第11章 生物学数据下载

第12章 生物学数据格式转换

第13章 同源序列分析

生命健康信息学院

解增言

第11章 生物学数据下载

- NCBI
- E-utilities
- Ensembl
- Ensembl API

NCBI

- 美国国家生物技术信息中心（National Center for Biotechnology Information, NCBI）是美国国家卫生研究所（NIH）下属的美国国家医学图书馆（NLM）的一部分，是生物信息学领域最常用的资源门户。

NCBI常用数据库

- Gene　　基因数据库
- Genome　　基因组数据库
- GEO　　基因表达数据库
- Nucleotide　　核酸数据库
- Protein　　蛋白质数据库
- PubMed　　生物学医学文献数据库
- Taxonomy　　物种分类数据库
- ……

E-utilities

- The Entrez Programming Utilities （E-utilities）是NCBI的Entrez数据库系统提供的应用程序接口（API），即为使用程序从NCBI数据库中下载数据提供的接口服务，包括E-search、E-fetch、E-summary等，其形式为带有定义数据库、数据编号和数据类型等信息的URL地址，如：

  https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=protein&id=NP_030436.1&rettype=fasta

一次下载多条数据

- wget-O -
  ”https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=protein&id=NP_191702.1,NP_182120.1,NP_171668.1,NP_194071.1,NP_567178.1,NP_190910.1,NP_181254.1,NP_181255.1,NP_191042.1,NP_181434.1,NP_200874.1,NP_195236.1,NP_179277.1,XP_001694120.1&rettype=fasta”

Ensembl

- Ensembl 是一个提供基因组数据、注释及分析工具的免费开源数据库与软件资源库，服务于生物医学和基因组学研究。
- 由欧洲分子生物学实验室（EMBL-EBI）和威康桑格研究所联合开发维护，1999 年为配合人类基因组计划首次发布。

Ensembl API

- Ensembl API 是 Ensembl 项目提供的编程接口，允许研究人员通过编程方式直接访问 Ensembl 数据库中的基因组数据，包括：

  Perl API

  REST API

  Java API （JEnsembl）

第12章 生物学数据格式转换

- 常用生物序列格式
- 利用shell脚本实现序列格式转换
- 利用生物信息学工具实现序列格式转换

常用生物序列格式

- Fasta
- Fastq
- GenBank
- EMBL

利用shell脚本实现序列格式转换

```
LOCUS       BAH04252                  173 aa            linear   PLN 05-AUG-2014
DEFINITION  LEAFY, partial [Cardamine alpina].
ACCESSION   BAH04252
VERSION     BAH04252.1
DBSOURCE    accession AB378290.1
......
ORIGIN
        1 wnptratvqa lppvppppqq qpattqtaaf gmrlgglegl fgaygirfyt aakiaelgft
       61 astlvgmrde eleemmnsls hifrwellvg erygikaavr aerrrlqeee eesskrrhll
      121 lsaagdsgth haldalsqed dwtglseepv qqqdnqtdaa gnnggyweag kgk
//
```

GenBank

```
>BAH04252 LEAFY, partial [Cardamine alpina].
WNPTRATVQALPPVPPPPQQQPATTQTAAFGMRLGGLEGLFGAYGIRFYTAAKIAELGFT
ASTLVGMRDEELEEMMNSLSHIFRWELLVGERYGIKAAVRAERRRLQEEEEESSKRRHLL
LSAAGDSGTHHALDALSQEDDWTGLSEEPVQQQDNQTDAAGNNGGYWEAGKGK
```

FastA

利用生物信息学工具实现序列格式转换

$ seqret seq.gb seq.fa

- seqret 是 EMBOSS 软件包中的一个核心工具。EMBOSS 是一个开源的、功能非常强大的生物信息学软件套件。
- seqret 的两大核心功能：序列提取与序列格式转换。

第13章 同源序列分析

- BLAST+
- HMMER

BLAST+

- BLAST（Basic Local Alignment Search Tool）是一组在蛋白质或核酸数据中搜索相似序列的分析工具，它能迅速将查询序列（Query）与序列数据库进行比较找出相似序列。
- BLAST有多个版本，其中最常用的是NCBI BLAST+。

