Моделирование данных бисульфитного секвенирования

Сергей Лебедев

Руководитель: Олег Шпынов

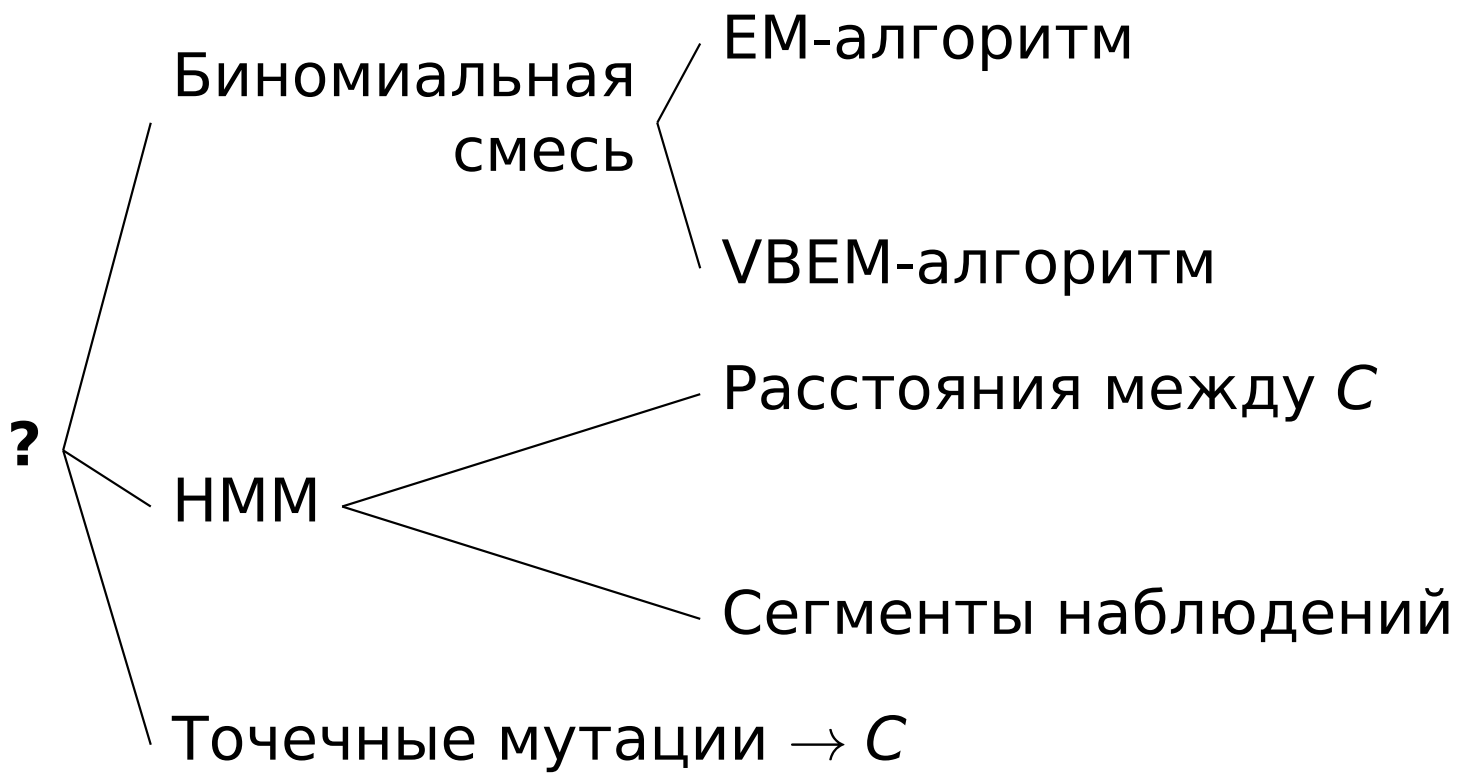
JetBrains

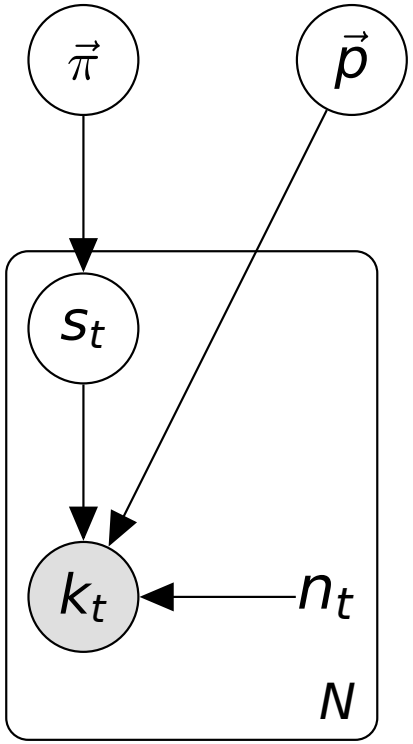
21 декабря, 2013

Я слышал(а) про **...**

- метилирование ДНК,
- бисульфитное секвенирование,
- порождающие вероятностные модели,
- функцию правдоподобия модели,
- метод максимального правдоподобия,
- EM-алгоритм,
- Томаса Байеса и байесовские модели.

