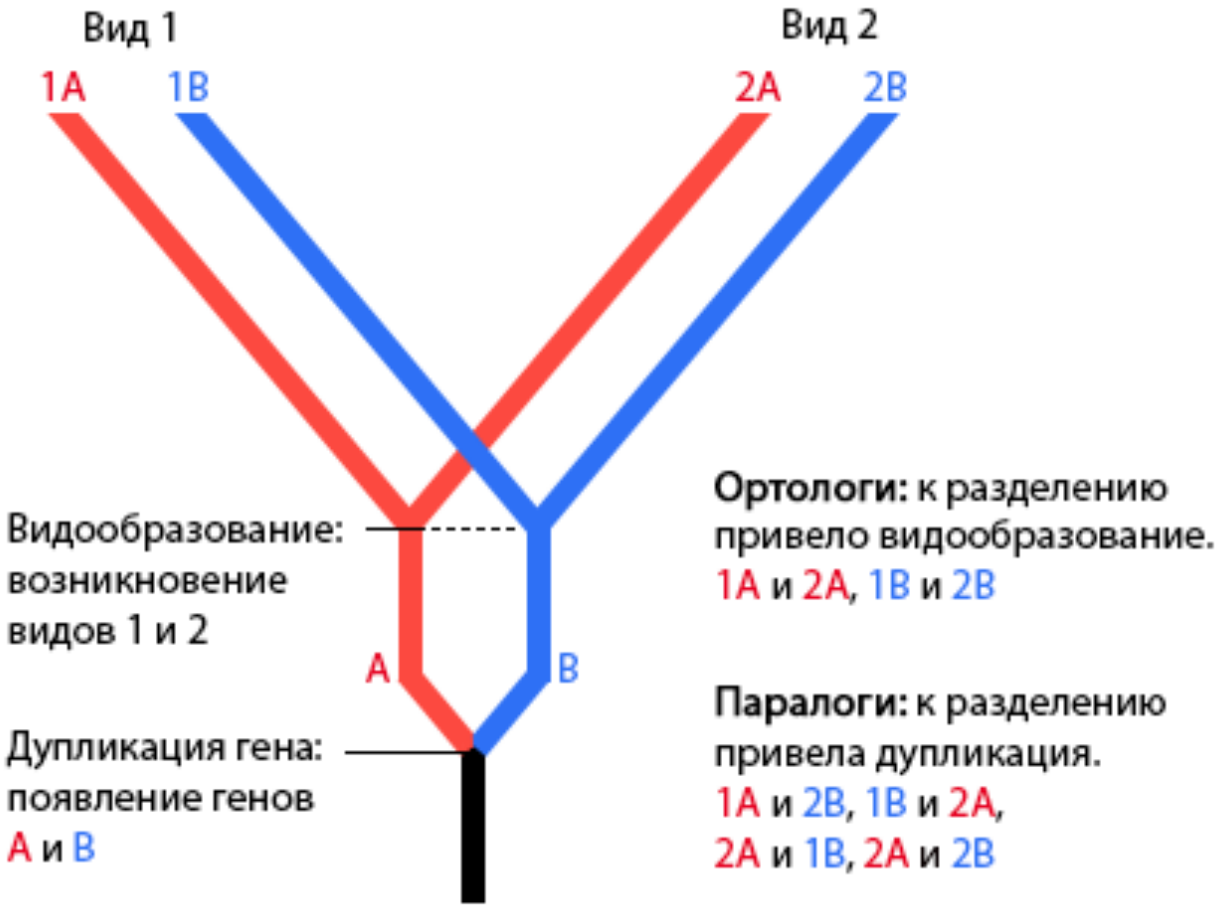
Разработка инструмента для аннотации генов через поиск групп ортологов

