**IPPH国外生物序列数据加工**

# 目的意义

本次工作的主要目的是基于[www.ebi.ac.uk](http://www.ebi.ac.uk)下载的国外生物序列数据，形成国外生物序列加工工具，利用此工具将原始数据转换成支持DI系统数据存储格式入库文件。

注：本工具目前只针对数据解析文档中定义为【EMBL数据库格式（1）】、【FASTA格式（1）】、【FASTA格式（2）】的数据进行加工。

# 流程示意图



# 概述

## 3.1 输入

数据资源管理中心从国外网站下载的生物序列（核苷酸序列、蛋白质序列）数据中定义为【EMBL数据库格式（1）】、【FASTA格式（1）】的数据，在本文档描述中，如未特别说明，EMBL代表【EMBL数据库格式（1）】，FASTA代表【FASTA格式（1）】，基本信息如下：

|  |  |
| --- | --- |
| 数据清单 | 内容 |
| 数据来源 | 数据资源管理中心从国外网站下载的数据 |
| 数据类型 | 生物序列数据 |
| 数据内容 | 生物序列（核苷酸序列、蛋白质序列）相关信息 |
| 数据格式 | txt |
| 使用语言 | 英文 |
| 数据容量 | 74.14G |
| 更新周期 | 未知 |

详细信息及示例参照附件：《生物序列（国外网站下载）数据解析-20150428.doc》

## 3.2 逻辑

在数据加工处理过程中，主要分为以下3个阶段：

1. 输入数据阶段—将EMBL、FASTA经过初始加工，基本保持原始情况保存到Mongo数据库中；本阶段的主要实现为：数据入库工具。
2. 数据处理阶段—将阶段1的输出成果根据《国外生物序列索引标签对应说明2015-5-20.xlsx》转换为阶段3输入要求类型并保存在Mongo数据库中；本阶段的主要实现为：数据整理工具。
3. 输出数据阶段—将阶段2的输出成果按照接口标准文档输出文件包并保存到服务器指定位置；本阶段的主要实现为：数据文件打包工具。

