Non Redundant Patent Sequence Database(s)

User Manual
1. Introduction

2. What are the Non-redundant Patent Sequence Databases?

3. Non-Redundant Level 1 database records
   3.1. Lines description
      3.1.1. The ID line
      3.1.2. The ED line
      3.1.3. The Cluster members
         3.1.3.1. The DR line (Database Reference)
         3.1.3.2. The DE line (Description Line)
         3.1.3.3. The PN (Publication Number) line
         3.1.3.4. The CC line
      3.1.4. The SQ line
      3.1.5. The Sequence block
   3.2. A record example

4. Non-Redundant Level 2 database records:
   4.1. Lines description
      4.1.1. The ID line
      4.1.2. The MF line
      4.1.3. The PR line
      4.1.4. The FT lines:
   4.2. A record example

5. Merging and Adding Features & Qualifiers from the cluster members in a single L2 database record
   5.1. Merging and Adding Concepts
   5.2. Master Entry and Priority concepts
   5.3. Rules for Merging/Adding features
      5.3.1. Merging Features
      5.3.2. Adding Features
   5.4 Rules for Merging / Adding Qualifiers.
      5.4.1. Merging Qualifiers
      5.4.2. Adding Qualifiers
   5.5. Rules for Merging / Adding source feature.
   5.6. An example of a NRL2 record with Merged / Added features and qualifiers

6. Search and public availability
1. Introduction

The existing patent sequence databases, show redundancy at the level of sequence and at the level of equivalent entries, originated from different members of the same patent family. This redundancy has an impact in the searches, resulting on more data to scan per sequence submission (bigger databases) leading to slower searches, and cumbersome results analysis (more hits to analyse).

Thereby, the European Bioinformatics Institute (EBI) and the European Patent Office (EPO) have worked together to develop Non-Redundant patent sequence database(s).

2. What are the Non-redundant Patent Sequence Databases?

There are two different levels of non-redundancy. The redundancy is removed from the public patent sequence repositories (EMBL-patents and Protein Patents) in two steps, and thereby, 2 types of non-redundant databases are generated (Figure 1):

I. Non-redundant patent sequence database(s) at Level 1: redundancy is removed based on sequences 100% identical over the same length. The results are clusters of identical sequences stemming from different patents, thus potentially having biological annotations in different contexts.

II. Non-redundant patent sequence database(s) at Level 2: this level works over the sequence clusters generated in Level 1 databases, and splits them according to simple patent families (see simple family definition below). The clusters have identical sequences, stemming from exactly the same invention (same family), thus the biological annotations are within the same context.

**Simple family**: set of documents that share exactly the same active priorities, in other words, the applicant has been filing exactly the same invention in different patent offices around the world.

![Figure 1: Overview, of the 2 levels of non-redundant patent sequence databases (same colour of squares represents equal sequences; 100% identical over the same length. Identical colour and pattern, represents identical sequences belonging to the same invention)](image)

The Non-redundant nucleotide databases (NRNL1 and NRNL2), stem from EMBL-patents ([http://www.ebi.ac.uk/patentdata/nucleotides/](http://www.ebi.ac.uk/patentdata/nucleotides/)), and the Non-redundant protein databases (NRPL1 and NRPL2) are originated from Protein Patents ([http://www.ebi.ac.uk/patentdata/proteins/](http://www.ebi.ac.uk/patentdata/proteins/)).
3. Non-Redundant Level 1 database records

3.1. Lines description

3.1.1. The ID line

ID <accession>; <molecule type>; <non-redundant level 1>; <cluster size L1>

This line indicates the entry accession number (starting with NRN for nucleotides and NRP for proteins), then the molecule type is given (DNA or PRT), the level of non-redundancy (NR1) and the cluster size (number of sequences forming the level 1 cluster).

An example of a complete ID (IDentification) line is shown below:

ID NRP_AX000635; PRT; NR1; 15 SQ

3.1.2 The ED line

The ED (Earliest Date) line indicates the earliest patent publication date within the cluster, the corresponding patent number (earliest) and its complete kind code (document type).

