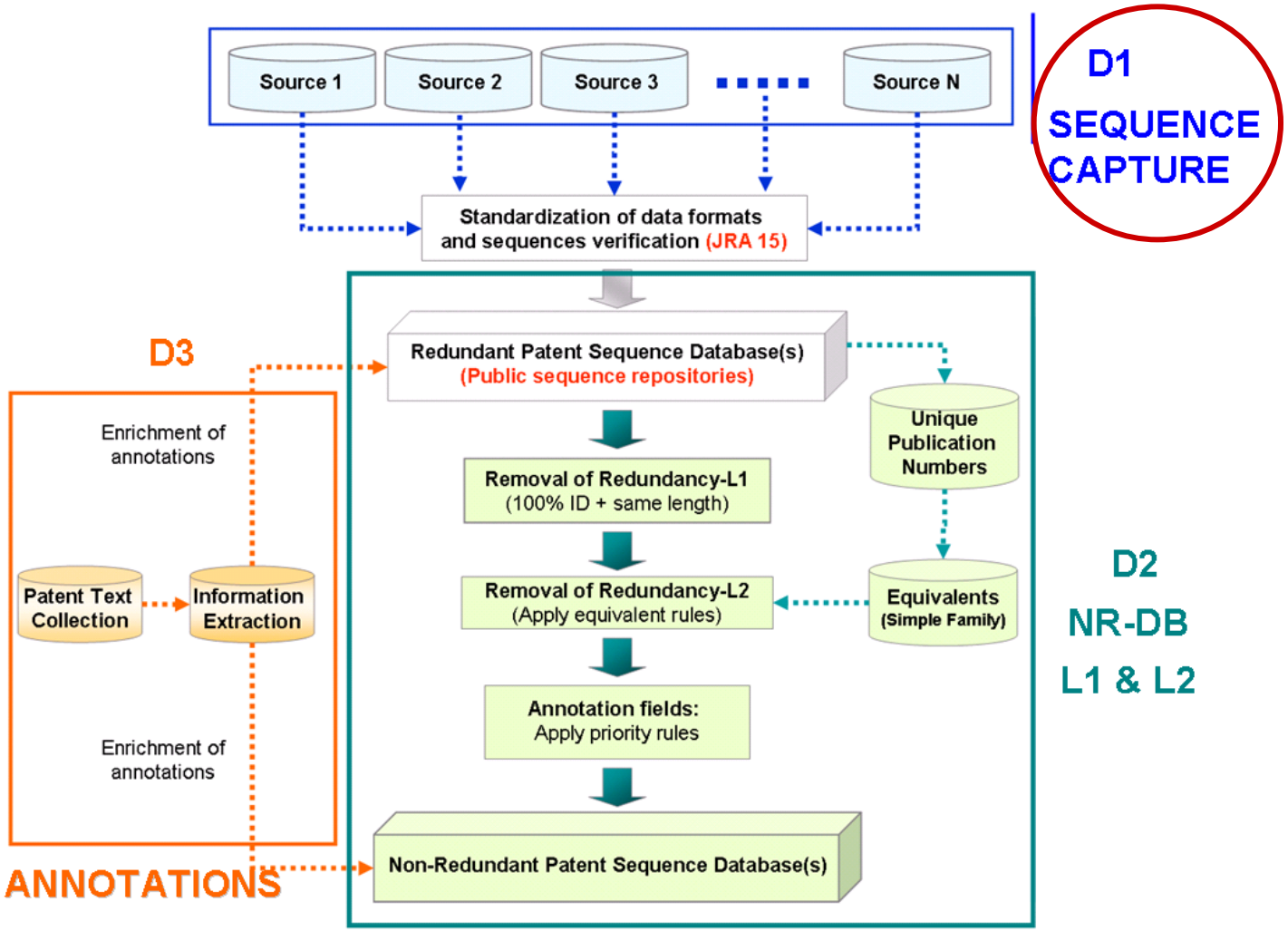
International Cooperation

The Trilateral patent offices exchange and publish their biological sequences, through the public database providers (INSDC)



