Отчёт по изучению Arabidopsis Thaliana

Документ создан 11 октября 2014 г..

- **Выбор родственного организма**

  Используется бобовое растение люцерна. По филогенетическому дереву двудольных, семейство бобовых находится ближе к семейству Капустных (к которому относится арабидопсис), чем также секвенированное семейство Паслёновые.

  Краткую русскую информацию можно почитать здесь:
  $https://ru.wikipedia.org/wiki/Medicago_truncatula$

  Ссылка на скачивание генома:
  $http://www.jcvi.org/medicago/display.php?pageName=General\&section=Download$

  Полная карта каждой из 8 хромосом:
  $http://www.medicagohapmap.org/downloads/mt40$

- **Blastn**

  Необходимо произвести выравнивание всех доступных Expressed sequence tags (ESTs) организма Mendicago truncatula на геном Arabidopsis Thaliana при помощи Blastn, чтобы найти гомологичные последовательности в геноме организма Arabidopis Thaliana (скрипт $run\_blastn.sh$).

  ESTs доступны для скачивания по адресу:
  $www.plantgdb.org/download/Download/PublicPlantSeq/Dump/M/Medicago_truncatula/FASTA/Medicago_truncatula.mRNA.EST.fasta.bz2$

  Выходной файл представляет собой *tab*-separated файл вида: 1-й и 2-й столбцы содержат начало и конец найденной гомологичной последовательности, а 3-й и 4-й - её *id* и представление, соответственно. Такой формат очень просто получить с использованием опции $outfmt$.

  В скрипте указан порог percent identity в 85%. Это сделано с целью исключения гомологичных последовательностей, содержащих значительное количество *gap*-ов и/или несовпадений.

- **Подготовка последовательностей для Augustus**

  Каждая последовательность из output'a blastn, полученная на предыдущем этапе, проверяется на наличие start, stop кодонов и ORF (скрипты $get\_homologous\_sequences$ и $filter\_seq.py$).

Если последовательность не является корректной, то она не входит в training set Augustus'a для поиска генов в Arabidopsis thaliana.

Training set сохраняется в формате genbank или gff.