**任务1**

处理PathwayCommons11.All.hgnc.sif文件，目的在于将原文件中第二列表示相互作用类型的关键字，从BioPAX体系转换为PMapp体系。也即进行简单的查询和替代，并给出每种替换发生的次数填在下表。转换对应关系见下表：

|  |  |  |
| --- | --- | --- |
| 转换前字符串 | 转换后字符串 | 转换数目（红字为答案） |
| controls-state-change-of | SR | **199,817** |
| controls-phosphorylation-of | SRp | **17,604** |
| controls-expression-of | ER | **126,396** |
| in-complex-with | CAI | **190,769** |
| controls-transport-of | TR | **7,806** |
| controls-transport-of-chemical | TRc | **3,562** |
| chemical-affects | ca | **607,820** |
| interacts-with | MI | **485,608** |
| used-to-produce | sp | **13,683** |
| consumption-controlled-by | sE | **22,401** |
| controls-production-of | Ep | **20,766** |
| catalysis-precedes | EF | **151,000** |
| reacts-with | rw | **3,775** |

注意（1）如何处理“controls-transport-of”和“controls-transport-of-chemical”这两个替换的先后关系，得到正确结果；（2）输出文件保持原有的制表符分隔列的方式。

**任务2**

接着任务1中输出的文件（这里重命名为formSIF\_1851007.txt）继续处理，目的在于将生物分子相互作用数据进行去冗余和证据合并。即有更为细节和高可信度的相互作用，则去掉粗糙或低可信度的相互作用，如保留SRp删除SR，保留CAI删除MI等。处理后请统计一下被删掉MI的数目。保留删除原则见下表：

|  |  |
| --- | --- |
| 如果有 | 则删除 |
| **X** SRp **Y** | **X** SR **Y**  **X** MI **Y** |
| **X** SR **Y** | **X** MI **Y** |
| **X** CAI **Y** | **X** MI **Y** |
| **X** TR **Y** | **X** MI **Y** |

为了保证CAI、MI这两种双向无差别的相互作用对不存在因为倒置节点而产生的冗余，请检查是否同时存在以下的作用对，如果有请给出数目，并保留其中一种形式（排在前面的那条记录）即可。

“**X** CAI **Y**” 和 “**Y** CAI **X**”

“**X** MI **Y**” 和 “**Y** MI **X**”

**部分答案**

* 17,604 SRp 完全包含在 SR 中，也即需要替换掉相同数目的SR。
* 和SRp、SR、CAI、TR重合的MI记录有44,920条被删除，保留MI记录440,688条。
* 最终记录数为1,788,483。

**任务3**

* 从文件中提取第1列中以字符串“hsa-”开头的行并计数，检查这些记录的第2列是否全部都为“INHIBITION”。
* 提取第1列字符串“regulates ”前的内容为新的第1列，新第1列列名为“Node1”。
* 新第2列列名为“Interaction”，内容均为“RNAi”。
* 提取第1列字符串“regulates ”后的内容为新的第3列，新第3列列名为“Node2”。