## 3.3 输出

本次数据加工的主要输出成果为：符合《生物序列数据更新文件定义》要求的目录结构及文件格式的文件包，文件包中包含以下文件种类：

1. 生物序列控制文件（xml）

2. 生物序列数据索引文件（xml）

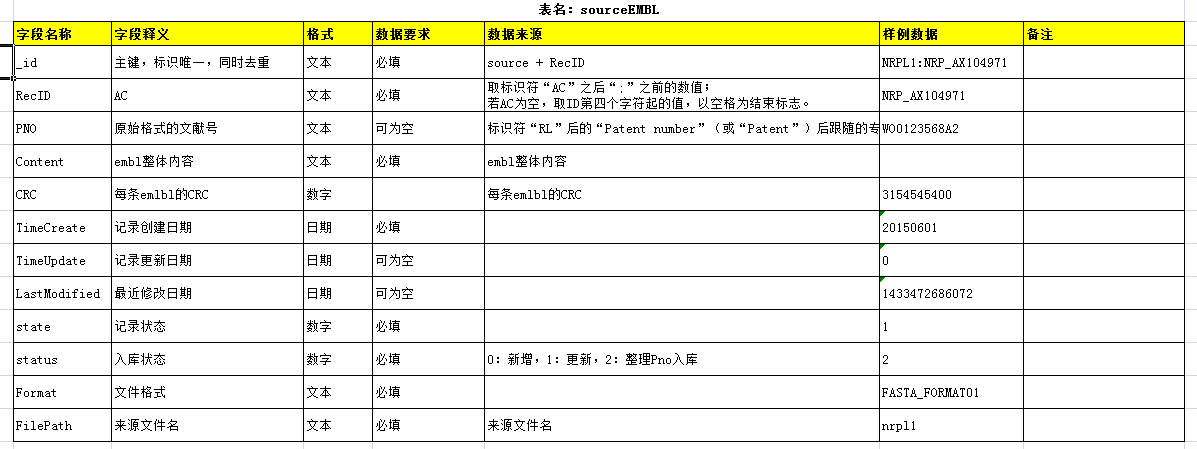
3. 生物序列数据包文件（批量txt的压缩包）

详细信息及示例参照附件：化学及生物序列接口流程说明书\_20150414.docx 章节4.5生物序列数据更新文件定义。

## 3.4 数据库定义

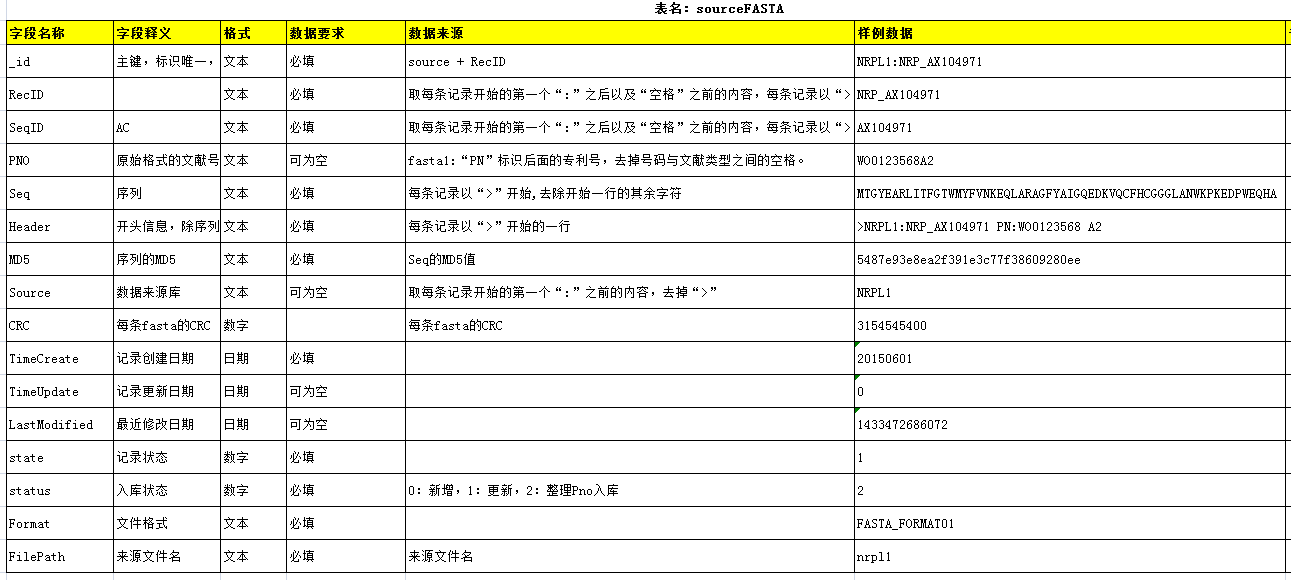
1. EMBL文件库：sourceEMBL

将原始文件中EMBL格式1的数据导入到此数据库，数据结构及示例数据如下：



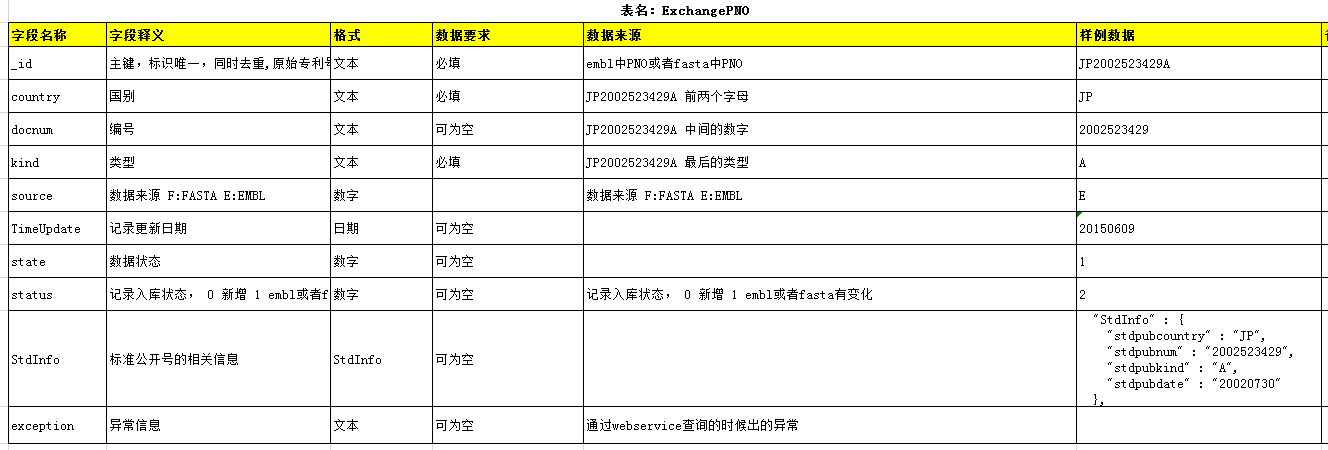
1. FASTA文件库:sourceFASTA