An example of a complete ED line is shown hereunder:

ED 17-FEB-1999 EP0897010 A2

3.1.3 The Cluster members

Each cluster member contains at least 3 lines (when Patent Number adaptations are done, or inconsistencies are detected, more than a PN line can appear and/or a CC line will always be shown explaining the adaptation, or failure to adapt). Between the cluster members there is always a XX line (spacer line) for separation.

An example of several cluster members within an entry, is shown below:

DR EPOP:AX000635;
DE Sequence 6 from Patent EP0897010.
PN EP0897010-A2/6, 17-FEB-1999
XX
DR USPOP:AAE81988;
DE Sequence 33 from patent US 6291221.
PN US6291221-A/33, 18-SEP-2001
XX
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
XX
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
XX
DR KPOP:DI578993;
DE Phytase Variants.
CC Patent Number could not be successfully verified
XX
3.1.3.1. The DR line (Database Reference)

This line displays the original accession number from the source database, EMBL patents (http://www.ebi.ac.uk/patentdata/nucleotides/) or Protein Patents (http://www.ebi.ac.uk/patentdata/proteins/).

Example:

DR  EPOP:AX000635;

3.1.3.2. The DE line (Description Line)

This line gives the same DE line as in the original entries, which contains general descriptive information about the sequence stored in EMBL/Protein Patents.

Example of the DE line corresponding to the DR line accession given in section 4.1.3.2. above:

DE  Sequence 6 from Patent EF0897010.

3.1.3.3. The PN (Publication Number) line

This line provides the patent number and kind code (document type), the sequence identification number within the patent document and the patent publication date.

Example:

PN  EP0897010-A2/6, 17-FEB-1999

More than one PN line will appear in some cluster members (PN ≥ 1), and this is due to the Patent Number, Kind Code and publication level adaptations supplied by the EPO. In these cases, the first PN line will correspond to the Patent Numbers as given in the original data sources (EMBL and Protein Patents) and the second PN line will provide the adapted number. A CC line will be furnished immediately below, explaining the adaptation or failure to adapt. If the Patent Number kind code and date given in the original databases are correct, only one PN line will appear in the cluster member and no CC line will be furnished.

3.1.3.4. The CC line

The CC (Comment) line provides a comment for the PN in the line immediately above. There are several kinds of comments available here:

(a)  First level of publication supplied by the EPO: the PN in the line immediately above is the first level of publication.

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

(b) Adapted Kind Code supplied by the EPO: The Kind Code in the line immediately above is adapted 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
(c) **Adapted Patent Number supplied by the EPO:** the PN in the line immediately above is adapted 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
```

(d) **Patent Number could not be successfully verified:** the PN in the line immediately above could not be successfully verified by EPO

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

### 3.1.4. The SQ line

The SQ (SeQuence) line contains the sequence length and the sequence checksum (MD5)

**Example:**

```
SQ  Sequence 465 AA; 3963407aa91d3a0d622fec679a4524e0; MD5;
```

### 3.1.5. The Sequence block

It contains the actual sequence string

```
MVTLTFLLSA AYLLSGRVSA APSSAGSKSC DTVDLGYQCS PATSHLMGQY SPFFSLEDEL
SVSSKLKDCC RITLVQVLSR HGARYPTSSK SSKYKCLKVA IQANATDFKQ KFAFLKTYNY
TLGADDLTFP GEQQQLVQNSG KFYQRYKALA RYVPFIRAS GSDRVIASSE KFIEGFOQAK
LAGPGATNRA APAISVIPE SETFVIMTLD SVCTKFQASQ LGDEVANFT ALPAPDIRAR
AEKHLPOVTL TSDVDVSLMD MCSFQTVART SDASQSLPPC QLPTNEWWK YNVLQSLGKY
YGGAGGPLG PQGIQFTTNE LIALLTRSPV QDIHTSN3TL VSNFATFFPN ATMVYDPSHD
NSWSISFFAL GLYNGTEPLS RTSVESAKEL DYGSA5VWVP FGARAYFETM CKSE8EPLV
RALINDRVVF LHHGCDVTKLG RCKLNDVFVKG LSWORDGNNW GECFS
```

### 3.2. A record example:

