第11章 生物学数据下载

生命健康信息学院

解增言

主要内容

- 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）

讨论

- Shell中的小括号是不是关键字?

单元测试3