将原始文件中FASTA格式1、FASTA格式2的数据导入到此数据库，数据结构及示例数据如下：



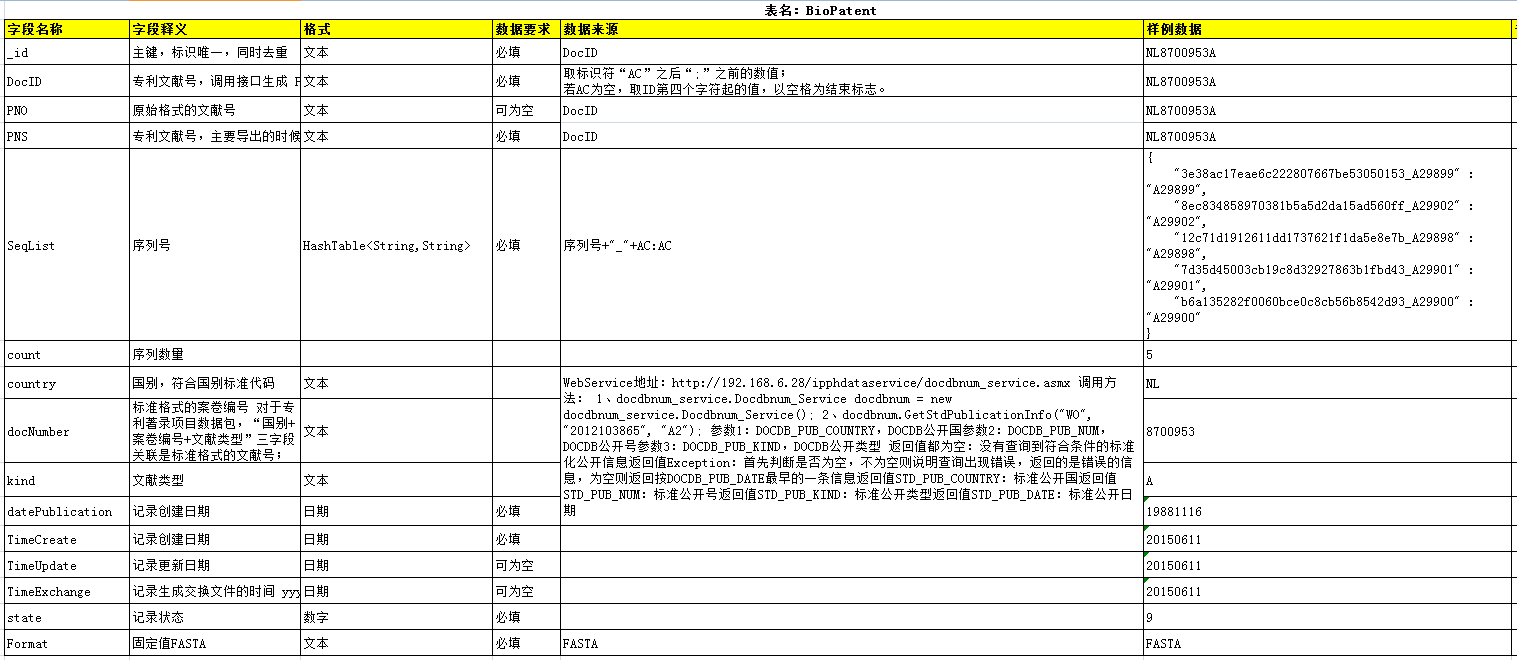
1. 专利号提取库：ExchangePNO

将sourceEMBL中对应的原始专利好导入此数据库，并拆分国别、号码、类型，然后通过标准专利接口获取标准信息并保存在本数据库，数据结构及示例数据如下：



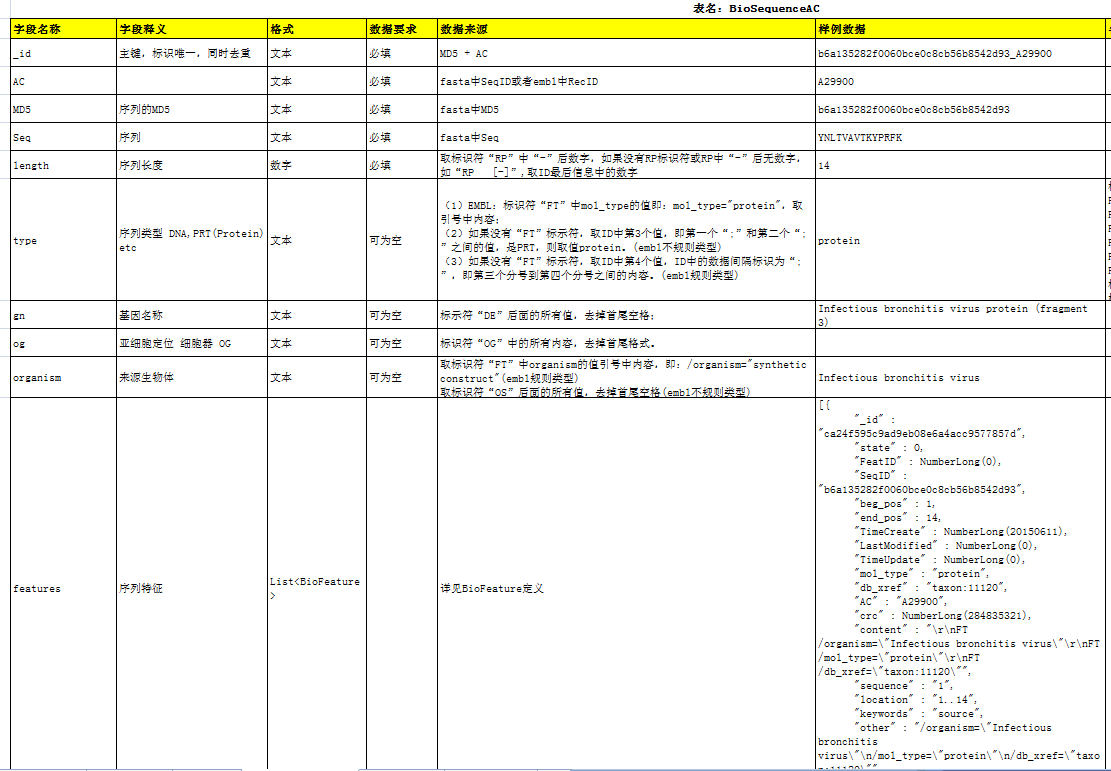
1. 生物专利库：BioPatent

专利信息表，并保存每件专利对应的序列列表，数据结构及示例数据如下：



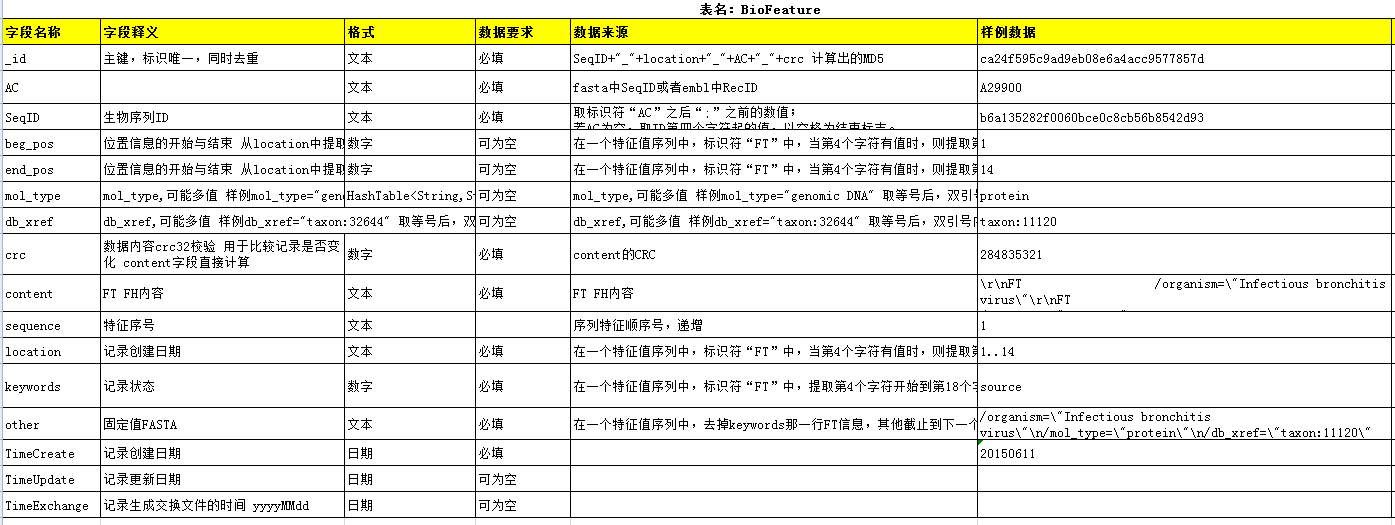
1. 生物序列AC库：BioSequenceAC

原始生物数据号码库，以EBI网站提供的序列号及系统通过序列字符串计算后的MD5作为主键，数据结构及示例数据如下：



1. 生物序列特征库：BioFeature

以随机生成的字符串为主键，数据结构及示例数据如下：



1. 生物序列库：BioSequence

生物序列排重库，序列字符串计算后的MD5作为主键,并保存序列长度，序列内容等基本序列信息，数据结构及示例数据如下：

# 输入数据阶段

## 4.1 输入

见章3.1输入

## 4.2 逻辑





## 4.3 输出

EMBL文件库（sourceEMBL)、FASTA文件库（sourceFASTA），数据结构及数据示例见3.4数据库定义。

## 4.5 问题

（1）epo\_prt.dat重复问题

1、生物序列\蛋白生物专利序列注解EMBL\epo\_prt.dat  
2、生物序列\核苷酸专利序列（冗余，EMBL格式）\epo\_prt.dat  
3、生物序列\patentdata\epo\_prt.dat  
其中 1、2完全一致：3361603条记录，每条记录内容也一致  
其中 3、3358986和1、2一致，2617有变化，新增：22066，看数据情况  
一致加变化=3361603，完全覆盖上面两个文件  
  
