

第12章 生物学数据格式转换



生命健康信息学院

解增言

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 lsaagdsqth haldalsqed dwtglseepv qqdnqtdaa gnnggyweag kgk

//



>BAH04252 LEAFY, partial [Cardamine alpina].

WNPTRATVQALPPVPPPPQQQPATTQTAAFGMRLGGLEGLFGAYGIRFYTAAKIAELGFT
ASTLVGMRDEELEEMMNLSHIFRWELLVGERYGIKAAVRAERRRLQEEEEESSKRRHLL
LSAAGDSGTHHALDALSQEDDWTGLSEEPVQQDNQTDAAAGNNGGYWEAGKGK

GenBank



FastA

EMBOSS

EMBOSS 是一个开源的、功能强大的生物信息学工具集。其中常用的工具有：

- seqret 序列格式转换
- transeq 核酸序列翻译成蛋白质
- revseq 序列反向互补
- needle 全局序列比对
- water 局部序列比对
- pepstats 蛋白质理化性质分析

EMBOSS程序路径： /usr/local/bin/bin, 如/usr/local/bin/bin/seqret



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。
- seqret 的两大核心功能：序列提取与序列格式转换。

单元测试

讨论：路径、绝对路径、相对路径

