Модели метилирования ДНК

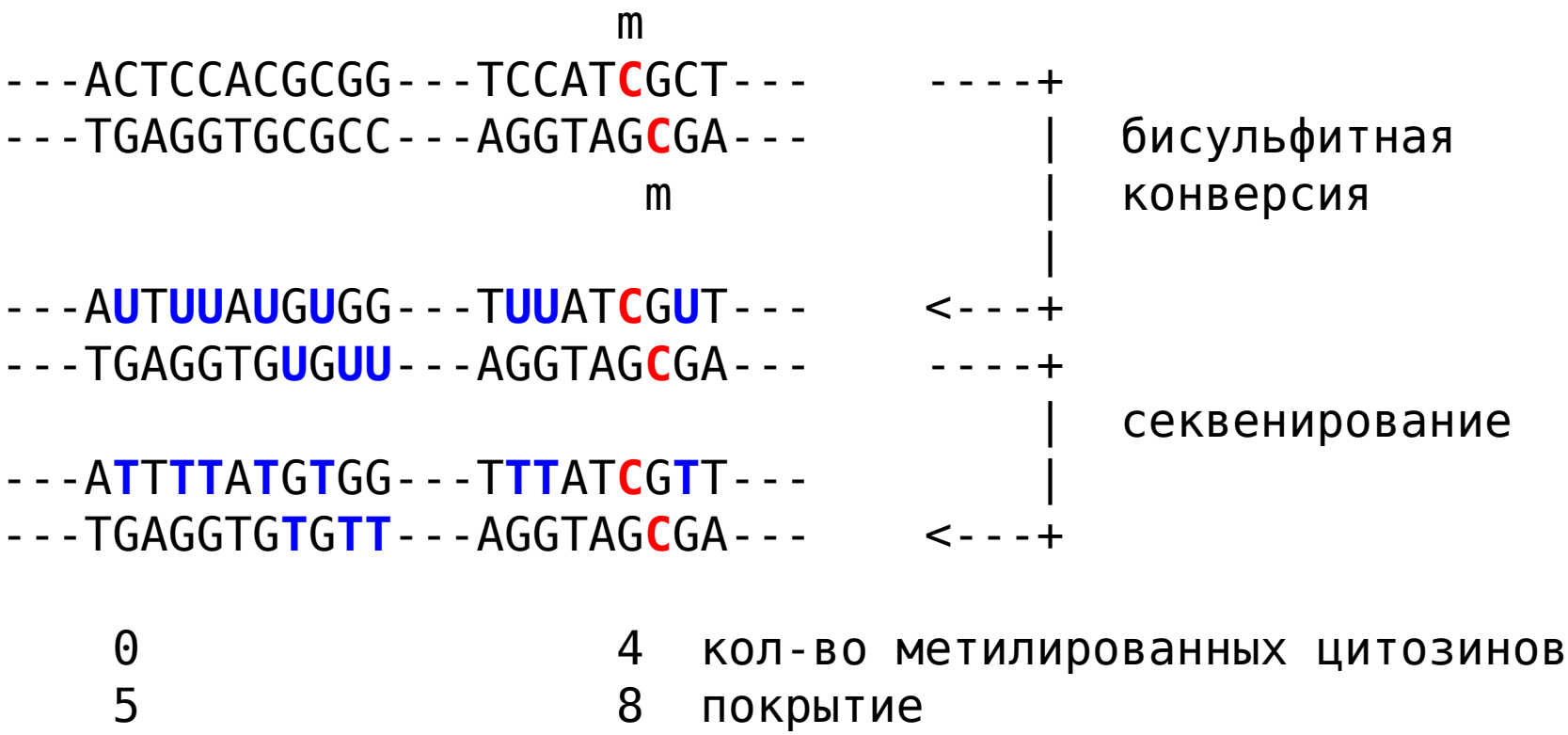
Сергей Лебедев

23 октября 2013 г.

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?