目前分析文件，应该属于下重复了，而且可能不同的文件先后顺序是不一致的，都入库，可能导致旧数据覆盖新数据。可以确认一下吗，目前来看，全量数据应该是保留一份。

见附件：[附件\epo\_prt.dat重复且不一致记录2617条.xls](file:///E:\workspace\BioFactory\document\附件\epo_prt.dat重复且不一致记录2617条.xls)

于海燕20150618：使用3、生物序列\patentdata\epo\_prt.dat

（2）个别序列长度超长、如HG670306，序列长度7.7亿，文本大小781M，目前由于数据库限制，此条数据无法入库，请确认一下这个长度区间的序列数据对目标系统是否有影响。

（3）发现标准号接口问题。如序列专利号PNO：US7435798A 欧专局US7435798 (B2) 接口返回 US6133024A

其他数据：PNS-接口 PNO-原始

US4801456A US4801456A

US4801456A US7135187A

US5449720A US5449720A

US5449720A US6489293A

US5468610A US5468610A

US5468610A US7427593A

US6133024A US6133024A

US6133024A US7435798A

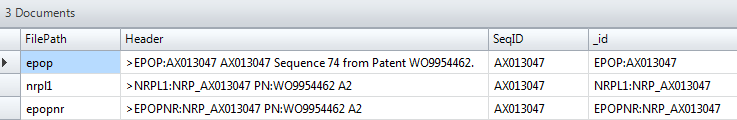
US6136585A US6136585A

US6136585A US7160698A

US6153420A US6153420A

US6153420A US7238498A

（4） 如图，根据序列号ID规则截取出同样的序列ID，在没有FASTA格式2时，通过分析数据将序列ID的NRP\_给截取了，现在格式2入库以后，有指定数据了，关于截取NRP\_的处理是否需要改动？



（5）个别序列长度超长、如HG670306，序列长度7.7亿，文本大小781M，目前由于数据库限制，此条数据无法入库

于海燕：20150629 需要反馈给DI项目应用组 我会把问题反馈的

（6）FASTA数据量超大，目前FASTA总记录2亿多条，具体应该达到什么数量级有大概的范围吗，目前不清楚这个数量级现在是否合理

于海燕：20150629 具体有多少数量级 现在也不清楚，我们看看有没有验证的地方

# 数据处理阶段

## 5.1 输入

数据库表：sourceEMBL、sourceFASTA、ExchangePNO

## 5.2 逻辑

1、根据查询条件查询PNO表，然后逐条处理  
2、取专利对应embl的\_id、RecID字段，不存在或大于1W条不处理。然后将这些embl去重  
3、根据embl的AC列表查出对应的fasta，去重  
4、去除embl中与fasta列表中对应不上的记录，更新embl中的状态  
5、如果去除之后embl列表为空，则更新PNO相应状态，此条专利处理结束  
6、根据剩下embl记录查出完整信息  
7、逐条处理对应的embl、fasta，对Patent、BioSequenceAC赋值  
8、查询Patent、BioSequenceAC记录是否存在，是否有更新、如果有则进行新增或更新，并且更新状态，包括PNO

## 5.3 输出

数据库表：BioPatent、BioSequenceAC

## 5.4 实现

## 5.5 问题

1、数据解析规则还在修正，如果变动还需要修改；

2、数据量过大，不清楚是否全面；

3、数据格式多种，还有未出解析规则的；

1. webservice提供查询的功能还不全面，部分标准专利号取不到。

# 输出数据阶段

## 6.1 输入

数据库表：BioPatent、BioSequenceAC

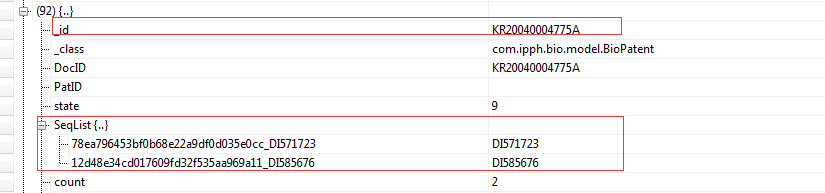
## 6.2 逻辑

1、取出未被处理过的专利

检索BioPatent中状态为0（新增）、1（更新）的专利

2、取出专利对应的序列信息及序列特征信息

专利与序列的对应关系为BioPatent.SeqList



上图表示专利KR20040004775A存在两条序列信息

3、导出文件

根据数据接口定义导出指定格式的控件文件（XML）索引文件（XML）专利序列文件（ZIP/TXT），详细信息参照：[附件\化学及生物序列接口流程说明书\_20150414.docx](file:///D:\\MyWork\\didoc\\生物序列\\IPPH国外生物序列数据加工流程\\附件\\化学及生物序列接口流程说明书_20150414.docx)