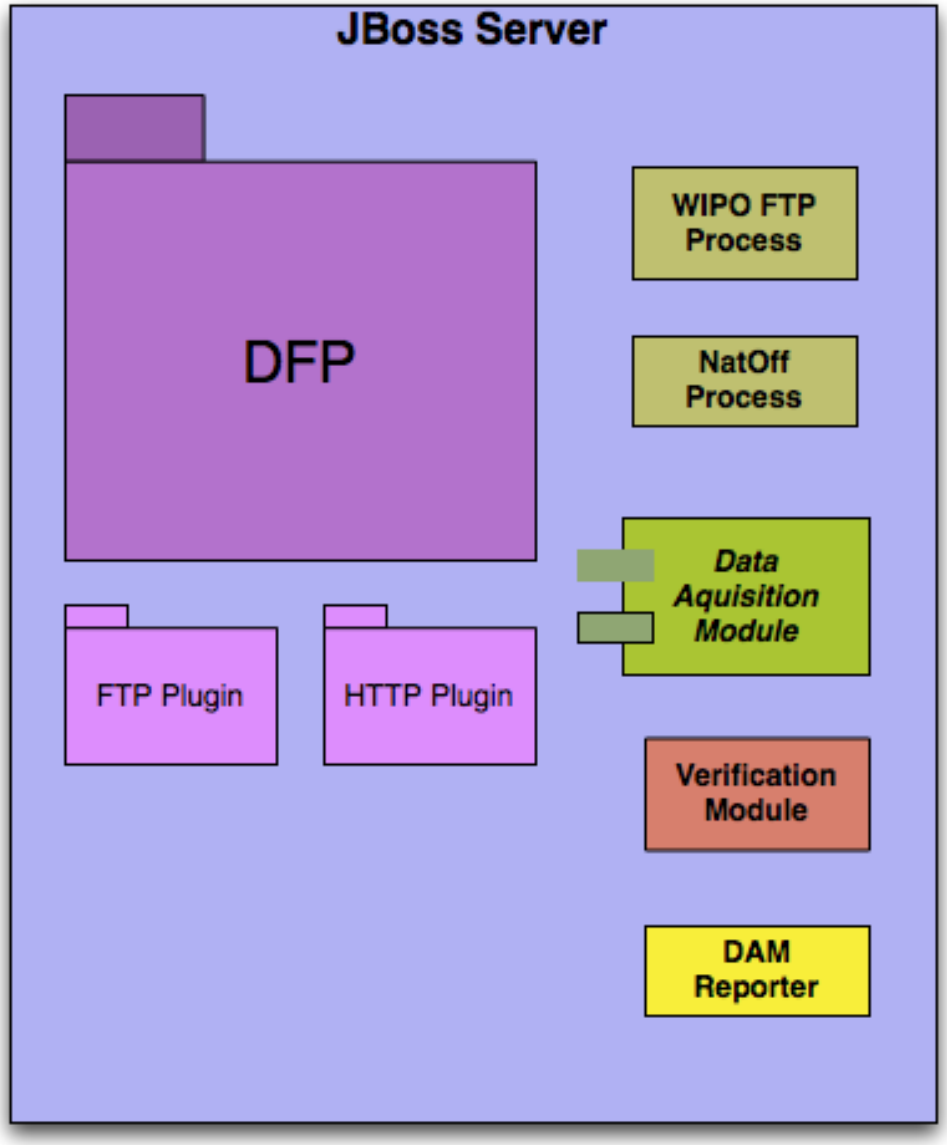
We expect more redundancy in the near future, since other National Offices will participate in the data exchange.

PROJECT OVERVIEW

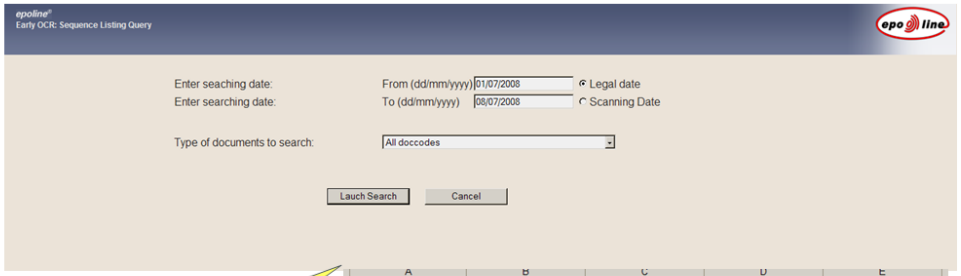


DATA CAPTURE

Architecture of the Sequence Data capture application



DATA CAPTURE



	A	B	C	D	E
1	Date Range				
2	From Date:01/06/2008				
3	To Date:08/07/2008				
4					
5	Applications containing SEQL				
6	Application Number	Document Code	Legal Date	Scan Date	EOCR Id
7	05811185	DESC	02/06/2008	02/06/2008	2914344
8	EP2008056756	DESC	02/06/2008	02/06/2008	2914774
9	06836632	CLMS	02/06/2008	02/06/2008	2914858
10	08104208	DESC	02/06/2008	02/06/2008	2915515
11	DK2008050107	DESC	14/05/2008	02/06/2008	2915613
12	DK2008050107	CLMS	14/05/2008	02/06/2008	2915614
13	DK2008050105	DESC	09/05/2008	02/06/2008	2915625
14	DK2008050105	CLMS	09/05/2008	02/06/2008	2915634
15	04722236	DESC	02/06/2008	02/06/2008	2915956
16	04722236	CLMS	02/06/2008	02/06/2008	2915967
17	07018988	CLMS	26/05/2008	02/06/2008	2916081

