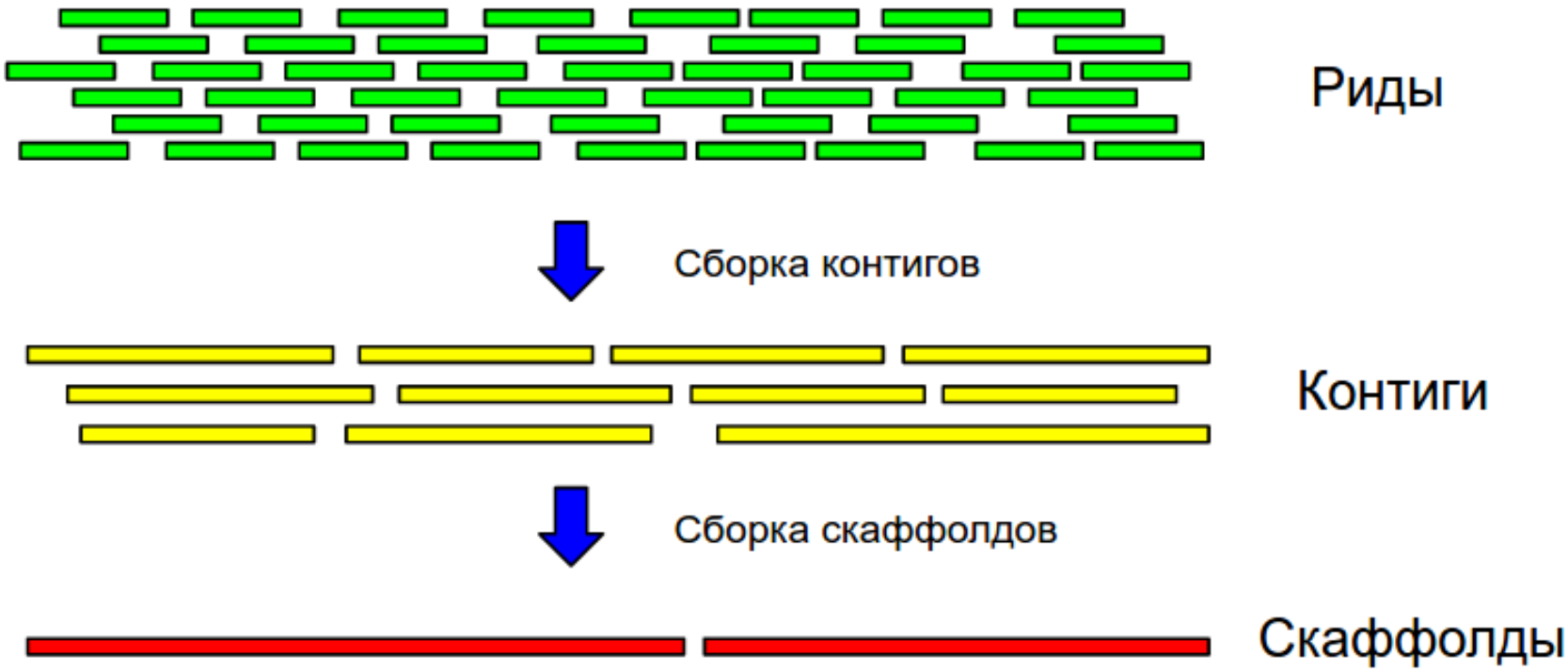
Построение скаффолдов на основе анализа сборок нескольких родственных геномов

Надия Ситдыкова

Руководитель: Алексеев М.А., PhD

СПбАУ РАН

Сборка генома



Существующие методы построения скаффолдов

Технологические решения (прыгающие библиотеки, длинные риды)
Недостатки:

- Дорого
- Неточно

Выравнивание контигов на референсный геном
Недостатки:

- Ошибки из-за структурных вариаций

Ragout
Недостатки:

- Рассчитан на бактерий
- Строит скаффолды только для одного генома

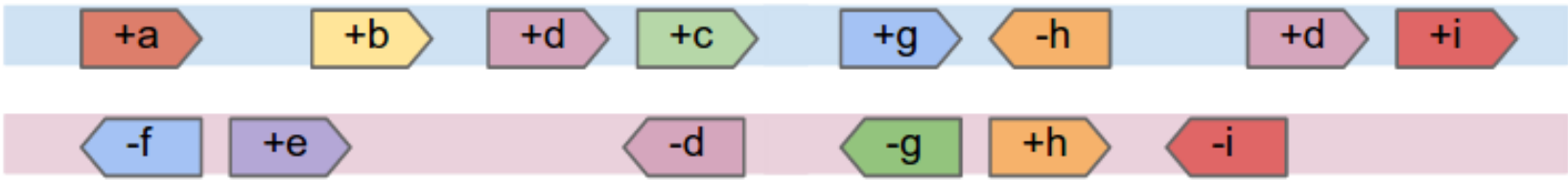
Постановка задачи

Цель:
Разработать алгоритм построения скаффолдов.
Задачи:

- Изучить свойства брейкпоинт графа для фрагментированных геномов
- Разработать алгоритм, опирающийся на свойства брейкпоинт графа
- Расширить алгоритм использованием информации о повторах

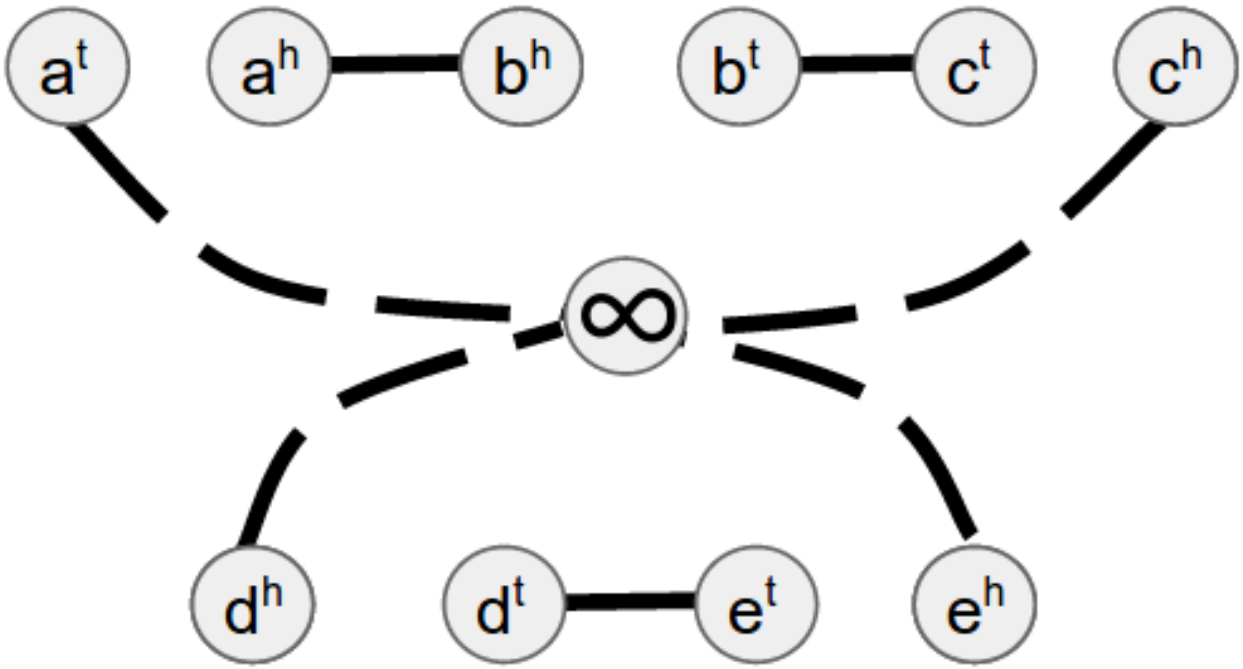
Геном в виде знаковых перестановок

Каждая хромосома – последовательность генов.



