**EMBL标准文件格式解析**

**在邮件提供的4个示例EMBL文件中，“patentdata--样例3(epo\_prt.dat).txt”属于该格式。**

**1.节点：Country**

**数据主题：国别，符合国别标准代码；**

**备注：**解析规则正确；

**2. 节点：docNumber**

**数据主题：标准格式的案卷编号**

**备注：**知识产权单位自有WebService

**3. 节点：kind**

**数据主题：文献类型**

**备注**：在EMBL标准文件类型中，RL行后的Patent编号部分格式示例为：

**Patent number EP0238993-A2/21, 30-SEP-1987.**

其中，Patent Number后，第一个逗号前的部分，为该专利的完整编号，如该示例中，专利完整编号为**EP0238993-A2/21。**该完整编号中，横杠前的部分为专利应用编号（patent application number），如该示例中，专利应用编号为**EP0238993。**在横杠后，斜线前的部分为专利类型（patent type），如该示例中，专利类型为**A2**。斜杠后的部分，为该序列在专利中的顺序编号（sequence's serial number），在该示例中，顺序编号为**21**。

如果定义的是要提取专利类型，则该解析规则正确。

**4. 节点：PNO**

**数据主题：原始格式的文献号**

**备注：**解析规则正确。

**5. 节点：id**

**数据主题：序列ID**

**备注：**AC标签后的为序列的登录号（accession number），解析规则基本正确。**需要特别注意的是，该登录号可能有多个，多个登录号间以分号分割。此外，登录号可能表征一个范围，如：X00001-X00005，即代表从X00001到X00005之间的所有登录号。如果没有AC标签，可以取ID中的第一项；**

**6. 节点：length**

**数据主题：序列长度**

**备注：从RP中提取序列长度是错误的。RP的含义是Reference Position，即参考位置。该属性标签不一定在每个序列中都有。RP属性中存储的是分段的序列位置，符合如下格式：i-j[, k-l...]（即可以有多个逗号分隔的分段值）。**

**提取序列长度要从ID标签中提取，提取其中最后一项（第7项）中的数字；**

**标准的EMBL文件中，ID的具体格式如下：**

**ID <1>; SV <2>; <3>; <4>; <5>; <6>; <7> BP.**

**其中，<1>代表：该序列的原始登录号（Primary accession number）；**

**<2>代表：该序列的版本号（Sequence version number）；**

**<3>代表：序列的拓扑形状，线性还是环形（Topology: 'circular' or 'linear'）；**

**<4>代表：分子类型（Molecule type ）；**

**<5>代表：数据类别（Data class）,如果是Protein，则为PRT。但该标签也存在标识为PAT，即代表Patent这个类别的可能性；**

**<6>代表：分类学上的来源（Taxonomic division），其实就是指来源于物种或者人工合成。这里是物种的三字符缩写，三字符缩写和详细的物种名称的对应关系如下表：**

|  |  |
| --- | --- |
| **三字符** | **全称** |
| **PHG** | **Bacteriophage** |
| **ENV** | **Environmental Sample** |
| **FUN** | **Fungal** |
| **HUM** | **Human** |
| **INV** | **Invertebrate** |
| **MAM** | **Other Mammal** |
| **VRT** | **Other Vertebrate** |
| **MUS** | **Mus musculus** |
| **PLN** | **Plant** |
| **PRO** | **Prokaryote** |
| **ROD** | **Other Rodent** |
| **SYN** | **Synthetic** |
| **TGN** | **Transgenic** |
| **UNC** | **Unclassified** |
| **VRL** | **Viral** |

**<7>代表：序列长度（Sequence length）。取其中的数字部分即代表序列长度。**

**另一种方法是从SQ标签的第一行提取。**

**7. 节点: type**

**数据主题：序列类型**

**备注：**如第6条所述，取ID行中的第四项即可；

**8. 节点：gn**

**数据主题：基因名称**

**备注：**解析规则正确，需要注意的是该标签不仅可能包含序列的基因名，也有可能表征其来自于哪个基因组，来自于其他专利的序列，或者是任何有助于识别该序列的描述。一个序列可能会有多个DE标签，一般情况下只要取第一个DE标签的描述即可。如果是来自于其他专利的序列，如“Sequence 139 from patent US 5994069.”，则需要到引用的专利中取DE标签的描述；

**9. 节点：og**

**数据主题：亚细胞定位**

**备注：**解析规则正确。需要注意的是该标签不仅代表亚细胞定位，也有可能是描述该序列来源于哪个质粒。该标签的内容可能为如下三种格式：

"Mitochondrion", "Plastid:Chloroplast" 以及 "Plasmid pBR322".

