Исследование неоднородности распределения триплетов в символьных последовательностях

Никитина Ксения Андреевна, ИКИТ СФУ
Садовский Михаил Георгиевич, ИВМ СО РАН

Летняя школа по биоинформатике, 2016

Основные термины

Модель: бесконечная символьная последовательность из алфавита $\aleph = \{A, C, G, T\}$, заданная Марковским процессом порядка t.

Слово — любая конечная подпоследовательность, содержащаяся в исходной последовательности.

Зафиксируем два слова ω_1 и ω_2, каждое длины k.

$$k = 3, \quad \omega_1 = ACC, \quad \omega_2 = AAA$$

$$\ldots TA\mathbf{ACC}ACTGT\mathbf{AAA}TG \ldots$$

$$n = 8$$

Расстояние n между словами ω_1 и ω_2 — количество символов от начала ω_1 до начала ω_2.

Если нигде ближе n слово ω_2 не встретилось, то в таком случае ω_2 назовем **ближайшим соседом** к ω_1.

Матрица перехода марковского процесса

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & p_{25} & p_{26} & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & p_{rr} \end{pmatrix}$$

p_{ij} — вероятность перехода из слова i в слово j.
$r = m^k = 4^3 = 64$ — количество всевозможных слов длины k.
$\forall i = \overline{1, r} \quad \sum_{j=1}^{r} p_{ij} = 1; \forall j = \overline{1, r} \quad \sum_{i=1}^{r} p_{ij} = 1.$

$$A\mathbf{AA} \to \mathbf{AA}C \qquad C\mathbf{AA} \to \mathbf{AA}A$$

$$A\mathbf{AA} \to \mathbf{AA}G \qquad G\mathbf{AA} \to \mathbf{AA}A$$

$$A\mathbf{AA} \nrightarrow \mathbf{AC}A \qquad A\mathbf{CA} \nrightarrow \mathbf{AA}A$$

Функция распределения расстояний между словами

$f_{\omega_1,\omega_2}(n)$ — вероятность встретить слово ω_2 на расстоянии n справа от ω_1 и не раньше (то есть вероятность того, что ω_2 будет являться ближайшим соседом ω_1 на расстоянии n):

$$f_{\omega_1,\omega_2}(n) = a_{\omega_1} \cdot P_{\omega_2}^{n-2} \cdot b_{\omega_2}\,, \tag{1}$$

Здесь P_{ω_2} — матрица P с обнуленным столбцом перехода в слово ω_2; a_{ω_1} — строка перехода из слова ω_1, выписанная из матрицы P или P_{ω_2} соответственно; b_{ω_2} — столбец перехода в слово ω_2, выписанный из матрицы P.

Составление матрицы перехода на основании конечной последовательности

Для вычисления теоретических значений функции $f_{\omega_1,\omega_2}(n)$ составлялась матрица перехода P для слов длины $k = 3$ по частотному словарю слов длины $k + 1 = 4$ для каждой исследуемой последовательности по следующему правилу:

$$p_{\omega_1\omega_2} = \frac{F_{\nu_1\nu_2\nu_3\nu_4}}{\sum\limits_{x\in\{A,C,G,T\}} F_{x\nu_1\nu_2\nu_3}},$$

$\omega_1 = \nu_1\nu_2\nu_3$, $\omega_2 = \nu_2\nu_3\nu_4$.

Материалы и методы

Геном *Drosophila melanogaster* из 6 последовательностей длины $\sim 2,3 \cdot 10^7$ из EMBL-банка.

Порядок Марковского процесса $t = 3$.
Длина слов $k = 3$.
Расстояние между словами $n \in \{1, 2, \ldots, 2000\}$.

