Анализ подходов к моделированию данных метилирования ДНК

Сергей Лебедев

Руководитель: Олег Шпынов

JetBrains

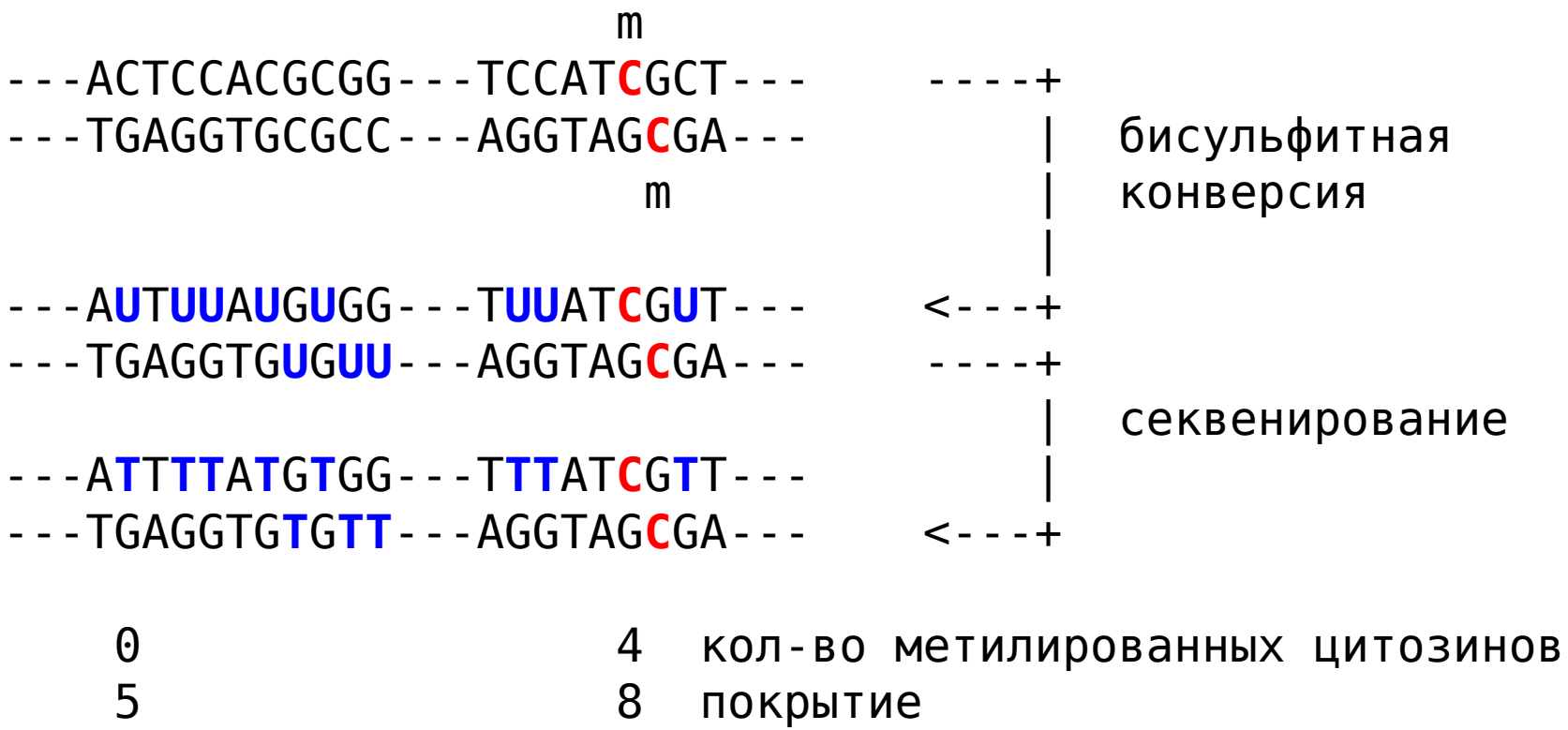
21 сентября, 2013

Эпигенетический минимум

Метилирование ДНК —

химическая модификация, добавляющая метильную группу к цитозину.