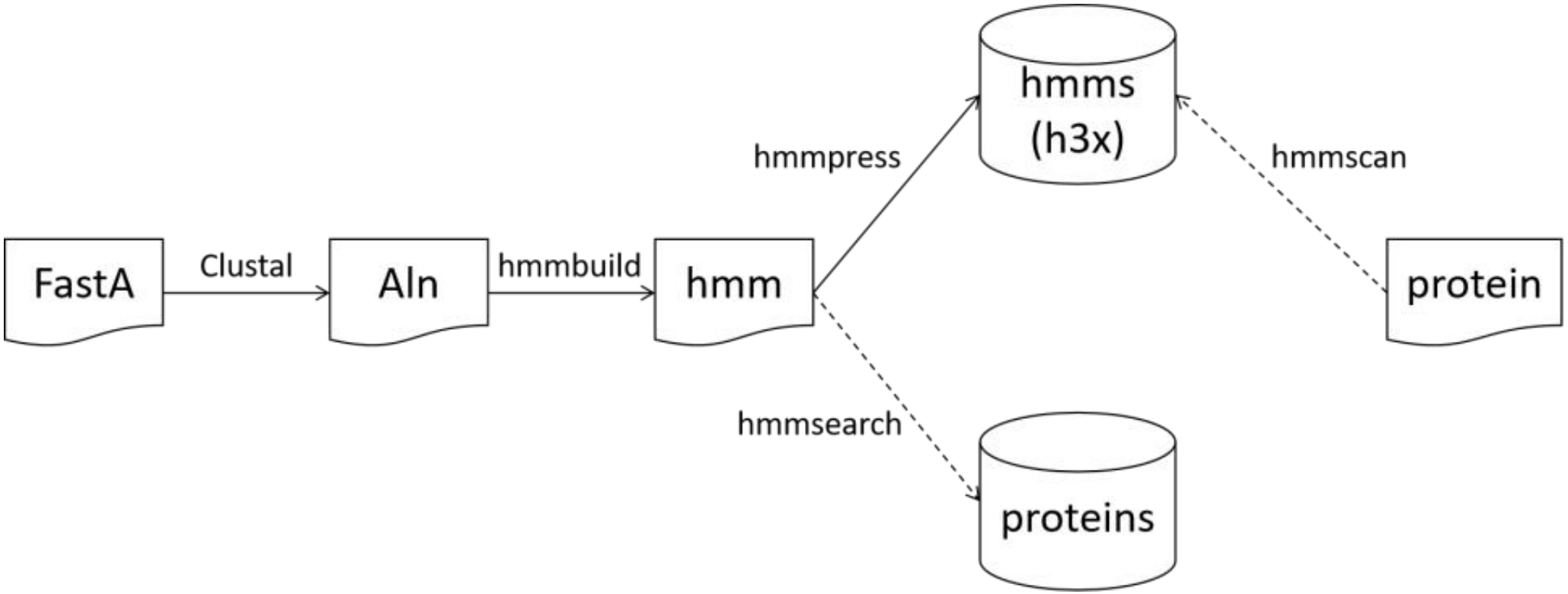
BLAST+

BLAST+程序	作用
makeblastdb	构建本地序列库
blastp	在蛋白质序列库中搜索蛋白质序列
blastn	在核酸序列库中搜索核酸序列
blastx	将给定的核酸序列按照六种阅读框翻译成蛋白质然后与蛋白质序列库中的序列进行比对
tblastn	将给定的蛋白质序列与核酸序列库中序列的六种阅读框进行比对
tblastx	将核酸序列和核酸序列库中的序列按不同的阅读框全部翻译成蛋白质序列，然后进行蛋白质序列比对
blastdbcmd	从序列库中取出指定序列

HMMER

- 蛋白质保守域（Protein Domain）是指蛋白质序列中的一段保守区，长短不一，有的蛋白质包含多个保守域。
- 蛋白质保守域分析最常用的工具是HMMER。HMMER是利用隐马尔科夫模型谱分析生物序列同源性的工具包。

HMMER分析流程


FastA
Clustal
Aln
hmmbuild
hmm
hmmpress
hmms
(h3x)
hmmscan
protein
hmmsearch
proteins

HMMER常用程序

HMMER程序	功能
hmmbuild	用多重比对序列构建HMM模型
hmmpress	格式化HMM数据库，供hmmscan搜索使用
hmmsearch	使用HMM模型搜索序列库
hmmscan	使用序列搜索HMM模型库

单元测试4

15分钟