Владислав Савельев

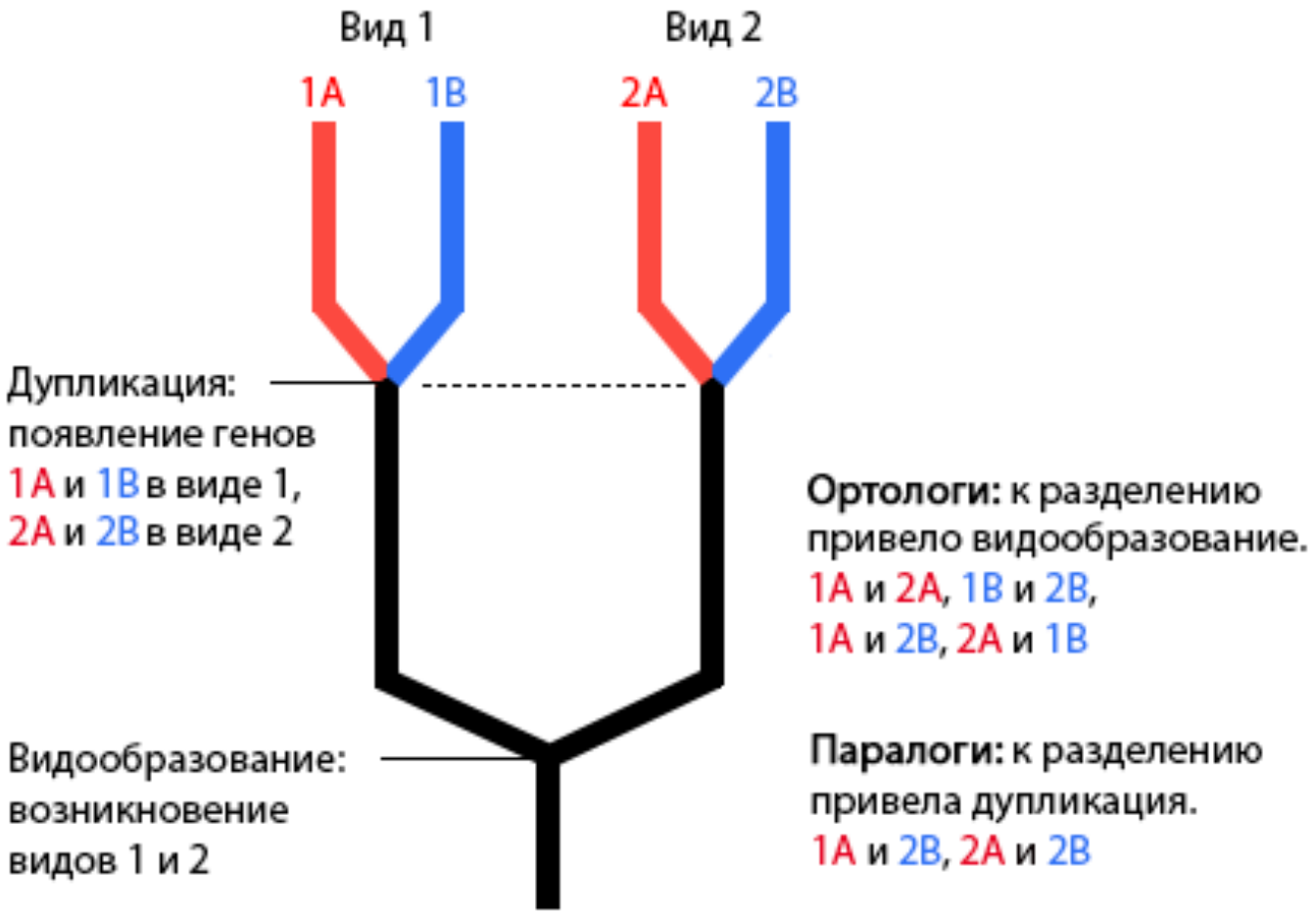
Научный руководитель: Алексей Гуревич

Академический университет, 2014

Гомология, ортология и паралогия



Гомология, ортология и паралогия



Аннотация генома

1. Структурная аннотация (предсказание генов).
2. Функциональная аннотация, маркирование генов биологической информацией:
   - кодируемый белок
   - биологическая функция
   - регулятивная функция и взаимодействие с другими генами
   - экспрессия

   Используется сопоставление с данными из биологических баз данных (Ensembl, RefSeq).

Цель и задачи

Цель: разработать программное обеспечение, помогающее эффективно аннотировать гены в большом наборе (50–200) бактериальных сборок через поиск групп ортологов совместно с изученными геномами.

Задачи

1. Изучить существующие подходы поиска ортологов
2. Выбрать метод, адаптировать к условиям задачи и создать инструмент для поиска групп ортологов
3. Добавить возможность расширения групп по новым геномным сборкам
4. Добавить функциональность автоматического поиска неидентифицированных генов в биологических базах
5. Оптимизировать программу для вычислений на кластере

Подходы

1. На основе *филогенетического дерева.*

   MetaPhOrs, EnsemblCompara, PhylomeDB.

