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

Эсаулова Екатерина Николаевна, 422 группа

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

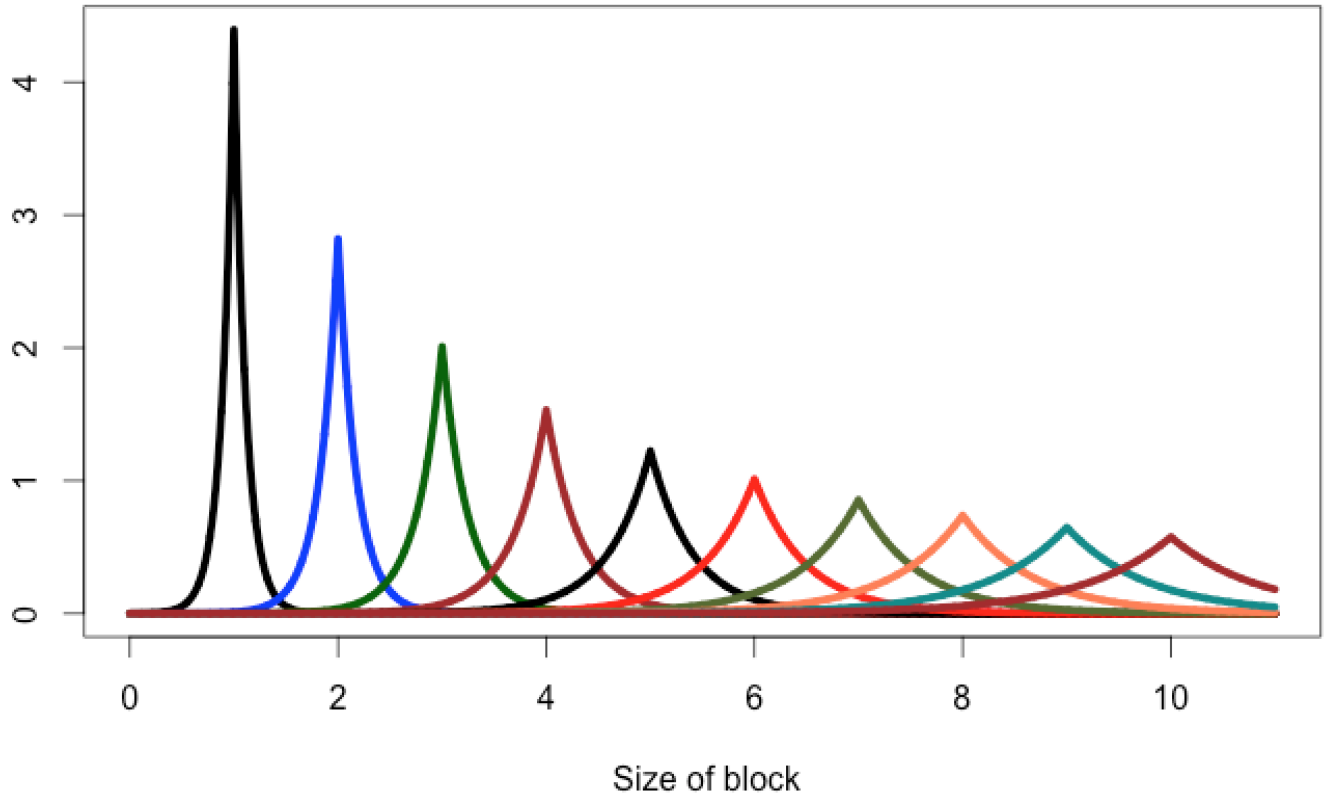
Научный руководитель: к. ф.-м. н., доц. Коробейников А. И.
Рецензент: ассистент Шлемов А. Ю.

Санкт-Петербург
2015 г.

Ion Torrent

Объект исследования — чтения ДНК, полученные технологией Ion Torrent — строки из $\{A, C, G, T\}$ длиной $200 - 600$ символов.

Суть технологии: строки читаются не побуквенно, а блоками по несколько идущих подряд одинаковых букв.



Ion Torrent

Объект исследования — чтения ДНК, полученные технологией Ion Torrent — строки над $\{A, C, G, T\}$ длиной $200 - 600$ символов.

Суть технологии: строки читаются не побуквенно, а блоками по несколько идущих подряд одинаковых букв.

Проблемы:

- С ростом количества одинаковых букв, встретившихся подряд, хуже распознается длина блока;
- Более 15 одинаковых букв подряд невозможно прочитать верно;
- Качество чтения падает с приближением к концу читаемой строки.