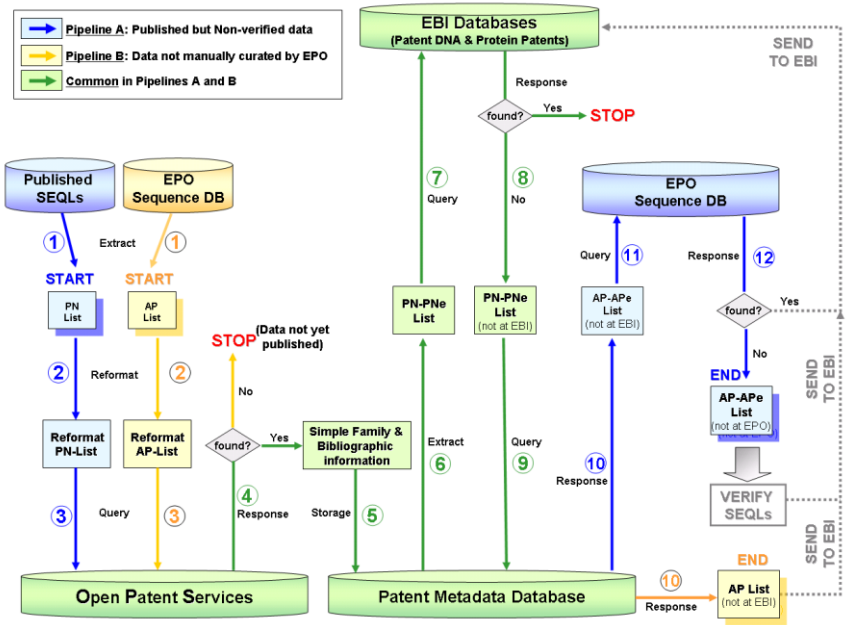
Avoids Misdetection of sequences

Sequence detection algorithm:

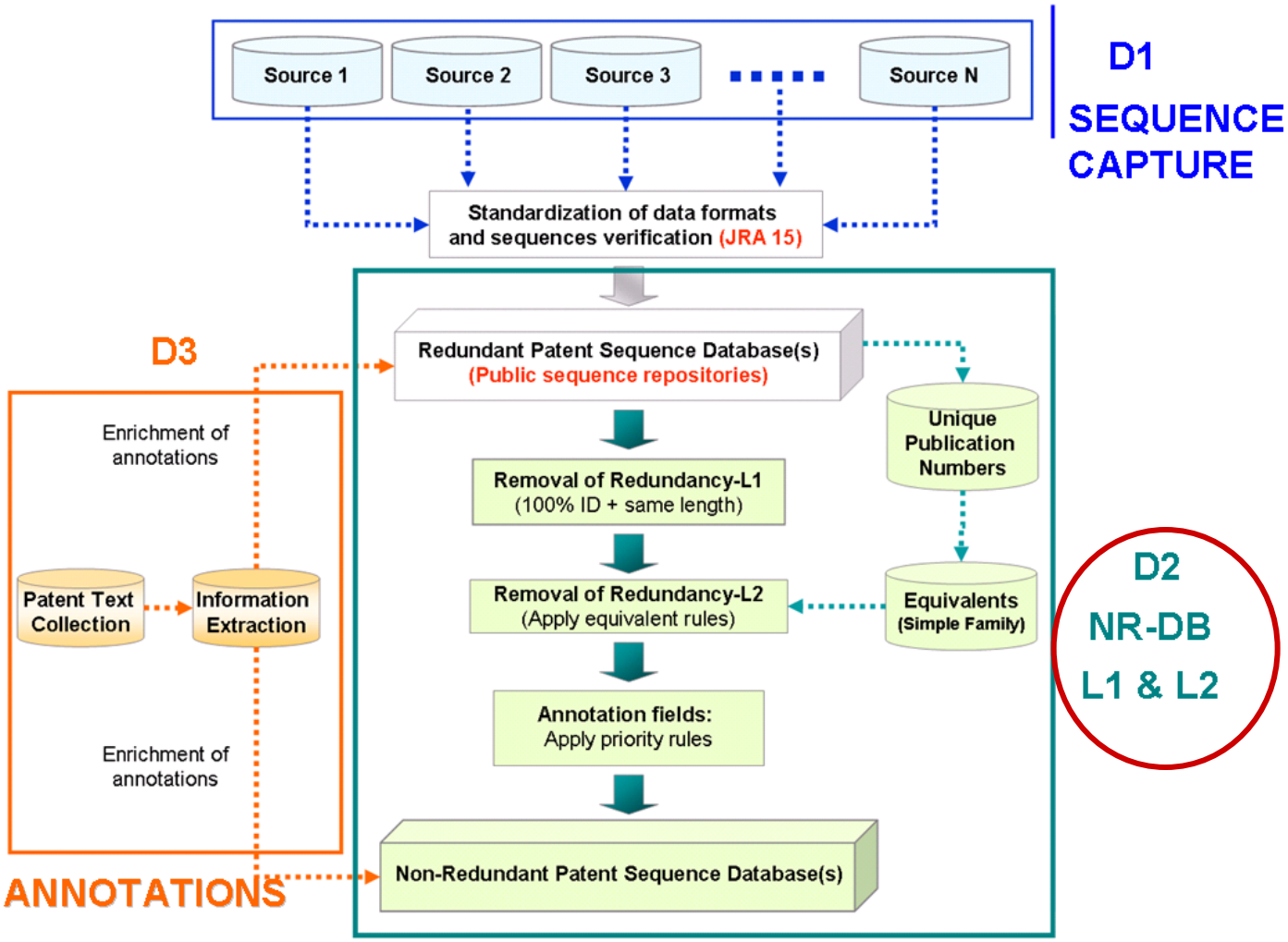
Detects the presence of sequences in the patent application, using a multi-scanning process with different detection levels