**10. 节点：organism**

**数据主题：来源生物体**

**备注：解析规则可能存在问题。**来源生物体的大类可以从ID行中取，是 ID行中的第六项。该项是一个缩写，缩写的三字符和全称如前述的第6条中的表格；来源生物体的具体信息可以取OS标签对应的值。该值的格式一般分为2种，第一种是Genus species (name)的格式，Genus放的是物种拉丁名中的”属名”，species放的是拉丁名中的“种名”，而name中就是物种的英文常规名称。第二种是该生物体为两个物种的杂交，一般格式为AA x BB (其中AA和BB是物种名称)。

**接下来的Sequence、keywords、location、other节点都属于序列的特征信息；**

**11. 节点：sequence**

**数据主题：特征序号**

**备注：**可以按规则定义方式做顺序编号，即第一个特征的特征序号为1；建议这个节点的名称改为feature，即说明存储的信息为这个序列的特征，以免混淆（仅为建议供参考，如相关程序已经编写则尽量不要改动）；

此外，在该部分中FH这个标签标识了特征信息的开始；第一个FH标签一般存储的是Key，Key中代表了接下去FT标签的格式。在标准的EMBL文件中，FH的Key标签一般为Location/Qualifiers，即接下去FT信息中的第一个FT标签的信息是Locations，即该序列特征指的是哪段序列范围，而“/Qualifiers”表示接下去的FT标签存储的是以“/”开头的特征详细信息。第二个FH的标签是空行，这个空行作为分割符使用，出现空行即标识接下去会有FT信息。

**12. 节点：keywords**

**数据主题：名称或关键词**

**备注：**解析规则可以，实际就是空行的FH下，第一个FT行的信息，代表了这个特征的名字，该字段是一个缩写。**最常见的Keyword**s缩写与全称对应关系如下表格：

|  |  |
| --- | --- |
| **缩写** | **全称** |
| CDS | Protein-coding sequence |
| RBS | ribosome binding site |
| rep\_origin | Origin of replication |
| protein\_bind | Protein binding site on DNA |
| tRNA | mature transfer RNA |

**13. 节点：location**

**数据主题：序列起始、终止位置**

**备注：需要特别注意Excel表中描述的是最简单情形下的序列起始和终止位置，该标签的信息规则比较复杂，常见Location信息的具体规则举例如下：  
（1）467 ： 单个位置；**

**（2）340..565： 从340位到565位；**

**（3）<345..500：该特征的起始位置不确定，但是可以确认是在345位之前，该特征的终止位置是500；**

**（4）<1..888：该特征的起始位置不在这个序列上，终止位置是888；**

**（5）1..>888：该特征的起始位置是1，终止位置在888之后；**

**（6）102.110： 该特征的位置不确定，但是可以确认是在102到110之间的某一个位点上；**

**（7）123^124：该特征是一个123位置到124位置间的位点，通常用来指剪切点特征，这个规则里一般前后两个数字是连号的，如果看到不连号的也不要紧，如1000^1，通常说明这个是环形的序列。**

**（8）join(12..78,134..202)：这个特征位置包括12-78位置的序列以及134-202位置的序列；这两段连在一起代表了这个特征所指的序列位置。**

**（9）complement(34..126)：指这个特征位置，是指的该序列的互补序列（DNA序列中A对T，C对G）的34到126位置；**

**（10）complement(join(2691..4571,4918..5163))：根据规则8先连接，再根据规则9互补；**

**（11）join(complement(4918..5163),complement(2691..4571))：根据规则9先互补，再根据规则8连接；**

**（12）J00194.1:100..202：指的是另一个登录号的序列，如J00194.1这个序列的100到202位置；**

**（13）join(1..100,J00194.1:100..202)：根据规则12先取到另一个序列的位置，在根据规则8连接；**

**EMBL非标准文件格式（去冗余序列文件）解析**

**在邮件提供的4个示例EMBL文件中，“patentdata--样例7(nrnl1.annot).txt”属于该格式中的Level 1格式。理论上还应该有个文件名字类似于“nrnl2.annot.txt”的，是这种格式中的Level 2格式。**

**该格式的具体解释和标签含义如下所述：**

**（1）这种格式的文件名为一般为nrnl开头，指的是“Non-redundant patent nucleotides level”，同理还有nrpl开头（蛋白序列）的文件。该格式又分为2种，一种是Level 1格式，另一种是Level 2格式。格式1和2的区别在于去冗余时的粒度不同。分辨是Level 1还是Level 2格式，一般从官网上下载的EMBL序列文件的文件名的最后一个数字去判断，如果数字是1，就是Level 1格式，如果数字是2，就是Level 2格式。**