4、将导出成功的专利记录更新为9（处理成功）

## 6.3 输出

## 6.4 实现

## 6.5 问题

1

文件大小的控制，由于专利对应的序列字符长度及序列个数的不确定性，导致导出文件时，大小不能够准确控制，目前的解决方案是：

情况1：当所需输出的专利对应的序列总个数加序列特征总个数超过100000时，输出文件，重新计数。

情况2：当专利个数达到1000时，即使不满足情况1，也输出文件，重新计数，目前索引文件大小大部分在30M左右，但是如果遇到极端情况，如某单件专利包含几十万上百万序列信息时，由于不可拆分，则导致索引文件变大



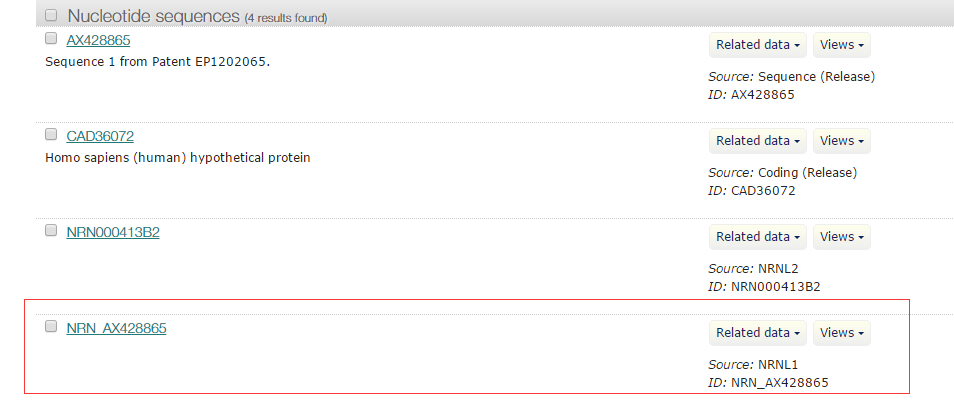


2、专利序列不全

例：EP1202065A1：没有序列AX428876，网站有，但是Source为Sequence (Release)

有AX428865，但是是从NRN\_ AX428865切取出来的，并没有原始ID为AX428865的记录。





# 运行维护

## 7.1 安装部署

本系统的执行程序为BioFactory.jar

部署方式为在有java运行环境的机器上执行一下命令：

java -Xmx4896m -Xms1024m -jar BioFactory.jar

内存大小示机器性能设置，部署后默认端口为8082，

访问[http://{serverip}:8082/record/](http://localhost:8082/record/) 如果返回描述性信息，则系统部署成功。

默认输出地址为：/soft/java/data/

注：如果是windows系统，则是系统所在盘符为根目录，如果是linux在是root为根。

## 7.2 运行

7.2.1 输入数据阶段-FASTA格式1文件入库

将数据文件xxx上传到服务器目录/soft/ipph/fasta1/

执行[http://{serverip}:8082/record/fasta/excute?filepath=/soft/ipph/fasta1/xxx](http://localhost:8082/record/fasta/excute?filepath=/soft/ipph/fasta1/xxx)

url请求可并发

7.2.2输入数据阶段-FASTA格式2文件入库

将数据文件xxx上传到服务器目录/soft/ipph/fasta2/

执行[http://{serverip}:8082/record/fasta2/excute?filepath=/soft/ipph/fasta2/xxx](http://localhost:8082/record/fasta2/excute?filepath=/soft/ipph/fasta2/xxx)

url请求可并发

7.2.3输入数据阶段-EMBL格式1文件入库

将数据文件xxx上传到服务器目录/soft/ipph/embl1/

执行http://{serverip}:8082/record/embl/excute?filepath=/soft/ipph/embl1/xxx

url请求可并发

7.2.4 数据处理阶段-专利号提取

执行http://{serverip}:8082/record/embl/pno

7.2.5 数据处理阶段-专利号提取标准化

执行http://{serverip}:8082/record/pno/std

7.2.6数据处理阶段-数据加工转换

执行http://{serverip}:8082/embltest/embl3

7.2.7输出数据阶段-专利序列关系数据导出到文件

执行http://{serverip}:8082/exchange/excute

# 数据情况

## 8.1 源数据

|  |  |  |  |
| --- | --- | --- | --- |
| 集合 | 总数据 | 未解析出原始号 | 语法 |
| SourceEMBL | 37030433 | 587742 | db.sourceEMBL.find({"PNO":null}).count() |
| SourceFasta | 格式1：37473014+格式2：190333444=  227806458 |  |  |
| 集合 | 总数据 | 无标准号 | 语法 |
| ExchangePNO | 226272 | 7 | db.ExchangePNO.find({"StdInfo":null}).count() |