Data management workflows:

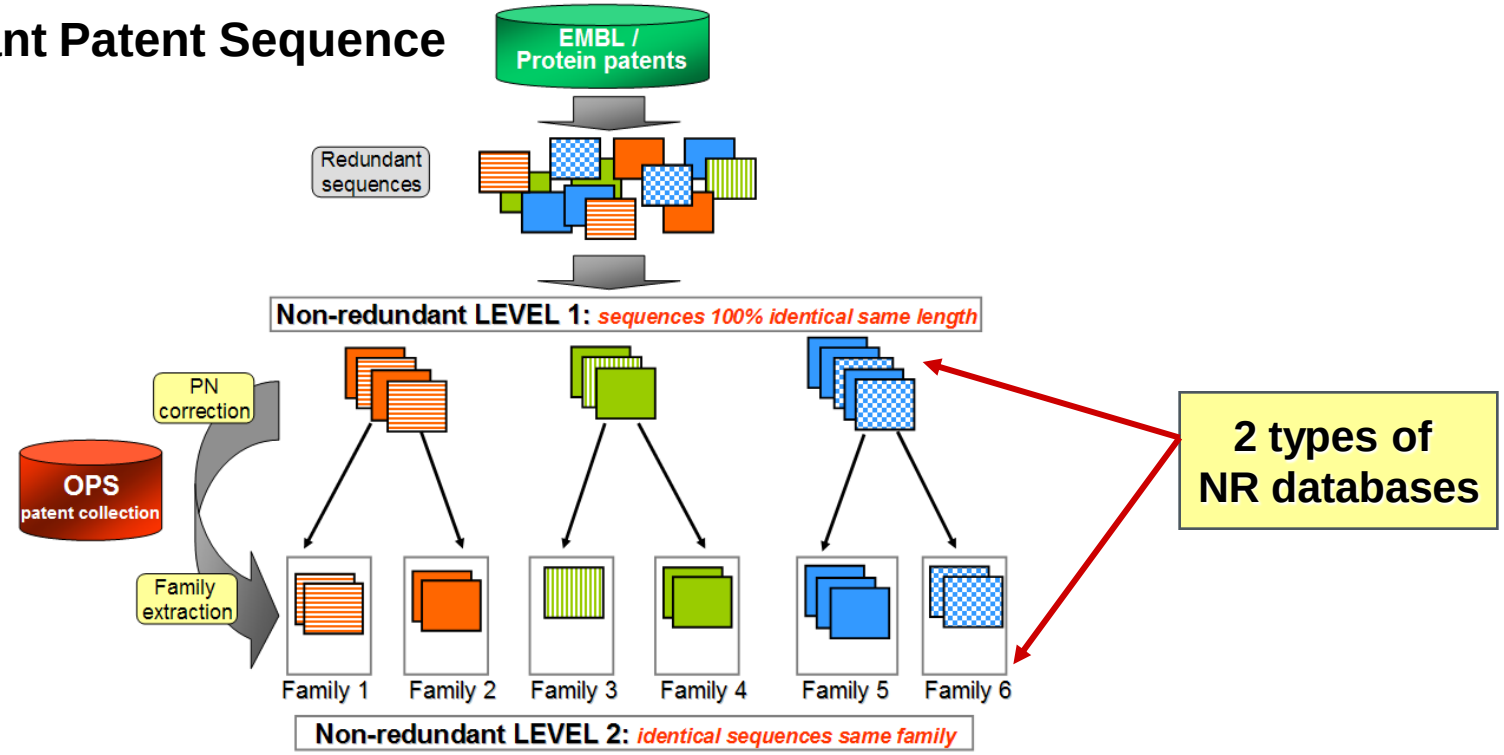
Increase the database coverage without creating more redundancy