   Точно, но долго.

2. На основе *попарного сравнения последовательностей* (идея: ортологи меньше разошлись в эволюции, чем паралоги).

   InParanoid, OrthoMCL, COG, EggNOG.

   Быстро, но не так точно.

3. На основе *порядка генов.*

   YGOB.

Методы на основе попарного сравнения последовательностей

1. InParanoid (O'Brien *et al.*, 2005 г.) — поиск ортологов только между двумя геномами.

2. OrthoMCL (Li *et al.*, 2008 г.), OrthoInspector (Linard *et al.*, 2011) — кластеризация для поиска в группе геномов. Нет полноценного или расширяемого ПО.

3. COG (Kristensen *et al.*, 2010 г.), EggNOG (Powell *et al.*, 2013 г.) — онлайн-базы заранее подсчитанных групп ортологов и веб-интерфейс для загрузки пользовательских геномов.

Разработанный инструмент

1. Входные данные

- Известные геномы: ID в базе RefSeq или имена штаммов.

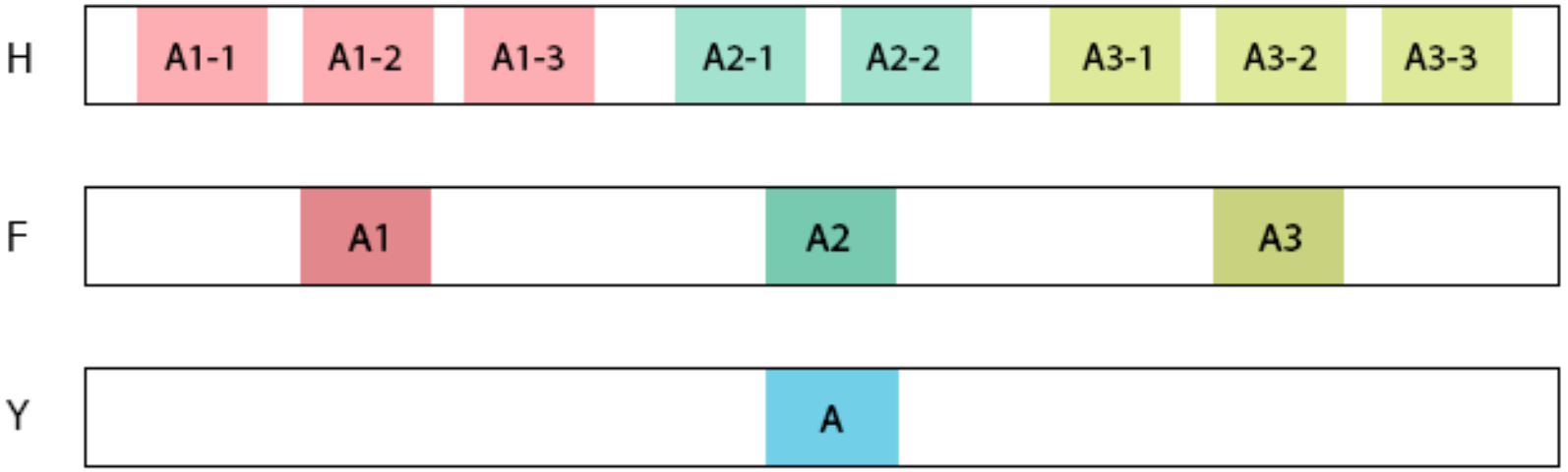
  Аннотации автоматически скачиваются из базы

- Новые сборки: контиги в FASTA-файле.

  Гены предсказываются с помощью Prodigal (Hyatt *et al.*, 2010 г.)

Гены транслируются в протеины и фильтруются по длине и проценту стоп-кодонов.