原始专利国别统计：

|  |  |  |
| --- | --- | --- |
|  | **country** | **count** |
| 1 | "AT" | 4 |
| 2 | "AU" | 15 |
| 3 | "BE" | 6 |
| 4 | "CA" | 7 |
| 5 | "CH" | 3 |
| 6 | "DE" | 173 |
| 7 | "EP" | 23112 |
| 8 | "FR" | 103 |
| 9 | "GB" | 114 |
| 10 | "JP" | 62621 |
| 11 | "KR" | 14171 |
| 12 | "LU" | 3 |
| 13 | "NL" | 4 |
| 14 | "NO" | 1 |
| 15 | "US" | 73489 |
| 16 | "USRE" | 1 |
| 17 | "WO" | 52445 |

## 8.2 数据转换情况

## 8.3 导出数据情况

EMBL导出情况统计：

|  |  |  |
| --- | --- | --- |
|  | 专利数 | EMBL数 |
| 原始总量 | 226272 | 37030433 |
| 超过30w | 11 | 8402316 |
| 未解析专利号 |  | 587742 |
| 未找到标准号 | 7 | 443 |
| 重复 | 393 | 28039932 |
| 正确导出 | 225861 |
| 合计 | 226272 | 37030433 |

导出序列数量 28039932

分国别统计：

|  |  |  |
| --- | --- | --- |
|  | **country** | **count** |
| 1 | "AT" | 23 |
| 2 | "AU" | 134 |
| 3 | "BE" | 35 |
| 4 | "CA" | 34 |
| 5 | "CH" | 44 |
| 6 | "DE" | 1962 |
| 7 | "EP" | 4223440 |
| 8 | "FR" | 1123 |
| 9 | "GB" | 1180 |
| 10 | "JP" | 6358735 |
| 11 | "KR" | 487013 |
| 12 | "LU" | 8 |
| 13 | "NL" | 18 |
| 14 | "NO" | 8 |
| 15 | "US" | 10055516 |
| 16 | "WO" | 6908381 |

导出专利数量 225861

导出专利起止日期 19640609 - 20150107

分国别情况：

|  |  |  |
| --- | --- | --- |
|  | **country** | **count** |
| 1 | "AT" | 4 |
| 2 | "AU" | 15 |
| 3 | "BE" | 6 |
| 4 | "CA" | 7 |
| 5 | "CH" | 3 |
| 6 | "DE" | 173 |
| 7 | "EP" | 22716 |
| 8 | "FR" | 103 |
| 9 | "GB" | 114 |
| 10 | "JP" | 62616 |
| 11 | "KR" | 14170 |
| 12 | "LU" | 3 |
| 13 | "NL" | 4 |
| 14 | "NO" | 1 |
| 15 | "US" | 73486 |
| 16 | "WO" | 52440 |

**其他：、**

专利重复（patentrepeat）：393

|  |  |  |
| --- | --- | --- |
|  | **country** | **count** |
| 1 | "KR" | 1 |
| 2 | "EP" | 391 |
| 3 | "WO" | 1 |

样例：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **PNO** | **kind** | **PNS** | **Date** |
| 1 | EP0116778A | A | EP0116778A1 |  |
| 2 | EP0116778A1 | A1 | EP0116778A1 |  |
| 3 | EP0117063A | A | EP0117063A1 |  |
| 4 | EP0117063A1 | A1 | EP0117063A1 |  |
| 5 | EP0117767A | A | EP0117767A1 |  |
| 6 | EP0117767A1 | A1 | EP0117767A1 |  |
| 7 | EP0119031A | A | EP0119031A2 |  |
| 8 | EP0119031A2 | A2 | EP0119031A2 |  |
| 9 | KR1019940700536A | A | KR20050084550A | 20050826 |
| 10 | KR1020057004382A | A | KR20050084550A | 20050314 |

单专利Embl记录大于30w（ExchangePNO.state==3）

|  |  |  |
| --- | --- | --- |
|  | **country** | **count** |
| 1 | "EP" | 1 |
| 2 | "JP" | 4 |
| 3 | "US" | 2 |
| 4 | "WO" | 4 |