PROJECT OVERVIEW



Non-Redundant Patent Sequence Databases



Statistics Sept 2010

NR Databases	Abbreviation	Coverage	Number of entries	Redundancy before
NR Patent Nucleotides Level1	NRNL1	EMBL-Bank patents (17,526,371 entries)	10,077,547	1.74
NR Patent Nucleotides Level2	NRNL2	EMBL-Bank patents (17,526,371 entries)	14,612,812	1.2
NR Patent Proteins Level1	NRPL1	EPO+JPO+KIPO+USPTO (4,947,423 entries)	2,124,798	2,33
NR Patent Proteins Level2	NRPL2	EPO+JPO+KIPO+USPTO (4,947,423 entries)	3,372,114	1,47

1. caggc .... gatcc
2. caggc .... gatcc
3. caggc .... gatcc
....
500. caggc .... gatcc

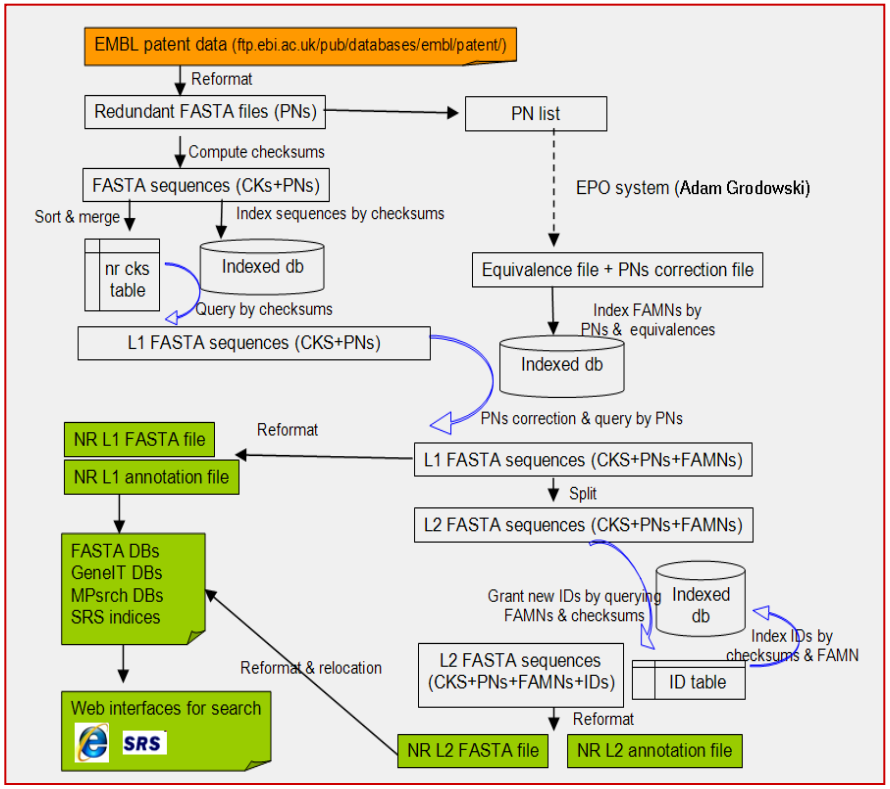
→ 00003f38f0619583f4a536583d92c240

1) We calculate a "fingerprint" per sequence (checksum), since it is faster to compare checksums than sequences.

00003f38f0619583f4a536583d92c240 →

A) caggc .... gatcc *from Umbrella Corp.*
B) caggc .... gatcc *from SuperGen Ltd.*
C) caggc .... gatcc *from GeneTech S.A.*

2) We merge in the same entry, all the sequences with the same 'fingerprint' and belonging to the same invention (simple family)



L1

```
ID  AX224178; PRT; NR1; 3 SQ
XX
ED  18-Feb-2000
XX
DR  EPOP:AX224178;
DE  Sequence 31 from Patent WO0161010.
PN  WO0161010-A2/31, 23-AUG-2001
XX
DR  USPOP:AAO98545;
DE  Sequence 31 from patent US 6509155.
PN  US6509155-A/31, 21-JAN-2003
XX
DR  JPOP:BD728677;
DE  GTPASE ACTIVATING PROTEINS.
PN  JP2004500822-A/31, 15-JAN-2004
XX
SQ  Sequence 433 AA; 001ec8b4ba930012dc11d34cf11f203b; MD5;
//
```