2. Попарное сравнение

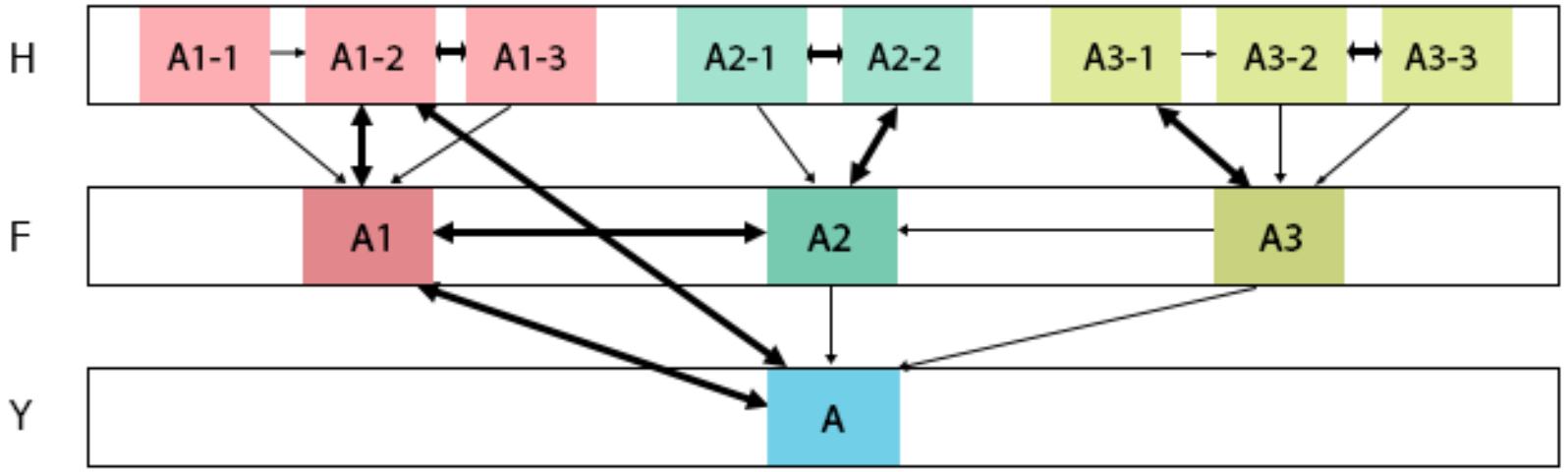


Протеины сравниваются попарно с помощью BLAST (Altschul *et al.*, 1990 г.)

GenomeID|ProtID	**GenomeID|ProtID**	**Score**	**E-value**
H|A1-1	F|A1	122	8e-40
H|A1-1	H|A1-2	194	5e-67
Y|A	F|A1	136	4e-44

...

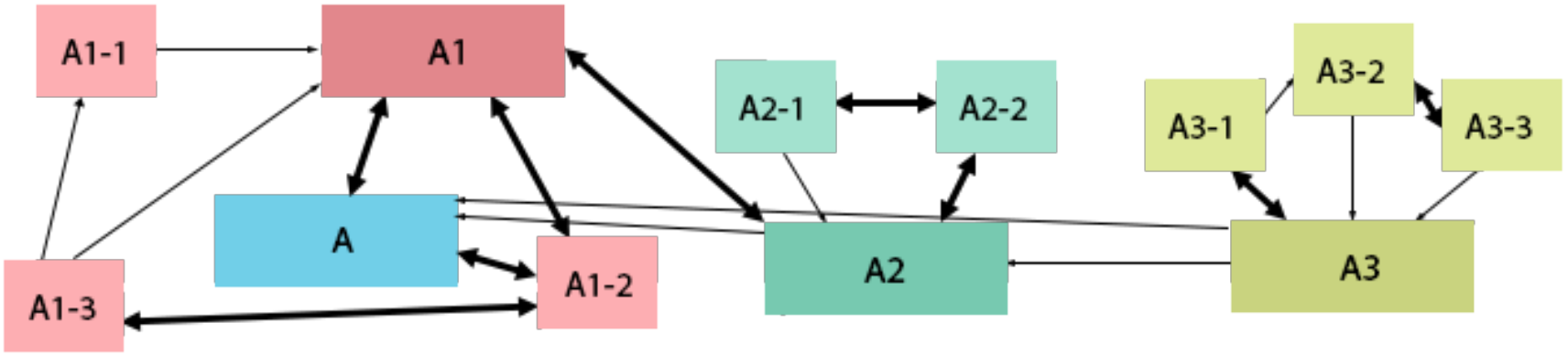
3. Выбор лучших взаимных соответствий



Между геномами — предполагаемые ортологи.

