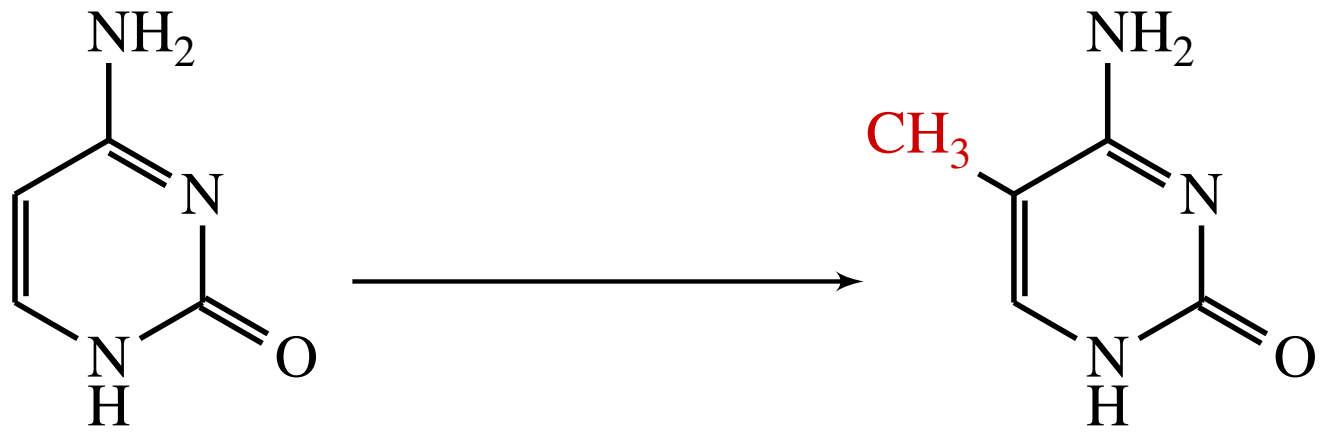
Моделирование данных бисульфитного секвенирования

Сергей Лебедев
Руководитель: О. Ю. Шпынов, науч. сотр.
Рецензент: С. В. Малов, ведущ. науч. сотр., доцент, к.ф.-м. н.

СПбАУ НОЦНТ РАН

5 июня 2014 г.