```
ID   NRP_AX000635; PRT; NR1; 15 SQ
XX
ED   17-FEB-1999 EP0897010 A2
XX
DR   EPOP:AX000635;
DE   Sequence 6 from Patent EP0897010.
PN   EP0897010-A2/6, 17-FEB-1999
XX
DR   EPOP:AX085196;
DE   Sequence 6 from Patent WO0112792.
PN   WO0112792-A1/6, 22-FEB-2001
XX
DR   USPOP:AAE81988;
DE   Sequence 33 from patent US 6291221.
PN   US6291221-A/33, 18-SEP-2001
PN   US6291221 B1 18-SEP-2001
CC   Adapted Kind Code supplied by the EPO
XX
DR   USPOP:AAE96575;
DE   Sequence 33 from patent US 6358722.
PN   US6358722-A/33, 19-MAR-2002
PN   US6358722 B1 19-MAR-2002
CC   Adapted Kind Code supplied by the EPO
XX
DR   USPOP:AAN97218;
```
DE   Sequence 6 from patent US 6475762.
PN   US6475762-A/6, 05-NOV-2002
PN   US6475762 B1 05-NOV-2002
CC   Adapted Kind Code supplied by the EPO
XX   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
XX   DR  USPOP:AAS33207;
DE   Sequence 8 from patent US 6689358.
PN   US6689358-A/8, 10-FEB-2004
PN   US2002127218 A1 12-SEP-2002
CC   First level of publication supplied by the EPO
XX   DR  USPOP:AAT17963;
DE   Sequence 33 from patent US 6699704.
PN   US6699704-A/33, 02-MAR-2004
PN   US6699704 B1 02-MAR-2004
CC   Adapted Kind Code supplied by the EPO
XX   DR  USPOP:AAU99375;
DE   Sequence 78 from patent US 6734004.
PN   US6734004-A/78, 11-MAY-2004
PN   US2003092155 A1 15-MAY-2003
CC   First level of publication supplied by the EPO
XX   DR  USPOP:ABE25759;
DE   Sequence 6 from patent US 7022371.
PN   US7022371-A/6, 04-APR-2006
PN   US2003124700 A1 03-JUL-2003
CC   First level of publication supplied by the EPO
XX   DR  USPOP:AB105657;
DE   Sequence 78 from patent US 7078183.
PN   US7078183-A/78, 18-JUL-2006
PN   US2004142424 A1 22-JUL-2004
CC   First level of publication supplied by the EPO
XX   DR  USPOP:ABZ212138;
DE   Sequence 4 from patent US 7309505.
PN   US7309505-A/4, 18-DEC-2007
PN   US2004126844 A1 01-JUL-2004
CC   First level of publication supplied by the EPO
XX   DR  USPOP:ABZ268249;
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
XX   DR  JPOP:BD555512;
DE   Phytase variants.
PN   JP2002507412-A/9, 12-MAR-2002
PN   JP2002507412 T 12-MAR-2002
CC   Adapted Patent Number supplied by the EPO
XX   DR  KPOP:DI578933;
DE   Phytase Variants.
CC   Patent Number could not be successfully verified
XX   SQ   Sequence 465 AA; 3963407aa91d3a0d622fec679a4524e0; MD5;
///  MVTLTFLLSA AYLLSGRVSA APSSAGSKSC DTVDLGYQCS PATSHLNGQY SPFFSLEDEL SVSSKLKDC RITLQVQLSR HGARYPTSSK SKKYKVLVTQ IQANATDFKG KFAFLKTYNY
4. Non-Redundant Level 2 database records:

All Non-redundant Level 2 database records, follow basically the same structure as Level 1 records (see section 3.1 above). There are only some minor line differences, and those are explained hereunder:

4.1. Lines description

4.1.1. The ID line

ID   <L2-accession>; <molecule type>; <non-redundant level 2>; <cluster size L2>

The ID (IDentification) line contains specific ID created for NRL2 database with prefix NRP for proteins or NRN for DNAs, then the molecule type (PRT or DNA / RNA), the non-redundant level (NRL2) and the level 2 cluster size (number of sequences forming the level 2 cluster)

An example of a complete ID (IDentification) line is shown below:

ID   NRP0000016E; PRT; NR2; 5 SQ

4.1.2. The MF line

The MF (Master Family) line contains the simple patent family identifier used by the EPO in Open Patent Services (http://ops.espacenet.com/)

Example:

MF   27341889

4.1.3. The PR line

The PR (PRIority) line provides the earliest active priority within the family. The priority number comes first, followed by the priority date.

The example hereunder gives the earliest priority of the family 27341889 provided in section 4.1.2. above:

PR   JP19990377484 16-DEC-1999

4.1.4. The FT lines:

The FT (FEaTure) lines follow the same format and conventions as provided by the original repositories (http://www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html). Some extra information is furnished, since features from all cluster member's original entries are added or merged, and the source accession numbers are specified (section 5 provides all the details in this respect)

Example:

FT   source 1..99
FT   /organism="Corynebacterium glutamicum"
FT   /mol_type="protein"
FT   /db_xref="taxon:1718"
4.2. A record example:

<table>
<thead>
<tr>
<th>ID</th>
<th>NRP0000016E; PRT; NR2; 5 SQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>MF</td>
<td>27341889</td>
</tr>
<tr>
<td>PR</td>
<td>JP1999037748A 16-DEC-1999</td>
</tr>
<tr>
<td>ED</td>
<td>20-JUN-2001 EP1108790 A2</td>
</tr>
<tr>
<td>DR</td>
<td>EPOP:AX124797</td>
</tr>
<tr>
<td>DE</td>
<td>Sequence 4713 from Patent EP1108790.</td>
</tr>
<tr>
<td>DR</td>
<td>USPOP:ACC04578</td>
</tr>
<tr>
<td>DE</td>
<td>Sequence 4713 from patent US 7332310.</td>
</tr>
<tr>
<td>PN</td>
<td>US7332310-A/4713, 19-FEB-2008</td>
</tr>
<tr>
<td>PN</td>
<td>US2006228712 A1 12-OCT-2006</td>
</tr>
<tr>
<td>CC</td>
<td>First level of publication supplied by the EPO</td>
</tr>
<tr>
<td>DR</td>
<td>JPPOP:BD572124</td>
</tr>
<tr>
<td>DE</td>
<td>Novel polynucleotide.</td>
</tr>
<tr>
<td>PN</td>
<td>JP2002191370-A/4771, 09-JUL-2002</td>
</tr>
<tr>
<td>DR</td>
<td>JPPOP:BD575624</td>
</tr>
<tr>
<td>DE</td>
<td>Novel polynucleotide.</td>
</tr>
<tr>
<td>PN</td>
<td>JP2002191370-A/8271, 09-JUL-2002</td>
</tr>
<tr>
<td>DR</td>
<td>KPOP:DI520601</td>
</tr>
<tr>
<td>DE</td>
<td>Novel polynucleotides.</td>
</tr>
<tr>
<td>PN</td>
<td>KR1020000077439-A/4713, 16-DEC-2000</td>
</tr>
<tr>
<td>PN</td>
<td>KR20010082585 A 30-AUG-2001</td>
</tr>
<tr>
<td>CC</td>
<td>Corrected Patent Number supplied by the EPO</td>
</tr>
<tr>
<td>FT</td>
<td>source 1..99</td>
</tr>
<tr>
<td>FT</td>
<td>/organism=&quot;Corynebacterium glutamicum&quot;</td>
</tr>
<tr>
<td>FT</td>
<td>/mol_type=&quot;protein&quot;</td>
</tr>
<tr>
<td>FT</td>
<td>/db_xref=&quot;taxon:1718&quot;</td>
</tr>
<tr>
<td>SQ</td>
<td>Sequence 99 AA; 018852aac650ff9b667216802250d612; MD5;</td>
</tr>
<tr>
<td></td>
<td>MLFDVVMDQK GCLLSPSNII RIAAVLPND DQQLCVRKE GTELFMFPGG KQELWETPAQ</td>
</tr>
<tr>
<td></td>
<td>AAANSRKSTS IFMGVFRHRQ QTNLASMWTA MCLAHLMCS</td>
</tr>
</tbody>
</table>

This Level 2 record has 5 cluster members, and two of them belong to the same patent application (PN JP2002191370-A/8271, 09-JUL-2002 and PN JP2002191370-A/4771, 09-JUL-2002) but represent a different sequence id number within the patent. This is due to sequence redundancy within the same patent application.

5. Merging and Adding Features & Qualifiers from the cluster members in a single L2 database record

Non Redundant Level 2 records can be made up of several cluster members. Each member has an entry with annotations in the original databases they stem from: EMBL patents (http://www.ebi.ac.uk/patentdata/nucleotides/) or Protein patents (http://www.ebi.ac.uk/patentdata/proteins/). In general, annotations are quite similar if not equal, amongst the members of the same Level 2 cluster. This fact matches with the family concept, since the applicant filed the same invention with different patent offices around the world and therefore, the sequences of the invention and their corresponding annotations are supposed to be the same. However, differences are detected in some cases, and we resolve them by applying features / qualifiers merging and adding rules.
Example:

This example shows a L2 record with two cluster members and the resulting FT lines made up of merged/added features and qualifiers from the two members (original EMBL entries CS00125 and CS008337). The annotations of the cluster member CS008337, are more complete. Therefore the three "variation" features and the "protein id" qualifier (CDS) were taken from this member and added to the final L2 record. The rest of features and qualifiers were merged from the two cluster members.