Earliest PD in all Families

Cluster Members (from SEQ-DB)

L2

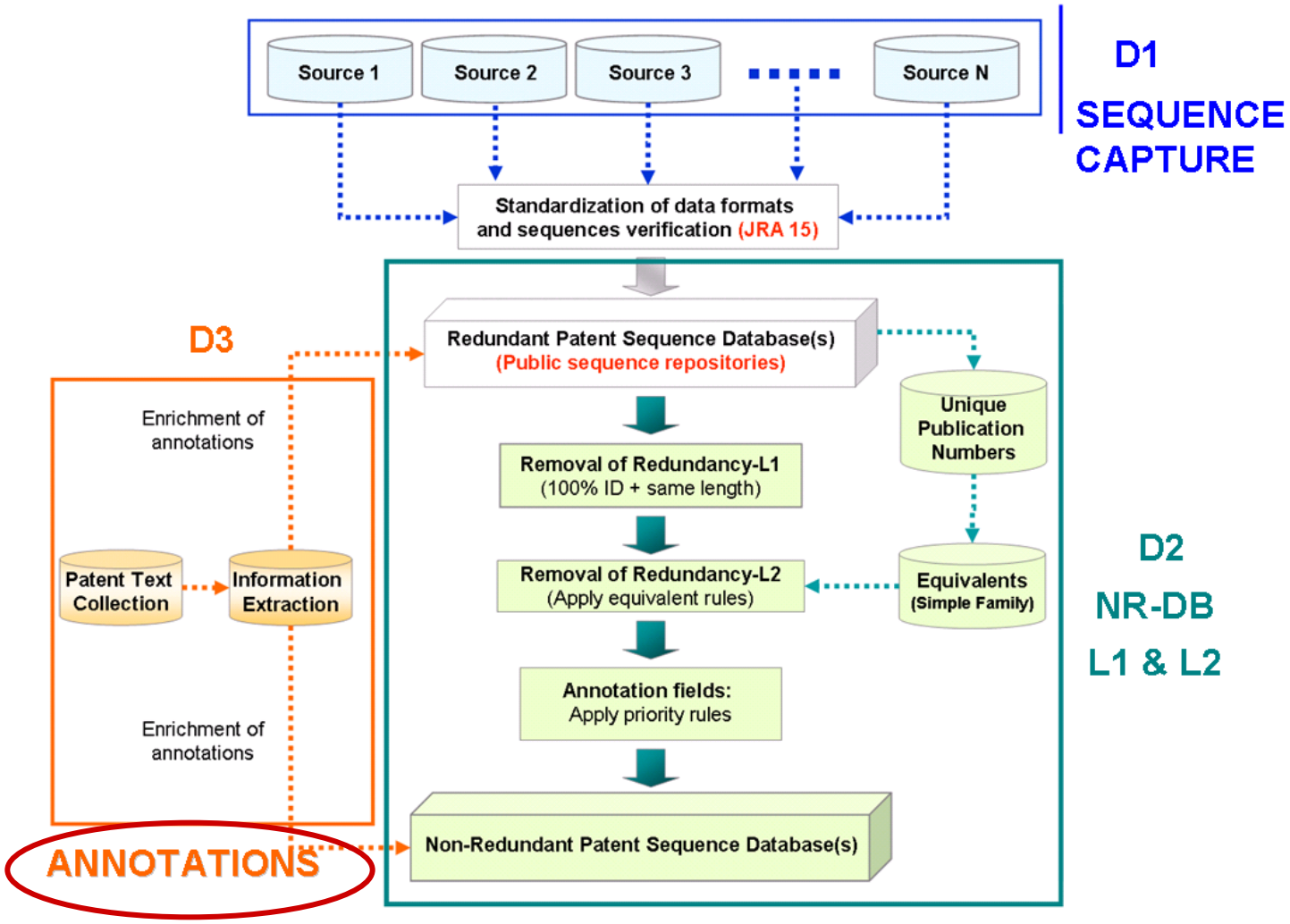
```
ID  NRP00000024; PRT; NR2; 3 SQ
XX
MF  24020039
PN  WO0161010
PR  US20000507765 20000218
ED  18-Feb-2000
XX
DR  EPOP:AX224178;
DE  Sequence 31 from Patent WO0161010.
PN  WO0161010-A2/31, 23-AUG-2001
XX
DR  USPOP:AAO98545;
DE  Sequence 31 from patent US 6509155.
PN  US6509155-A/31, 21-JAN-2003
XX
DR  JPOP:BD728677;
DE  GTPASE ACTIVATING PROTEINS.
PN  JP2004500822-A/31, 15-JAN-2004
XX
SQ  Sequence 433 AA; 001ec8b4ba930012dc11d34cf11f203b; MD5;
//
```

