准备结构库并跑通批量结构搜索和结果筛选

完整操作教程 · 参考固定 · 往返恢复验收

固定Foldseek Release10的单链结构搜索模式，将查询与目标结构建库、运行搜索并导出明确字段，再筛选候选和检查无命中情况。

运行边界：示例代码需按安装版本和真实项目输入复核；未提供真实生物数据，语法或小样本检查不代表科学结果和性能验收。

批量结构搜索需要先明确查询对象、目标库和比对模式。本篇固定Foldseek Release10，处理已有PDB/mmCIF单链结构，以3Di+AA局部比对寻找候选相似结构。它不等于复合物界面搜索，也不证明功能相同；输出解释与硬件需求都依赖所选任务。

先让结构ID和数据库来源可追溯

把已核对的查询结构放在queries目录，目标结构放在targets目录，保留文件来源、版本和链标识。多链文件可能产生多个链条目，不能把文件数量直接当查询数。结构缺失、截短和低置信区域需要记录，避免把局部相似片段解释为整蛋白同源或功能一致。

步骤	材料或输出	验收重点
输入整理	queries与targets结构目录	链身份、组装、来源与重复ID
createdb	查询DB与目标DB多个文件	条目数、头部及所需坐标信息
search	结果DB与临时目录	版本、alignment-type及阈值
convertalis	自定列TSV	明确列名，无命中允许为空
候选复核	序列/结构/注释对照	局部与全长关系及研究证据

在新目录建库并搜索

以下需要兼容的Foldseek和已准备结构目录。示例固定alignment-type=2、4线程及E值阈值0.001；这些是本次协议，不是通用生物学门槛。数据库会生成一组配套文件，保存和迁移时保留完整前缀文件集，不只复制一个主文件。

```
#!/usr/bin/env bash
set -euo pipefail
query_dir=/data/queries
target_dir=/data/targets
test -d "$query_dir" && test -d "$target_dir" || exit 2
mkdir foldseek-search-001
cd foldseek-search-001
foldseek version > version.txt
foldseek createdb "$query_dir" queryDB --threads 4 > createdb-query.log 2>&1
foldseek createdb "$target_dir" targetDB --threads 4 > createdb-target.log 2>&1
foldseek search queryDB targetDB resultDB tmp-search \
  --alignment-type 2 --threads 4 -e 0.001 > search.log 2>&1
foldseek convertalis queryDB targetDB resultDB matches.tsv \
  --format-output query,target,evalue,bits,alnlen,qstart,qend,qlen,tstart,tend,tlen \
  > export.log 2>&1
test -f matches.tsv
```

需要反复查询同一目标集合时，可按该版本文档另行测试createindex，建索引时间、体积和峰值单独记录；一次搜索不应默认必须先做所有索引。目标库更新后用新版本前缀，不能让旧索引与新数据混用。错误日志需保留，空输入不能被记成快速完成的有效任务。

按明确列定义筛选，而非盲目取前十

下面仅演示以E值和查询对齐跨度筛出候选。跨度比例包含起止范围，不等同于逐残基覆盖；有插入缺失时尤其应复核。阈值0.5仅为示例，长短结构域与研究目标需要不同解释，不能据此宣布所有命中为高置信同源。

```
import csv, math
fields=
["query","target","evalue","bits","alnlen","qstart","qend","qlen","tstart","tend","tlen"
]
count=kept=0
with open("matches.tsv",newline="") as source, open("candidates.tsv","x",newline="") as
target:
    writer=csv.DictWriter(target,fieldnames=fields+["query_span"],delimiter="\t")
    writer.writeheader()
    for row in csv.DictReader(source,fieldnames=fields,delimiter="\t"):
        count+=1
        length=int(row["qlen"]); start=int(row["qstart"]); end=int(row["qend"])
        e=float(row["evalue"])
        assert length>0 and 1<=start<=end<=length and math.isfinite(e) and e>=0
        span=(end-start+1)/length
        if e<=0.001 and span>=0.5:
            row["query_span"]=span; writer.writerow(row); kept+=1
print("■■■■",count,"■■■■",kept)
```

结果无命中时TSV可以为空，需先确认查询和目标DB确有正确条目、搜索正常完成且模式与阈值适用。不要把缺文件当零命中，也不要随意放宽阈值后仍按原门槛报告。保存全部原始搜索结果与筛选脚本，便于重新解释边界候选。

模式变化需要重新解释分数字段

Release10的TMalign模式会改变排序和部分输出字段的含义，不能把type=1的evalue列继续按本例type=2的E值筛选。只有预测3Di而没有Cα坐标的序列建库，也不能默认提供TM-score等坐标相关输出。结构、序列语言模型与复合物入口要分别确认。

最后抽查候选结构叠合、对齐区域、序列和注释，核对是否由低质量或很短片段驱动。资源试跑同时记录库残基量、查询量、RAM和临时盘；压缩结构文件大小不能直接换算任务内存，功能解释也不能由服务器配置代替。

配套练习包

本PDF对应的ZIP包含本文代码块、主题练习模板或合成数据、README和SHA256校验文件。请从同一教程页面下载练习包；先核对校验值，再在独立目录复现。

参考资料

Foldseek Release10完整接口与模式 · 来源资料

Foldseek Release10发布说明 · 来源资料

Foldseek命令签名实现 · 来源资料

操作建议：先在合成小样本上验证流程，再依据项目参考、标签合同、存储策略与获配资源验收真实数据。