```
ID NRHOOOCO20D; DNA; NR2; 2 SQ
XX
MF 34079046
PN WO2005007891
FR 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="MITFLYIFFSILIMVLGFLGNGFIALVNFDWVKKKISSAD"
FT QILTVAVSKVFLNVTLPFAYVLSFRDLRTLTSVNAWVTNHFMSWLLAALISI
FT FYLLKIANFSNLLFLHLKRKRVRKVSIVSILVPLTVNLVCHLNVANMDESWAEYEGNMT
FT GMKMLNTVLVSVLTVTFLPPFLSFLSFLMLLCFLVQYTVKRRFLQMKMKLRNTVHLSYLTFTLSDRFRNVPVSVKAVGNYALFDFILIW
FT RTKKLHTFILLICQIRC"
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;
```
5.1. Merging and Adding Concepts

The main goal of annotation merge is to provide complete biological information about NR L2 clusters. This is achieved by collecting all feature entries from all sequences (of the same NR L2 cluster), merging identical features into one feature entry, merging identical qualifiers, and adding all the non-common and compatible qualifiers pointing to the accession number they come from. Whenever the cluster members do not share a feature, it will be added in the NR L2 record, stating its origin (source database accession number).

5.2. Master Entry and Priority concepts

Each Level 2 cluster member, has an original accession number stemming from the original data source it was taken from (EMBL patents / Protein Patents). In case of conflicts, the information is taken from the original entry with highest priority.

The master accessions within a L2 cluster are chosen based on annotations quality, and Publication numbers correctness, therefore, the election follows the priority rule: EPO>USPTO>JPO>KIPO. In case of conflict within a group, the earliest publication is chosen as a master.

5.3. Rules for Merging/Adding features

5.3.1. Merging Features

Two features are considered identical, if both have the same name and the same location. Identical features are merged and then the qualifiers are looked at carefully, to follow merging/adding qualifier rules into the feature.

Example: this example shows a NR L2 entry with a merged feature. This record has five cluster members (original EMBL entries CQ112748, CQ151620, CQ234997, CQ272553 and CQ346829). The source feature was merged from the five members (them all sharing it). However, some of the qualifiers ("note" qualifiers) were not present in all the entries, therefore they were added to the feature source, but pointing to their origin as {EMBL accession}(section 5.3.2 for more details).