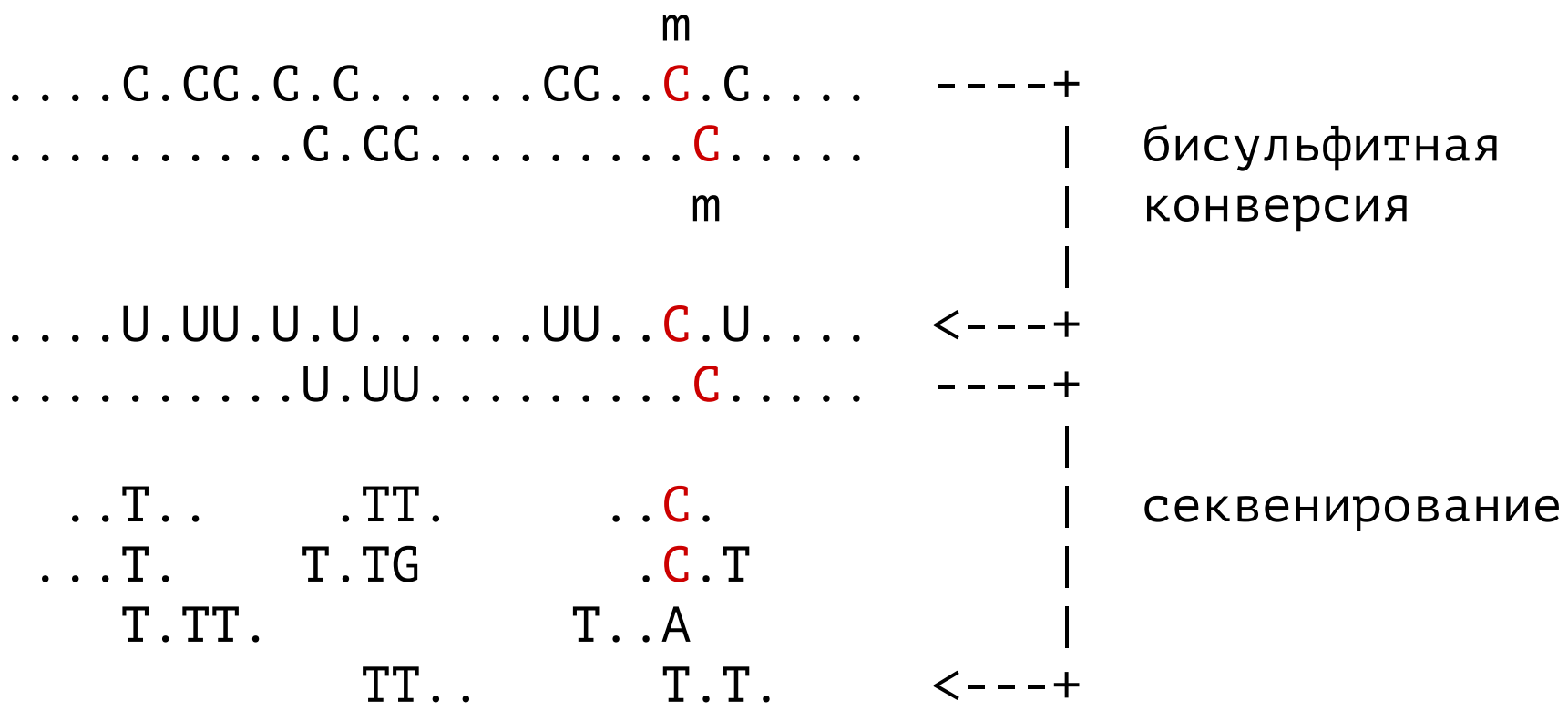


цитозин

5-метилцитозин

- Цитозин — один из четырех нуклеотидов, образующих ДНК.
- В составе ДНК цитозин может быть «метилирован», то есть содержать метильную группу.
- Изучение метилирования ДНК важно для понимания механизмов дифференцировки клетки и изучения рака.

Бисульфитное секвенирование ДНК



Представление результатов бисульфитного секвенирования

```
                              m
....C.CC.C.C......CC..C.C....
..........C.CC.........C.....
                            m


  ..T..      .TT.       ..C.
 ...T.     T.TG         .C.T
    T.TT.             T..A
            TT..        T.T.
                        ↑
```

Возможные представления результатов:

- пара $(\#C, \#T) = (2, 1)$,
- уровень метилирования ${}^{\#C}/_{\#C+\#T} = {}^{2}/_{3}$,
- $(\#A, \#T, \#C, \#G) = (1, 1, 2, 0)$.

Цель — построить математическую модель, позволяющую анализировать результаты бисульфитного секвенирования.

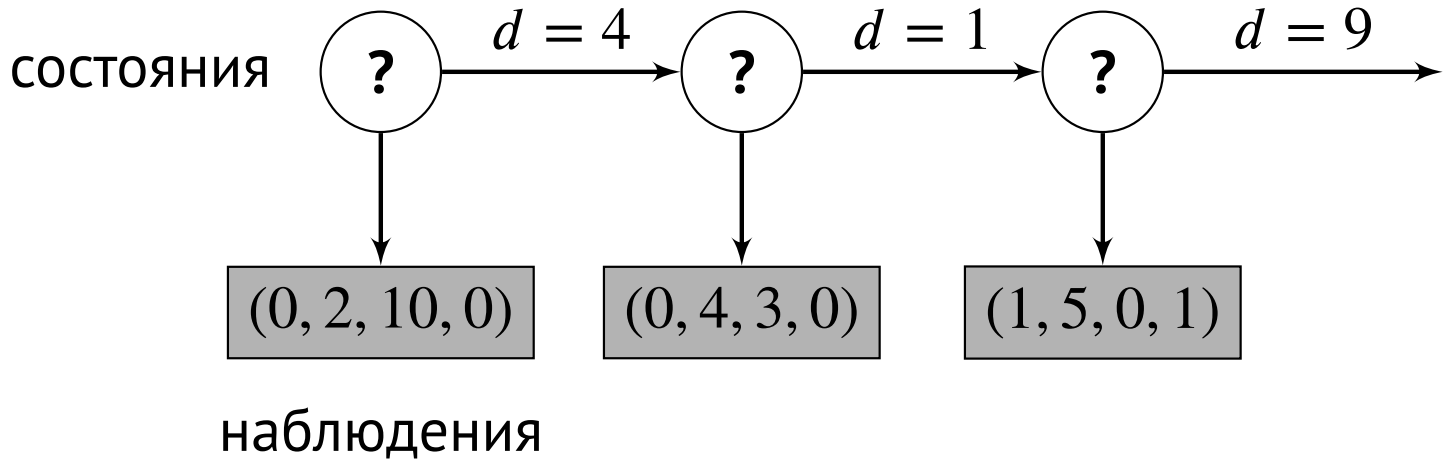
Задачи

- Проанализировать существующие подходы.
- Предложить, обосновать и реализовать несколько вероятностных моделей.
- Сформулировать критерии отбора и выбрать наиболее подходящую модель.
- Сравнить её с уже существующими.

