Non-Redundant Patent Sequence Databases

Ana Richart de la Torre
& Irina Benediktovich
Current Situation: Search process needs to be accelerated

Expert Fasta Summary Table

<table>
<thead>
<tr>
<th>Alignment</th>
<th>DB:ID</th>
<th>Source</th>
<th>Length</th>
<th>Identity%</th>
<th>Ungapped%</th>
<th>Overlap</th>
<th>Date of entry</th>
<th>E()</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EPO:CS255624</td>
<td>Sequence 80 from Patent EP1621</td>
<td>282</td>
<td>100.000</td>
<td>100.000</td>
<td>282</td>
<td>29-FEB-2006</td>
<td>8.6e-118</td>
</tr>
<tr>
<td>2</td>
<td>EPO:CS112270</td>
<td>Sequence 8 from Patent EP14453</td>
<td>282</td>
<td>100.000</td>
<td>100.000</td>
<td>282</td>
<td>22-JUN-2005</td>
<td>8.6e-118</td>
</tr>
<tr>
<td>3</td>
<td>EPO:CS110541</td>
<td>Sequence 208 from Patent WO200</td>
<td>282</td>
<td>100.000</td>
<td>100.000</td>
<td>282</td>
<td>22-JUN-2005</td>
<td>8.6e-118</td>
</tr>
<tr>
<td>499</td>
<td>GSP:ABMC7998</td>
<td>Human secreted polypeptide HKU12</td>
<td>282</td>
<td>100.000</td>
<td>100.000</td>
<td>282</td>
<td>29-SEP-2003</td>
<td>8.6e-118</td>
</tr>
<tr>
<td>500</td>
<td>GSP:ABMC8300</td>
<td>Human secreted polypeptide PRO12</td>
<td>282</td>
<td>100.000</td>
<td>100.000</td>
<td>282</td>
<td>20-SEP-2003</td>
<td>8.6e-118</td>
</tr>
</tbody>
</table>

500 identical results. Too much to analyze!
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
We expect more redundancy in the near future, since other National Offices will participate in the data exchange.
PROJECT OVERVIEW

D1
SEQUENCE CAPTURE

D2
NR-DB L1 & L2

D3
ANNOTATIONS

Source 1  Source 2  Source 3  Source N

Standardization of data formats and sequences verification (JRA 15)

Redundant Patent Sequence Database(s)
(Public sequence repositories)

Removal of Redundancy-L1
(100% ID + same length)

Removal of Redundancy-L2
(Apply equivalent rules)

Annotation fields:
Apply priority rules

Unique Publication Numbers

Equivalents
(Simple Family)

Non-Redundant Patent Sequence Database(s)
Architecture of the Sequence Data capture application
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.

Avoids Misdetection of sequences.
### Non-Redundant Patent Sequence Databases

<table>
<thead>
<tr>
<th>NR Databases</th>
<th>Abbreviation</th>
<th>Coverage</th>
<th>Number of entries</th>
<th>Redundancy before</th>
</tr>
</thead>
<tbody>
<tr>
<td>NR Patent Nucleotides Level1</td>
<td>NRNL1</td>
<td>EMBL-Bank patents (17,526,371 entries)</td>
<td>10,077,547</td>
<td>1.74</td>
</tr>
<tr>
<td>NR Patent Nucleotides Level2</td>
<td>NRNL2</td>
<td>EMBL-Bank patents (17,526,371 entries)</td>
<td>14,612,812</td>
<td>1.2</td>
</tr>
<tr>
<td>NR Patent Proteins Level1</td>
<td>NRPL1</td>
<td>EPO+JPO+KIPO+USPTO (4,947,423 entries)</td>
<td>2,124,798</td>
<td>2.33</td>
</tr>
<tr>
<td>NR Patent Proteins Level2</td>
<td>NRPL2</td>
<td>EPO+JPO+KIPO+USPTO (4,947,423 entries)</td>
<td>3,372,114</td>
<td>1.47</td>
</tr>
</tbody>
</table>

**Non-Redundant LEVEL 1:** sequences 100% identical same length

**Non-Redundant LEVEL 2:** identical sequences same family

**2 types of NR databases**

*Statistics Sept 2010*
1) We calculate a "fingerprint" per sequence (checksum), since it is faster to compare checksums than sequences.

