МАТЕМАТИЧЕСКАЯ БИОЛОГИЯ И БИОИНФОРМАТИКА, 2007, том 2, №1, с.20-35,
http://www.matbio.org/downloads/Chaley2007(2_20).pdf

==========МАТЕРИАЛЫ I МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ==========
==========«МАТЕМАТИЧЕСКАЯ БИОЛОГИЯ И БИОИНФОРМАТИКА»==========

УДК: 577.2:519.23

Совместное использование различных критериев проверки однородности для выявления скрытой периодичности в биологических последовательностях

М.Б. Чалей[1*], Н.Н. Назипова[1], В.А. Кутыркин[2]

[1]*Институт математических проблем биологии, Российская академия наук, Пущино, Московская область, 142290, Россия*
[2]*Московский Государственный Технический Университет им. Н.А. Баумана, Москва, 107005, Россия*

Аннотация. В настоящей работе для выявления скрытой периодичности в биологических последовательностях используется модель дополнительных статистических экспериментов. Эта модель, включающая понятие нечетких тандемных повторов (Fuzzy Tandem Repeats), позволила предложить оригинальные статистические методы для оценки паттерна периодичности в размытых тандемных повторах (Approximate Tandem Repeats). В ряде случаев, при существенном проценте вставок и делеций в последовательности размытого тандемного повтора, выравнивание на основе полученной оценки размера паттерна периодичности оказывается более оптимальным по сравнению с выравниванием на основе известного метода Tandem Repeats Finder (TRF). Предложенные оригинальные статистические методы обладают значительно большей мощностью по сравнению с существующими аналогами. Основное достоинство этих методов состоит в возможности их применения в практических условиях нерепрезентативных выборок.

Ключевые слова: *скрытая периодичность, тест-период, профильная матрица, спектр относительных* χ^2 *-амплитуд*

1. ВВЕДЕНИЕ

В настоящее время под скрытой периодичностью в нуклеотидных (аминокислотных) последовательностях понимают размытые тандемные повторы (ATR) [1,2], в которых идентичность последовательно повторяющихся копий паттерна периодичности нарушена заменой оснований (аминокислотных остатков), и также вставками и делециями. В частности, скрытые тандемные повторы без вставок и делеций получили название нечетких тандемных повторов (FTR) [3,4].

Проблема выявления такой скрытой периодичности в биологической последовательности (строке) актуальна для поиска древних микро- и минисателлитов, генотипирования микроорганизмов с помощью VNTR (Variable Number Tandem Repeats) районов в их геномах [5–9], определения участков генетической нестабильности и анализа их влияния на проявление заболеваний [10-12], а также для исследования эволюции и

* maramaria@yandex.ru

поиска функциональных регулярных структур в белках [13-16], и для других подобных задач.

Для выявления периодичности в биологических строках ранее широко использовались методы Фурье- и вейвлет-анализа [17-21], комбинаторные методы и методы динамического программирования [1,3,4,22-24], различные статистические критерии проверки однородности строк [25-29] и др., см., например, обзор [30]. Однако по сравнению со статистическими критериями Фурье-анализ имеет плохую чувствительность к паттернам периодичности, размер которых в несколько раз превышает размер алфавита анализируемой строки [26]. Комбинаторные методы и динамическое программирование, как будет показано ниже, не всегда оптимально оценивают размер паттерна периодичности.

В настоящей работе рассматриваются стандартные статистические критерии проверки однородности строк [31,32], как наиболее простые, но эффективные при выявлении скрытой периодичности в биологических последовательностях. Однако в большинстве случаев биологическое значение анализируемой последовательности обусловлено небольшим количеством (<30) копий паттерна. Например, среди известных тандемных повторов в базе данных TRDB (http://tandem.bu.edu/cgi-bin/trdb/trdb.exe) тандемные повторы с количеством копий паттерна более 90 (т.е. с репрезентативной выборкой) составляют не более 0,1%. В остальных случаях объем выборки может не позволить непосредственного применения стандартных критериев.

Проблему создания критерия проверки однородности строк в случае нерепрезентативной выборки можно решить, если корректно выбрать модель дополнительных статистических экспериментов и использовать метод Монте-Карло для достоверного получения однозначного решения. Именно модель позволяет оценивать и сравнивать вероятностные характеристики достоверности принятия решения, в частности, ошибки I-го и II-го рода для эмпирически создаваемых нестандартных критериев.

Ранее предпринимались попытки создания эмпирических нестандартных критериев проверки однородности строк [26,27,29,33]. Строго говоря, эти критерии требуют для метода Монте-Карло генерации громадного числа перестановок ($\sim 10^{120}$ и более), так как не опираются на какую-либо теоретическую модель статистических экспериментов.

В настоящей работе для случаев недостаточного объема выборки разработан эмпирический нестандартный критерий (*NEP* критерий). Он удобен для предварительного автоматизированного выявления скрытой периодичности. С одной стороны, *NEP* критерий (подробно он будет описан далее) использует стандартные статистики, с другой стороны – метод Монте-Карло, подтверждающий достоверность принимаемого решения. Использование метода Монте-Карло в *NEP* критерии опирается на результаты предварительных статистических экспериментов.

2. МАТЕРИАЛЫ И МЕТОДЫ

Настоящий раздел содержит формулы статистик стандартных критериев проверки однородности строк и спектральные характеристики строк, используемые для предварительной оценки размера паттерна периодичности размытого тандемного повтора. Кроме того, здесь описаны модели статистических экспериментов, необходимые для создания нестандартных эмпирических критериев проверки однородности строк в условиях нерепрезентативной выборки.

2.1. Статистики стандартных критериев

Рассматриваемые далее статистические методы выявления скрытой периодичности, по

существу, направлены на выделение значимой неоднородности на тест-периодах анализируемой строки. Для этого анализируемая строка, состоящая из букв алфавита $A=\langle a_1,...a_K\rangle$, на каждом тест-периоде L последовательно разбивается на подстроки длины L (последняя подстрока может иметь меньшую длину). Такое разбиение на подстроки называется горизонтальным L-профилем строки. Например, горизонтальный 5-профиль нуклеотидной строки **(gcgctgggagccggagcg)** имеет вид

$$\textbf{(gcgct) (gggag) (ccgga) (gcg)}\ . \tag{1}$$

Если n – длина анализируемой строки, то $R_L=n/L$ – её тест-экспонент для тест-периода L. Каждый горизонтальный L-профиль строки можно представить в виде вертикального L-профиля (и наоборот). Для этого подстроки горизонтального L-профиля последовательно располагаются друг под другом, образуя строки вертикального L-профиля. Так для рассмотренного выше горизонтального 5-профиля (1) нуклеотидной строки её вертикальный 5-профиль имеет вид:

$$\begin{array}{ccccc}\textbf{(g} & \textbf{c} & \textbf{g} & \textbf{c} & \textbf{t)}\\ \textbf{(g} & \textbf{g} & \textbf{g} & \textbf{a} & \textbf{g}\\ \textbf{(c} & \textbf{c} & \textbf{g} & \textbf{g} & \textbf{a)}\\ \textbf{(g} & \textbf{c} & \textbf{g)} & & \end{array} \tag{2}$$