Программный комплекс

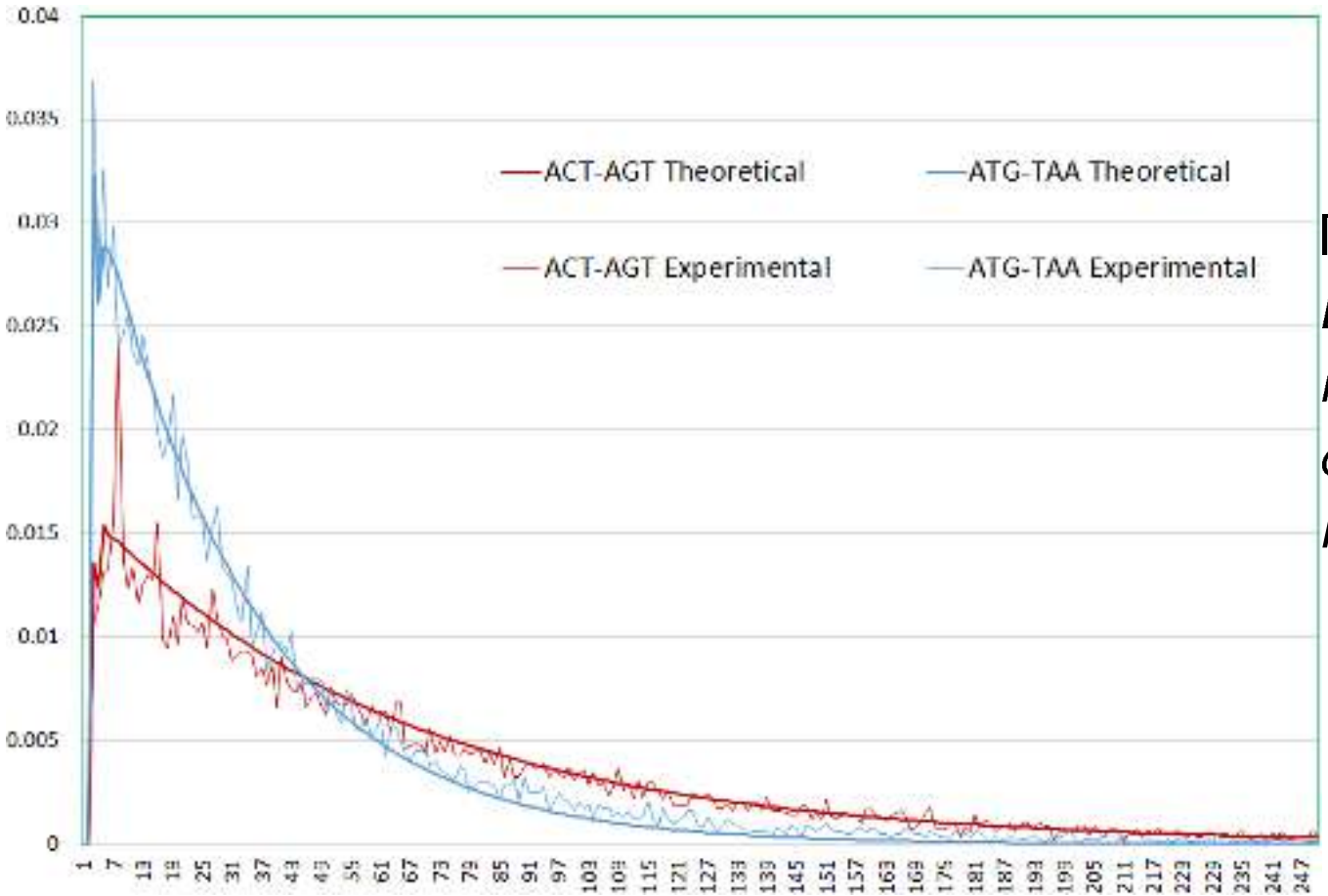
1. Программа для вычисления теоретических значений $f_{\omega_1,\omega_2}(n)$ с помощью матрицы перехода.
2. Программа для вычисления реальных значений частот расстояний до ближайшего соседа $\tilde{f}_{\omega_1,\omega_2}(n)$ в конечных последовательностях.
3. Программа для вычисления максимальных расстояний между словами в последовательности, называемых **мостами**:

   Правый мост $n^{max}_{\omega_2}$ — без ω_2 внутри;

   Левый мост $n^{max}_{\omega_1}$ — без ω_1 внутри;

