

# 为什么使用本地数据库？

对数据库的检索主要是通过两个方式：数据库提供的API和本地访问。使用本地数据库的优点就是使用数据自由度高，响应快，不依赖网络；缺点是如果使用爬虫，合法性上可能有问题。

接下来的部分对每个数据库使用爬虫来构建本地数据库的合法性和必要性进行讨论。

## BRENDA Database

### 合法性

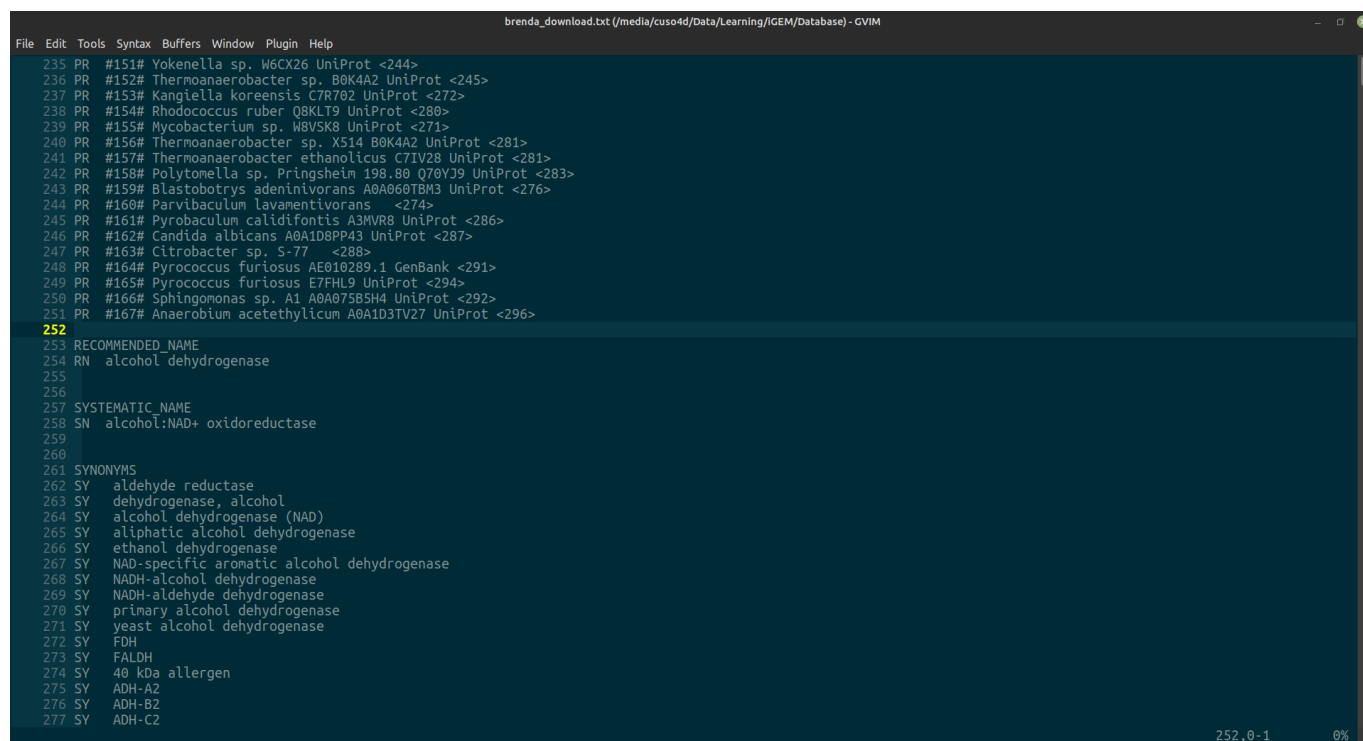
协议：[Copyright and license - BRENDA Enzyme Database](#)