Поскольку выравнивание строк в вертикальном L-профиле не производится, то его можно рассматривать как простейший вариант профиля [34], в столбцах которого обычно находятся буквы из множественного выравнивания последовательностей, расположенных в его строках. Созданный вертикальный L-профиль позволяет вычислить частоту π_j^i встречаемости i-той буквы алфавита A в j-той позиции (j-том столбце) этого профиля анализируемой строки для $i=1,...,K$ и $j=1,...,L$. Матрица $\boldsymbol{\pi}=(\pi_j^i)_L^K$ называется выборочной L-профильной матрицей для анализируемой строки. Например, для 5-профиля (2) такая матрица имеет вид:

$$\boldsymbol{\pi}=\begin{pmatrix} 0 & 0 & 0 & 1/3 & 1/3\\ 3/4 & 1/4 & 1 & 1/3 & 1/3\\ 1/4 & 3/4 & 0 & 1/3 & 0\\ 0 & 0 & 0 & 0 & 1/3 \end{pmatrix}. \tag{3}$$

Частота p^i встречаемости i-той буквы алфавита A в анализируемой строке определяется по *профильной матрице* $\boldsymbol{\pi}=(\pi_j^i)_L^K$:

$$p^i=\frac{1}{L}\sum_{j=1}^{L}\pi_j^i,\quad i=1,...,K\,. \tag{4}$$

Статистика ν_S стандартного статистического критерия (S критерия) позволяет с помощью матрицы $\boldsymbol{\pi}$ проверить статистическую гипотезу об однородности строки на тест-периоде L.

В настоящей работе рассматриваются следующие стандартные критерии: χ^2-критерий Пирсона (P критерий), нормализованный χ^2-критерий Пирсона (NP критерий) [31] и информационный критерий (IC критерий) [32]. В случае репрезентативной выборки

($R_L \sim 100$ для нуклеотидных строк) статистика ν_S ($S \in \{P, NP, IC\}$) имеет χ^2-распределение с $N = (K-1)(L-1)$ степенями свободы, т.е.

$$\nu_S(L,n) \sim \chi^2_N \ , \tag{5}$$

где n – длина анализируемой строки.

Статистики стандартных критериев *P*, *NP* и *IC* имеют вид:

$$\nu_P(L,n) = R_L \sum_{j=1}^{L}\sum_{i=1}^{K}(\pi_j^i - p^i)^2 \big/ p^i \tag{6}$$

$$\nu_{NP}(L,n) = R_L \sum_{j=1}^{L}\sum_{i=1}^{K}(\pi_j^i - p^i)^2 \big/ p^i(1-p^i) \tag{7}$$

$$\nu_{IC}(L,n) = 2R_L \sum_{j=1}^{L}\sum_{i=1}^{K}(\pi_j^i \ln(\pi_j^i) - p^i \ln(p^i)) \tag{8}$$

Заметим, что статистика ν_{NP} ранее не использовалась для выявления скрытой периодичности в биологических последовательностях. Кроме того, в литературе по поиску скрытой периодичности (см., например, [26]) статистика ν_{IC} приводится в её оригинальном виде [32]:

$$\nu_{IC} = 2\Big(\sum_{j=1}^{L}\sum_{i=1}^{K} f_{ij}\ln(f_{ij}) - \sum_{j=1}^{L} f_{\bullet j}\ln(f_{\bullet j}) - \sum_{i=1}^{K} f_{i\bullet}\ln(f_{i\bullet}) + n\ln(n)\Big), \tag{9}$$

где f_{ij} - количество *i*-той буквы в *j*-той позиции тест-периода *L*, $f_{i\bullet} = \sum_{j=1}^{L} f_{ij}$ и $f_{\bullet j} = \sum_{i=1}^{K} f_{ij}$ для $i = 1,...,K$ и $j = 1,...,L$. Представление статистики (9) в виде (8) более наглядно раскрывает её смысловое содержание. Вид (8) показывает, что *IC*-статистика суммирует по позициям тест-периода отклонения энтропии реального распределения букв алфавита в каждой позиции тест-периода от ожидаемой энтропии.

2.2. Спектральные характеристики анализируемой строки

Поскольку анализируются базы данных нуклеотидных и белковых последовательностей (строк) большого объема, для стандартных критериев проверки однородности строк выбирается уровень значимости (ошибка I-го рода) $\alpha = 10^{-7}$. Для выбранного тест-периода *L* этому уровню соответствует критическое значение $\chi^2_{crit}(\alpha, N)$ с $N = (K-1)(L-1)$ степенями свободы. Следовательно, если для репрезентативной выборки значение статистики $\nu_S(L,n)$ ($S \in \{P, NP, IC\}$) анализируемой строки длины *n* на тест-периоде *L* удовлетворяет условию

$$\nu_S(L,n) \big/ \chi^2_{crit}(\alpha,(K-1)(L-1)) \le 1 \, , \tag{10}$$

то на тест-периоде *L* принимается гипотеза об однородности строки, в противном случае – строка признаётся неоднородной. Поэтому для стандартного *S* критерия в качестве спектральной характеристики анализируемой строки в работе предлагается функция H_S вида

$$H_S(L) = \nu_S(L,n) \big/ \chi^2_{crit}(\alpha,(K-1)(L-1)), \quad (11)$$

где $L = 1,...,L_{\max}$ $(L_{\max} \sim n/5K)$.

Функцию H_S (11) будем называть относительной χ^2-амплитудой *S* критерия. График функции H_S наглядно демонстрирует проявление значимых неоднородностей анализируемой строки на тех тест-периодах L, где $H_S(L) > 1$. Далее такой график называется *H*-спектром *S* критерия анализируемой строки (или H_S-спектром).

В настоящей работе в качестве дополнительных спектральных характеристик предлагается использовать значения уровня сохранности буквы и уровня совпадения букв в рассматриваемых вертикальных *L*-профилях анализируемой строки. Для этого на каждом тест-периоде *L* строки по выборочной *L*-профильной матрице $\boldsymbol{\pi} = (\pi^i_j)^K_L$ вычисляются два параметра $pl(L)$ и $cl(L)$:

$$pl(L) = \frac{1}{L}\sum_{j=1}^{L} \max\{\pi^i_j : i \in 1,...,K\}, \quad (12)$$

$$cl(L) = \frac{1}{L}\sum_{j=1}^{L}\sum_{i=1}^{K} \pi^i_j \cdot \pi^i_j. \quad (13)$$

Параметр $pl(L)$ называется *уровнем сохранности буквы* на тест-периоде *L*, параметр $cl(L)$ – *уровнем совпадения букв* на на тест-периоде *L*. Формулы (12) и (13) определяют функции pl и cl для тест-периодов $L = 1,...,L_{\max}$ $(L_{\max} \sim n/5K)$. По аналогии с *H*-спектрами графики функций pl и cl будут называться pl- и cl-спектрами анализируемой строки, соответственно.