Бисульфитное секвенирование ДНК



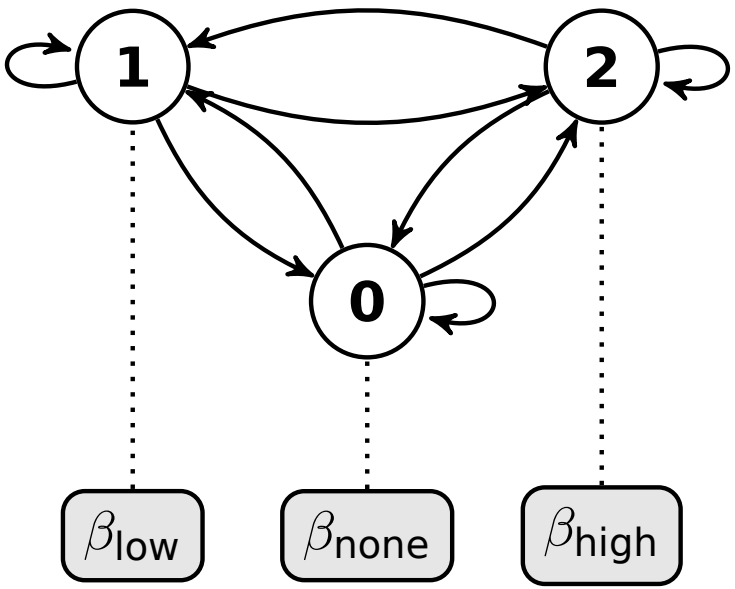
Долины метилирования

- DNA Methylation Valleys [XSL+13] — длинные слабо метилированные участки генома.
- Что особенного в **DMV**:
  - ассоциированы с транскрипционными факторами и генами развития,
  - пересекают некоторые промоторы lncRNA,
  - консервативны между различными видами,
  - остаются слабо метилированными при дифференциации клетки и
  - становятся сильно метилированными в раковых клетках.
- Как локализовать DMV, используя данные бисульфитного секвенирования?
- Как сегментировать метилом на слабо- и сильно- метилированные участки?

Уровень метилирования

$$\beta = \frac{mC}{C} \in [0, 1]$$

Скрытая марковская модель с Гауссовыми вероятностями испусканий [SMB+11]



$$\beta_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

Сегментирование метилома с GHMM

- С помощью модели мы можем приписать каждому цитозину одну из трех меток:
  - 0 метилирование отсутствует,
  - 1 низкий уровень метилирования,
  - 2 высокий уровень метилирования.
- Всем остальным нуклеотидам припишем метку, соответствующую ближайшему цитозину.
- Пример:

```
                          m
---ACTCCACGCGG---TCCATCGCT---
   00000000000   00001110 0
                         ^
                         |
                    почти DMV
```

Проблемы GHMM

- Уровень метилирования – это величина из $[0, 1]$, в то время как GHMM описывает величину из $\mathbb{R}$. Можно:
  - моделировать испускания с помощью бета-распределения,
  - преобразовать уровень метилирования:

$$\mathrm{M} = \log \frac{\beta}{1-\beta} \in \mathbb{R}$$

$$\Gamma = -\log(1-\beta) \in [0, +\infty)$$

- При переходе от пары (mC, C) к уровню метилирования теряется информация о покрытии, например:

$$\beta_i = {}^{1}/_{2} = \beta_j = {}^{10}/_{20} = 0.5$$

Альт. подход к сегментированию метилома

- Будем моделировать не уровень метилирования, а пары (mC, C) напрямую:

```
                          m
---ACTCCACGCGG---TCCATCGCT---
                          ^
                          |
                       (4, 8)
```

- Таким образом, значение mC в каждой позиции определяется покрытием в этой позиции C и некоторой вероятностью “метилирования” p_i:

$$mC \sim \mathcal{B}(C, p_i) \qquad \text{for } i \in \{\text{high}, \text{low}, \text{none}\}$$

- Покрытие в каждой позиции C мы считаем известным с точностью до распределения.

Как моделировать покрытие?

- Категориальным распределением на множестве $\{1, \ldots, k\}$, где k — максимальное покрытие.
- Смесью распределений Пуассона:
  - Если предположить, что каждый нуклеотид в геноме длины N покрыт одинаковым числом фрагментов n, то вероятность получить k ридов в некоторой позиции:

$$\mathcal{B}(k; n, p = {}^{1}/_{N}) \underset{N \to \infty}{\approx} \text{Pois}(k; \frac{n}{N})$$

  - В реальном мире покрытие почти никогда не равномерно, поэтому разумно использовать смесь распределений Пуассона.

Биномиальная смесь для парных наблюдений

- Моделирует независимые пары (mC, C), полученные в результате бисульфитного секвенирования.
- Предполагает распределение покрытия в каждой позиции известным: категориальное или смесь распределений Пуассона.
- Использует набор компонент, аналогичный GHMM [SMB+11]: $\{\text{high}, \text{low}, \text{none}\}$.

Дальнейшие планы

- Провести сравнительный анализ результатов PBMM и GHMM на данных [SMB+11, XSL+13].
- Обобщить модель для поиска различно метилированных регионов в результатах нескольких экспериментов.

Вопросы?[1]

[1] http://www.flickr.com/photos/maggie_oc/9210581429

Ссылки

Michael B Stadler, Rabih Murr, Lukas Burger, Robert Ivanek, Florian Lienert, Anne Schöler, Erik van Nimwegen, Christiane Wirbelauer, Edward J Oakeley, Dimos Gaidatzis, et al.

Dna-binding factors shape the mouse methylome at distal regulatory regions.

Nature, 2011.

Wei Xie, Matthew D Schultz, Ryan Lister, Zhonggang Hou, Nisha Rajagopal, Pradipta Ray, John W Whitaker, Shulan Tian, R David Hawkins, Danny Leung, et al.

Epigenomic analysis of multilineage differentiation of human embryonic stem cells.

Cell, 2013.