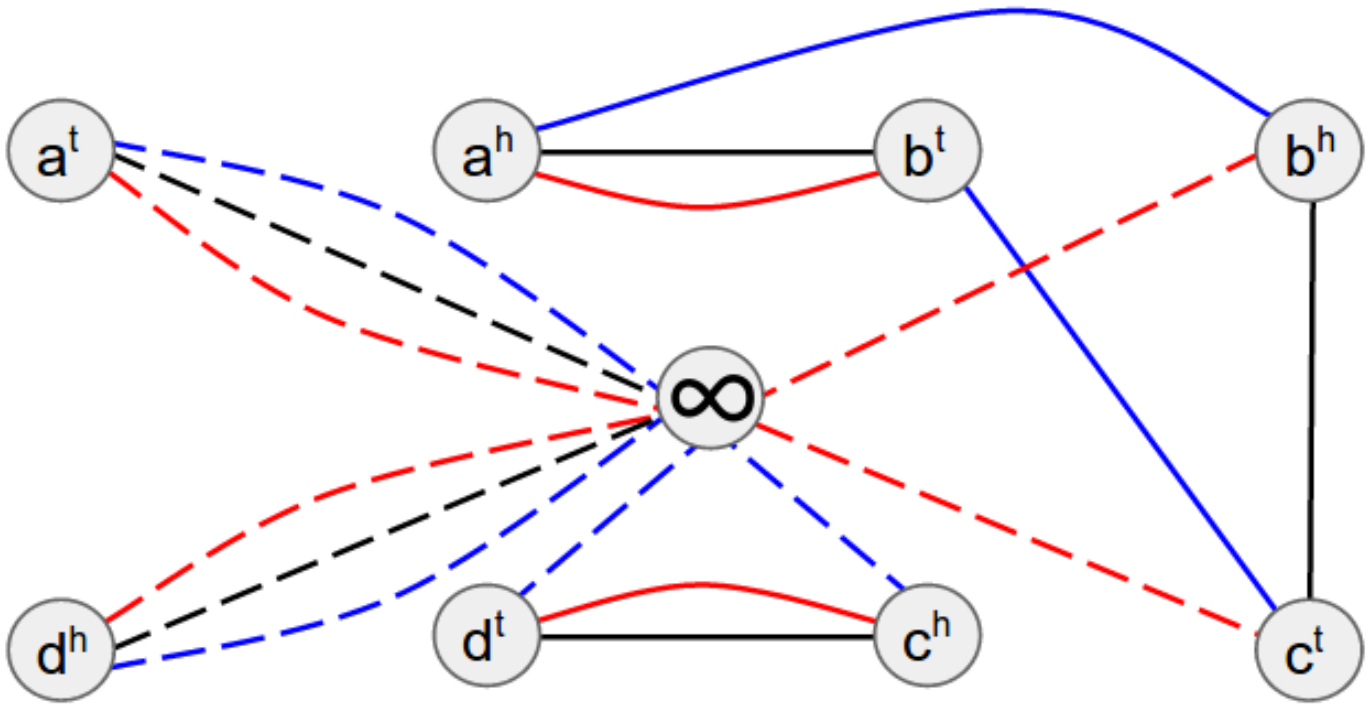
$$G = [a, b, d, c, g, -h, d, i], [-f, e, -d, -g, h, -i)]$$