2.3. Оценка размера паттерна периодичности размытого тандемного повтора

В H_S-спектре размытого тандемного повтора, как правило, выделяется ряд тест-периодов, на которых проявляется значимая неоднородность последовательности. Поэтому для оценки размера паттерна периодичности этого повтора применяются дополнительные спектральные характеристики – pl- и cl-спектры (см. формулы (12) и (13)).

Как правило, при низком проценте вставок и делеций букв в последовательности размытого тандемного повтора оценка размера паттерна периодичности по методу спектральных характеристик совпадает с оценкой, полученной другими известными методами, например, методом Tandem Repeats Finder (TRF) [1]. Однако, совпадение оценок размера паттерна тандемных повторов, полученных методом TRF и с помощью *H*-спектров стандартных критериев (P, NP, IC), происходит не всегда. Рассмотрим один из таких случаев на примере тандемного повтора *Caenorhabditis elegance* с размером паттерна 19 оснований и числом копий 37.4 (база данных TRDB (http://tandem.bu.edu/cgi-bin/trdb/trdb.exe), *C. elegance* (March 2004) Full Genome). Профиль этого тандемного повтора, полученный выравниванием каждой копии относительно консенсус-паттерна, соответствующего паттерну периодичности 19 оснований (осн.) по методу TRF [1], показан на рисунке 1(а). Вид консенсус-паттерна совпадает со строкой 1 на рис. 1(а). Консенсус-паттерн, согласно [1], выводится по правилу большинства из выравнивания копий повтора (строк профиля) с паттерном. На рисунке 1(б) приведен *H*-спектр *P* критерия для последовательности рассматриваемого тандемного повтора. Согласно этому

спектру тест-период 19 осн. не является значимым, и наибольшей значимостью обладает тест-период 35 осн. Оценка размера паттерна в 35 осн. также поддерживается pl- и cl-спектрами (см. рис. 1(в)). Профиль последовательности повтора, соответствующий оценке паттерна в 35 осн., показан на рис. 1(г). Выравнивание каждой копии было выполнено относительно консенсус-паттерна, вид которого совпадает со строкой 1 на рис. 1(г). Из сравнения профилей на рис. 1(а) и рис. 1(г) видно, что профиль на основе паттерна 35 осн. более оптимален. Кроме того, на рис. 1(г) можно заметить, что строки с номерами 3-7 и строки с номерами 8-12 образуют две идентичные последовательности.

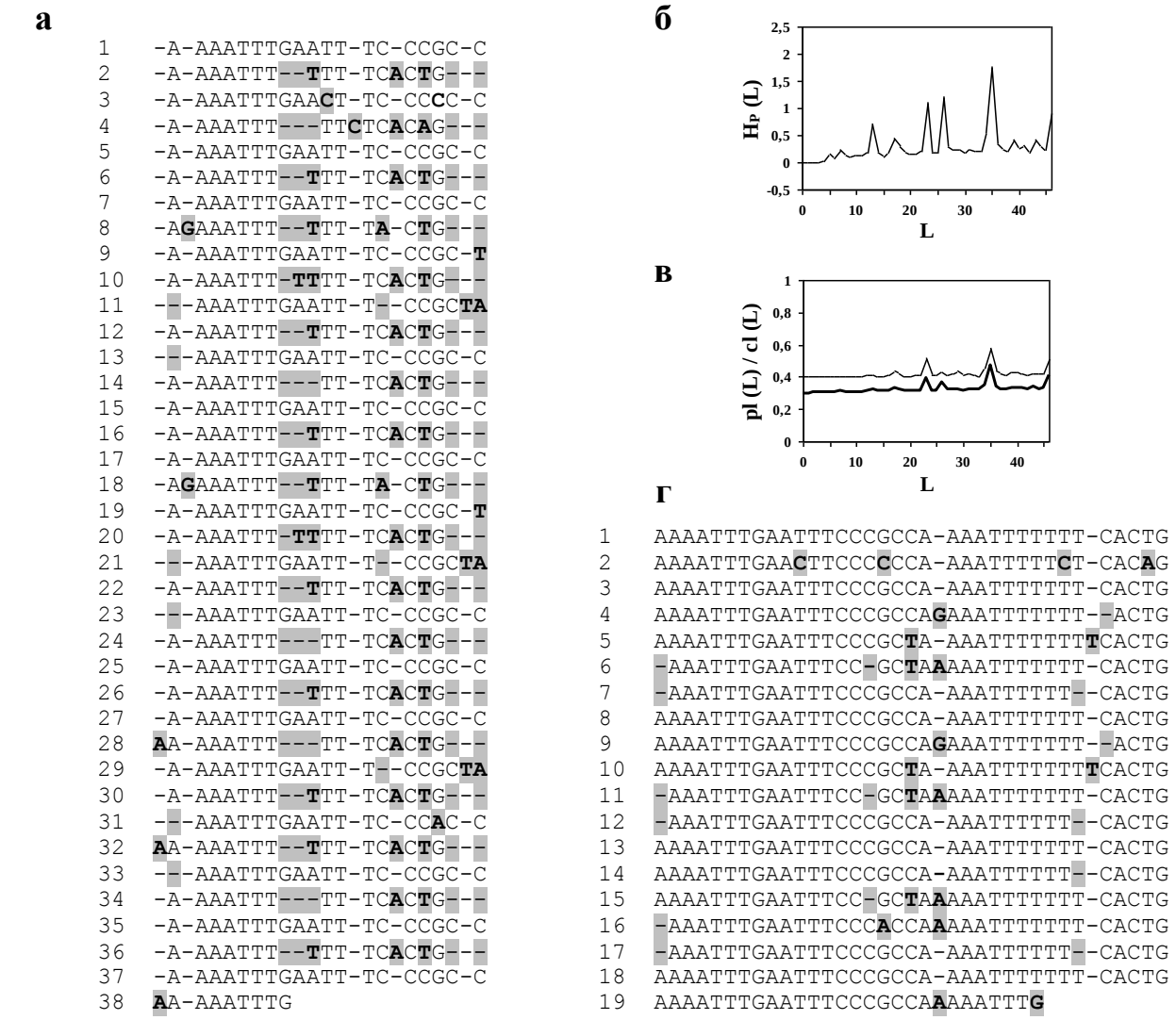
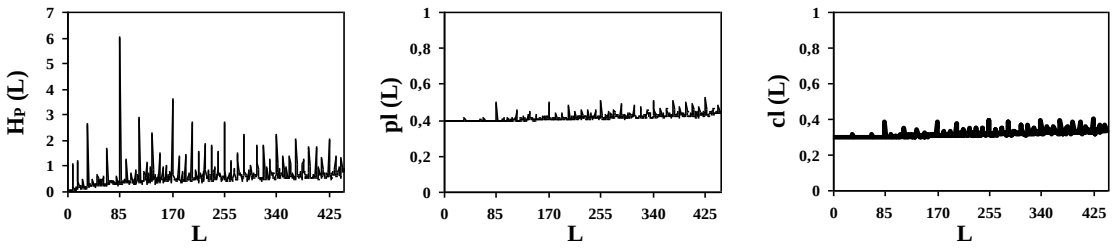