2) We merge in the same entry, all the sequences with the same 'fingerprint' and belonging to the same invention (simple family).
<table>
<thead>
<tr>
<th>ID</th>
<th>AX224178; PRT; NR1; 3 SQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED</td>
<td>18-Feb-2000</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>DR</td>
<td>EPOP:AX224178;</td>
</tr>
<tr>
<td>DE</td>
<td>Sequence 31 from Patent WO0161010.</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>DR</td>
<td>USPOP:AAO98545;</td>
</tr>
<tr>
<td>DE</td>
<td>Sequence 31 from patent US 6509155.</td>
</tr>
<tr>
<td>PN</td>
<td>US6509155-A31, 21-JAN-2003</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>DR</td>
<td>JPOP:BD728677;</td>
</tr>
<tr>
<td>DE</td>
<td>GTPASE ACTIVATING PROTEINS.</td>
</tr>
<tr>
<td>PN</td>
<td>JP2004500822-A31, 15-JAN-2004</td>
</tr>
<tr>
<td>SQ</td>
<td>Sequence 433 AA; 8b4ba9300012dc11d34cfl1f203; MD5;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>NRP00000024; PRT; NR2; 3 SQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>24020039</td>
</tr>
<tr>
<td>PN</td>
<td>WO0161010</td>
</tr>
<tr>
<td>PR</td>
<td>US20000507765 20000218</td>
</tr>
<tr>
<td>ED</td>
<td>18-Feb-2000</td>
</tr>
<tr>
<td>DR</td>
<td>EPOP:AX224178;</td>
</tr>
<tr>
<td>DE</td>
<td>Sequence 31 from Patent WO0161010</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>DR</td>
<td>USPOP:AAO98545;</td>
</tr>
<tr>
<td>DE</td>
<td>Sequence 31 from patent US 6509155.</td>
</tr>
<tr>
<td>PN</td>
<td>US6509155-A31, 21-JAN-2003</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>DR</td>
<td>JPOP:BD728677;</td>
</tr>
<tr>
<td>DE</td>
<td>GTPASE ACTIVATING PROTEINS.</td>
</tr>
<tr>
<td>PN</td>
<td>JP2004500822-A31, 15-JAN-2004</td>
</tr>
<tr>
<td>SQ</td>
<td>Sequence 433 AA; 8b4ba9300012dc11d34cfl1f203; MD5;</td>
</tr>
</tbody>
</table>
Correction of Publication Numbers and kind Codes

A) 22.54% Correct, 77.46% Total Wrong

B) 22% KC only, 11% KC completeness only, 13% KC + PN, 6% PN only, 0.16% Publication level only, 13% Correct, 33% pending

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:AA099687;
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 12-MAR-2002
CC Adapted Patent Number supplied by the EPO

DR KPOP:DI578933;
DE Phytase Variants.
CC Patent Number could not be successfully verified
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.
### Earliest PR

First publication in the Sequence Databases

- **PR**: JP19990377848
- **ED**: 20-JUN-2001
- **EP1108790 A2**

### First publication in the Sequence Databases

- **DR**: EP0:AX124797
- **DE**: Sequence 4713 from Patent EP1108790.

### Biological annotations

```
ID    NRP0000016; PRT; NR2; 5 SQ
XX
MF    27341889
PR    JP19990377848 16-DEC-1999
ED    20-JUN-2001
XX
DR    EP0:AX124797;
DE    Sequence 4713 from Patent EP1108790.
XX
DR    USP0:ACC04578;
DE    Sequence 4713 from patent US 7332310.
PN    US7332310-A/4713, 15-FEB-2008
PN    US20062200712 A1 12-OCT-2006
CC    First level of publication supplied by the EPO
XX
DR    JP0:BP572124;
DE    Novel polynucleotide.
PN    JP2002191370-A/4771, 09-JUL-2002
XX
DR    JP0:BP575624;
DE    Novel polynucleotide.
PN    JP2002191370-A/8271, 09-JUL-2002
XX
DR    KE0:BP520601;
DE    Novel polynucleotides.
PN    KE10200000077439-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="Corynbacterium glutamicum"
FT    /mol_type="protein"
FT    /db_xref="taxon:1718"
XX
SQ    Sequence 99 A2; 018852aac650ff9b667216802230d612; MD5;
     //
     MLFDVYMDQR SCLLSAPSNII RIAAVLVLPND QQIQICQFKE RTELVMFPGS RQELWETPAQ
     AAAANGRRRTS IFMGYFRHIQ OTNLAAWITAA MCLAHILMC
```

### 5 cluster members with publication corrections

- Example: The user would have to analyze 5 entries
- Only 1 ENTRY has to be checked with the Non-redundant database!!!

### Sequence and checksum (MD5)

- Sequence: MLFDVYMDQR SCLLSAPSNII RIAAVLVLPND QQIQICQFKE RTELVMFPGS RQELWETPAQ AAAANGRRRTS IFMGYFRHIQ OTNLAAWITAA MCLAHILMC
- MD5: 018852aac650ff9b667216802230d612
The Non-Redundant databases are publicly available through the EBI

<table>
<thead>
<tr>
<th>Sequence Similarity Search</th>
<th>FASTA: <a href="http://www.ebi.ac.uk/Tools/sss/fasta/">http://www.ebi.ac.uk/Tools/sss/fasta/</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>SRS query</td>
<td>SRS: <a href="http://srs.ebi.ac.uk/">http://srs.ebi.ac.uk/</a></td>
</tr>
<tr>
<td>Web services</td>
<td>WSFASTA, etc: <a href="http://www.ebi.ac.uk/Tools/webservices/">http://www.ebi.ac.uk/Tools/webservices/</a></td>
</tr>
</tbody>
</table>
Non-redundant patent sequence databases with value-added annotations at two levels

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

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

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 contain 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

Ana Richart de la Torre
irichart@epo.org

Irina Benediktovich:
ibenediktovich@epo.org