Исправление ошибок в чтениях, полученных с помощью технологии IonTorrent

Ершов Василий Алексеевич

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: к.ф.-м.н. Коробейников А. И.
Рецензент: разработчик ПО Тарасов А. Л.

2017

Задача исправления ошибок

- Неизвестный геном $g \in \mathbb{A}^*$ — строка над алфавитом $\mathbb{A} = \{A,\ C,\ G,\ T\}$
- Дано множество $\mathbb{S} = \{r_i\}_{i=1}^{N}$ (коротких) подстрок g, содержащих ошибки — чтения.
- Всем $r_i \in \mathbb{S}$ соответствует (неизвестные) подстроки r_i^* генома g.
- Требуется по $\mathbb{S}$ восстановить $\mathbb{S}^* = \{r_i^*\}_{i=1}^{N}$.

```
Геном    | AGCTTTCATTAAGCGGCGAAAAGACTCAAGAAATTAGTTGCA
Чтения   |  GC_TTCA       GG_GAA__GAC      AAATTAGT_G
Чтения   |   CTTTCATAAAGCG                GAA_TTAGT_GC
Чтения   |           TAAGCGCGGAAA_GAACCAAG
```

Обозначения

Определение

- Нуклеотиды (полимеры): $\mathbb{A} = \{A,\ C,\ G,\ T\}$
- Гомополимеры: $\mathbb{H} = \{(a, n) | a \in \mathbb{A}, n \in \mathbb{N}\}$
- k-мер — последовательность из полимеров длины k — строка над алфавитом $\mathbb{A}$ длины k
- hk-мер — последовательность из гомополимеров длины k — элемент пространства $\mathbb{H}^k$

Полимеры	Гомополимеры	k-мер ($k = 4$)	hk-мер ($k = 4$)
A, C, G, T	AAAA, CC, T	AACG	AACGGGTT

Задача исправления ошибок

- В общем случае решить практически невозможно.
- В практических задачах есть априорная информация о чтениях:
  - Чтения получены в результате работы некоторого прибора, про который известен профиль ошибок.
  - Строчек много — каждый нуклеотид в геноме покрыт достаточно много раз.

```
Геном   | AGCTTTCATTAAGCGGCGAAAAGACTCAAGAAATTAGTTGCA
Чтения  |  GC_TTCA       GG_GAA__GAC      AAATTAGT_G
Чтения  |   CTTTCATAAAGCG                GAA_TTAGT_GC
Чтения  |          TAAGCGCGGAAA_GAACCAAG
```

Покрытие каждого нуклеотида приблизительно 2.

Hammer/BayesHammer

Алгоритмы коррекции.

- Hammer (Medvedev et al., 2011)
- BayesHammer (Nikolenko et al., 2013)

Идея алгоритмов

- Переход от чтений к подстрокам длины k.
- Оценка множества геномных k-меров:
  - Кластеризация множества k-меров.
  - Фильтрация «ошибочных» кластеров.
- Алгоритм коррекции на основе оценки k-мерного спектра генома.

IonHammer — метод коррекции ошибок для технологии IonTorrent

IonHammer

- Обобщение BayesHammer на ошибки вида «вставки» и «удаления».
- Переход от алфавита из полимеров к алфавиту из гомополимеров.

Недостатки

- Медленный шаг кластеризации.
- Необходимость подбора параметров для работы алгоритма.
- Низкое качество коррекции.