Рис. 1. Профили последовательности тандемного повтора *C. elegance* на хромосоме I из базы данных TRDB (индексы: 1429679-1430330), построенные в соответствии с выбранной оценкой размера паттерна повторяющихся копий в 19 осн. (а) и в 35 осн. (г). **а.** Профиль тандемного повтора для размера паттерна 19 осн. (% несовпадений между копиями – 7, % вставок и делеций между копиями – 14). **б.** H_P-спектр (*H*-спектр стандартного *P* критерия) анализируемого тандемного повтора. **в.** *pl*-спектр уровня сохранности буквы (тонкая линия) и *cl*-спектр уровня совпадения букв (жирная линия), см. формулы (12) и (13). **г.** Профиль последовательности анализируемого тандемного повтора, соответствующий оценке размера паттерна в 35 осн., полученной из спектров на рис. 1(б,в) (подробности см. в тексте).

Необходимо отметить, что с точки зрения статистики выборку, использованную для оценки размера паттерна в 35 осн. (тест-экспонент $R_L = 18.6$ для тест-периода L=35), нельзя считать репрезентативной. Поэтому, строго говоря, без дополнительной информации нет оснований для того, чтобы принять или отвергнуть гипотезу об однородности рассматриваемой последовательности на тест-периоде L=35. На получение необходимой дополнительной информации и направлены предлагаемые далее нестандартные эмпирические критерии.

В биологических последовательностях может наблюдаться эффект дупликаций фрагмента периодической последовательности, содержащего несколько идентичных копий. В этом случае при наличии в тандемном повторе вставок и делеций значимая неоднородность не всегда выявляется на тест-периоде, совпадающем с размером паттерна периодичности (*основным периодом*). Значимая неоднородность может выявляться на тест-периоде близком к одному из обертонов основного периода. Приведем пример проявления такого эффекта.

а



б

```
1........10........20........30........40........50........60........70........80

TRF PATTERN

AGCACCAAATAGCACTTTAGCACCAAATAGCACCAATATAGCACCAATGTAGCACCAAATAGCACCAATTTAGCACCAATG

PATTERN

AGCACCAATAGCACCAATAGCACCAATAGCACCAATAGCACCAATAGCACCAATAGCACCAATAGCACCAATTT
```

Рис. 2. а. Спектральные характеристики размытого тандемного повтора на I хромосоме *C. elegance* (база данных TRDB, индексы: 10953995-10961073). **б.** Паттерны периодичности рассматриваемого тандемного повтора, полученные методом TRF [1] и методом оценки размера паттерна по спектральным характеристикам последовательности (см. формулы (11)-(13)).

На рис. 2(а) показаны спектральные характеристики размытого тандемного повтора на I хромосоме *C. elegance*. Из анализа этих характеристик следует, что тест-период $L = 85$ осн. является оценкой размера паттерна периодичности для этой последовательности, на тест-периоде $L = 32$ осн. также выявляется значимая неоднородность, а для тест-периода $L = 53$ осн. принимается гипотеза об однородности последовательности. Тест-периоды $L = 84, 53, 32$ осн. предлагаются методом TRF [1] в качестве оценок размера паттерна периодичности для этой последовательности. Параметры выравнивания по методу TRF для этих тест-периодов различаются несущественно. На рис. 2(б) показаны паттерн периодичности для тест-периода $L = 84$ осн., полученный методом TRF, и паттерн периодичности, соответствующий выравниванию копий согласно оценке в 85 осн. по спектральным характеристикам последовательности на рис. 2(а). Выравнивание анализируемой последовательности, соответствующее оценке в 85 осн., и его консенсус-паттерн приведены на рисунке 3.

