Анализ ошибок в чтениях, полученных в результате секвенирования технологией Ion Torrent

Эсаулова Екатерина

Институт биоинформатики

Научный руководитель: Коробейников А. И.

Санкт-Петербург
2015 г.

Выравнивание строк

Σ — алфавит, Σ^+ — пространство строк над алфавитом.
$\widetilde{\Sigma} = \Sigma \cup \ '-'$, где $'-'$ — специальный символ.

Рассмотрим задачу построения выравнивания для двух строк $r, s \in \widetilde{\Sigma}^+$.

На позиции i произошла ошибка, если $s[i] \neq r[i]$.

Гомополимеры

$\Sigma = \{A,\ C,\ G,\ T\},\ \Sigma^+$ — пространство строк.

Гомополимер — последовательность одинаковых букв в строке, идущих подряд. Обозначение: $AAA \to \langle A, 3\rangle$.

Тогда для $s \in \Sigma^+$ существует эквивалентное представление s^h, где s^h — последовательность гомополимеров.

Пример:

$s = AAAGCTTGG \Leftrightarrow s^h = \langle A, 3\rangle\langle G, 1\rangle\langle C, 1\rangle\langle T, 2\rangle\langle G, 2\rangle$

Ion Torrent

Объект исследования — последовательности, полученные технологией Ion Torrent — строки из Σ^+ длиной $200 - 600$ символов.

Суть технологии: чтение происходит по гомополимерам.

Проблемы:

- С ростом длины гомополимера хуже распознается его длина $\rightarrow$ возникают ошибки вставки-удаления;
- Гомополимер длиной более 15 невозможно прочитать верно;
- Качество чтения падает с приближением к концу последовательности.

Hidden Markov Models

Скрытые марковские модели (Hidden Markov Models, HMM) используются для моделирования ошибок, происходящих в процессе чтения строк над $\Sigma = \{A,\ C,\ G,\ T\}$ (Durbin, 1998).

Задача проекта:

- Формализация HMM для строк из гомополимеров;
- Упрощение полученной HMM;
- Построение процедуры оценки параметров;
- Реализация полученной модели, ее проверка.

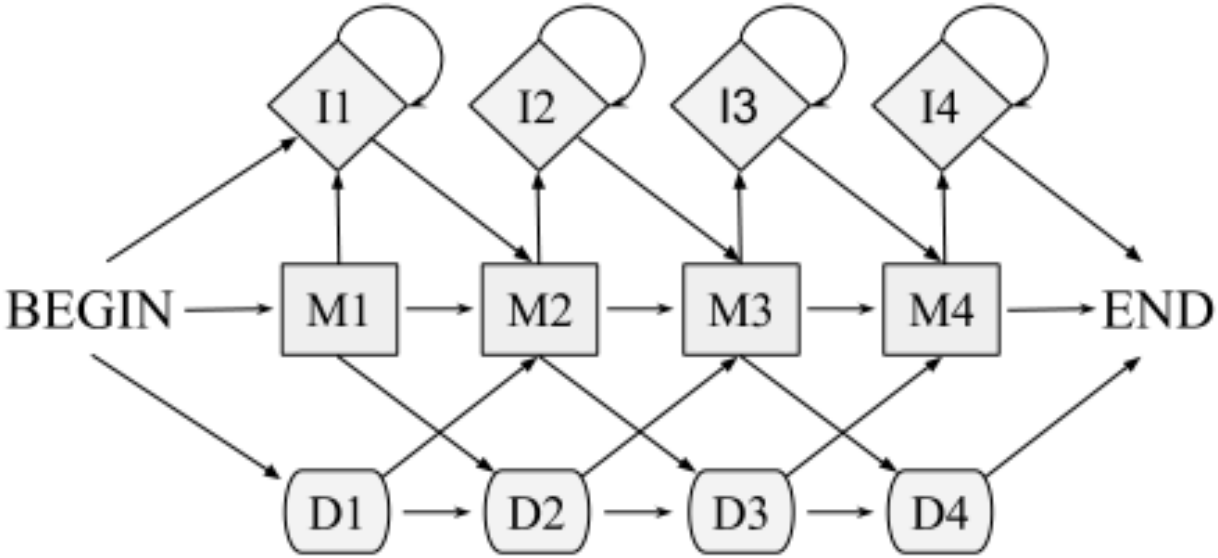
Hidden Markov Models

(Ω, F, P) — вероятностное пространство, X, Y — множества.
$\xi_i : \Omega \to X,\ \eta_i : \Omega \to Y,\ i = 1, 2, \ldots$ — случайные величины.

$\{\xi_i,\ \eta_i\}_{i=1,2,\ldots}$ — скрытая марковская модель, если:

- $p(\xi_t|\xi_{t-1}, \xi_{t-2}, \ldots, \xi_1) = p(\xi_t|\xi_{t-1})$, т.е. ξ_i образуют марковскую цепь;
- $p(\eta_t|\xi_t, \xi_{t-1}, \xi_{t-2}, \ldots, \xi_1, \eta_{t-1}, \eta_{t-2}, \ldots, \eta_1) = p(\eta_t|\xi_t)$;
- наблюдаются только реализации $\eta_i,\ i = 1, 2, \ldots$

HMM



Оцениваемые параметры:

- Матрица переходных вероятностей: P;
- Распределение вероятностей для наблюдений: $p_{\eta|\xi}$.