Чтение строк, типы ошибок

Рассмотрим побуквенное чтение строк.
$\Sigma = \{A, C, G, T\}, \widetilde{\Sigma} = \Sigma \cup '-'$.
Пусть s — читаемая строка, r — результат чтения, $s, r \in \widetilde{\Sigma}$.

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

- Замена (M, mismatch): $s[i] \neq r[i],\ s[i],\ r[i] \in \Sigma$;
- Вставка (I, Insertion): $r[i] = '-',\ s[i] \in \Sigma$;
- Удаление (D, Deletion): $s[i] = '-',\ r[i] \in \Sigma$.

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:

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

Hidden Markov Models

Скрытые марковские модели (Hidden Markov Models, HMM) используются для выравнивания строк над $\Sigma = \{A,\ C,\ G,\ T\}$ (Durbin, 1998).

Задача бакалаврской работы:

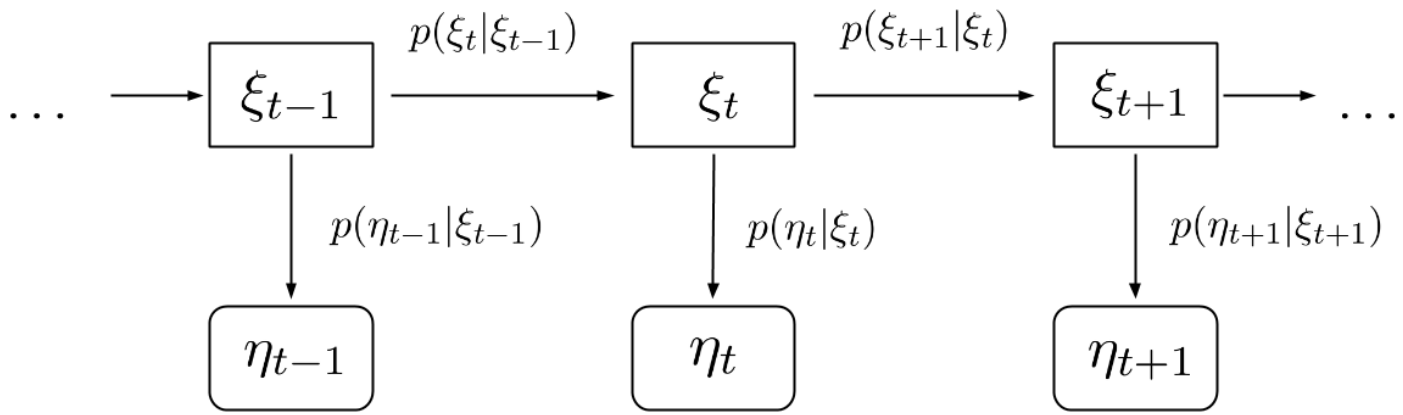
- Формализация HMM для строк из гомополимеров;
- Адаптация HMM для моделирования ошибок, происходящих при чтении строк технологией Ion Torrent;
- Построение процедуры оценки параметров;
- Реализация полученной модели, ее проверка.