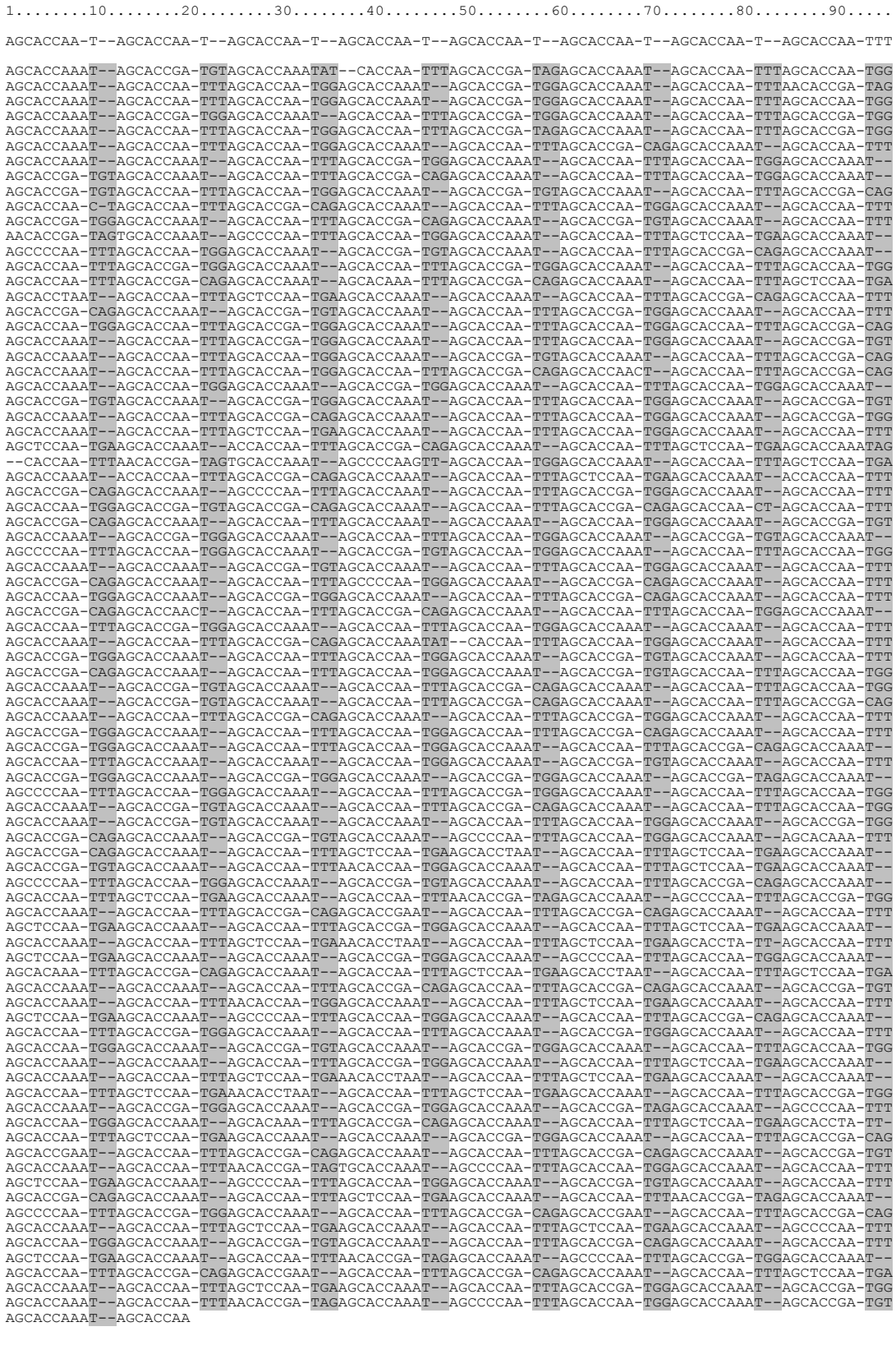


Рис. 3. Консенсус-паттерн и выравнивание копий в последовательности размытого тандемного повтора на I хромосоме *C. elegance* (база данных TRDB, индексы: 10953995-10961073), соответствующие оценке размера паттерна периодичности в 85 осн. согласно спектральным характеристикам последовательности (см. рис. 2(а)). Вертикальные серые линии выделяют структуру размытого тандемного повтора.

Из представленных на рисунках 2 и 3 данных можно сделать вывод, что в рассматриваемой последовательности происходили дупликации некоторого исходного размытого тандемного повтора длиной около 85 осн. с паттерном AGCACCAAT. Заметим, что процесс размытия тандемного повтора, в основном, происходил на стыках восьми последовательно повторяющихся копий этого паттерна.

2.4. Модели статистических экспериментов

Как отмечалось ранее, статистическое моделирование необходимо для обоснования метода Монте-Карло, встроенного в процедуру предлагаемого в работе нестандартного эмпирического критерия проверки однородности строк, применимого в условиях нерепрезентативной выборки. Фактически, при применении статистических критериев проверки однородности анализируемую строку можно рассматривать как реализацию случайной модельной строки с простой структурой периодичности. Такая модельная строка однозначно индуцируется некоторой L-профильной матрицей $\boldsymbol{\pi} = (\pi_j^i)_L^K = (\boldsymbol{\pi}_1,...,\boldsymbol{\pi}_L)$ в заданном алфавите $A = \langle a_1,...,a_K \rangle$ (см., например, (3)). В свою очередь, столбцы этой матрицы являются выборкой объема L в схеме независимых испытаний случайного столбца $\mathbf{P}(F)$ частот букв заданного алфавита. При этом предполагается, что в пространстве столбцов частот букв алфавита A задано равномерное (или близкое к нему) распределение F. С помощью датчика случайных чисел можно осуществить такое моделирование случайной матрицы $\boldsymbol{\Pi}(F)$ размера $K \times L$, реализацией которой и является указанная ранее матрица $\boldsymbol{\pi}$.

В дальнейшем выборку единичного объема из рассматриваемой случайной величины (генеральной совокупности) будем называть квази-случайной величиной из этой генеральной совокупности. Например, $\boldsymbol{\pi}$ – квази-случайная матрица из $\boldsymbol{\Pi}(F)$.

По заданному квази-случайному j-тому столбцу частот $\boldsymbol{\pi}_j$ ($j = 1,...L$) L-профильной матрицы $\boldsymbol{\pi}$ с помощью датчика случайных чисел создается j-тый квази-случайный столбец вертикального L-профиля периодичности создаваемой квази-случайной строки длины n. Затем, из полученного таким образом вертикального L-профиля периодичности восстанавливается квази-случайная строка str из генеральной совокупности $\mathbf{Str}(n, A, \boldsymbol{\pi})$, являющейся случайной строкой с простой структурой периодичности.

Создав с помощью датчика случайных чисел выборку большого объема квази-случайных строк из генеральной совокупности $\mathbf{Str}(n, A, \boldsymbol{\pi})$ с квази-случайной матрицей $\boldsymbol{\pi}$ из генеральной совокупности $\boldsymbol{\Pi}(F)$, эту выборку можно отождествить с квази-случайной строкой длины n из генеральной совокупности $\mathbf{STR}^{(1)}(n, A, L)$, являющейся случайной строкой со случайной L-профильной матрицей.

2.4.1. Модельные квази-случайные строки различных серий

Пусть $\mathbf{p}^{(1)} = (p^{1(1)},..., p^{K(1)})^T$ – квази-случайный столбец из $\mathbf{P}(F) = \mathbf{P}^{(1)}(F)$. Тогда для каждого $i=1,...,K$ и некоторого натурального числа $d > 1$ вычисляется число $x^{i(d)} = p^{i(1)} \cdot ... \cdot p^{i(1)}$ – d-кратное произведение числа $p^{i(1)}$, и полагается $x^{(d)} = \sum_{i=1}^{K} x^{i(d)}$, $p^{i(d)} = x^{i(d)} / x^{(d)}$. В этом случае $\mathbf{p}^{(d)} = (p^{1(d)},..., p^{K(d)})^T$ – квази-случайный столбец из случайного столбца $\mathbf{P}^{(d)}(F)$ серии d. Если среди компонент столбца $\mathbf{p}^{(1)}$ нет

одинаковых, то в пределе, при $d \to +\infty$, будет получен столбец частот $\mathbf{p}^{(\infty)}$, в котором все компоненты, кроме одной, – нулевые. В настоящей работе использовались случайные столбцы частот $\mathbf{P}^{(d)}(F)$ первых 12-ти серий ($d = 1,...,12$).

Пусть $\boldsymbol{\pi} = (\pi_j^i)_L^K = (\boldsymbol{\pi}_1,...,\boldsymbol{\pi}_L)$ – квази-случайная L-профильная матрица, в которой столбцы образуют выборку объема L из генеральной совокупности $\mathbf{P}^{(d)}(F)$. Как и ранее, по этой матрице датчиком случайных чисел создается выборка большого объема из генеральной совокупности $\mathbf{Str}(n, A, \boldsymbol{\pi})$. Эту выборку можно отождествлять с квази-случайной строкой серии d из генеральной совокупности $\mathbf{STR}^{(d)}(n, A, L)$.