Брейкпоинт граф



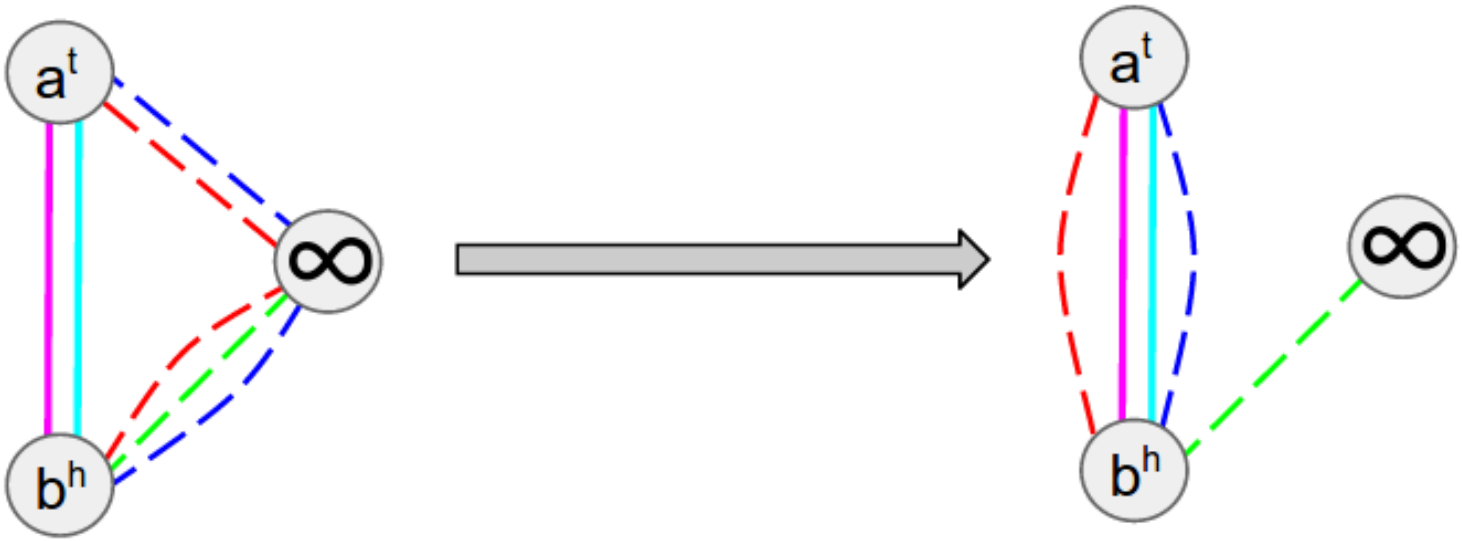
Брейкпоинт граф для генома из двух линейных хромосом
$G = \{(+a\ -b\ +c), (-d\ +e)\}$

Множественный брейкпоинт граф



Множественный breakpoint граф для $G_1 = \{(+a\ +b\ +c\ +d)\}$, $G_2 = \{(+a\ +b), (+c\ +d)\}$, $G_3 = \{(+a\ -b\ +c), (+d)\}$.
Цвета ребер соответсвуют цветам геномов: C_1 — черный, C_2 — красный, C_3 — синий

Базовый алгоритм



Фрагмент брейкпоинт графа до и после применения операции $scaffold(e_1, e_2, Q)$ для иррегулярных мультиребер $e_1 = (a^t, \infty), e_2 = (b^h, \infty)$ и мультицвета $Q = \{\text{красный, синий}\}$