```
ID  NR00008941E; DNA; NR2; 5 SQ
XX
MF 27562579
FN WO01572751
PR US20000180312P 04-FEB-2000
ED 09-AUG-2001 WO0157276 A2
XX
DR EM_PAT:CQ112748;
DE Sequence 21607 from Patent WO0157272.
FN WO0157272-A2/21607, 09-AUG-2001
XX
DR EM_PAT:CQ151620;
DE Sequence 21642 from Patent WO0157276.
FN WO0157276-A2/21642, 09-AUG-2001
XX
DR EM_PAT:CQ234997;
DE Sequence 21836 from Patent WO0157273.
FN WO0157273-A2/21836, 09-AUG-2001
XX
DR EM_PAT:CQ272553;
DE Sequence 20814 from Patent WO0157277.
FN WO0157277-A2/20814, 09-AUG-2001
XX
DR EM_PAT:CQ346829;
DE Sequence 20923 from Patent WO0157275.
FN WO0157275-A2/20923, 09-AUG-2001
XX
FT source 1..118
   /organism="Homo sapiens"
FT /mol_type="unassigned DNA"
FT /note="MAP TO AL139001.3"
```
5.3.2. Adding Features

All the features that were not “found” in the master entry (in other words - which were not merged with any feature of the master entry) are added to the result set and get an additional "note" qualifier, which stores the original accession (EMBL patents or Protein Patents) of this sequence with quotation marks (" "): 

Example: This example shows a NR L2 record with added features. This record has 2 cluster members (original EMBL entries are AX384394 and AX473364). Most of the features of this record were merged, since they were present in both members, but the three “gene” features were taken from the member AX473364 and added to the NR L2 record.

| ID   | NRN002584A3; DNA; NR2; 2 SQ |
| MF   | 27499429                    |
| PN   | WO0214486                   |
| FR   | US20000226422P 18-AUG-2000   |
| ED   | 21-FEB-2002 WO0214524 A2    |
| XX   |                             |
| DR   | EM_PAT:AX384394;            |
| DE   | Sequence 3 from Patent WO0214524. |
| PN   | WO0214524-A2/3, 21-FEB-2002 |
| XX   |                             |
| DR   | EM_PAT:AX473364;            |
| DE   | Sequence 1 from Patent WO0214486. |
| PN   | WO0214486-A2/1, 21-FEB-2002 |
| XX   |                             |
| FT   | source 1..9359              |
| FT   | /organism="synthetic construct" |
| FT   | /mol_type="unassigned DNA" |
| FT   | /db_xref="taxon:32630"     |
| FT   | promoter 2941..4920         |
| FT   | /note="Ubi-promoter from maize" |
| FT   | misc_feature 4921..6400     |
| FT   | /note="AtH1 gene from Arabidopsis thaliana" |
| FT   | polyA_signal 6401..6672     |
| FT   | /note="Poly-A signal from the nopaline synthetase gene from Agrobacteriu m tumefaciens" |
| FT   | misc_feature 7434..8084     |
| FT   | /note="First exon-intron combination from Ubi-maize" |
| FT   | CDS 839..1699               |
| FT   | /transl_table=11            |
| FT   | /note="Beta-lactamase gene (AmpR)" |
| FT   | /protein_id="CAD28571.1"   |
| FT   | /translation="MSIQHFRVALIPFAAACFLPVFAPHEPPKLTVKDAEQDLGRARGVYI ELDNJSKIGEFLRFEPEERFFMSMTFFPKVLCLGAVLRSIDAQGELQGRIRNYSDLVEYS FVTENKJLDGTMVRELCISAAITMSDNTAANLLTITDGGPELTALFLHWNMDHIVTDLRW EPLNEKAIKNQDEXTTMPPVAMATTTPKLIXGELL7LASRQQ1IDMEADVAGFJLLRSA LPGWFPIADSKSGAGERSGRIIAALPGDGPXSIVVIIYTGATQTMDDERNRQIAEIGAS LIKHW" |