В пределах генома — предполагаемые недавние паралоги.

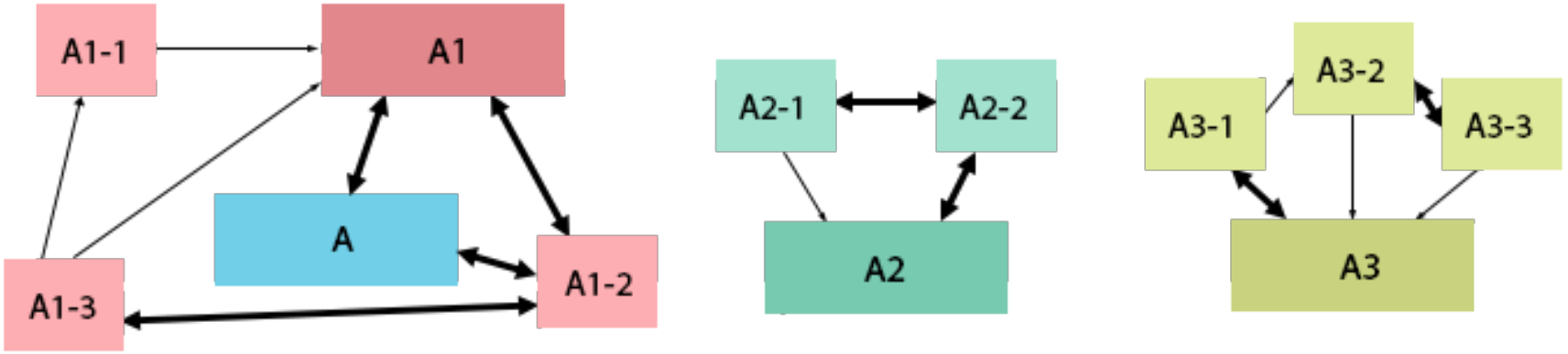
4. Построение графа



Узлы — протеины из пар «лучших взаимных соответствий».

Ребра взвешены значениями $-\log_{10}$(E-value)

5. Кластеризация



Алгоритм марковской кластеризации MCL (van Dongen *et al.*, 2000 г.)