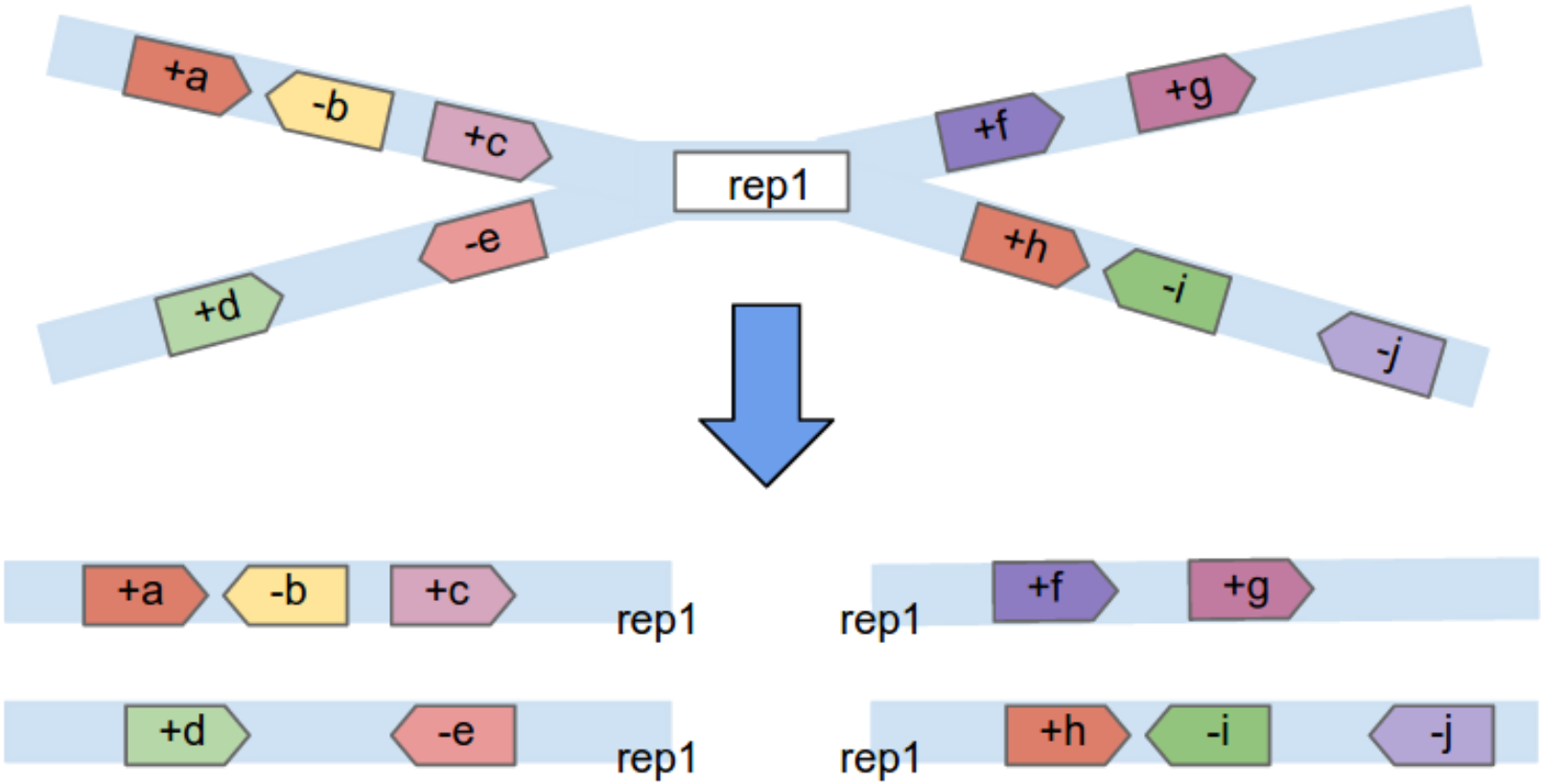
Базовый алгоритм

Операция *scaffold* применяется к ребрам $e_1 = (u, \infty), e_2 = (v, \infty)$ цвета Q, если выполнены условия:

- В графе BG уже есть ребро (u, v)
- $Score(u, v, Q) > 1$
- $\forall e = (x, \infty)\ Scaffold(u, x, Q) < Score(u, v, Q)$
  $\forall e = (x, \infty)\ Scaffold(v, x, Q) < Score(u, v, Q)$