3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

В этом разделе приводятся результаты статистических экспериментов по применению статистик стандартных критериев в условиях недостаточного объема выборки. На их основе предложены оригинальные статистические нестандартные эмпирические критерии проверки однородности строк, применимые в условиях нерепрезентативной выборки.

3.1. Результаты статистических экспериментов по применению статистик стандартных критериев в условиях недостаточного объема выборки

- Для оценки неизвестного номера серии d генеральной совокупности случайных строк $\mathbf{STR}^{(d)}(n, A, L)$ можно использовать её средний уровень сохранности буквы $\overline{pl}^{(d)}$ и средний уровень совпадений букв $\overline{cl}^{(d)}$, которые вычисляются согласно формулам (12) и (13). Эти средние уровни достаточно вычислить для генеральной совокупности случайных строк $\mathbf{STR}^{(d)}(n, A, L = 1)$. Проведенные статистические эксперименты показали, что для нуклеотидных строк (K=4) – $\overline{pl}^{(1)} \sim 0.45$, $\overline{pl}^{(3)} \sim 0.65$ и $\overline{pl}^{(6)} \sim 0.85$, и для аминокислотных ($K$=20) – $\overline{pl}^{(1)} \sim 0.1$, $\overline{pl}^{(3)} = 0.18$, $\overline{pl}^{(6)} = 0.26$ и $\overline{pl}^{(12)} = 0.40$.
- В общем случае для выявления значимой неоднородности на рассматриваемом тест-периоде L анализируемой строки в алфавите из K букв длины n необходимо иметь тест-экспонент $R_L \geq 3.5 \cdot K$.
- Для достижения уровня значимости $\sim 10^{-6}$ в условиях недостаточного объема выборки ($R_L \sim 5K$) гипотезу о неоднородности анализируемой строки на тест-периоде L можно принимать, если выполняется условие: $H_{NP}(L) > \Omega_{NP} = 2$ ($H_P(L) > \Omega_P = 1.5$ или $H_{IC}(L) > \Omega_{IC} = 1.5$).
- Статистики стандартных критериев проверки однородности анализируемой строки в условиях недостаточного объема выборки имеют различную чувствительность. Рисунок 4 на примере модельных строк с частотами букв близкими к равновероятным (рис. 4(а,г)) иллюстрирует выявленную общую закономерность: *NP* статистика наиболее чувствительна и *IC* статистика наименее чувствительна к неоднородности анализируемых строк, *P* статистика занимает некоторое промежуточное положение. Поскольку *H*-спектры для стандартных *NP*, *P* и *IC* критериев аналогичны, на рис. 4(б,д) приведены только H_{NP}-спектры для модельных квази-случайных строк с λ-профильными матрицами: 50-профильной матрицей для нуклеотидной строки серии d=1 и 10-профильной матрицей для аминокислотной строки серии d=6. Объем выборки R_λ является недостаточным в обоих случаях. На графиках нижнего ряда рис. 4(в,е) показано различие значений

соответствующих H_S амплитуд S критериев ($S \in \{NP, P, IC\}$) только на обертонах основного тест-периода $L = \lambda$.

- В условиях недостаточного объема выборки, при заданном теоретическом уровне значимости α, практическая ошибка I-го рода у *NP* критерия больше, чем у *P* и *IC* критериев. При этом ошибка II-го рода β, определяющая мощность (чувствительность) $(1-\beta)$ критерия, у *NP* критерия меньше, чем у *P* и *IC* критериев. Следовательно, чувствительность *NP* критерия выше. В дальнейшем совместное использование статистик *NP* и *IC* критериев в *NEP* критерии позволит сохранить преимущества и компенсировать недостатки каждой из указанных статистик.
- В условиях недостаточного объема выборки можно предполагать *некоррелированность* статистик *NP* и *IC* критериев. Это свойство позволяет значительно сократить количество проводимых испытаний в предлагаемом далее *NEP* критерии проверки однородности строк, использующем в своей процедуре метод Монте-Карло.

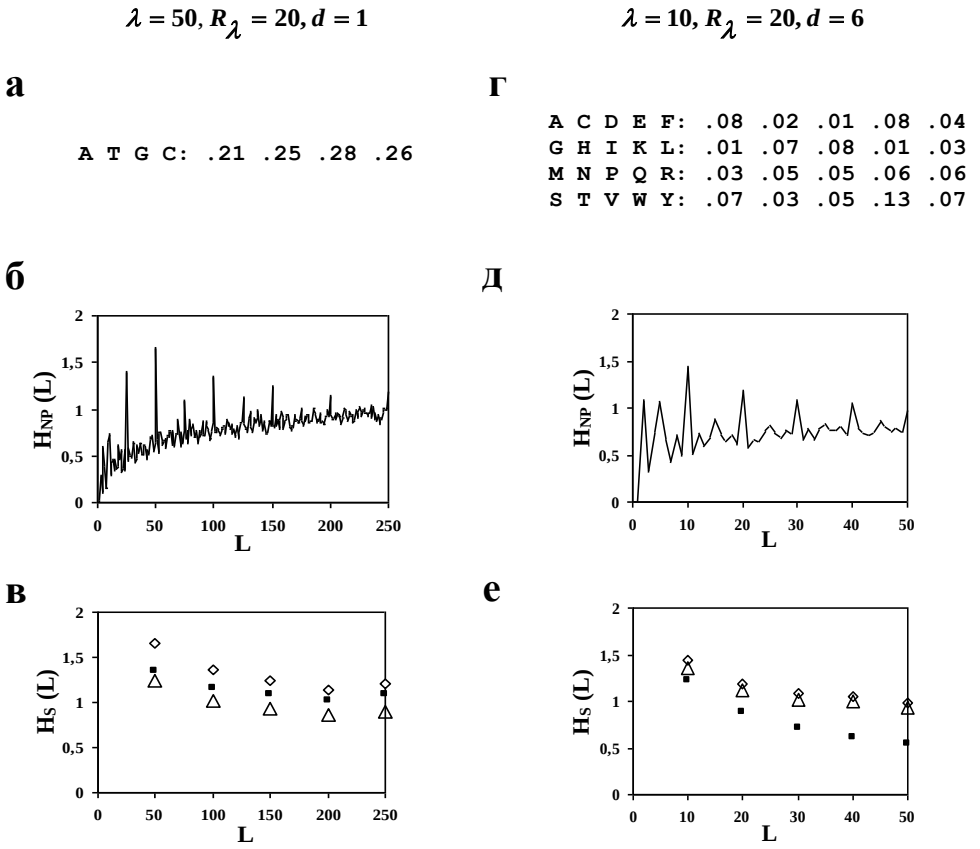


Рис. 4. Сравнение чувствительности статистик стандартных критериев проверки однородности (*NP*,*P*,*IC*) на модельных строках с λ-профильными матрицами в условиях недостаточного объема выборки ($R_\lambda = 20$). (**а**,**б**,**в**) Модельная нуклеотидная строка с 50-профильной матрицей серии *d*=1. (**г**,**д**,**е**) Модельная аминокислотная строка с 10-профильной матрицей серии *d*=6. (**а**,**г**) Средние частоты встречаемости букв в модельных последовательностях. (**б**,**д**) *H*-спектры *NP* критерия (H_{NP}-спектры) анализируемых строк. (**в**,**е**) Значения H_S амплитуд соответствующих *S* критериев ($S \in \{NP, P, IC\}$) на обертонах основного тест-периода $L = \lambda$ анализируемых строк. ◊ – *NP* критерий, △ – *P* критерий, ▪ – *IC* критерий.