Скрытая марковская модель (1)

- $\vec{x} = \{x_1, x_2, \ldots, x_T\} \in \mathcal{V}^T$ последовательность наблюдений, например, $\{0.31, 0.29, 0.1\}$,
- $\vec{z} = \{z_1, z_2, \ldots, z_T\} \in \mathcal{S}^T$ последовательность состояний, например, $\{\text{low}, \text{low}, \text{none}\}$,
- параметры модели:
  - $\mathbf{A}_{ij}$ вероятность перейти из состояния i в состояние j,
  - $\mathbf{B}_i$ параметры испускания в состоянии i, например, $\{\mu_{\text{none}}, \sigma^2_{\text{none}}\}$.
- Как оценить
  1. вероятность последовательности наблюдений $\vec{x}$ при условии модели?
  2. наиболее вероятную последовательность состояний модели $\vec{z}$, породившую наблюдаемые данные $\vec{x}$?
  3. параметры модели $\mathbf{A}$, $\mathbf{B}$ по последовательности наблюдений $\vec{x}$?

Вычисление правдоподобия наблюдений (1)

$$
\begin{aligned}
P(\vec{x}\,;\,\mathbf{A},\mathbf{B}) &= \sum_{\vec{z}} P(\vec{x},\vec{z};\mathbf{A},\mathbf{B}) \\
&= \sum_{\vec{z}} P(\vec{x}\,|\,\vec{z};\mathbf{A},\mathbf{B})P(\vec{z};\mathbf{A},\mathbf{B}) \\
&= \sum_{\vec{z}} \Big(\prod_{t=1}^{T} P(x_t\,|\,z_t;\mathbf{B})\Big)\Big(\prod_{t=1}^{T} P(z_t\,|\,z_{t-1};\mathbf{A})\Big) \\
&= \sum_{\vec{z}} \Big(\prod_{t=1}^{T} P(x_t;\mathbf{B}_{z_t})\Big)\Big(\prod_{t=1}^{T} \mathbf{A}_{z_{t-1}\,z_t}\Big)
\end{aligned}
$$

- Прямое вычисление правдоподобия потребует $O(|\mathcal{S}|^T)$ операций, т. к. мы суммируем по всем возможным последовательностям состояний.

Вычисление правдоподобия наблюдений (2)

- Можно ускорить вычисление с помощью динамического программирования:
  1. Обозначим $\alpha_i(t) = P(x_1, x_2, \ldots, x_t, z_t = s_i; \mathbf{A}, \mathbf{B})$ полную вероятность того, что модель породила последовательность наблюдений $\{x_1, \ldots, x_t\}$ и находится в состоянии s_i на шаге t.
  2. Тогда правдободобие можно переписать как:

$$\begin{aligned} P(\vec{x}; \mathbf{A}, \mathbf{B}) &= P(x_1, x_2, \ldots, x_T; \mathbf{A}, \mathbf{B}) \\ &= \sum_{i=1}^{|\mathcal{S}|} P(x_1, x_2, \ldots, x_T, z_T = s_i; \mathbf{A}, \mathbf{B}) \\ &= \sum_{i=1}^{|\mathcal{S}|} \alpha_i(T) \end{aligned}$$

  3. Алгоритм “вперед” (Forward Procedure) позволяет посчитать $\alpha_i(t)$ за время $O(|\mathcal{S}| \cdot T)$.

Вычисление правдоподобия наблюдений (3)

- Алгоритм “вперед”:

1: **for** $i = 1 \dots |\mathcal{S}|$ **do**
2: $\quad \alpha_i(1) \leftarrow \mathbf{A}_{0\,i}\, P(x_1; \mathbf{B}_i)$
3: **end for**
4: **for** $t = 2 \dots T$ **do**
5: $\quad$ **for** $j = 1 \dots |\mathcal{S}|$ **do**
6: $\quad\quad \alpha_j(t) \leftarrow \sum_{i=0}^{|\mathcal{S}|} \alpha_i(t-1)\mathbf{A}_{i\,j}\, P(x_t; \mathbf{B}_j)$
7: $\quad$ **end for**
8: **end for**

- Аналогичным образом с помощью алгоритма “назад” (Backward Procedure) можно посчитать $\beta_i(t) = P(x_T, x_{T-1}, \dots, x_{t+1} \,|\, z_t = s_i; \mathbf{A}, \mathbf{B})$

Восстановление последовательности состояний (1)

- Переформулируем задачу:

$$\arg\max_{\vec{z}} P(\vec{z} \mid \vec{x}; \mathbf{A}, \mathbf{B}) = \arg\max_{\vec{z}} \frac{P(\vec{x}, \vec{z}; \mathbf{A}, \mathbf{B})}{\sum_{\vec{z}} P(\vec{x}, \vec{z}; \mathbf{A}, \mathbf{B})}$$
$$= \arg\max_{\vec{z}} P(\vec{x}, \vec{z}; \mathbf{A}, \mathbf{B})$$

- Вычисление $\arg\max$ в лоб, как и в случае задачи вычисления правдободобия наблюдений, требует $O(|\mathcal{S}|^T)$ времени.