- Алгоритм MSC (Methylation Status Calling) разделяет все цитозины на две группы: цитозины, содержащие метильную группу, и не содержащие её.
- Позволяет контролировать количество ошибок в полученном разделении.
- Не учитывает, что
  - для метилирования ДНК характерен эффект кластеризации,
  - секвенаторы делают ошибки при чтении последовательности ДНК.

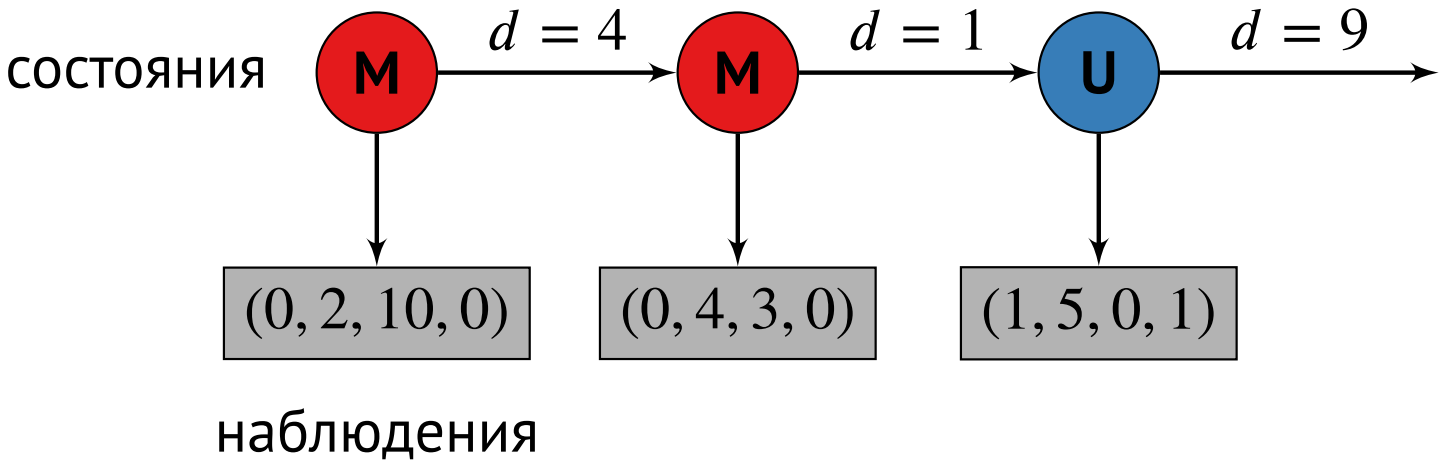
1. Биномиальная **смесь**:
   - моделирует $(\#C, \#T)$ с помощью биномиального распределения;
   - считает состояния всех цитозинов независимыми.
2. Биномиальная **скрытая марковская модель**: учитывает зависимость состояний соседних цитозинов.
3. **Переключающаяся** биномиальная скрытая марковская модель: учитывает геномное расстояние между цитозинами.
4. Переключающаяся **мультиномиальная** скрытая марковская модель:
   - моделирует $(\#A, \#T, \#C, \#G)$ с помощью мультиномиального распределения;
   - учитывает наличие ошибок секвенирования.

Переключающаяся скрытая марковская модель



- Скрытые состояния образуют марковский процесс с вероятностями перехода, зависящими от расстояния d.
- Наблюдения $(\#A, \#T, \#C, \#G)$ подчиняются мультиномиальному распределению, параметры которого зависят от скрытого состояния.

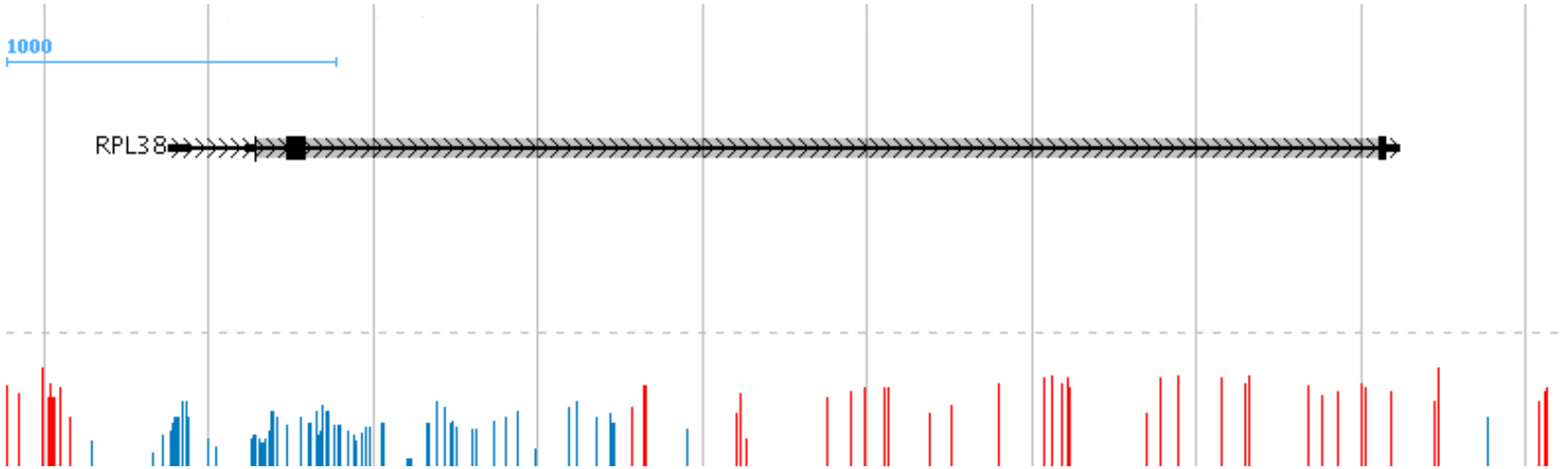
Обучение переключающейся скрытой марковской модели