3.2. Нестандартные эмпирические поли-критерии проверки однородности строк

Можно сказать, что применяемые ранее эмпирические критерии проверки однородности строк [25,26,28], фактически, использовали только результаты предварительных статистических экспериментов для одной стандартной статистики. Такие критерии можно назвать одноэтапными моно-критериями. Можно предложить двухэтапные моно-критерии, которые, наряду с результатами предварительных экспериментов, на втором этапе используют метод Монте-Карло, встроенный в процедуру критерия. При этом существенно повышается мощность критерия. Такие моно-критерии на втором этапе должны использовать большое ($\sim 10^6$) количество испытаний для достижения уровня значимости не более 10^{-6} Совместно используя некоррелирующие статистики *NP* и *IC* критериев, количество испытаний можно значительно сократить (до 10^3).

Двухэтапный би-критерий, совместно использующий статистики *NP* и *IC* критериев, был назван *NEP* критерием (Нормализованным Эмпирическим χ^2-критерием Пирсона), поскольку в своей процедуре он опирается на высокую чувствительность статистики *NP* критерия в условиях недостаточного объёма выборки. Ниже описана пошаговая процедура *NEP* критерия.

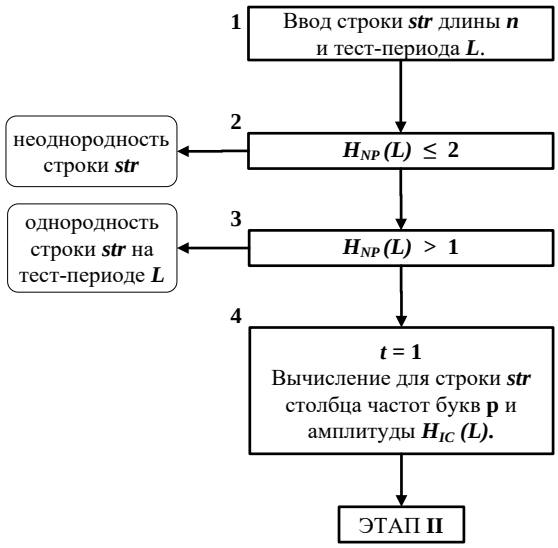


Рис. 5. Блок-схема I-го этапа алгоритма *NEP* критерия для анализируемой строки *str* . Арабские цифры последовательно нумеруют шаги алгоритма.

Этап 1. Блок-схема первого этапа алгоритма *NEP* критерия приведена на рисунке 5.

Шаг 1. Ввод анализируемой строки *str* длины *n* и тест-периода *L*.

Шаг 2. Если для строки *str* значение относительной χ^2-амплитуды $H_{NP}(L)$ на уровне значимости $\alpha = 10^{-7}$ удовлетворяет условию: $H_{NP}(L) > \Omega_{NP} = 2$, то для строки принимается гипотеза о её неоднородности на тест-периоде *L*.

Шаг 3. Если $H_{NP}(L) \le \omega_{NP}$, то принимается гипотеза об однородности строки на тест-периоде *L*. В настоящей работе используется значение $\omega_{NP} = 1$.

Шаг 4. При $\omega_{NP} < H_{NP} \le \Omega_{NP}$ для проверки гипотезы об однородности строки *str* на тест периоде *L* необходимо проводить не менее $T = 10^3$ испытаний модельной случайной

строки $\mathbf{Str}(n, A, \mathbf{p})$ с тем же столбцом $\mathbf{p}$ частот букв, что и в анализируемой строке str. Поэтому здесь выполняется подготовка к проведению дополнительных испытаний: вычисление относительной χ^2-амплитуды $H_{IC}(L)$ и столбца частот букв $\mathbf{p}$ для строки str и присвоение номеру t счетчика количества испытаний начального значения $t = 1$. Затем переходят к следующему этапу.

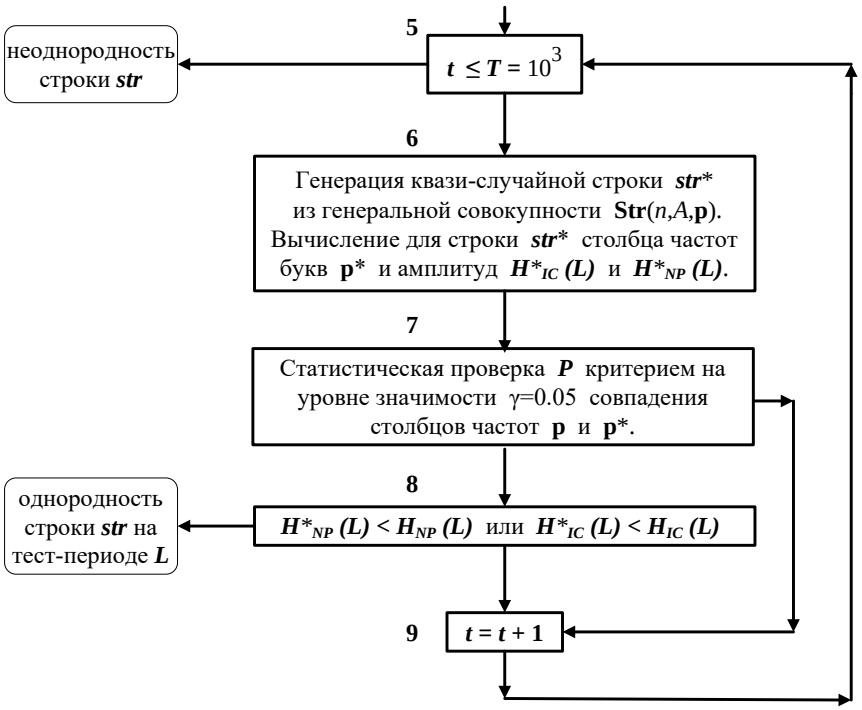


Рис. 6. Блок-схема II-го этапа алгоритма *NEP* критерия для анализируемой строки *str*. Арабские цифры нумеруют шаги алгоритма. **Str**(*n*,*A*,**p**) – случайная строка длины *n* (в алфавите *A*) со столбцом **p** частот букв строки *str* (см. рис. 5, *Шаг 4*).