Hidden Markov Models

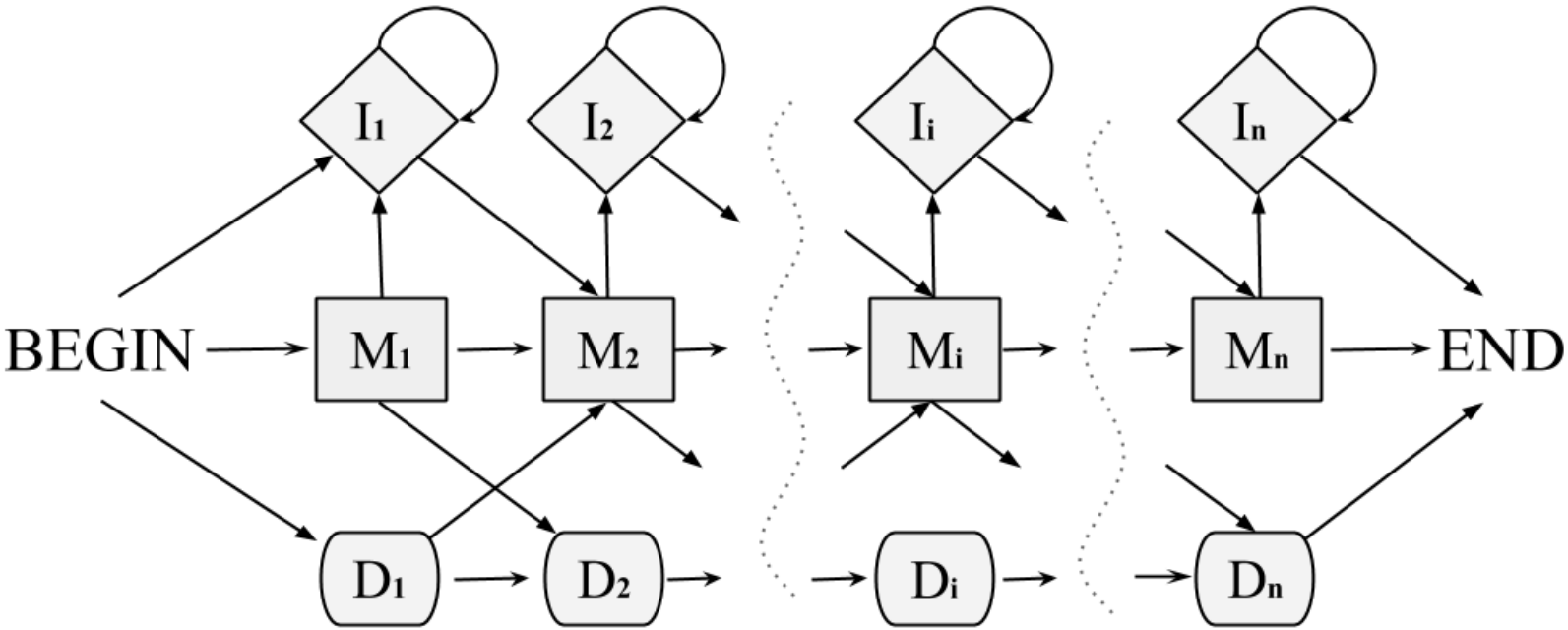
(Ω, F, P), $\xi_i : \Omega \to X$, $\eta_i : \Omega \to Y$, $i = 1, 2, \ldots$ — случайные величины.

$\{\xi_i,\ \eta_i\}_{i=1,2,\ldots}$ — 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;
- $p(\eta_t|\xi_t, \xi_{t-1}, \xi_{t-2}, \ldots, \xi_1, \eta_{t-1}, \eta_{t-2}, \ldots, \eta_1) = p(\eta_t|\xi_t)$.



