

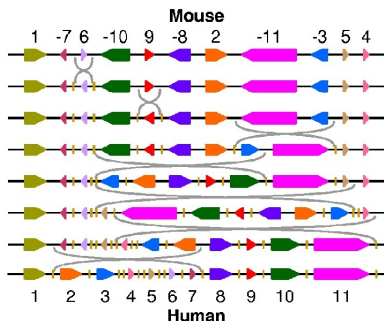
Разработка алгоритмов реконструкции и линеаризации предковых геномов с учетом вставок и удалений генов.

Авдеев Павел Вадимович

научный руководитель: PhD, М.А. Алексеев

СПб АУ НОЦНТ РАН

16 июня 2015 г.



Структурные изменения:

- Стандартные перестроечные операции - DCJ (инверсия, транслокация, разделение, слияние)
- События вставки и удаления генов
- Транспозиции
- Дупликации (полное дублирование генома, тандемные дупликации, ретротранспозиции)

Зачем все это нужно?

- 1 Понимание механизмов эволюции
- 2 Радиационная биология
- 3 Скаффолдинг с множественным референсом

Три главных задачи с точки зрения биологии:

- 1 Восстановление предковых геномов
- 2 Восстановление эволюционной истории между геномами
- 3 Восстановление филогенетического дерева

Линейная проблема медианы [2014 год]

Для линейных геномов G_1, G_2, G_3 и построить линейный медианный геном M' , такой что $\sum_{i=1}^3 d(M', G_i)$ была минимальна.

Малая филогенетическая проблема

Для *филогенетического дерева* и *существующих геномов* (листья дерева) найти **предковые геномы** (внутренние узлы) так, чтобы сумма расстояний по каждому ребру дерева была минимальна.

- **Вероятностная модель**

InferCARs (2006), InferCARsPro (2010), DupCARs (2008), PMAG (2013), PMAG+ (2014), GapAdj (2012).

- **Модель, основанная на анализе перестроек**

- **Решение проблемы медианы**

GASTS (2011), GRAPPA (2001), MGR (2002), BPAAnalisys (1998), SCJ (2013), EMRAE (2008)

- **Методы, использующие множественный брейкпоинт граф**

MGRA (2009)

- EGCHL (2011)

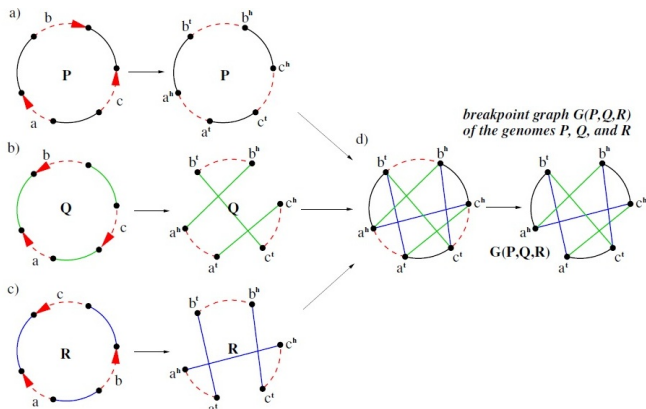
- **Модель, основанная на выравнивании**

MultiOrthoAlign (2014), progressiveCactus (2014)

- Решить малую филогенетическую проблему с учетом событий вставок и удалений генов на основе модели, предложенной в MGRA.
- Реализовать алгоритм линейаризации медианного генома, предложенный Шуаи Джиангом и Максом Алексеевым в 2014 году, и расширить его на события вставок и удалений.

- Исследовать модель, предложенную в MGRA
- Разработать алгоритм, позволяющий работать с данными, которые подверглись большому количеству структурных изменений
- Исправить доказательство основной теоремы в линейаризации медианного генома
- Расширить модель и алгоритм линейаризации для событий вставок и удалений генов
- Реализовать применение алгоритма линейаризации к малой филогенетической проблеме

Множественный брейкпоинт граф



a) $P = (+a + b - c)$ – черно-красный цикл.

b) $Q = (+a - b + c)$ – зелено-красный цикл.

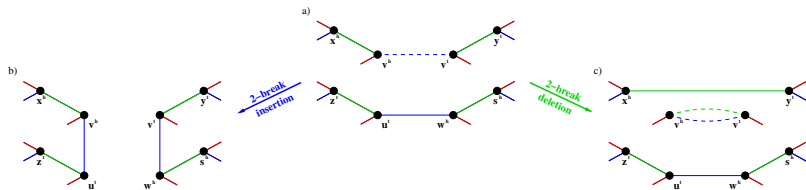
c) $R = (+a - c - b)$ – сине-красный цикл.

d) $G(P, Q, R)$ с и без наблюдаемых ребер.

Балансировка графа

Вершина *несбалансированна*, если её степень меньше количества геномов.