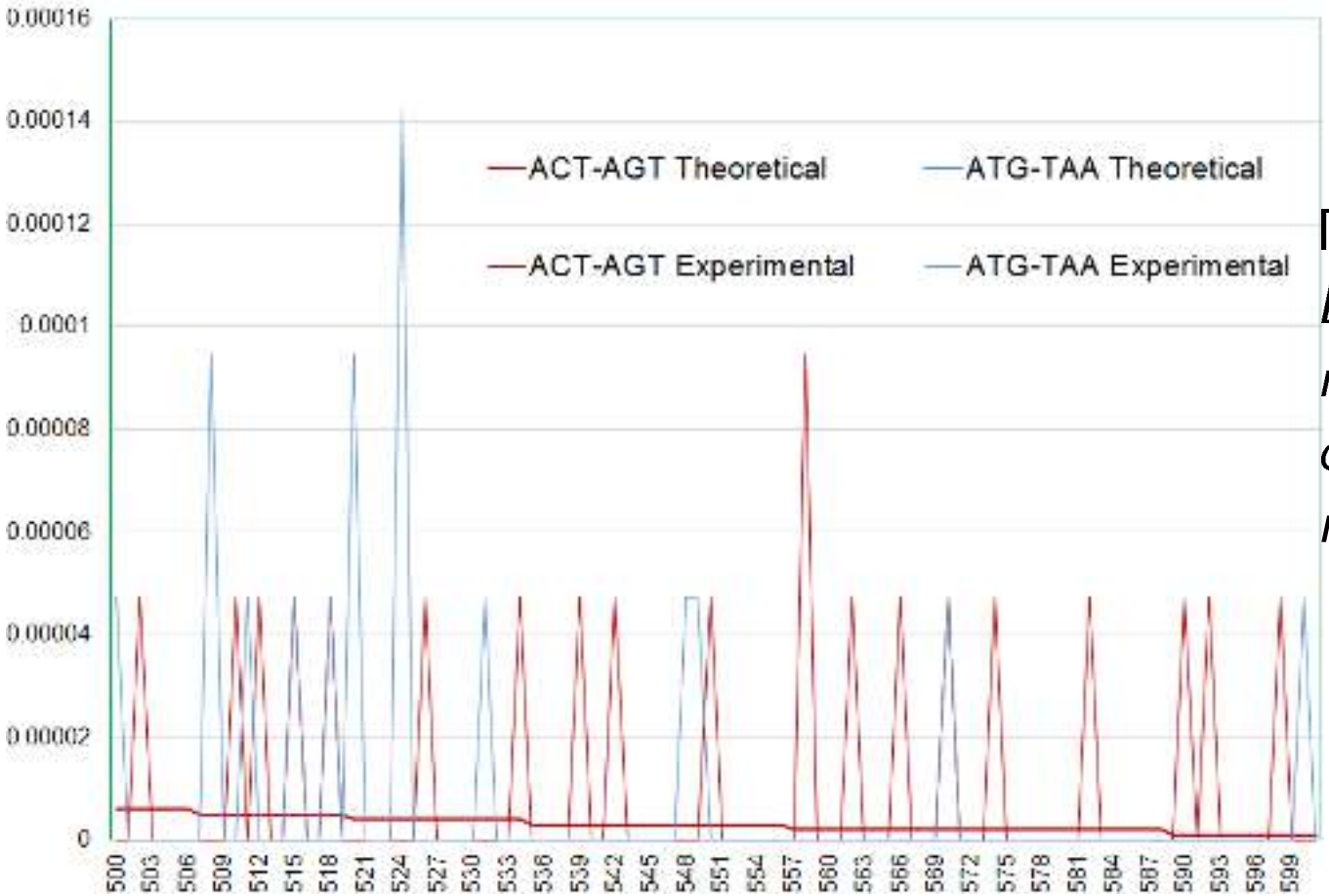
   Двусторонний мост $n^{max}_{\omega_1,\omega_2}$ — без ω_1 и ω_2.

