需求

- ①通过给定的 原基团、目标基团 确定反应类型（生物组确定大类，后端确定小类）
- ②通过反应类型获得酶的`EC`编码（通过`BRENDA`的`.txt`文件，可以手写解析器把它解析成`.sql`文件）
- ③通过`EC`编码获得底物（`ExploreEnz`+类似于`PubChem`、`Rhea`的数据库 或 `KEGG`）

Rhea的链接：Rhea- Annotated reactions database (rhea-db.org)

已知可以下载到本地的数据库：`BRENDA`、`ExploreEnz`、Rhea（可能可以）

思路

- 可以**通过反应类型在从`BRENDA`获取的`TextFile`中找到能催化这类反应的`EC`编码**，`TextFile`中有很多缩写，但好在反应类型那一项并没有缩写（②）
- 可以通过`EC`编码在`ExploreEnz`里查到底物，但是底物是以文字形式给出而非`SMILES`格式（③）
- 而在`KEGG`里可以查到对应底物的结构式，但是不知道`KEGG`的可用性（③）
- 关于酶对应的有机体，可以在`BRENDA`里查到
- （*）能不能通过比对反应的相似度，来代替比对化合物的相似程度（调研R包Molecules | Free Full-Text | A Structural Hierarchy Matching Approach for Molecular Similarity/Substructure Searching | HTML (mdpi.com)）（这个问题可以长远考虑）

主要问题

- 如何把`BRENDA`的`TextFile`解析成方便处理的类型，比如`.sql`文件
- `KEGG`能不能直接用（指直接爬取所有信息或者无限制查询，花钱也行），这个问题最好请教一下往年的同学（③）
- （*）`KEGG`这种数据库查询结果的每个条目是什么意思，`ExplreEnz`里表示底物的方式是否规范，`BRENDA`提供的反应类型是否全面，给这些反应类型分类是否现实之类的（需要生物组科普
- （*）如何将`ExploreEnz`“a primary alcohol” 这类模糊的底物类型转化成我们需要的`SMILES`或者与用户给出的结构式相比对（③），或者有没有数据库能把名称转化成`SMILES`（比如`PubChem`，数据库最好支持将整个数据库下载到本地进行查询）
- （*）**如何通过基团确定反应类型**（这里的反应类型需要多详细才满足需求）（①）
- （*）反应类型过多，是后端判断还是用户输入，**能不能让生物组分一些大类出来，用户选择时指定大类，随后由后端在大类中挑选小类进行比对**（①）

（*）为重点需要考虑的问题。

有关②的一些链接

通过反应类型获取`EC`编码的官方`API`，但是存在问题SOAP access help - BRENDA Enzyme Database (brenda-enzymes.org)

反应类型列举（点进反应类型的链接可以直接获取能催化这类反应酶的`EC`编码）Search result - BRENDA Enzyme Database (brenda-enzymes.org)

任务

- 调研`Rhea`，能否下到本地，能否通过别名确定`SMILES`（当然如果`KEGG`可以用也行）
- 调研R包
- 研究把`.txt`转到`.sql`解析器怎么写