数据：

|  |  |  |  |
| --- | --- | --- | --- |
|  | **PNO** | **kind** | **PNS** |
| 1 | EP2213738A2 | A2 | EP2213738A2 |
| 2 | JP2006507841A | A | JP2006507841A |
| 3 | JP2007536925A | A | JP2007536925A |
| 4 | JP2008500039A | A | JP2008500039A |
| 5 | JP2008518610A | A | JP2008518610A |
| 6 | US7374927A | A | US7374927B2 |
| 7 | US7618814A | A | US7618814B2 |
| 8 | WO2005116204A | A | WO2005116204A1 |
| 9 | WO2005116250A2 | A2 | WO2005116250A2 |
| 10 | WO2005116265A2 | A2 | WO2005116265A2 |
| 11 | WO2007106407A2 | A2 | WO2007106407A2 |

对应序列数量：{PNO:{$in:["JP2006507841A","EP2213738A2","WO2007106407A2","WO2005116204A","WO2005116265A2","JP2008518610A","US7374927A","JP2008500039A","JP2007536925A","WO2005116250A2"]}

未匹配标准号（ExchangePNO.StdInfo==null）：

|  |  |  |
| --- | --- | --- |
|  | **country** | **count** |
| 1 | "JP" | 1 |
| 2 | "EP" | 4 |
| 3 | "US" | 2 |

数据：

|  |  |  |  |
| --- | --- | --- | --- |
| EMBL专利号 | EMBL日期 | PNS | 序列号 |
| JP03074534T | 20010119 |  | BD004377 |
| EP2607483A1 | 20130626 |  | JB754055 |
| EP2589606A1 | 20130508 |  | JB420465 |
| EP2543365A1 | 20130109 |  | JB343810 |
| EP2639585A1 | 20130918 |  | JC011835 |
| US7138514A | 20061121 |  | EA028895 |
| US7135460A | 20061114 |  | ABL32763 |

统计代码：

db.ExchangePNO.group({

key : {country:true},

cond:{StdInfo:null},

initial : {count:0},

reduce : function Reduce(doc, out) {

out.count += 1;

},

finalize : function Finalize(out) {

return out;

}

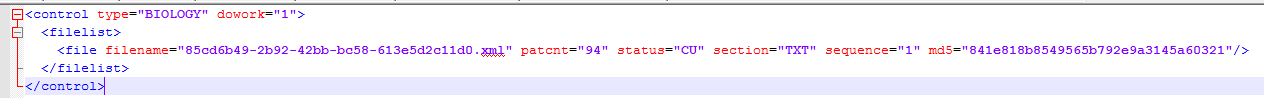
});

文件大小（压缩后）

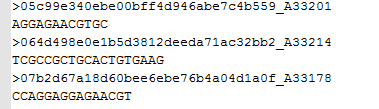
20.2G（压缩前）

导出时间：约5.5小时

控制文件结构：

索引文件结构：

FASTA文件结构：



总结：

国外生物序列记录总数：227806458  
输出专利数：225861  
输出序列数：28039932  
输出的文件数量：1554  
存储规模：  
输出文件大小：20G  
原始文件解压后大小：450G左右  
原始文件压缩包大小：75G左右  
数据库存储：210G

# 附件

1、

[附件\生物序列（国外网站下载）数据解析-20150428.doc](E:\\workspace\\BioFactory\\document\\附件\\生物序列（国外网站下载）数据解析-20150428.doc)

2、

[附件\国外生物序列索引标签对应说明2015-6-9.xlsx](file:///E:\workspace\BioFactory\document\附件\国外生物序列索引标签对应说明2015-6-9.xlsx)

3、

[附件\化学及生物序列接口流程说明书\_20150414.docx](file:///D:\MyWork\didoc\生物序列\IPPH国外生物序列数据加工流程\附件\化学及生物序列接口流程说明书_20150414.docx)

[附件\国外生物序列数据入库标准-20150612.xls](file:///E:\workspace\BioFactory\document\附件\国外生物序列数据入库标准-20150612.xls)

5、

[附件\新接口未返回标准号共61939.xls](file:///E:\workspace\BioFactory\document\附件\新接口未返回标准号共61939.xls)

6、

[附件\epo\_prt.dat重复且不一致记录2617条.xls](file:///E:\workspace\BioFactory\document\附件\epo_prt.dat重复且不一致记录2617条.xls)

7、

[附件\生物序列标准文件抽检-2015-6-16.docx](file:///E:\workspace\BioFactory\document\附件\生物序列标准文件抽检-2015-6-16.docx)

8、

[附件\国外生物序列索引标签对应说明2015-6-24.xlsx](file:///E:\workspace\BioFactory\document\附件\国外生物序列索引标签对应说明2015-6-24.xlsx)

9、

[附件\文献类型取值建议---核查-2015-6-25.docx](file:///E:\workspace\BioFactory\document\附件\文献类型取值建议---核查-2015-6-25.docx)