Links to Family members

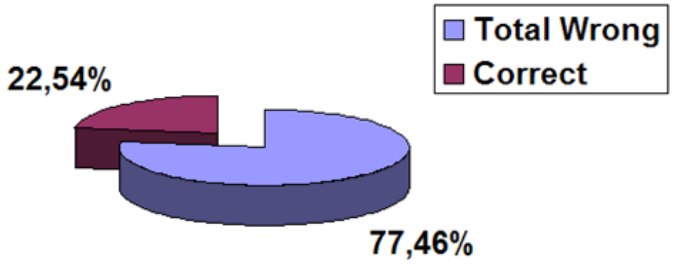
Earliest Priority in Family

Earliest PD in Family

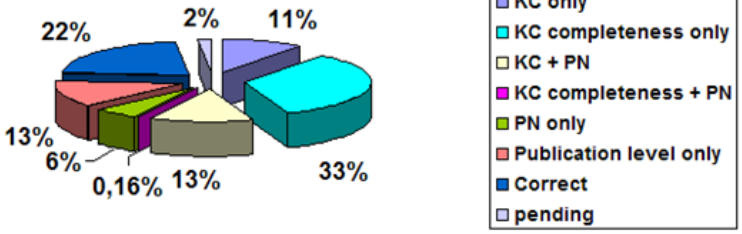
PROJECT OVERVIEW



A)



B)



Correction of Publication Numbers and kind Codes

```
DR   USPOP:ABZ68249;
DE   Sequence 8 from patent US 7326554.
PN   US7326554-A/8, 05-FEB-2008
PN   US2004175376 A1 09-SEP-2004
CC   First level of publication supplied by the EPO
```

```
DR   USPOP:AAO99687;
DE   Sequence 8 from patent US 6514495.
PN   US6514495-A/8, 04-FEB-2003
PN   US6514495 B1 04-FEB-2003
CC   Adapted Kind Code supplied by the EPO
```

```
DR   JPOP:BD555512;
DE   Phytase variants.
PN   JP2002507412-A/9, 12-MAR-2002
PN   JP2002507412T T 12-MAR-2002
CC   Adapted Patent Number supplied by the EPO
```

```
DR   KPOP:DI578933;
DE   Phytase Variants.
PN   KR1020007010543-A/8, 23-SEP-2000
CC   Patent Number could not be successfully verified
```

```
ID   NRN000C020D; DNA; NR2; 2 SQ
XX
MF   34079046
PN   WO2005007891
PR   US20030480035P 19-JUN-2003
ED   27-JAN-2005 WO2005007891 A2
XX
DR   EM_PAT:CS008125;
DE   Sequence 43 from Patent WO2005007891.
PN   WO2005007891-A2/43, 27-JAN-2005
XX
DR   EM_PAT:CS008337;
DE   Sequence 255 from Patent WO2005007891.
PN   WO2005007891-A2/255, 27-JAN-2005
XX
FT   source          1..900
FT                   /organism="Homo sapiens"
FT                   /mol_type="unassigned DNA"
FT                   /db_xref="taxon:9606"
FT   CDS             1..900
FT                   /protein_id="CAI53514.1"
FT                   /translation="MITFLYIFFSILIMVLFVLGNFANGFIALVNFIDWVKRKKISSAD
FT                   QILTALAVSRIGLLWALLLNWYLTVLNPAFYSVELRITSYNAWVVTNHFSMWLAANLSI
FT                   FYLLKIANFSNLLFLHLKRRVRSVILVILLGTLIFLVCHLLVANMDESMWAEEYEGNMT
FT                   GKMKLRNTVHLSYLTVTTLWSFIPFTLSLISFLMLICSLYKHLKKMQLHGEGSQDLSTK
FT                   VHIKALQTLISFLLLCAIFFLFLIVSVWSPRRLRNDPVVMVSKAVGNIYLAFDSFILIW
FT                   RTKKLKHTFLLILCQIRC"
FT                   /protein_id="CAI53620.1 {CS008337}"
FT   variation       181
FT                   /note="AAMTv0.9:CS008337"
FT                   /note="SNP"
FT   variation       608
FT                   /note="AAMTv0.9:CS008337"
FT                   /note="SNP"
FT   variation       155
FT                   /note="AAMTv0.9:CS008337"
FT                   /note="SNP"
XX
SQ   Sequence 900 BP; 2d845b295beed3bf3b4dda32e753c189; MD5;
```

Features only present in one member of the cluster:

CS008337

Identical Sequences stemming from the same invention (same family), very often have different annotations.
In the NR databases at Level 2, we have merged all the annotations in a single record, but still keeping the links to the original entries.