- В процессе обучения модели оцениваются вероятности перехода и параметры мультиномиального распределения.
- У модели два возможных состояния: **M** — цитозин содержит метильную группы и **U** — не содержит её.

Хромосома	MSC	MSHMM	Пересечение
chr1	1291624	1136756	1136756
chr2	1114488	1316179	1114488
chr3	896377	1040715	896377
⋮			

- Данные бисульфитного секвенирования стволовых клеток мыши.
- В ячейках таблицы — количество метилированных цитозинов.
- Время обучения предлагаемой модели на всех хромосомах генома мыши ≈ 25 минут.



Ген «домашнего хозяйства» RPL38[1]

[1] Синий цвет соответствует состоянию **U**, красный — состоянию **M**.

- Проведён анализ существующих подходов к моделированию данных бисульфитного секвенирования.
- Реализованы несколько вероятностных моделей разной степени сложности, в качестве итоговой модели выбрана переключающаяся мультиномиальная CMM.
- Результаты работы модели согласуются с результатами алгоритма MSC, но предлагаемая модель также учитывает:
  - зависимость состояний соседних цитозинов в геноме,
  - наличие ошибок секвенирования.

Благодарности

- JetBrains Biolabs
- СПбАУ
  - Михаил Колмогоров
  - Павел Яковлев
- СПбГУ
  - Алексей Кладов
  - Екатерина Лебедева

Контакты

sergei.a.lebedev@gmail.com

Спасибо за внимание!

- Пусть d_t — количество нуклеотидов в геноме между $(t-1)$-ым и t-ым цитозином,
- тогда вероятность перехода из состояния i в состояние j на шаге t

$$P(s_t = j | s_{t-1} = i, d_t) = \mathbf{A}_{\mathrm{w}(d_t)ij},$$

где $\mathrm{w}(d_t) : \mathbb{N}_0 \to \{1, \ldots, D\}$ — функция, кластеризующая расстояния.

- Тогда функция правдоподобия для переключающейся СММ

$$P(\mathbf{x}; \boldsymbol{\pi}, \mathbf{A}, \boldsymbol{\theta}) = \sum_{\mathbf{s} \in \{1,\ldots,S\}^T} \pi_{s_1} \prod_{t=2}^{T} \mathbf{A}_{\mathrm{w}(d_t)s_t s_{t-1}} \prod_{t=1}^{T} P(x_t; \theta_{s_t}),$$

где $\mathbf{x} = (x_1, \ldots, x_T)$ — результаты бисульфитного секвенирования, а $\mathbf{s} = (s_1, \ldots, s_T)$ — скрытые состояния модели.

0.990	0.002	0.966	0.034	0.285	0.715
0.023	0.987	0.014	0.986	0.033	0.967

$\mathrm{w}(d_t) = 1, d_t \in [1, 2]$ $\mathrm{w}(d_t) = 5, d_t \in [6, 9]$ $\mathrm{w}(d_t) = 9, d_t \geq 118$

- Матрицы вероятностей перехода для переключающейся СММ обученной.
- Метки столбцов и строчек слева-направо и сверху вниз: UNMETHYLATED, METHYLATED.
- С увеличением расстояния вероятность остаться в том же состоянии уменьшается.

- FDR (false discovery rate) — частота ложных предсказаний.
- Ограничить количество ошибок можно, контролируя FDR на некотором уровне α.
- В контексте бисульфитного секвенирования:
  0. цитозин не содержал метильной группы,
  1. обратное.
- При сравнении с алгоритмом MSC FDR контролировался на уровне $\alpha = 0.01$.
- Детали процедуры для контроля FDR можно найти в статье Cheng, Zhu, «A classification approach for DNA methylation profiling with bisulfite next-generation sequencing data», 2014.