Аннотация генами геномов *de novo*

**GLIMMER**

**Загрузите**

wget http://ccb.jhu.edu/software/glimmer/glimmer302b.tar.gz

Выполните последовательно следующие команды (объясните их смысл)

tar xzf glimmer302b.tar.gz

cd glimmer3.02/

cd src/

Make

**Необходимые нам тестовые варианты находятся в папке**

./glimmer3.02/sample-run

Это каталог, содержащий образец прогона Glimmer3. В нем содержится последовательность генома Treponema pallidum (файл tpall.fna) и

список аннотированных генов для него (файл tpall.nh), оба загружены из GenBank.

Этапы: Сперва нам необходимо натренировать модель. Для этого в папке ./glimmer3.02/sample-run уже есть пример с fasta (from-training.train)

**###---Запустим генерацию модели---###**

mkdir -p /home/dmitry/soft/glimmer3.02/sample-run\_new

Используем утилиту long-orfs

bin/long-orfs --help

USAGE: long-orfs [options] <sequence-file> <output-file>

/home/dmitry/soft/glimmer3.02/bin/long-orfs -n -t 1.15 /home/dmitry/soft/glimmer3.02/sample-run/tpall.fna /home/dmitry/soft/glimmer3.02/sample-run\_new/orf.positions

Найденные на предыдущем этапе ORF для тренировки - это результат работы готовой модели. Мы воспользовались уже существующей моделью, только для того, чтобы сгенерировать для себя обучающую выборку.

**### Трансформируем отобранные позиции orf в fasta**

Используем утилиту extract

bin/extract --help

USAGE: extract [options] <sequence-file> <coords>

/home/dmitry/soft/glimmer3.02/bin/extract -t /home/dmitry/soft/glimmer3.02/sample-run/tpall.fna /home/dmitry/soft/glimmer3.02/sample-run\_new/orf.positions > /home/dmitry/soft/glimmer3.02/sample-run\_new/orf.fasta

**###---Теперь обучим марковскую модель---###**

USAGE: build-icm [options] output\_file < input-file

/home/dmitry/soft/glimmer3.02/bin/build-icm -r /home/dmitry/soft/glimmer3.02/sample-run\_new/run1.icm < /home/dmitry/soft/glimmer3.02/sample-run\_new/orf.fasta

**###---Используем обученную модель, чтобы предсказать гены в геноме---###**

USAGE: glimmer3 [options] <sequence-file> <icm-file> <tag>

/home/dmitry/soft/glimmer3.02/bin/glimmer3 -o50 -g110 -t30 /home/dmitry/soft/glimmer3.02/sample-run/tpall.fna /home/dmitry/soft/glimmer3.02/sample-run\_new/run1.icm /home/dmitry/soft/glimmer3.02/sample-run\_new/run1

Ваши результаты находятся в двух файлах

run1.predict

run1.detail

Изучите координаты полученных ORF. Используйте /home/dmitry/soft/glimmer3.02/bin/extract, чтобы найти их fasta последовательность

Как бы вы нашли соответствуют ли данные последовательности генам?

Подсказка: Используйте BLAST, BLAT. Посмотрите на координаты этих последовательностей в геномном браузере.

**EXTRA:**

|  |  |
| --- | --- |
| **GeneMark** |  |

<http://opal.biology.gatech.edu/GeneMark/>

**Prodigal**

<https://github.com/hyattpd/Prodigal>