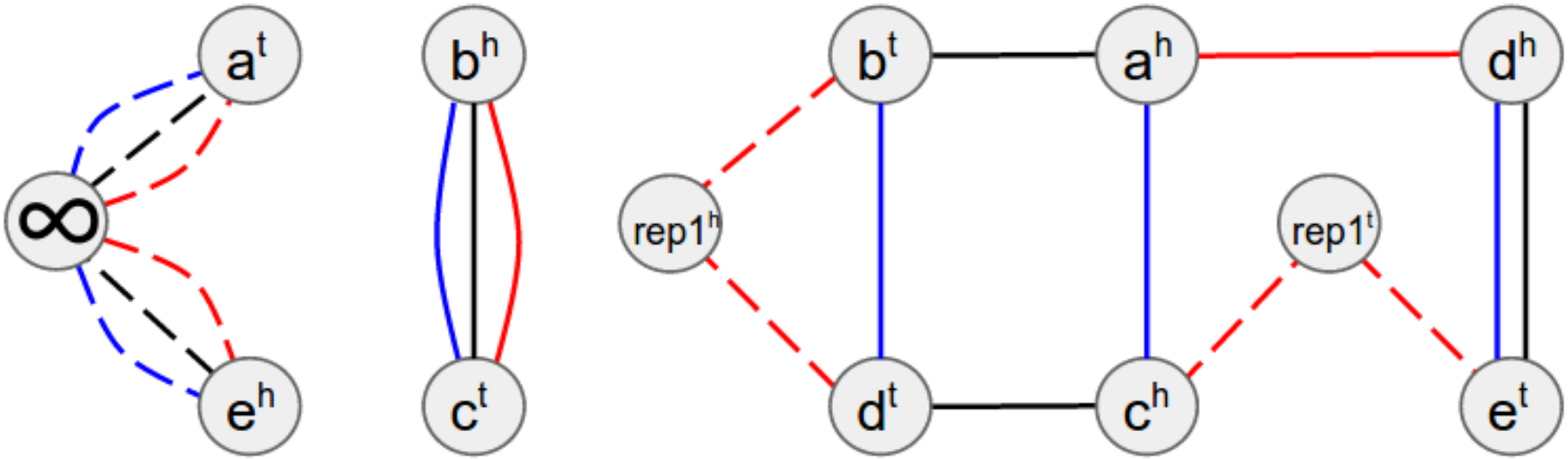
Расширенный алгоритм

Повторы — причина фрагментации генома



Используя информацию о повторах от ассемблера, помечаем концы иррегулярных ребер соответсвующими повторами

Расширенный алгоритм



Расширенный множественный брейкпоинт граф для полных геномов $G_1 = \{(+a\ +b\ +c\ +d\ +e)\}$, $G_2 = \{(+a\ -c\ -b\ +d\ +e)\}$ и сборки $G_3 = \{(+a\ -d)_{rep1},\ _{rep1}(-c\ -b)_{rep1},\ _{rep1}(+e)\}$

Результаты: характеристики

$$TP = \frac{correct}{all} \times 100\%$$

$$FP = \frac{incorrect}{all} \times 100\%$$

где

all — число пар контигов, которые смежны в геноме

correct — число пар контигов, которые смежны в построенных скаффолдах и смежны в геноме

incorrect — число пар контигов, которые смежны в построенных скаффолдах, но не смежны в геноме

Результаты: сборка одного генома

	Базовый алгоритм		Расширенный алгоритм		Ragout	
	TP(%)	FP(%)	TP(%)	FP(%)	TP(%)	FP(%)
6 M	6.74	7.79	24.98	10.49	4.47	13.21
4 M	6.78	7.05	24.58	9.83	4.82	13.52
5 П	7.77	5.99	27.91	9.52	6.38	15.86
3 П	7.96	4.73	28.03	8.23	7.03	16.53

Результаты построения скаффолдов для генома шимпанзе при использовании разных наборов геномов в качестве референсных

Результаты: сборка нескольких геномов

	Базовый алгоритм		Расширенный алгоритм		Ragout	
	TP(%)	FP(%)	TP(%)	FP(%)	TP(%)	FP(%)
Человек	1.57	1.57	2.45	1.41	0.93	3.21
Шимпанзе	5.75	3.53	12.88	3.35	6.03	15.35
Горилла	5.35	3.45	12.33	5.72	6.65	15.14

Результаты построения скаффолдов для человека, шимпанзе и гориллы из датасета «Приматы»

Выводы

- Разработан алгоритм построения скаффолдов на основе анализа сборок нескольких родственных геномов
- Показано, что информаци о повторах может быть очень полезна для решения данной задачи
- Реализовано программное обеспечение для построения скаффолдов, имеющее совместимость с MGRA2

Спасибо за внимание!