Восстановление последовательности состояний (2)

- Алгоритм Витерби:

1: **for** $i = 1 \dots |\mathcal{S}|$ **do**
2: $\quad V_i(1) \leftarrow \mathbf{A}_{0\,i}\, P(x_1; \mathbf{B}_i)$
3: $\quad BackPtr_i(1) \leftarrow 0$
4: **end for**
5: **for** $t = 2 \dots T$ **do**
6: $\quad$ **for** $j = 1 \dots |\mathcal{S}|$ **do**
7: $\quad\quad V_j(t) \leftarrow \max\limits_{i \in \{1 \dots |\mathcal{S}|\}} V_i(t-1)\mathbf{A}_{ij}\, P(x_t; \mathbf{B}_j)$
8: $\quad\quad BackPtr_j(t) \leftarrow \arg\max\limits_{i \in \{1 \dots |\mathcal{S}|\}} V_i(t-1)\mathbf{A}_{ij}$
9: $\quad$ **end for**
10: **end for**

- Наиболее вероятную последовательность состояний можно получить обратным обходом матрицы $BackPtr$, начиная с $\arg\max\limits_{i \in 1 \dots |\mathcal{S}|} V_i(T)$.

- Пусть (или [Blu02])
  - $\vec{\theta}$ вектор параметров модели,
  - $\vec{x} = \{x_1, x_2, \ldots, x_T\}$ вектор наблюдений и
  - $\vec{z} = \{z_1, z_2, \ldots, z_T\}$ вектор скрытых случайных величин
- и пусть для каждого i задано $Q_i(z)$ – некоторое распределение над z,
- тогда логарифм функции правдоподобия можно записать как

$$\ell(\vec{\theta}) = \sum_{i=1}^{T} \log P(x_i; \vec{\theta}) = \sum_{i=1}^{T} \log \sum_{z_i} P(x_i, z_i; \vec{\theta})$$
$$= \sum_{i=1}^{T} \log \sum_{z_i} Q_i(z_i) \frac{P(x_i, z_i; \vec{\theta})}{Q_i(z_i)}$$

- Так как log вогнутая функция, то можно применить нервенство Йенсена

$$\begin{aligned}\ell(\vec{\theta}) &= \sum_{i=1}^{T} \log \mathbb{E}_{z_i \sim Q_i(z)} \left[\frac{P(x_i, z_i; \vec{\theta})}{Q_i(z_i)} \right] \\ &\geq \sum_{i=1}^{T} \mathbb{E}_{z_i \sim Q_i(z)} \left[\log \frac{P(x_i, z_i; \vec{\theta})}{Q_i(z_i)} \right] \\ &= \sum_{i=1}^{T} \sum_{z_i} Q_i(z_i) \log \frac{P(x_i, z_i; \vec{\theta})}{Q_i(z_i)}\end{aligned}$$

- Неравенство Йенсена становится равенством, если $\mathbb{E}X = X$, т. е. X константная случайная величина.

- Для некоторого фиксированного $\vec{\theta}$ выберем $Q_i(z_i)$ таким образом, чтобы

$$\frac{P(x_i, z_i; \vec{\theta})}{Q_i(z_i)} = \text{const}$$

- причем const не должно зависеть от z

$$Q_i(z_i) = \frac{P(x_i, z_i; \vec{\theta})}{\sum_z P(x_i, z_i; \vec{\theta})} = \frac{P(x_i, z_i; \vec{\theta})}{P(x_i, \vec{\theta})} = P(z_i | x_i; \vec{\theta})$$

- Получили, что при специально выбранных $Q_i(z_i)$

$$\ell(\vec{\theta}) = \sum_{i=1}^{T} \sum_{z_i} Q_i(z_i) \log \frac{P(x_i, z_i; \vec{\theta})}{Q_i(z_i)}$$

- Сформулируем алгоритм

repeat
  for $i = 1 \dots T$ **do**
    $Q_i(z_i) \leftarrow P(z_i \mid x_i; \vec{\theta})$ ▷ E-step
  end for
  $\theta \leftarrow \arg\max_{\theta} \sum_{i=0}^{T} \sum_{z_i} Q_i(z_i) \log \frac{P(x_i, z_i; \vec{\theta})}{Q_i(z_i)}$ ▷ M-step
until converegence

- Проблемы EM-алгоритма:
  - сходимость к ОМП не гарантируется,
  - для мультимодальных распределений результат зависит от начальных параметров.