Теоретические $f_{\omega_1,\omega_2}(n)$ и реальные $\tilde{f}_{\omega_1,\omega_2}(n)$ значения для малых расстояний между словами



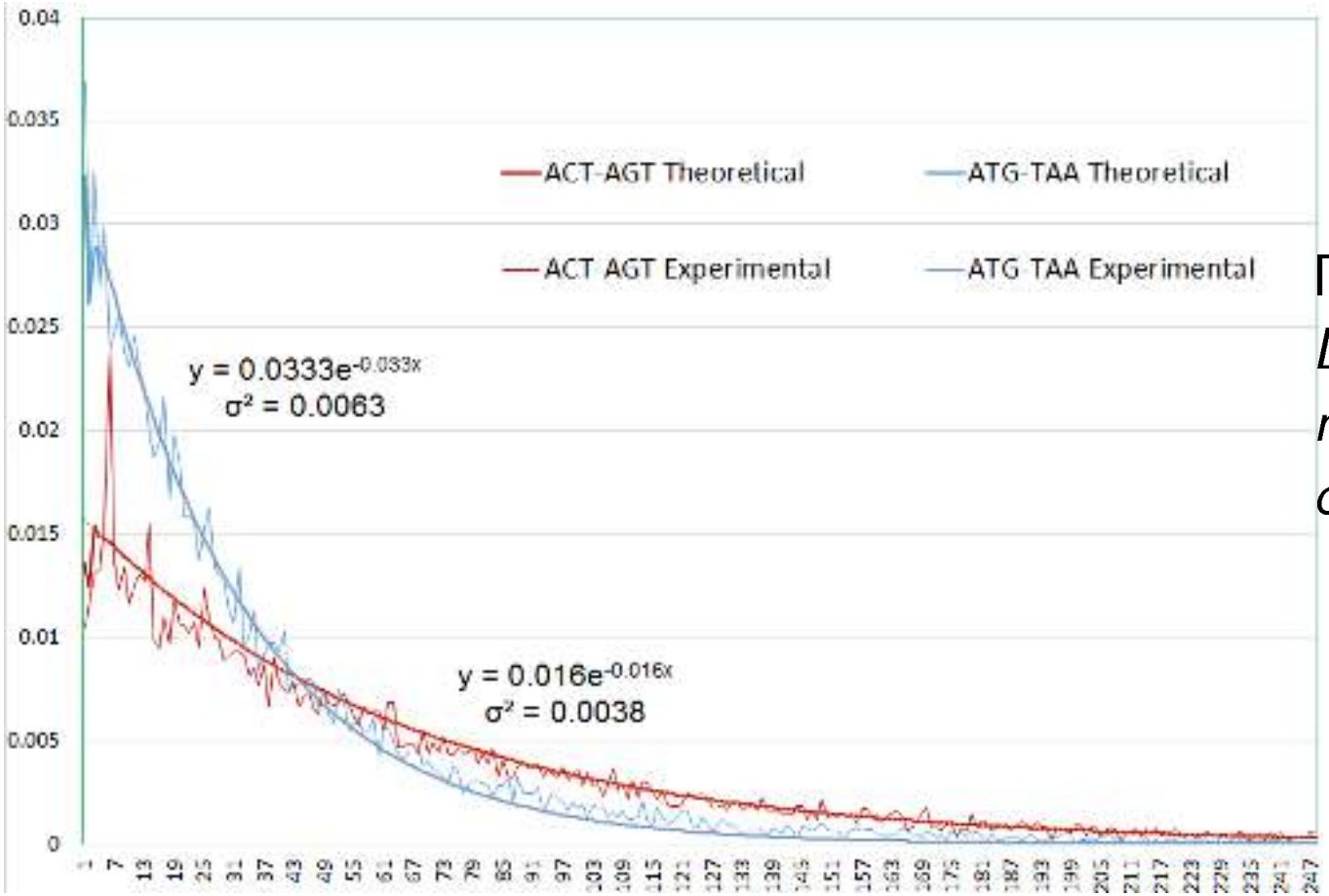
Последовательность: *Drosophila melanogaster chromosome 4*.
$n = 1, \ldots, 250$.

Теоретические $f_{\omega_1,\omega_2}(n)$ и реальные $\tilde{f}_{\omega_1,\omega_2}(n)$ значения для дальних расстояний



Последовательность: *Drosophila melanogaster chromosome 4*.
$n = 500, \ldots, 600$.