| FT   | polyA_signal 9120..9359     |
| FT   | misc_feature 8085..9119     |
| FT   | /note="Hygromycin resistance gene from Escherichia coli" |
| FT   | gene 8085..9119             |
| FT   | /note="AAMTv0.9:AX473364"   |
| FT   | /note="Hygromycin resistance gene from Escherichia coli" |


5.4 Rules for Merging / Adding Qualifiers.

5.4.1. Merging Qualifiers

Two qualifiers are considered equal if they belong to identical features, have equal names and identical values. Then they will be merged. The values are compared ignoring cases and trailing brackets (quotes or other characters). The following qualifiers are considered identical:

/function="RTX-TOXIN"
/function="#RTX-TOXIN#"

(note="PAGE 123.")

5.4.2. Adding Qualifiers

When two identical features are merged, one is always considered as "leading" (having highest priority, according to Master entry rules). All qualifiers, that are not found in the leading feature, will be appended with their corresponding accessions.

Example: this example shows a NRNL2 record with added qualifiers. This record has 2 cluster members (original EMBL accessions A58653 and A58658), sharing the same set of features (features merged). However, some qualifiers sets were not exact, and some of them were added from A58658, into the features. "Protein id" and "translation" were added into CDS, and "number" qualifiers were added into "intron" features.
### 5.5. Rules for Merging / Adding source feature.

The most often merged feature is source. Normally, all source features of one NR L2 cluster should be merged. If this is not possible in a first step, then such situation is considered as a conflict and should be solved following priority rules (section 5.2.)

The qualifiers organism and db_xref are processed together (in pairs). Only one organism:db_xref pair is allowed per source feature. The other pairs are saved as note qualifiers, keeping reference to the original accessions (EMBL patents and Protein Patents):

**Example:**

The "organism" and "db_xref" qualifiers are only taken from one of the entries (AX122418 entry, considered master in this cluster). The "organism" and "db_xref" qualifiers of the other 2 members of the cluster (EA430621 and BD164535) are stored as "note" qualifiers.

<table>
<thead>
<tr>
<th>ID</th>
<th>NRN001C7D79</th>
</tr>
</thead>
<tbody>
<tr>
<td>XX</td>
<td></td>
</tr>
<tr>
<td>FT</td>
<td>source</td>
</tr>
<tr>
<td>FT</td>
<td>/organism=&quot;Corynebacterium glutamicum&quot;</td>
</tr>
<tr>
<td>FT</td>
<td>/mol_type=&quot;unassigned DNA&quot;</td>
</tr>
<tr>
<td>FT</td>
<td>/db_xref=&quot;taxon:1718&quot;</td>
</tr>
<tr>
<td>FT</td>
<td>/note=&quot;genomic DNA (EA430621)&quot;</td>
</tr>
<tr>
<td>FT</td>
<td>/note=&quot;unidentified, taxon:32644 (BD164535,EA430621)&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>AX122418</th>
</tr>
</thead>
<tbody>
<tr>
<td>XX</td>
<td></td>
</tr>
<tr>