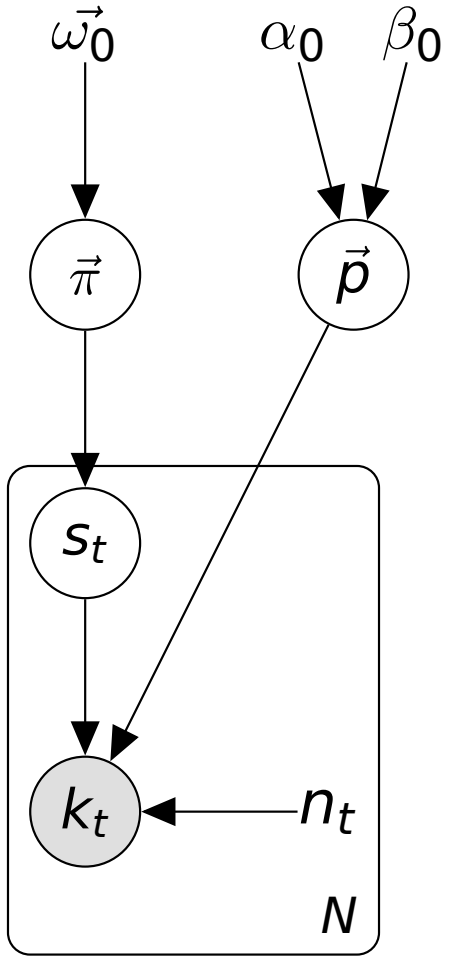
В конце было дерево ...





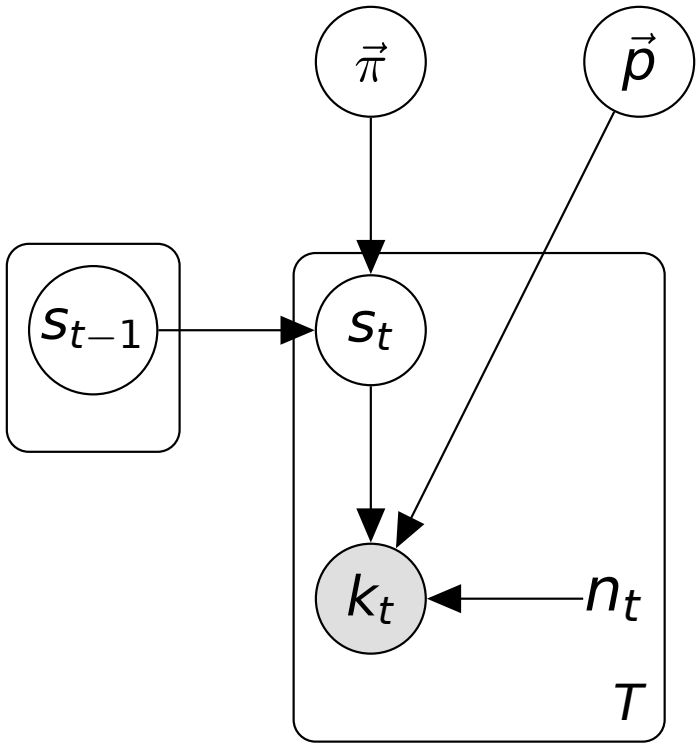
$$\vec{\theta}^* = \arg\max_{\vec{\theta}} \sum_{t=1}^{T} \log \sum_{c=1}^{C} \pi_c \mathcal{B}(k_t; n_t, p_c)$$

Байесовская биномиальная смесь



$$P(\vec{\theta}|\vec{k},\vec{n}) \propto \left(\sum_{t=1}^{T}\sum_{c=1}^{C}\pi_c P(k_t;n_t|p_c)\right) P(\vec{\pi})P(\vec{p})$$

HMM с биномиальными испусканиями



Гипотеза

Близко расположенные цитозины с бóльшей вероятностью будут в одинаковом состоянии.

- Пусть d_t — некоторое расстояние между $t-1$-м и t-м наблюдениями,
- тогда можно определить вероятность перехода из состояния i в состояние j на шаге t как

$$P(s_t = j; d_t | s_{t-1} = i) \stackrel{?}{=} (A^{d_t})_{ij} \stackrel{?}{=} A_{ijd_t}$$

1. Результаты поточечной разметки сложно интерпретировать, поэтому хочется моделировать сразу **последовательности** наблюдений в каждом состоянии (см. Hidden semi-Markov Model).
2. Точечную мутацию $C \to T$ можно ошибочно принять за неметилированный цитозин.
   - Текущий ”state of the art” подход: исключить из рассмотрения **все** позиции с известными SNV.
   - Альтернативный подход: использовать данные полногеномного секвенирования.