Значения коэффициентов в приближенной формуле $f_{\omega_1,\omega_2}(n) \approx A \cdot e^{-\lambda n}$



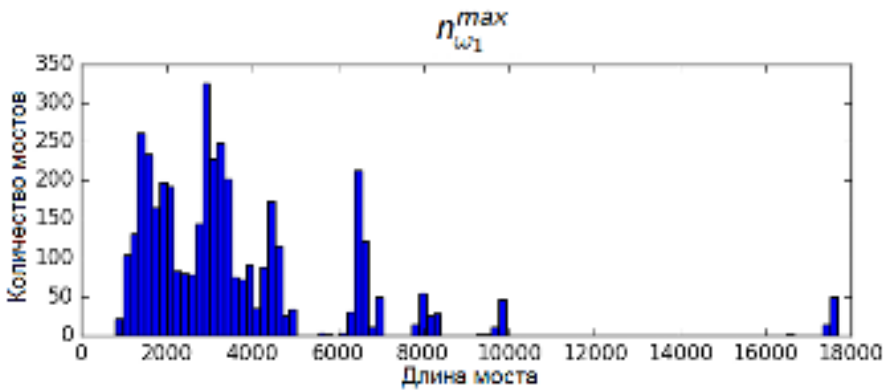
Последовательность: *Drosophila melanogaster chromosome 4*.

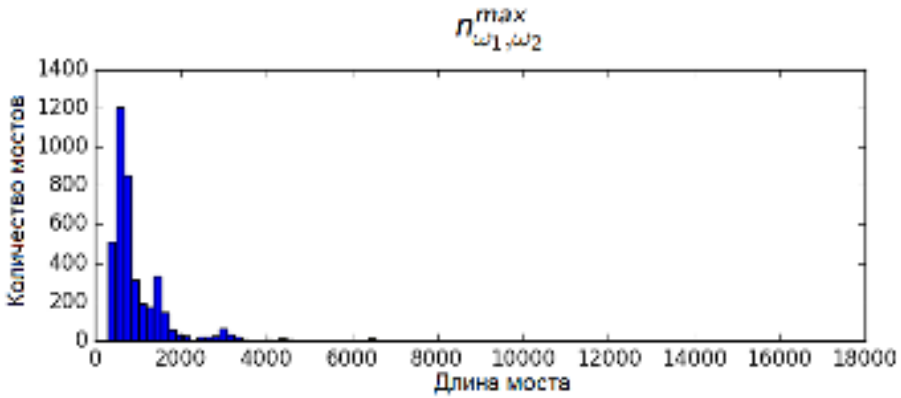
Максимальные расстояния $n^{max}_{\omega_2}$, на которых можно впервые встретить слово ω_2 справа от ω_1 (Правые мосты)

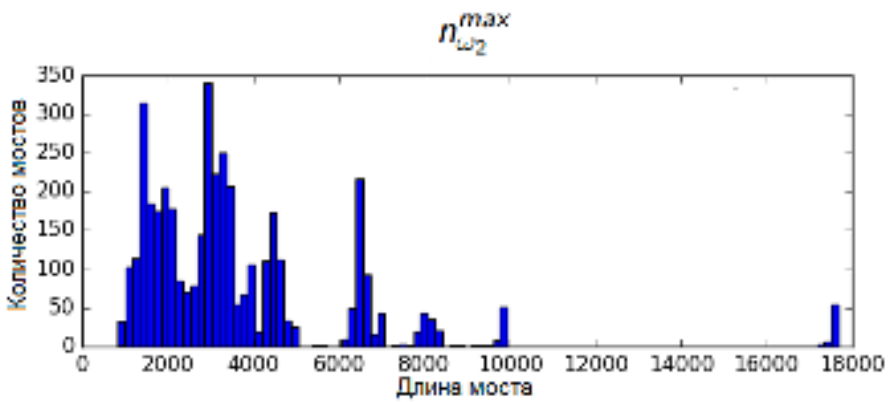
$\omega_1 - \omega_2$	$n^{max}_{\omega_2}$	$\omega_{1\,coord}$	$\omega_{2\,coord}$
agc-aac	836	14161324	14162160
gga-aac	877	14161283	14162160
tct-ttc	878	21232619	21233497
ccc-ttc	890	20669056	20669946
...	...	...	...
ggg-ggg	17679	22986728	23004407
tgg-ggg	17680	22986727	23004407
ttg-ggg	17681	22986726	23004407