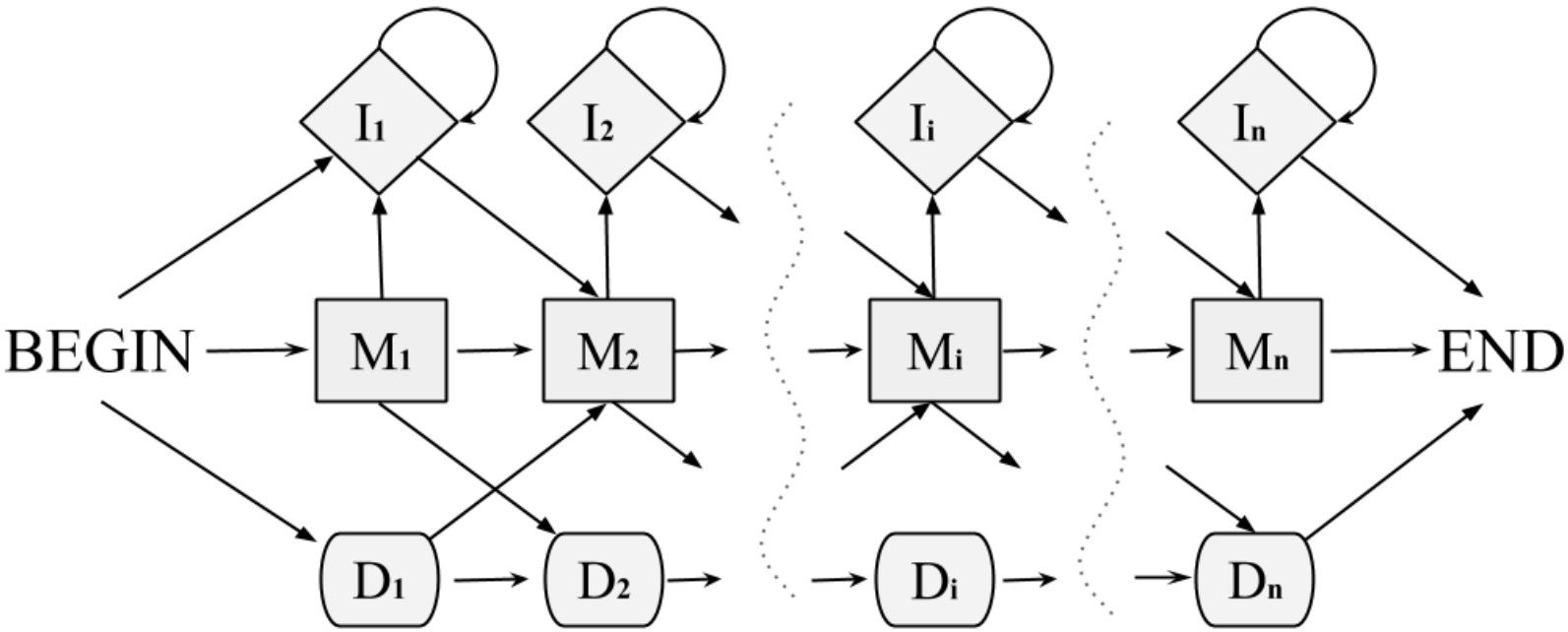
HMM, полная модель для чтения



Скрытыми состояниями являются:

- M (Match): правильное прочтение;
- I (Insertion): вставка;
- D (Deletion): удаление;
- Begin, End: обозначение начала/конца выравнивания.

HMM, полная модель для чтения



Оцениваемые параметры:

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

Адаптация HMM

Для полученной модели, $i : 1 \leq i \leq |s|$, s — читаемая строка:

- $\xi = M_i$: ~ 1000 параметров;
- $\xi = I_i$: ~ 20 параметров;

Строка s длиной до 600 символов $\rightarrow 600 \cdot 3 = 1800$ состояний HMM.

Таким образом, получается $\sim \underbrace{600 \cdot 1000}_{Match} + \underbrace{600 \cdot 20}_{Insertion} > 500000$ параметров.