BRENDA是免费、公开的数据库，使用CC BY 4.0协议。

### 必要性

不必要。

BRENDA提供了此数据库的full text版本下载。我们可以获得一个txt格式的文件，其中使用了大量缩写以减小文件的大小，这是它可能不方便的地方。但在[GitHub上有大量的parser](#)，如果真的需要，应该可以通过小修改一个parser的方式读取甚至生成一个完整版本的本地数据库。



```
brenda_download.txt (/media/cuso4d/Data/Learning/IGEM/Database) - GVIM
File Edit Tools Syntax Buffers Window Plugin Help
235 PR #151# Yokenella sp. W6CX26 UniProt <244>
236 PR #152# Thermoanaerobacter sp. B0K4A2 UniProt <245>
237 PR #153# Kangiella koreensis C7R702 UniProt <272>
238 PR #154# Rhodococcus ruber Q8KLT9 UniProt <280>
239 PR #155# Mycobacterium sp. W8VSK8 UniProt <271>
240 PR #156# Thermoanaerobacter sp. X514 B0K4A2 UniProt <281>
241 PR #157# Thermoanaerobacter ethanolicus C7IV28 UniProt <281>
242 PR #158# Polytomella sp. Pringsheim 198.80 Q70YJ9 UniProt <283>
243 PR #159# Blastobotrys adeninivorans A0A060TBM3 UniProt <276>
244 PR #160# Parvibaculum lavamentivorans <274>
245 PR #161# Pyrobaculum calidifontis A3MVR8 UniProt <286>
246 PR #162# Candida albicans A0A1D8PP43 UniProt <287>
247 PR #163# Clitrobacter sp. S-77 <288>
248 PR #164# Pyrococcus furiosus AE010289.1 GenBank <291>
249 PR #165# Pyrococcus furiosus E7FHL9 UniProt <294>
250 PR #166# Sphingomonas sp. A1 A0A07585H4 UniProt <292>
251 PR #167# Anaerobium acetethylicum A0A1D3TV27 UniProt <296>
252
253 RECOMMENDED_NAME
254 RN alcohol dehydrogenase
255
256
257 SYSTEMATIC_NAME
258 SN alcohol:NAD+ oxidoreductase
259
260
261 SYNONYMS
262 SY aldehyde reductase
263 SY dehydrogenase, alcohol
264 SY alcohol dehydrogenase (NAD)
265 SY aliphatic alcohol dehydrogenase
266 SY ethanol dehydrogenase
267 SY NAD-specific aromatic alcohol dehydrogenase
268 SY NADH-alcohol dehydrogenase
269 SY NADH-aldehyde dehydrogenase
270 SY primary alcohol dehydrogenase
271 SY yeast alcohol dehydrogenase
272 SY FDH
273 SY FALDH
274 SY 40 kDa allergen
275 SY ADH-A2
276 SY ADH-B2
277 SY ADH-C2
252,0-1 0%
```

# KEGG DataBase

---

## 合法性

在合法性上，对KEGG使用爬虫是模糊的。KEGG的版权协议提到：学术使用者可以免费使用KEGG的网页版和镜像网页，但KEGG的FTP是付费使用的。虽然使用爬虫爬取数据也算是属于使用"website"，但实质上几乎相当于在构建一个本地的FTP。而且它也提到“Non-academic users must understand that KEGG is **not a public database**.”。所以还是应当注意。

## 可行性

如果我们确实需要KEGG的本地数据，KEGG的网页组织是比较规整的，应该是可以实现在一些伪装策略下爬取的。但是如果想要这样使用，应该还是需要想办法圆过合法性上的问题。

另外，下面说到的ExplorEnz似可作为KEGG的平替。所以如果后面我们确实有对KEGG完整本地数据库的必要需求，有了相关的完整的讨论之后再行爬取吧。

## ExplorEnz Database 介绍

---

KEGG数据库的数据是在ExplorEnz数据库的基础上进行字段拓展后的结果。其中KEGG中增加的字段包括：

- **Reaction(KEGG)**：在IUBMB数据中没有完全给出反应式，在KEGG中收录了附加的、更全面的反应式，包括 由IUBMB给出的反应导出的亲子反应 和 与IUBMB无关但KEGG收录的附加反应。
- **Substrate**, **Product**：反应物、生成物
- **Pathway**：指向KEGG pathway的链接
- **Orthology**：指向KEGG Orthology的链接
- **Genes**：指向KEGG GENES的链接

ExplorEnz数据库是[完全公开下载sql格式的](#)，并且是目前看到最易处理的。我的想法是，当前我们还没有确定使用各数据库的哪些字段，如果我们接下来讨论的结果发现不需要使用KEGG中增加的字段，那么我们可以直接使用ExplorEnz作平替，因为KEGG不能像这样轻松地本地访问。

我尝试下载了ExplorEnz在2022.05.08更新的数据库，它的信息：

- 格式：MySQL
- 大小：25.9 MiB

- 数据导入：约 6 分钟

数据信息示例：

```
mysql> show tables;
+-----+
| Tables_in_EnzymeData |
+-----+
| cite                  |
| class                 |
| entry                 |
| hist                  |
| html                  |
| refs                  |
+-----+
6 rows in set (0.02 sec)
```

```
mysql> select ec_num, accepted_name from entry
-> where reaction like '%NAD+%';
+-----+-----+
| ec_num      | accepted_name      |
+-----+-----+
| 1.1.1.1     | alcohol dehydrogenase |
| 1.1.1.4     | (R,R)-butanediol dehydrogenase |
| ...         | ...                 |
| 2.7.1.236   | NAD+ 3'-kinase      |
+-----+-----+
476 rows in set (0.02 sec)
```

## References

[LICENSE AGREEMENT FOR USERS OF BRENDA](#)

[KEGG - Copyright](#)

[ExplorEnz Database - The Enzyme Database](#)

[Can anyone suggest a way to download "KEGG" pathway database?](#)