1. A, A1, A1-1, A1-2, A1-3
2. A2, A2-1, A2-2
3. A3, A3-1, A3-2, A3-3

Вывод программы

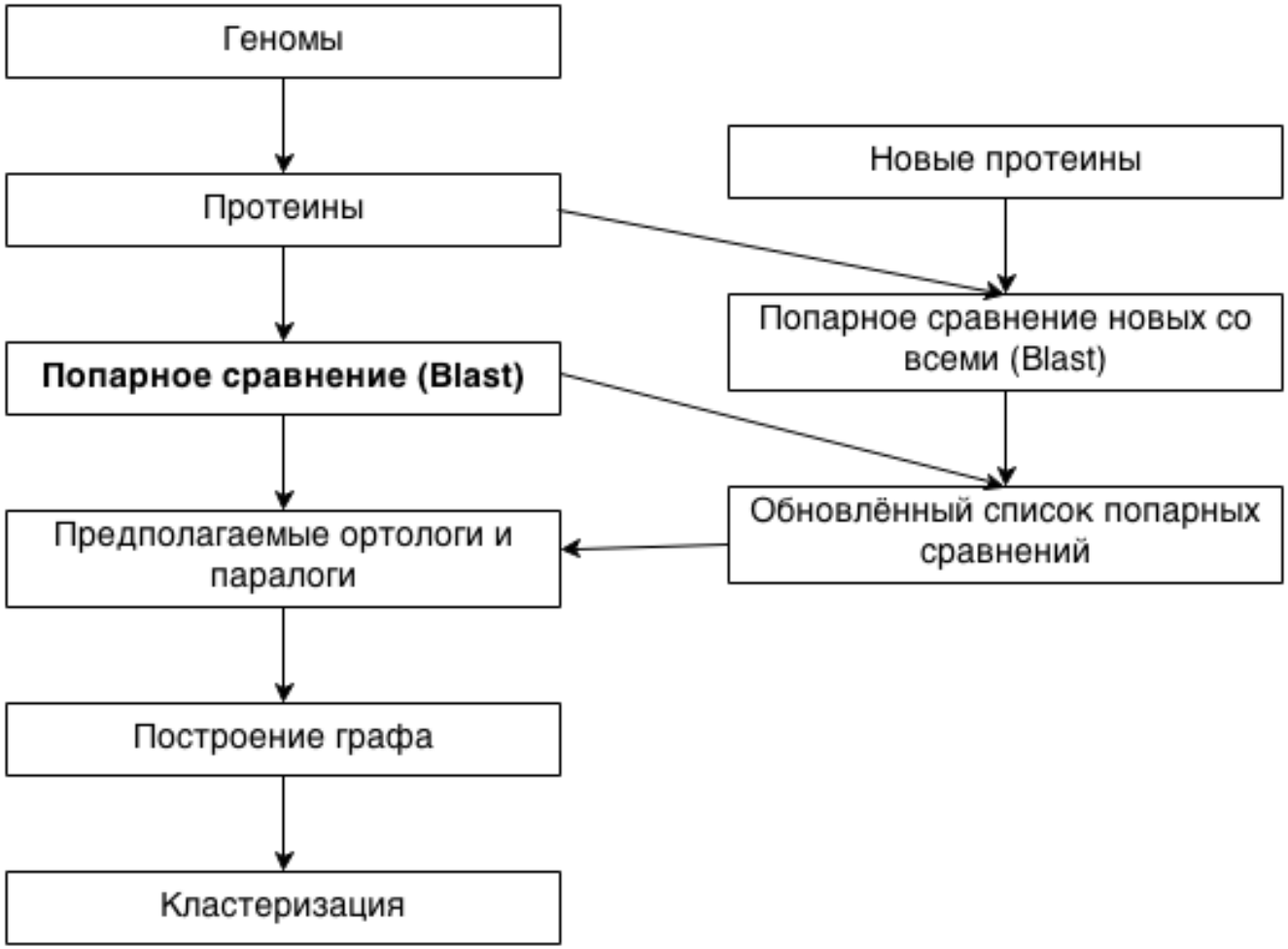
Файл со списком групп ортологов:

```
1
Human, Chr X                    Hsapiens          NP_003819    NA       MTMR1        myotubularin-related protein 1
Human, Chr 11                   Hsapiens          NP_057240    NA       MTMR2        myotubularin-related protein 2 isoform 1
Human, Chr 16                   Hsapiens          A1L3X4.1     NA       MT1DP        metallothionein 1D, pseudogene
Drosophila melanogaster         Dmelanogaster     Q9VMI9       CG9115   Dmel_CG9115  CG9115-PA (LD28822p)
Saccharomyces cerevisiae S288c  ScerevisiaeS288c  YJR110W      YJR110W  YMR1         Phosphoinositide 3-phosphatase

2
Human, Chr 13                   Hsapiens          NP_066576    NA       MTMR3        myotubularin related protein 3 isoform c
Human, Chr 17                   Hsapiens          NP_004678    NA       MTMR4        myotubularin related protein 4
Drosophila melanogaster         Dmelanogaster     Q7YU03       CG3632   Dmel_CG3632  LD11744p

3
Human, Chr 13                   Hsapiens          NP_004676    NA       MTMR6        myotubularin related protein 6
Human, Chr 8                    Hsapiens          NP_004677    NA       MTMR7        myotubularin related protein 7
Human, Chr X                    Hsapiens          NP_060147    NA       MTMR8        myotubularin related protein 8
Drosophila melanogaster         Dmelanogaster     Q8MLR7       CG3530   Dmel_CG3530  CG3530-PA, isoform A
```

Расширение групп



Вычисление на кластере

Распределяется самый вычислительно долгий шаг — попарное сравнение с помощью BLAST.

Пример: попарное сравнение трёх протеинов [1, 2, 3]:

- Разбить протеины на 3 части: [1], [2], [3].
- Запустить 3 задачи:
  - [1] сравнить с [1, 2, 3],
  - [2] сравнить с [1, 2, 3],
  - [3] сравнить с [1, 2, 3].
- Объединить результаты этих задач.

BLAST в 4 потоках — 46 минут

4 задачи на кластере — 17 минут

Результат

На основе идеи попарного сравнения и кластеризации создан инструмент для поиска групп ортологов в бактериальных геномах.