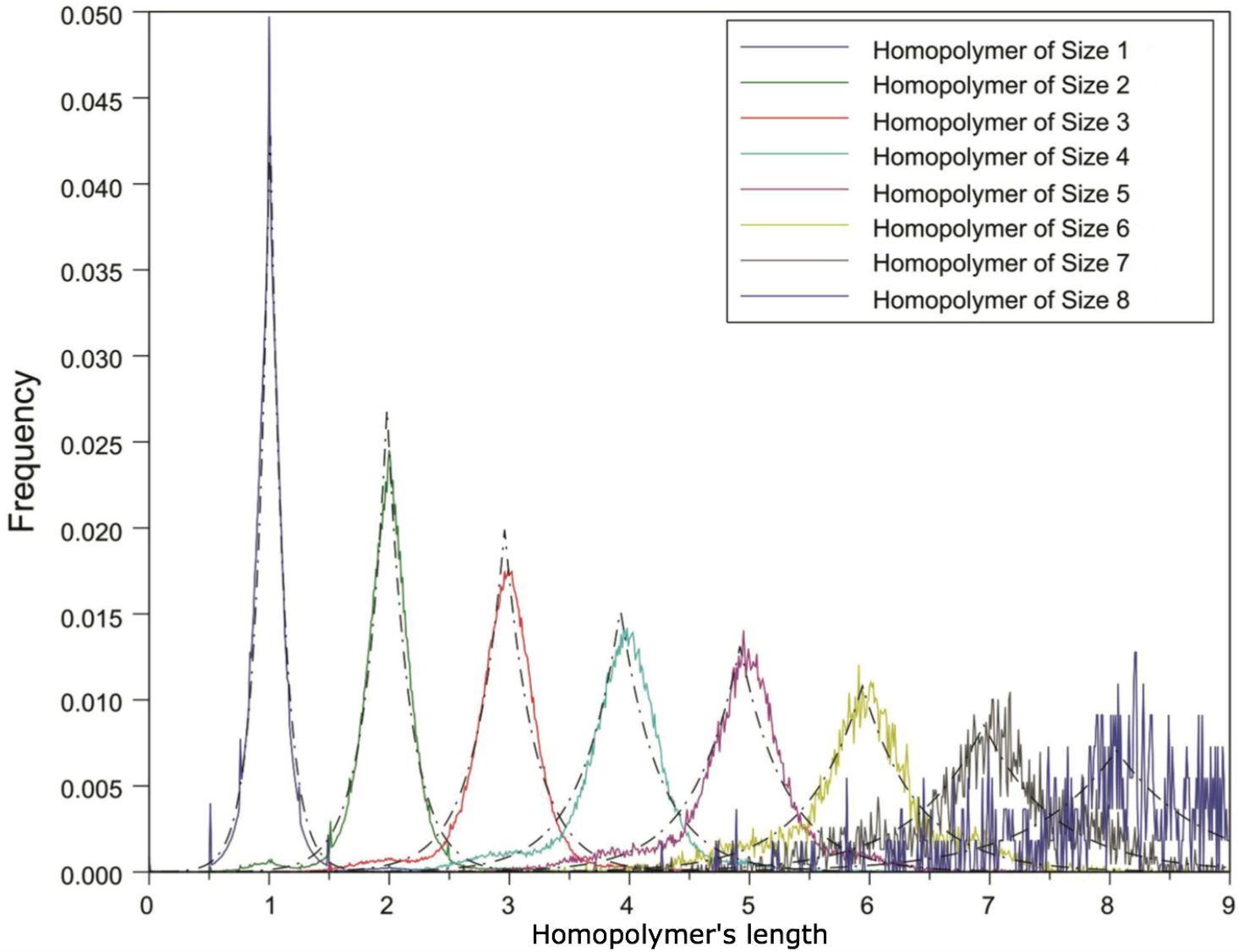
Адаптация HMM

Параметрическая модель:

- $\xi_i = M_i$ $(l > 0)$: $p(k|l, \alpha, M_i)$ моделируется с использованием распределения Лапласа;
- $\xi_i = I_i$ $(l = 0)$: $p(k|0,' -', I_i)$ моделируется с использованием лог-нормального распределения.

Получается ~ 10000 параметров.

Адаптация 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).

В силу ряда особенностей построенной модели нельзя пользоваться готовыми программными реализациями данных алгоритмов.

Оценка параметра перехода из Deletion в Deletion:

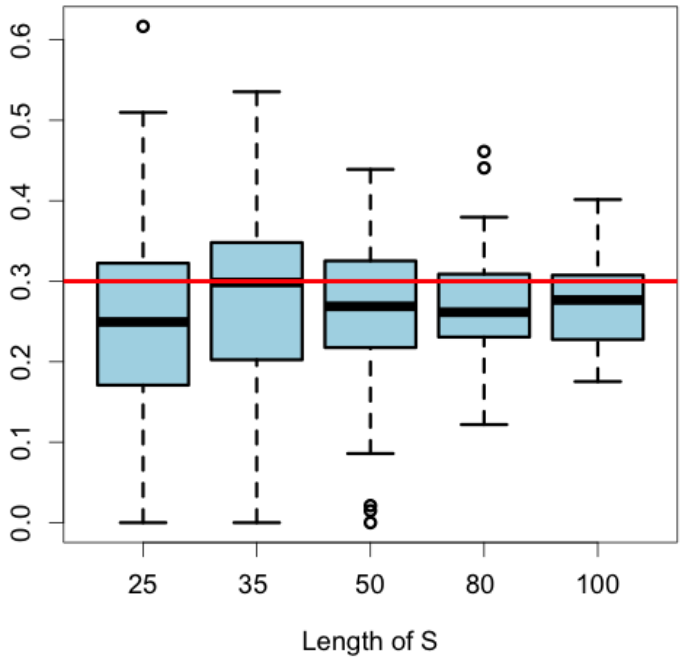


Рис.: Распределение оценки при разных параметрах выборки

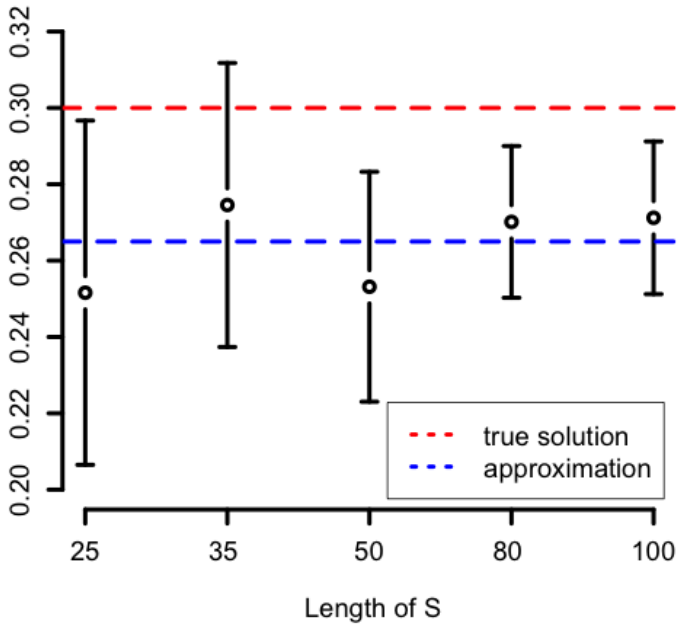


Рис.: Доверительные интервалы для уровня 0.95

Результаты

- Формализована HMM;
- Построена процедура оценки параметров;
- Реализована HMM;
- Написаны все алгоритмы для оценки параметров.
- Получены оценки HMM для реальных и модельных данных.