Оценка множества геномных *hk*-меров

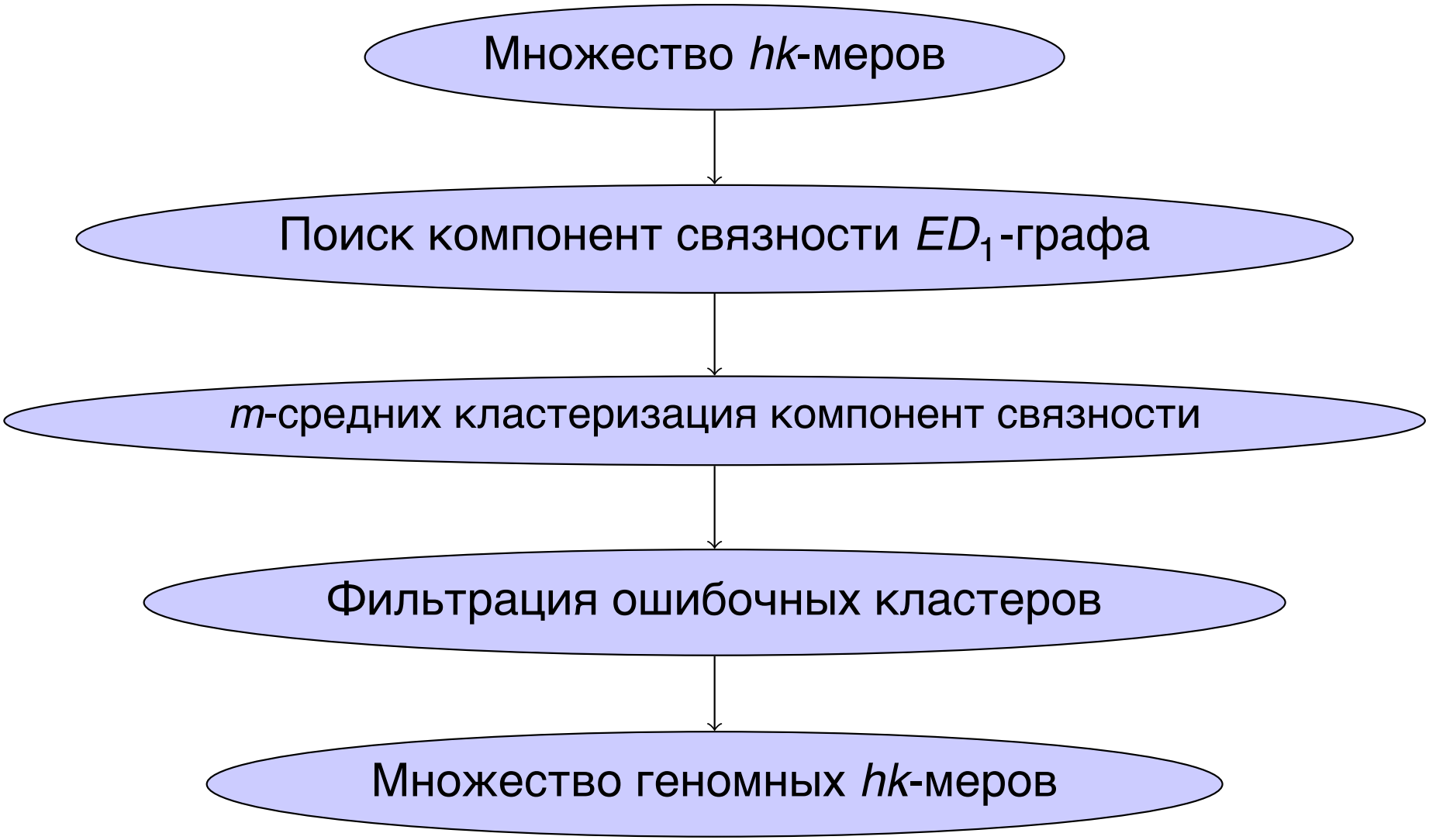
Входные данные и параметры

- Множество *hk*-меров $\mathscr{H}$, встретившихся в чтениях.
- Статистики, посчитанные для этих чтений:
  - $C(h)$ — количество раз, которое *hk*-мер встретился в чтениях.
  - $Q(h)$ — некоторая оценка качества *hk*-мера.
- $d(x, y) : \mathbb{H}^k \times \mathbb{H}^k \to \mathbb{R}_{\geq 0}$ — некоторая мера схожести *hk*-меров.

ED_1-граф

Граф с вершинами из $\mathscr{H}$. Ребра проведены между всеми вершинами x и y, для которых $d(x, y) \leq 1$.

Оценка множества геномных *hk*-меров



Оценка множества геномных *hk*-меров

В рамках ВКР были предложены

1. Эффективный алгоритм построения компонент связности для частного случая ED_1-графа.
2. Метод фильтрации ошибочных кластеров, параметры для которого оцениваются автоматически с помощью EM-алгоритма.
3. Набор эвристик для консервативного определения числа кластеров m с автоматической оценкой параметров.

Качество оценки множества геномных *hk*-меров

Организм	*E. coli str. DH10B*	*E. coli str. DH10B*
Чтений	7247730	22749163
Всего *hk*-меров	206560123	403698029
Геномные, оценные геномными	6562088 (99.78%)	6550682 (99.6%)
Геномные, оцененные негеномными	14325 (0.22%)	25861 (0.4%)
Негеномные, оцененные геномным	11971	39002

Количество геномных и негеномных центров. Для геномных центров указана доля от общего числа геномных *hk*-меров в чтениях.

Алгоритм коррекции

- Основная идея — согласовано заменить все *hk*-меры в чтении на оценки геномных.
- Различные исправления ранжируются с помощью функции штрафа.
- Предложен и реализован алгоритм, минимизирующий специальный вид функций штрафа.
- В общем случае временная сложность предложенного алгоритма для коррекции чтения длины n:

$$\log T(n) = \Theta(n^{\alpha}), \alpha \geq 1$$

Применен ряд эвристик, позволяющий перебирать только «наиболее важные» коррекции.

Сравнение новой и старой версий IonHammer

	Новый	Старый
Время работы (мин., 32 потока)	**1:22**	2:55
Кол-во геномных *hk*-меров	6553831 (99.65%)	**6567815** (99.86%)
Кол-во негеномных *hk*-меров	**716331**	3661340
Испорченные или выравненные на другую геномную позицию чтения	**33676** (4.92%)	40037 (5.86%)

Качество коррекции *E. coli str. DH10B*-C24. Всего в данных 10^7 негеномных *hk*-меров.

Качество коррекции

Алгоритм	Геномные *hk*-меры	Негеномные *hk*-меры
Coral (Salmela, Schoder, 2011)	6557985 (99.71%)	1578382
Fiona (Schulz et al., 2014)	6572520 (99.93%)	745943
Pollux (Marinier et al., 2015)	6548650 (99.57%)	827297
IonHammer	6553831 (99.65%)	716331

Качество коррекции *E. coli str. DH10B*-C24. Всего в данных 10^7 негеномных *hk*-меров.

Скорость работы

	IonHammer	Pollux	Coral	Fiona
Количество потоков	32	1	8	32
Процессорное время	**29298s**	65866s	439084s	∞
Время работы	**0:17:33**	18:19:48	15:46:34	∞
Максимальный расход памяти (GB)	**15**	20	33	NA

Време работы на наборе данных *E. coli DH10B*-520.

Результаты

Разработан алгоритм коррекции IonHammer 2.0

1. Реализована эфективная модификация алгоритма кластеризации.
2. Предложен и реализован метод автоматической оценки параметров для алгоритма кластеризации.
3. Метод коррекции чтений из BaysHammer обобщен на ошибки технологии IonTorrent и реализован в новой версии IonHammer.
4. Проведено сравнение алгоритма с существующим аналогами и предыдущей версией алгоритма.
5. Подготовлена статья в журнал Bioinformatics