ОМП для параметров скрытой марковской модели (1)

- Пусть
  - $\gamma_i(t) = P(z_t = s_i \,|\, x_t; \mathbf{A}, \mathbf{B})$ вероятность того, что на шаге t модель находится в состоянии s_i

$$\gamma_i(t) = \frac{\alpha_i(t)\beta_i(t)}{\sum\limits_{j=1}^{|\mathcal{S}|} \alpha_j(t)\beta_j(t)} \equiv Q_t(z_t)$$

  - $\xi_{ij}(t) = P(z_t = s_i, z_{t+1} = s_j \,|\, x_t; \mathbf{A}, \mathbf{B})$ вероятность того, что модель перешла из состояния s_i в s_j на шагах t и $t+1$

$$\xi_{ij}(t) = \frac{\alpha_i(t)\mathbf{A}_{ij}\,\beta_j(t+1)P(x_t; \mathbf{B}_j)}{\sum\limits_{k=1}^{|\mathcal{S}|}\sum\limits_{l=1}^{|\mathcal{S}|} \alpha_k(t)\mathbf{A}_{kl}\,\beta_l(t+1)P(x_t; \mathbf{B}_l)}$$

ОМП для параметров скрытой марковской модели (2)

- Тогда формулы обновления параметров на M шаге (алгоритм Баума-Велша):

$$\mathbf{A}_{0j} = \mathbb{E}\{\#\text{ of transitions from state } i\} = \gamma_i(1)$$

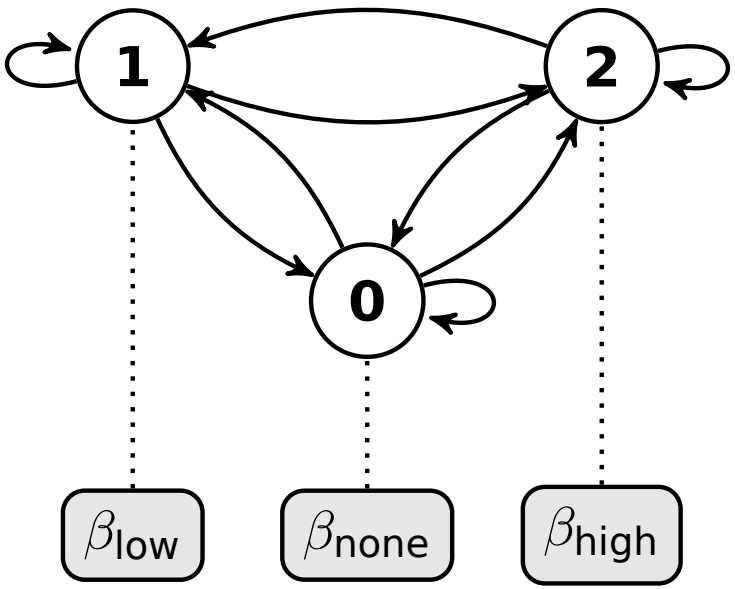
$$\mathbf{A}_{ij} = \mathbb{E}\{\#\text{ of transitions from state } i \text{ to state } j\} = \frac{\sum_{t=1}^{T} \xi_t(i,j)}{\sum_{t=1}^{T} \gamma_i(t)}$$

- Детали вывода алгоритма Баума-Велша можно посмотреть, например, в лекциях Andrew Ng.