Упрощение HMM: параметрическая модель

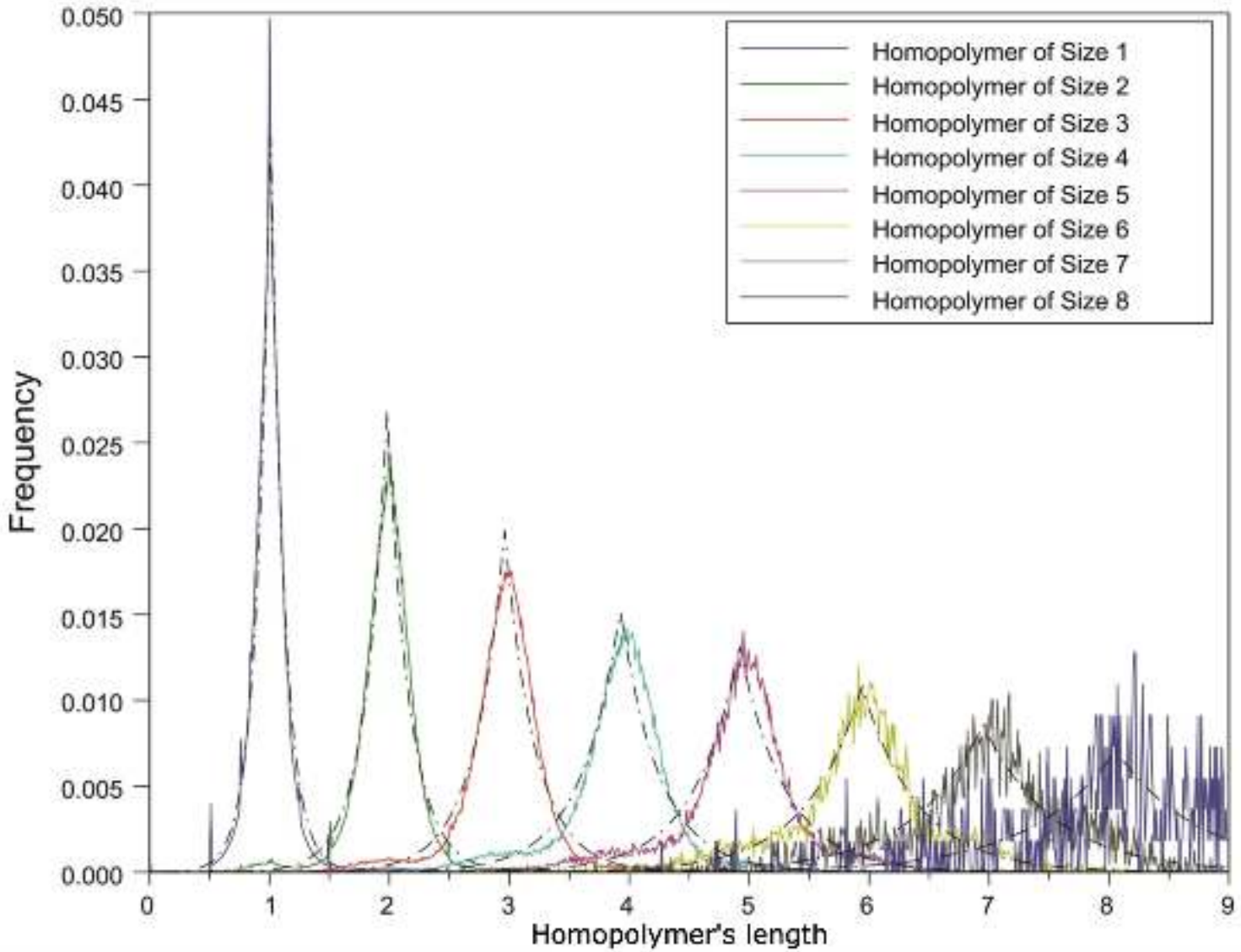
Пусть $s[i] = \langle \alpha, l \rangle$. Распределение вероятностей наблюдения $\langle \beta, k \rangle$:

$$p_{\eta|\xi} = p(\eta = \langle \beta, k \rangle \,|\, \xi) = \underbrace{p(\beta|\alpha, \xi)}_{base\ call} \underbrace{p(k|l, \alpha, \xi)}_{length\ call}.$$

Моделирование $length\ call$:

- $\xi_i = M_i$: распределение Лапласа;
- $\xi_i = I_i$: лог-нормальное распредление с $\mu = 0$.

Упрощение HMM: параметрическая модель



HMM: оценивание параметров

Цель: нахождение параметров модели, при которых вероятность наблюдать имеющиеся данные будет наибольшей.

- Оценка вероятности наблюдения данных в условиях модели — алгоритм Витерби;
- Оценка параметров — алгоритм Баума-Уэлча (частный случай EM-алгоритма).

Результаты

- Формализована HMM;
- Построена процедура оценки параметров;
- Реализована HMM — https://github.com/kesaulova/university;
- Осуществляется проверка HMM на реальных данных.

Спасибо за внимание!