Последовательность: *Drosophila melanogaster chromosome 2L.*

Длины левых, правых и двусторонних мостов

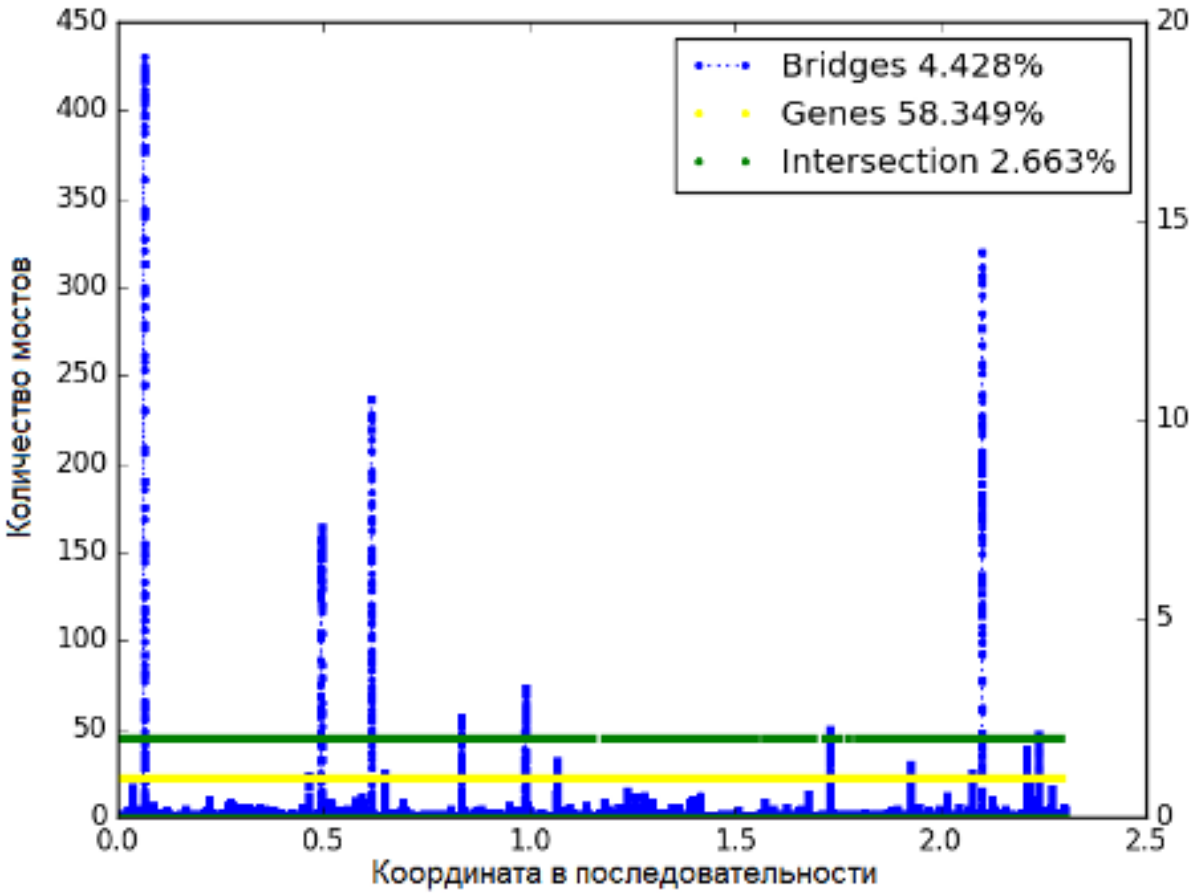






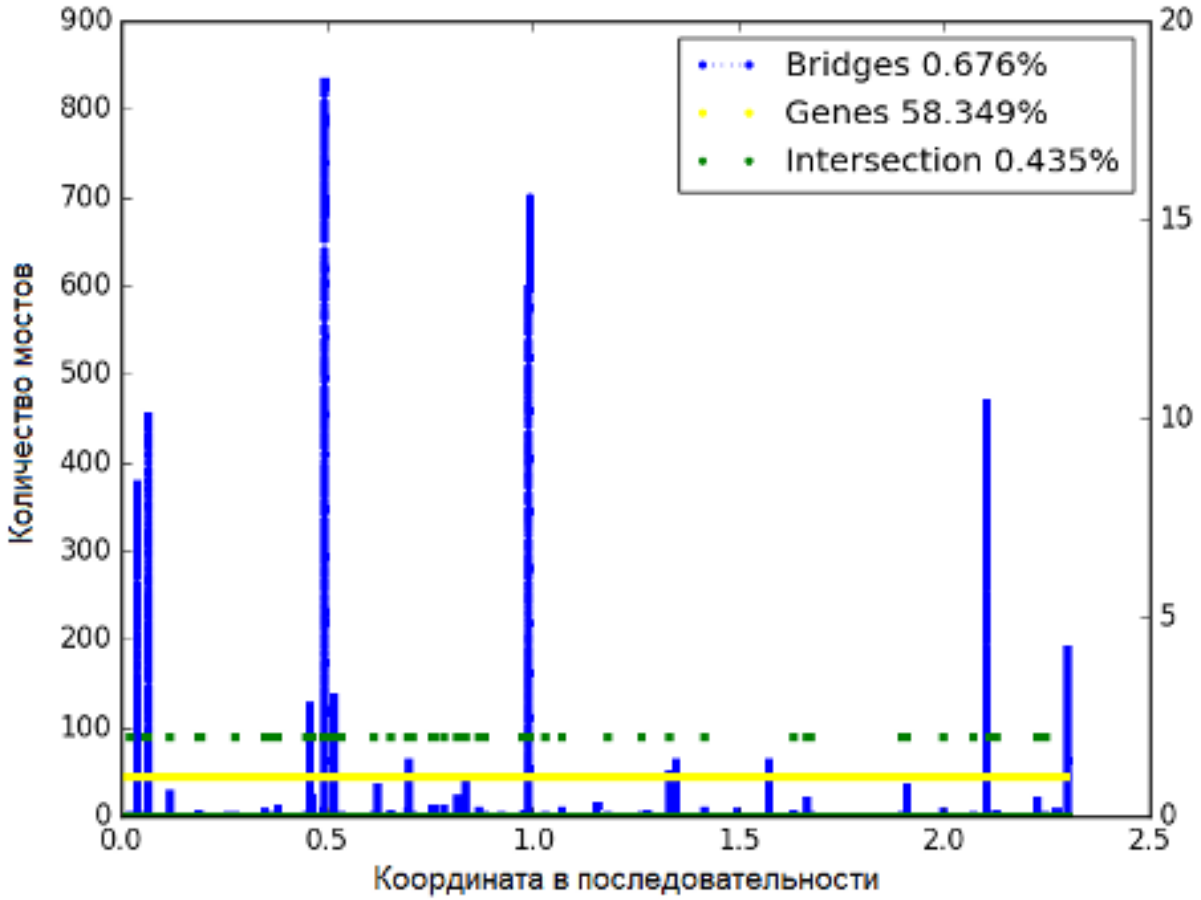
Последовательность:
Drosophila melanogaster chromosome 2L.

Расположение двусторонних $n^{max}_{\omega_1,\omega_2}$ мостов в последовательности



Последовательность: *Drosophila melanogaster chromosome 2L*.

Расположение правых $n_{\omega_2}^{max}$ мостов в последовательности



Последовательность: *Drosophila melanogaster chromosome 2L*.

Выводы

1. Функция распределения $f_{\omega_1,\omega_2}(n)$ сильно зависит от слов ω_1 и ω_2, выбранных в качестве соседей (частот этих слов).
2. Реальные значения $\tilde{f}_{\omega_1,\omega_2}(n)$ хорошо приближаются Марковским процессом третьего прядка, что, в свою очередь, можно аппроксимировать функцией $y = A \cdot e^{-\lambda n}$.
3. Характер распределения $\tilde{f}^*_{\omega_1,\omega_2}(n)$ на больших расстояниях не укладывается в коридор погрешности, вызванный конечностью последовательности («Тяжелые хвосты»).
4. Двусторонние мосты занимают больше места в последовательности, чем правые или левые, несмотря на то, что двусторонние мосты в среднем короче.
5. Наблюдается значительное пересечения мостов с кодирующими участками последовательности.

Спасибо за внимание!