**（2）这种格式不是存了一个具体的序列，而是专门为了去除序列冗余而做的。原因是由于一个序列可能存在于多个子专利中，为了避免序列重复冗余对于序列检索效率造成影响，故而由欧洲生物信息学中心牵头对各个序列做了去冗余处理后形成的文件格式。**

**（3）这种格式中，会由一系列三行（或者三行以上）专利信息形成的Cluster Members，一般在最后一个ED标签后，每个member用XX分割，如下示例：**

**DR EM\_PAT:DJ170560**

**DE Method for identification of useful proteins derived from brewery yeast.**

**PN JP2007228956-A/41178, 13-SEP-2007**

**XX**

**DR EM\_PAT:DL304111**

**DE METHODS FOR ANALYZING GENES OF INDUSTRIAL YEASTS.**

**PN JP2008511305-A/41178, 17-APR-2008**

**XX**

**（4）该格式文件中的ED标签，表示的是这些专利cluster中，最早发表的专利的编号和时间信息。**

**（5）该格式文件中的ID标签，一般由四项组成，即**

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

**含义如下:**

**Accession: 专利编号；**

**molecule type： 分子类型；**

**non-redundant level 1：去冗余的粒度，这里是Level 1，一般填NR1.**

**cluster size L1： 是指该专利的cluster中，专利的数量（也就是共同提到该序列的专利的个数）；**

**（6）该格式文件不一定会提供具体的序列，一般会提供一个序列的MD5校验值。具体的序列可以在这个文件包含的专利Cluster Members中的各个Member专利中查到。**

**EMBL非标准格式文件解析**

**在邮件提供的4个示例EMBL文件中，“patentdata--样例6(kipo\_prt.dat).txt”和“patentdata--样例5(jpo\_prt.dat).txt属于该格式。**

**1. 节点：country**

**数据主题：国别，符合国别标准代码；**

**备注：**解析规则正确；

**2. 节点：docNumber**

**数据主题：标准格式的案卷编号**

**对于专利著录项目数据包，“国别+案卷编号+文献类型”三字段关联是标准格式的文献号；**

**备注：**知识产权单位自有web service

**3. 节点：Kind**

**数据主题：文献类型**

**备注：**在EMBL标准文件类型中，RL行后的Patent编号部分格式示例为：

**Patent number EP0238993-A2/21, 30-SEP-1987.**

其中，Patent Number后，第一个逗号前的部分，为该专利的完整编号，如该示例中，专利完整编号为**EP0238993-A2/21。**该完整编号中，横杠前的部分为专利应用编号（patent application number），如该示例中，专利应用编号为**EP0238993。**在横杠后，斜线前的部分为专利类型（patent type），如该示例中，专利类型为**A2**。斜杠后的部分，为该序列在专利中的顺序编号（sequence's serial number），在该示例中，顺序编号为**21**。

如果定义的是要提取专利类型，则该解析规则正确。

**4. 节点：PNO**

**数据主题：原始格式的文献号**

**备注：**如果定义的“原始格式文献号”规则为标识符“RL”后的“Patent number”（或“Patent”）后跟随的专利号，去掉号码与文献类型之间的“-”。则该解析规则正确。

**5. 节点：id**

**数据主题：序列ID**

**备注：**解析规则正确；

**6. 节点：length**

**数据主题：序列长度**

**备注：取ID行中的最后一个数字；不要从RP标签中解析。另一种方法是从SQ标签的第一行提取。**

**7. 节点：type**

**数据主题：序列类型**

**备注：取ID中的第二项即可；**

**8. 节点：gn**

**数据主题：基因名称**

**备注：**解析规则正确，需要注意的是该标签不仅可能包含序列的基因名，也有可能表征其来自于哪个基因组，来自于其他专利的序列，或者是任何有助于识别该序列的描述。一个EMBL文件可能会有多个DE标签，一般情况下只要取第一个DE标签的描述即可。如果是来自于其他专利的序列，如“Sequence 139 from patent US 5994069.”，则需要到引用的专利中取DE标签的描述；

**9. 节点：og**

**数据主题：亚细胞定位**

**备注：**解析规则正确。需要注意的是该标签不仅代表亚细胞定位，也有可能是描述该序列来源于哪个质粒。该标签的内容可能为如下三种格式：

"Mitochondrion", "Plastid:Chloroplast" 以及 "Plasmid pBR322".

**需注意不是每个专利序列都有该标签。**

**10. 节点：organism**

**数据主题：来源生物体**

**备注：**从OS标签中的解析规则正确。该值的格式一般分为2种，第一种是Genus species (name)的格式，Genus放的是物种拉丁名中的”属名”，species放的是拉丁名中的“种名”，而name中就是物种的英文常规名称。第二种是该生物体为两个物种的杂交，一般格式为AA x BB (其中AA和BB是物种名称)。

**11. sequence、keywords、location和other这四个节点的解析方式请参见本文的第一部分。**