<td>FH</td>
<td>Key Location/Qualifiers</td>
</tr>
<tr>
<td>FH</td>
<td></td>
</tr>
<tr>
<td>FT</td>
<td>source</td>
</tr>
<tr>
<td>FT</td>
<td>/organism=&quot;Corynebacterium glutamicum&quot;</td>
</tr>
<tr>
<td>FT</td>
<td>/mol_type=&quot;unassigned DNA&quot;</td>
</tr>
<tr>
<td>FT</td>
<td>/db_xref=&quot;taxon:1718&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>EA430621</th>
</tr>
</thead>
<tbody>
<tr>
<td>XX</td>
<td></td>
</tr>
<tr>
<td>FH</td>
<td>Key Location/Qualifiers</td>
</tr>
<tr>
<td>FH</td>
<td></td>
</tr>
<tr>
<td>FT</td>
<td>source</td>
</tr>
<tr>
<td>FT</td>
<td>/organism=&quot;unidentified&quot;</td>
</tr>
<tr>
<td>FT</td>
<td>/mol_type=&quot;genomic DNA&quot;</td>
</tr>
<tr>
<td>FT</td>
<td>/db_xref=&quot;taxon:32644&quot;</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>BD164535</th>
</tr>
</thead>
<tbody>
<tr>
<td>XX</td>
<td></td>
</tr>
<tr>
<td>FH</td>
<td>Key Location/Qualifiers</td>
</tr>
<tr>
<td>FH</td>
<td></td>
</tr>
<tr>
<td>FT</td>
<td>source</td>
</tr>
<tr>
<td>FT</td>
<td>/organism=&quot;unidentified&quot;</td>
</tr>
</tbody>
</table>
5.6. An example of a NRL2 record with Merged / Added features and qualifiers

ID    NRN001DEAFE; DNA; NR2; 4 SQ
XX   
MF    26146343
FN    EP0875574
ED    04-NOV-1998 EP0875574 A2
XX   
DR    EM_PAT:AX002405;
DE    Sequence 1 from Patent EP0875574.
FN    EP0875574-A2/1, 04-NOV-1998
XX   
DR    EM_PAT:AX002403;
DE    Sequence 5 from Patent EP0875574.
FN    EP0875574-A2/5, 04-NOV-1998
XX   
DR    EM_PAT:CS056137;
DE    Sequence 1 from Patent EP1518558.
FN    EP1518558-A2/1, 30-MAR-2005
XX   
DR    EM_PAT:CS056141;
DE    Sequence 5 from Patent EP1518558.
FN    EP1518558-A2/5, 30-MAR-2005
XX   
FT    source    1..6736
XX
FT    /mol_type="unassigned DNA"
FT    /organism="Actinobacillus pleuropneumoniae"
FT    /strain="4074 (SEROTYPE 1 REFERENCE STRAIN)"
FT    /mol_type="unassigned DNA"
FT    /clone="PROK7"
FT    /db_xref="taxon:715"
FT   CDS             1576..6549
FT    /transl_table=11
FT    /gene="APXIV_VAR1"
FT    /product="APXIV_VAR1"
FT    /function="RTX-TOXIN"
FT    /protein_id="CAB77143.1"
FT    /number=1 {CS056137}
FT    /protein_id="CAI77270.1 {CS056137}
FT   -10_signal      617..623
FT    /note="AAMTv0.9:AX002409"
FT    /standard_name="#-10# {CS056141}"
-35_signal 594..599

promoter 454..1131

CDS 1132..6549

protein_id="CA877145.1"

/tranl_table=11

/gene="APXIV"

/product="APXIV"

/function="RTX TOXIN"

/experiment="experimental evidence, no additional details recorded"

/protein_id="CAI77272.1"

/tranl_table=11

/gene="MRP"

/product="MET-G"

/protein_id="CAI77273.1"

/translation="MDMPGPGTDQGISSLQIPVTGAVVVTDPDIALLDVKSGQMDQVYEDTNDNNDARARDITLKDFTVDNY"
6. Search and public availability

The databases of NRPL1, NRPL2, NRNL1 and NRNL2 are available in the EBI environment:

<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>
<tr>
<td>FTP download</td>
<td>ftp://ftp.ebi.ac.uk/pub/databases/fastfiles/patent</td>
</tr>
</tbody>
</table>