Инструмент обладает следующими свойствами:

- Автоматическое скачивание аннотаций известных геномов
- Эффективное расширение групп
- Параллелизация на кластере
- Автоматический поиск неидентифицированных генов в биологических базах данных.

С использованием инструмента было проаннотировано более 500 новых бактериальных сборок в компании AstraZeneca.

Использованные технологии

- Язык программирования Python
- Библиотека BioPython
- Инструменты BLAST, MCL, Prodigal
- Oracle Grid Engine для распределенных вычислений
- База NCBI RefSeq и её веб-API Entrez
- SQLite для хранения промежуточных результатов

Спасибо за внимание.

Вывод программы

Файл со списком групп ортологов:

1

Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_286725.1	Z1190	NA	glucosyl transferase
Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_287133.1	Z1629m	NA	glycosyl transferase
Escherichia coli 536	NC_008253.1	YP_668240.1	ECP_0306	NA	glycosyl transferase family protein

2

Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_286727.1	Z1192	NA	IS1 protein InsB
Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_287135.1	Z1632	NA	IS1 protein InsB
Escherichia coli 55989	NC_011748.1	YP_002401510.1	EC55989_0390	insB	IS1 transposase InsAB'
Escherichia coli 55989	NC_011748.1	YP_002401628.1	EC55989_0512	insB	IS1 transposase InsAB'

3

Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_287481.1	Z2012	galU	UTP-glucose-1-phosphate uridylyltransferase
Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_288547.1	Z3205	galF	UTP-glucose-1-phosphate uridylyltransferase
Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_290420.1	Z5300	rffH	glucose-1-phosphate thymidylyltransferase
Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_289975.1	Z4792	glgC	glucose-1-phosphate adenylyltransferase

4

Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_285756.1	Z0068	polB	DNA polymerase II
Escherichia coli 536	NC_008253.1	YP_668000.1	ECP_0061	NA	DNA polymerase II
Escherichia coli 55989	NC_011748.1	YP_002401197.1	EC55989_0058	polB	DNA polymerase II

5

Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_285757.1	Z0069	araD	L-ribulose-5-phosphate 4-epimerase
Escherichia coli 536	NC_008253.1	YP_668001.1	ECP_0062	araD	L-ribulose-5-phosphate 4-epimerase
Escherichia coli 55989	NC_011748.1	YP_002401198.1	EC55989_0059	araD	L-ribulose-5-phosphate 4-epimerase

Вывод программы

Файл со списком групп ортологов:

1

Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_286725.1	Z1190	NA	glucosyl transferase
Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_287133.1	Z1629m	NA	glycosyl transferase
Escherichia coli 536	NC_008253.1	YP_668240.1	ECP_0306	NA	glycosyl transferase family protein

2

Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_286727.1	Z1192	NA	IS1 protein InsB
Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_287135.1	Z1632	NA	IS1 protein InsB
Escherichia coli 55989	NC_011748.1	YP_002401510.1	EC55989_0390	insB	IS1 transposase InsAB'
Escherichia coli 55989	NC_011748.1	YP_002401628.1	EC55989_0512	insB	IS1 transposase InsAB'

3

Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_287481.1	Z2012	galU	UTP-glucose-1-phosphate uridylyltransferase
Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_288547.1	Z3205	galF	UTP-glucose-1-phosphate uridylyltransferase
Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_290420.1	Z5300	rffH	glucose-1-phosphate thymidylyltransferase
Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_289975.1	Z4792	glgC	glucose-1-phosphate adenylyltransferase
A235	A235	A235_P1331	-	-	-
A235	A235	A235_P1325	-	-	-

4

Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_285756.1	Z0068	polB	DNA polymerase II
Escherichia coli 536	NC_008253.1	YP_668000.1	ECP_0061	NA	DNA polymerase II
Escherichia coli 55989	NC_011748.1	YP_002401197.1	EC55989_0058	polB	DNA polymerase II

5

Escherichia coli O157:H7 str. EDL933	NC_002655.2	NP_285757.1	Z0069	araD	L-ribulose-5-phosphate 4-epimerase
Escherichia coli 536	NC_008253.1	YP_668001.1	ECP_0062	araD	L-ribulose-5-phosphate 4-epimerase
Escherichia coli 55989	NC_011748.1	YP_002401198.1	EC55989_0059	araD	L-ribulose-5-phosphate 4-epimerase
A235	A235	A235_P1223	-	-	-