Уровень метилирования

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

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

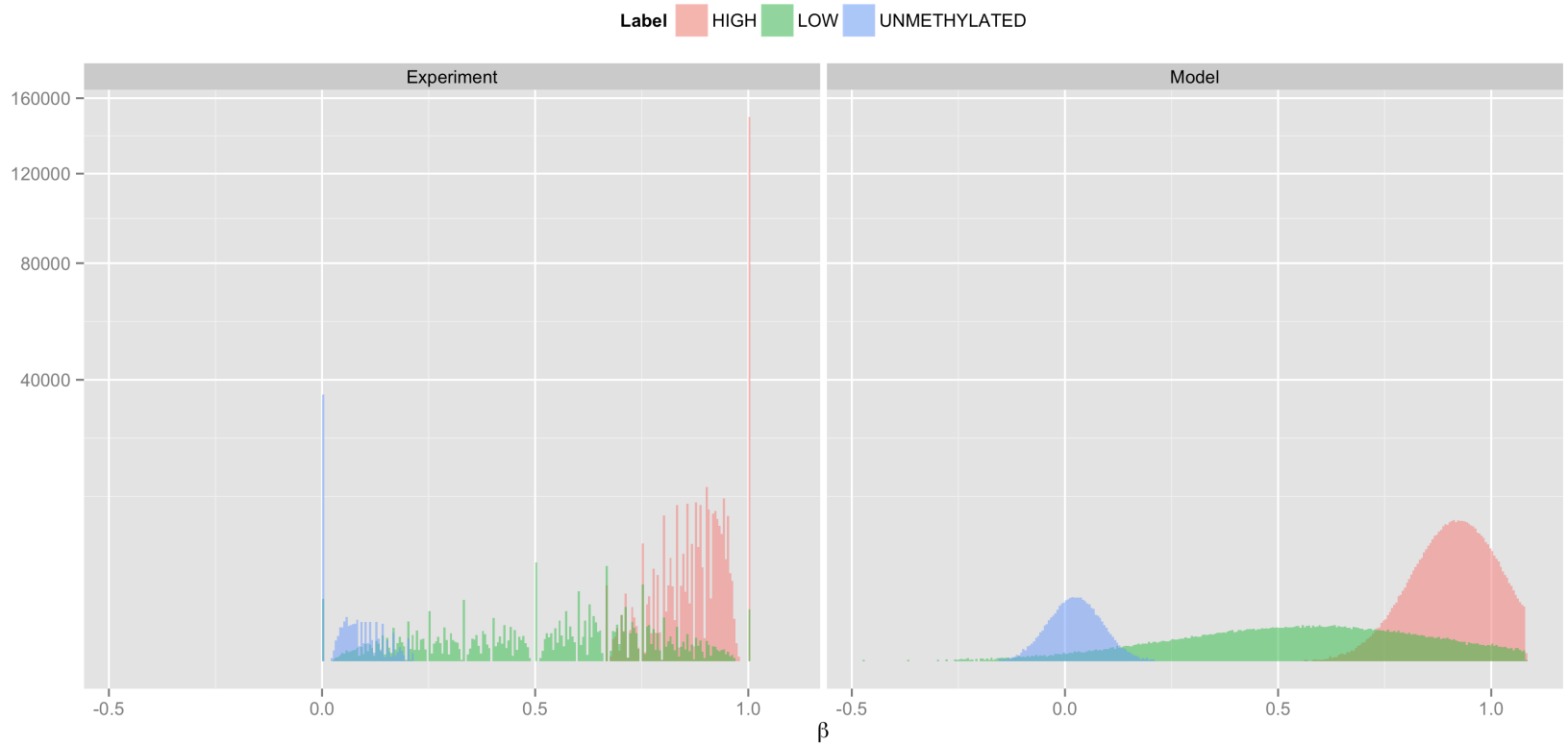


$$\beta_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

- 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
```

Результаты обучения GHMM на GSE30202 (1)



Гистограмма уровней метилирования для 19 хромосомы генома стволовых клеток мыши, каждой метке соответствует свой цвет

- Уровень метилирования – это величина из $[0, 1]$, в то время как GHMM описывает величину из $\mathbb{R}$. Можно:
  - (но сложно) 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.
- Распределением Пуассона
  - Если предположить, что каждый нуклеотид в геноме длины N покрыт одинаковым числом фрагментов n, то вероятность получить k ридов в некоторой позиции:

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

  - Проблемы:
    - покрытие почти никогда не равномерно,
    - дисперсия покрытия часто больше, чем теоретически возможная для распределения Пуассона.

Что делать?

- Использовать отрицательное биномиальное распределение aka Gamma-Poisson mixture:

$$\lambda \sim \text{Gamma}(r, \frac{p}{1-p})$$

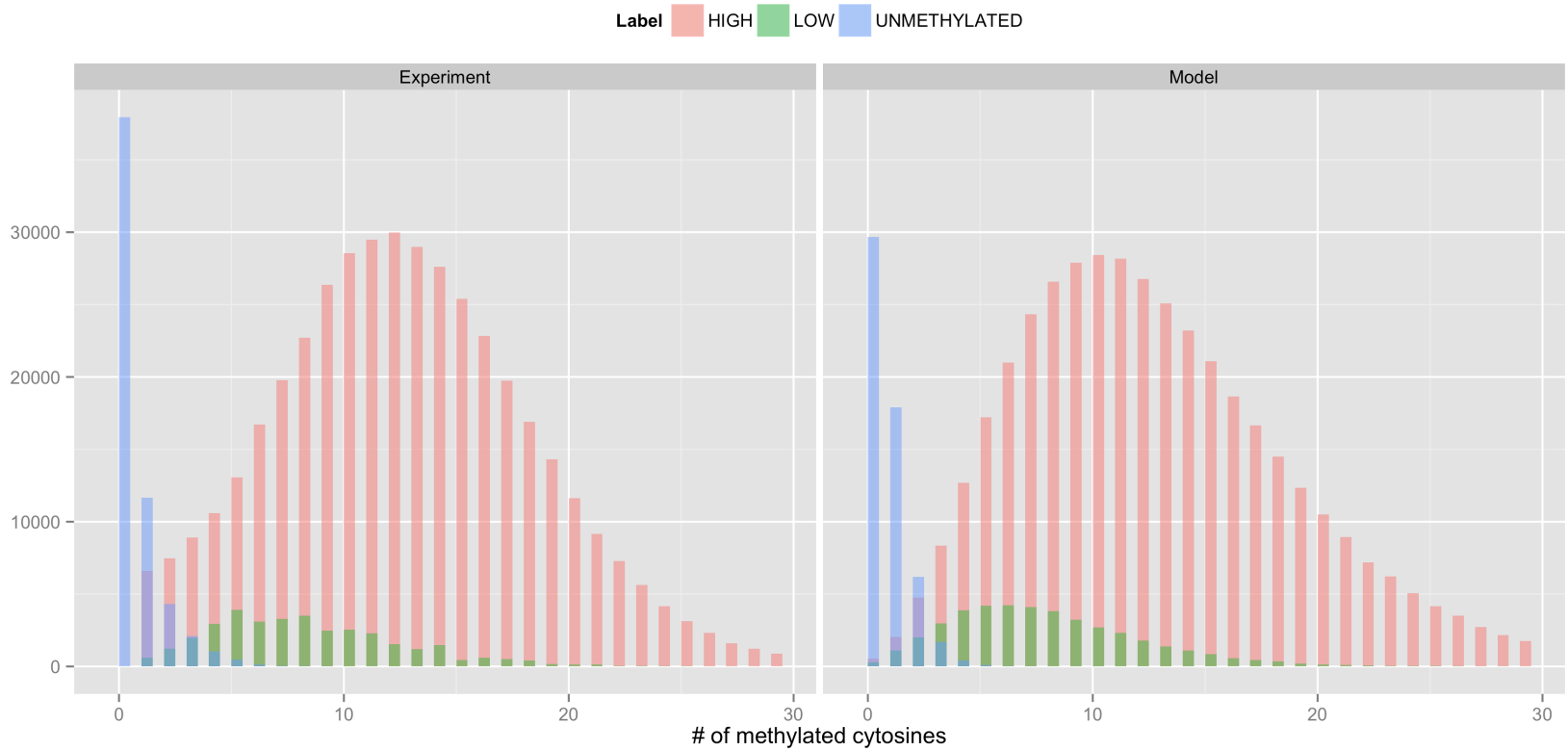
$$n \sim \text{Pois}(\lambda)$$

- Проблемы:
  - не учитывается неравномерность покрытия,
  - сложно выбрать “хорошие” начальные параметры для EM-алгоритма.

Биномиальная смесь для парных наблюдений

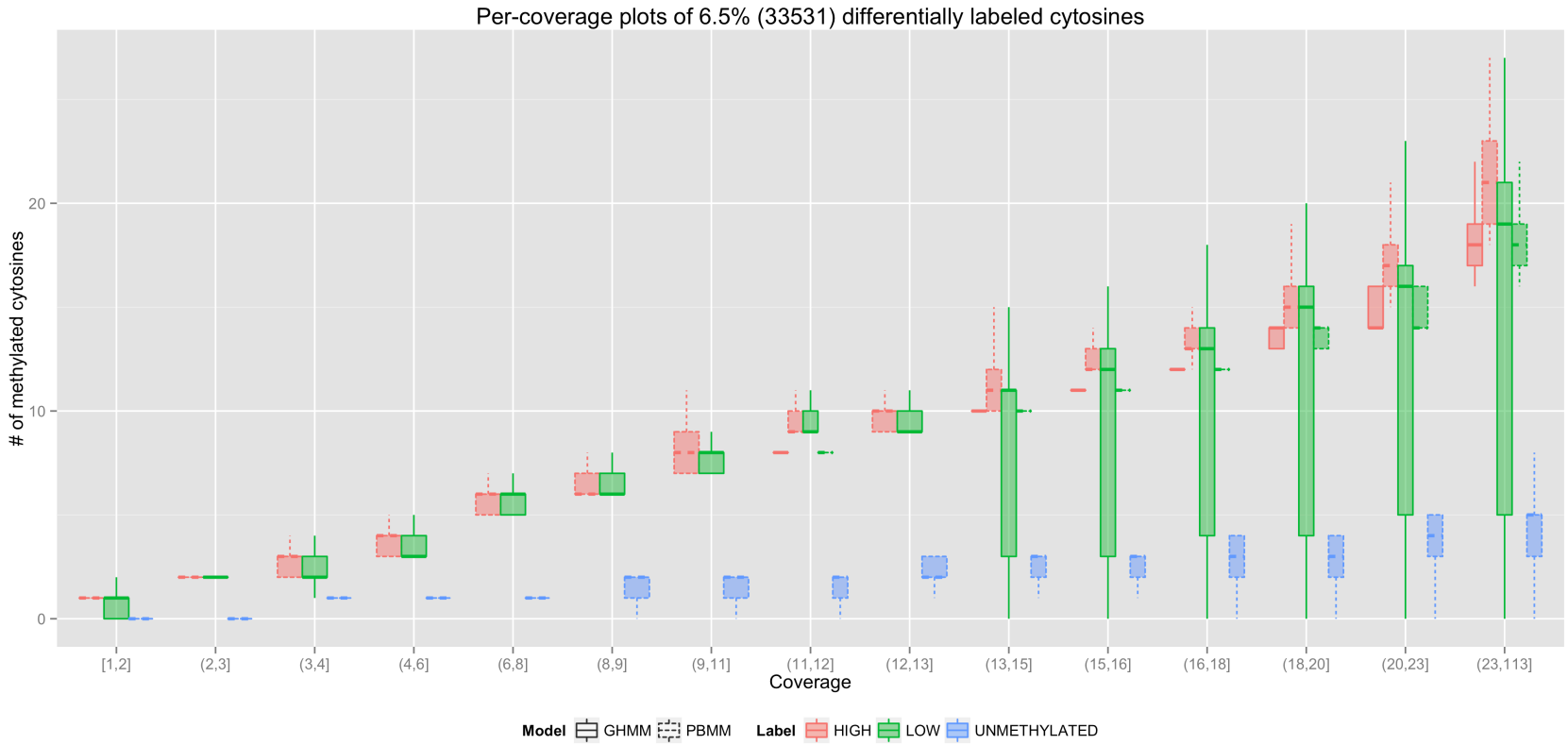
- Моделирует независимые пары (mC, C), полученные в результате бисульфитного секвенирования.
- Предполагает, что покрытие в каждой позиции подчиняется отрицательному биномиальному распределениею.
- Не предполагает зависимость между метилированием последовательных цитозинов.
- Использует набор меток аналогичный GHMM [SMB+11]: $\{\text{high}, \text{low}, \text{none}\}$.

Результаты обучения PBMM на GSE30202 (1)



Гистограмма количества метилированных цитозинов для 19 хромосомы генома стволовых клеток мыши, каждой метке соответствует свой цвет

Сравнение результатов GHMM и PBMM на GSE30202



“Ящик с усами” для количества метилированных цитозинов в зависимости от диапазона покрытия для 19 хромосомы генома стволовых клеток мыши, каждой метке соответствует свой цвет

Ссылки

Moritz Blume.

Expectation maximization: A gentle introduction, 2002.

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.