Ребро, называется *протезным*, если оно соединяет две вершины, соответствующие концам одного гена.



(Avdeyev P., Jiang S., et. al., 2015)

- 1 Обработка медианных адекватных подграфов**
Обобщение понятия хороших путей и адекватных подграфов.
- 2 Обработка немобильных ребер**
Обобщение и переосмысление понятия «справедливых» ребер.
- 3 Обработка немобильных ребер с «вилками»**
Новый тип событий – «клонирование» ребер.
- 4 Увеличение компонент связности**
Обобщение поиска DCJ при двух геномах.
- 5 Раунды для борьбы с брейкпоинт-переиспользованием**
Введение понятие раундов и дробление ребер.
- 6 Метод грубой силы**
Присвоение каждому ребру веса и поиск минимального паросочетания.

Линеаризация без учета событий вставок и удалений

Теорема

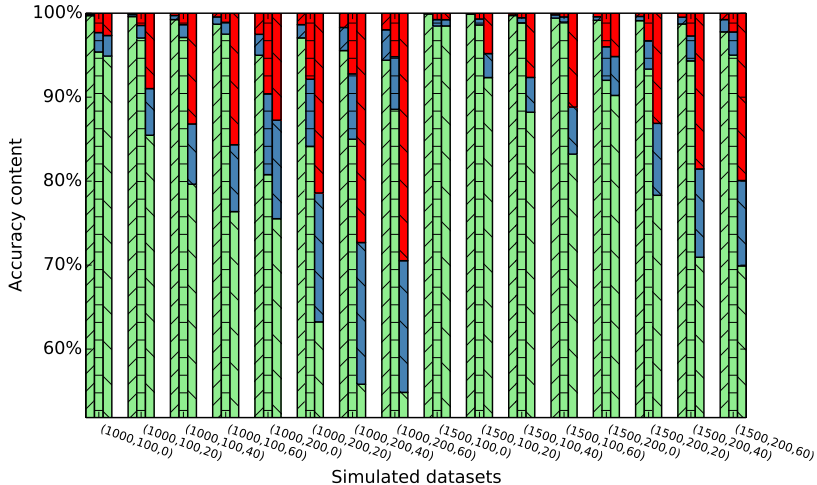
Пусть $P \xrightarrow{z} Q$ - трансформация между геномами P и Q с $c(P) > c(Q)$. Тогда существует трансформация $P \xrightarrow{r} P' \xrightarrow{z'} Q$, такая что r - единственная DCJ, $c(P') = c(P) - 1$, и $|z'| = |z| - 1$.

Доказательство

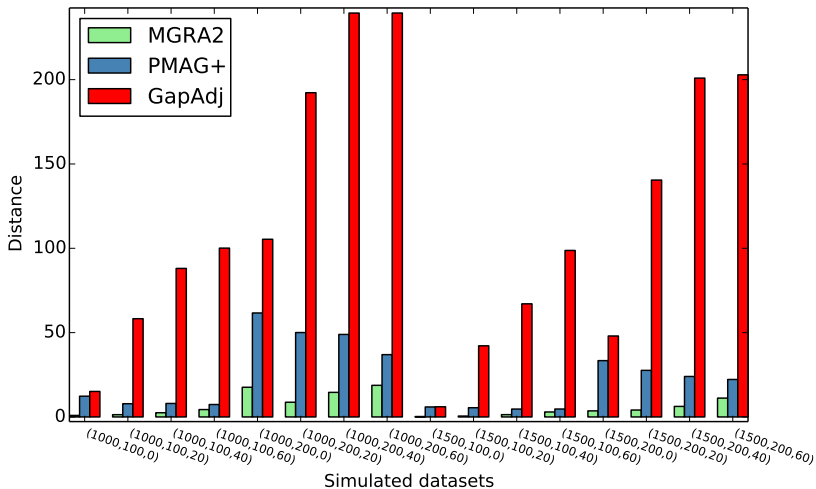
По индукции. Существует три случая:

- $c(P_0) > c(P_1)$
- $c(P_0) = c(P_1) > c(P_2)$
- $c(P_0) < c(P_1) > c(P_2) > c(P_3)$ (не было доказано в статье)

Количество правильных смежностей для шести геномов



Расстояние между геномами для шести геномов



- 1 Исследована модель брейкпоинт графа и доказаны несколько свойств
- 2 Разработан набор эвристик, дающих хорошие результаты
- 3 Исправлено доказательство основной теоремы для алгоритма линеаризации
- 4 Придуман способ учитывать события вставок и удалений в брейкпоинт графе и в алгоритме линеаризации
- 5 Все алгоритмы реализованы на C++. Они получили название **MGRA2**
github.com/ablab/mgra, mgra.cblab.org

Спасибо за внимание