Final Result

```
ID   NRP0000016E; PRT; NR2; 5 SQ
XX
MF   27341889
PR   JP19990377484 16-DEC-1999
ED   20-JUN-2001 EP1108790 A2
XX
DR   EPOP:AX124797;
DE   Sequence 4713 from Patent EP1108790.
PN   EP1108790-A2/4713, 20-JUN-2001
XX
DR   USPOP:ACC04578;
DE   Sequence 4713 from patent US 7332310.
PN   US7332310-A/4713, 19-FEB-2008
PN   US2006228712 A1 12-OCT-2006
CC   First level of publication supplied by the EPO
XX
DR   JPOP:BD572124;
DE   Novel polynucleotide.
PN   JP2002191370-A/4771, 09-JUL-2002
XX
DR   JPOP:BD575624;
DE   Novel polynucleotide.
PN   JP2002191370-A/8271, 09-JUL-2002
XX
DR   KPOP:DI520601;
DE   Novel polynucleotides.
PN   KR1020000077439-A/4713, 16-DEC-2000
PN   KR20010082585 A 30-AUG-2001
CC   Corrected Patent Number supplied by the EPO
XX
FT   source          1..99
FT                   /organism="Corynebacterium glutamicum"
FT                   /mol_type="protein"
FT                   /db_xref="taxon:1718"
XX
SQ   Sequence 99 AA; 018852aac650ff9b667216802250d612; MD5;
//
     MLFDVVMDQK GCLLSPSNII RIAAVLIPND QDQILCVRKE GTELFMFPGG KQELWETPAQ
     AAANSRKKTS IFMGVFRHRQ QTNLASMWTA MCLAHLMCS
//
```

Earliest PR

First publication in the Sequence Databases

5 cluster members with publication corrections

Biological annotations

Sequence and checksum (MD5)

Example: The user would have to analyze 5 entries

Only 1 ENTRY has to be checked with the Non-redundant database!!!

The Non-Redundant databases are publicly available through the EBI

Sequence Similarity Search	FASTA: http://www.ebi.ac.uk/Tools/sss/fasta/
SRS query	SRS: http://srs.ebi.ac.uk/.
Web services	WSFASTA, etc: http://www.ebi.ac.uk/Tools/webservices/
FTP download	ftp://ftp.ebi.ac.uk/pub/databases/fastafiles/patent

For more Information:

D52–D56 Nucleic Acids Research, 2010, Vol. 38, Database issue *Published online 1 November 2009*
doi:10.1093/nar/gkp960

Non-redundant patent sequence databases with value-added annotations at two levels

Weizhong Li[1], Hamish McWilliam[1], Ana Richart de la Torre[2], Adam Grodowski[2], Irina Benediktovich[2], Mickael Goujon[1], Stephane Nauche[2] and Rodrigo Lopez[1,*]

[1]European Bioinformatics Institute, EMBL Outstation, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK and [2]European Patent Office, IQ Life Sciences, Patentlaan 3-9, 2288 EE Rijswijk, The Netherlands

Received August 25, 2009; Revised September 22, 2009; Accepted October 13, 2009

ABSTRACT

The European Bioinformatics Institute (EMBL-EBI) provides public access to patent data, including abstracts, chemical compounds and sequences. Sequences can appear multiple times due to the filing of the same invention with multiple patent offices, or the use of the same sequence by different inventors in different contexts. Information relating to the source invention may be incomplete, and biological information available in patent documents elsewhere may not be reflected in the annotation

modified microbes) and agriculture (e.g. GMO and cultivars). Thus, the patent data are a valuable resource, not only for the intellectual-property world but also for the scientific community. Information in patent data can be more detailed (1), appears earlier or is not available in the scientific literature (2). The European Bioinformatics Institute (EMBL-EBI) provides public access to patent data resources, including abstracts, chemical compounds and sequences (http://www.ebi.ac.uk/patentdata/). Patent abstracts contains abstracts of biology-related patent applications derived from data products of the European Patent Office (EPO). Chemical compounds appearing

CONCLUSIONS

• Similarity and Homology sequence searches against a Non-redundant database, are faster and more sensible, since less entries need to be scanned in the search process.

•These databases are the first non-redundant collection that takes both, sequence and family concepts into consideration.

•The Publication data corrections, significantly increases the data quality. The earliest publication date availability, provides a direct link to track the patent history.

•The collation of all the biological features in a single record, provides a significant improvement for the proper understanding of the biological context the sequence is being used.

•The joint efforts and collaboration of the patent offices and the applicants, on providing sequences with high quality biological annotations, is beneficial for all the users of the public services.

Thank you

Irina Benediktovich: *ibenediktovich@epo.org*