Этап 2. Блок-схема второго этапа алгоритма *NEP* критерия приведена на рисунке 6.

Шаг 5. Проверка выполнения условия: $t \le T = 10^3$ (проверка исполнения заданного количества T дополнительных испытаний); если $t > T$, то принимается гипотеза о неоднородности строки str на тест-периоде L и алгоритм завершает свою работу.

Шаг 6. Генерация квази-случайной строки $str*$ из генеральной совокупности $\mathbf{Str}(n, A, \mathbf{p})$. Вычисление для строки $str*$ столбца частот $\mathbf{p}*$ и относительных χ^2-амплитуд $H*_{NP}(L)$ и $H*_{IC}(L)$ на уровне значимости $\alpha = 10^{-7}$.

Шаг 7. Далее с помощью P критерия на уровне значимости $\gamma = 0.05$ проверяется совпадение столбцов частот $\mathbf{p}*$ и $\mathbf{p}$. При значимом отличии переходят к Шагу 9.

Шаг 8. Для строки $str*$ проверяется совместное выполнение условий: $H*_{NP}(L) < H_{NP}(L)$ и $H*_{IC}(L) < H_{IC}(L)$. Если оба эти условия не выполняются, то для строки str принимается гипотеза об однородности на тест-периоде L.

Шаг 9. Увеличение номера счетчика t количества испытаний на единицу с последующим переходом к Шагу 5, т.е. переход к следующему дополнительному испытанию.

Таким образом, если строка *str* выдержит все $T = 10^3$ испытаний, то для неё принимается гипотеза о неоднородности на тест-периоде L. В противном случае, принимается гипотеза об однородности на тест-периоде L.

Для дальнейшего снижения количества дополнительных испытаний до $T = 10^2$ (и менее) можно предложить 2-х этапные поли-критерии, например, 3-критерий, использующий на втором этапе спектры H_{NP}, H_{IC} и pl (или cl).

4. ВЫВОДЫ

- Исследование баз данных TRDB и Genbank показало, что при низких процентах вставок и делеций размер паттерна, полученый методом TRF, как правило, совпадает с его оценкой, полученной с помощью эмпирического критерия совместно с анализом pl-спектра рассматриваемой последовательности. Тем не менее, как было показано выше, возможны и несовпадения, когда выравнивание на основе статистической оценки размера паттерна является более оптимальным.

Поэтому в процессе поиска тандемного повтора следует рассматривать тест-периоды, на которых выявлена значимая неоднородность анализируемой последовательности. Кроме того, выбор среди таких тест-периодов должен быть ориентирован на максимальное значение pl-спектра последовательности с учетом оптимальности получаемого консенсус-паттерна.

- Предлагаемые 2-х этапные поли-критерии могут предоставлять полезную предварительную информацию для комбинаторных методов и методов динамического программирования при поиске размытых тандемных повторов, поскольку они эффективны при оценке размера паттерна периодичности. Эти поли-критерии достоверны в часто встречающихся практических условиях недостаточного объема выборки. Обладая высокой чувствительностью (мощностью), они позволяют добиться ошибки I-го рода не более 10^{-6} на каждом тест-периоде.
- Статистические эксперименты на модельных нуклеотидных и аминокислотных последовательностях показали, что 2-х этапные поли-критерии наиболее эффективны при выявлении скрытых микро- и миниперио дов ($L = 2,...,30$). Например, в этой области тест-периодов для умеренных повторов ($R_L \le 30$) на нуклеотидных последовательностях в случае применения *NEP* критерия выигрыш составляет более 70% и, в частности, для микропериодов ($L = 2,...,6$) – более 80%, если сравнивать с нестандартными эмпирическими 2-х этапными моно-критериями, основанными только на статистиках *IC* или *P* критериев. Эффективность в этих областях важна для поиска потенциальных районов тандемных повторов, связанных с возможными генетическими заболеваниями или с древними полиморфными районами геномов.

Работа была выполнена при частичной поддержке грантов № 06-07-89274, № 06-01-08039 Российского Фонда Фундаментальных Исследований.

СПИСОК ЛИТЕРАТУРЫ

1 Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucl. Acids Res.* 1999. **27**. 573-580.

2 Benson G. A new distance measure for comparing sequence profiles based on path length along an entropy surface. *Bioinformatics.* 2002. **18**. S44-S53.

3 Kolpakov R., Bana G., Kucherov G. mreps: efficient and flexible detection of tandem repeats in DNA. *Nucl. Acids Res.* 2003. **31**. 3672-3678.
4 Boeva V., Regnier M., Papatsenko D., Makeev V. Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. *Bioinformatics*. 2006. **22**. 676-684.
5 Krishnan A., Tang F. Exhaustive whole-genome tandem repeats search. *Bioinformatics*. 2004. **20**. 202702-2710.
6 Collins J.R., Stephens R.M., Gold B., Long B., Dean M., Burt S.K. An exhaustive DNA micro-satellite map of the human genome using high performance computing. *Genomics*. 2003. **82**. 10-19.
7 Denoeud F., Vergnaud G. Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains: a web-based resource. *BMC Bioinformatics*. 2004. **5.** 4.
8 Le Fleche P., Hauck Y., Onteniente L., Prieur A., Denoeud F., Ramisse V., Sylvestre P., Benson G., Ramisse F., Vergnaud G. A tandem repeats database for bacterial genomes: application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*. *BMC Microbiol*. 2001. **1.** 2.
9 Naslund K., Saetre P., von Salome J., Bergstrom T.F., Jareborg N., Jazin E. Genome-wide prediction of human VNTRs. *Genomics*. 2005. **85**. 24-35.
10 Boby T., Patch A.M., Aves S.J. TRbase: a database relating tandem repeats to disease genes for the human genome. *Bioinformatics*. 2005. **21**. 811-816.
11 Missirlis P.I., Mead C.L., Butland S.L., Ouellette B.F., Devon R.S., Leavitt B.R., Holt R.A. Satellog: a database for the identification and prioritization of satellite repeats in disease association studies. *BMC Bioinformatics*. 2005. **6**. 145.
12 P. Siwach, S.D. Pophaly, Ganesh S. Genomic and evolutionary insights into genes encoding proteins with single amino acid repeats. *Mol. Biol. Evol.* 2006. **23**. 1357-1369.
13 Katti M.V., Sami-Subbu R., Ranjekar P.K., Gupta V.S. Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. *Protein Sci*. 2000. **9**. 1203-1209.
14 Tompa P. Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays*. 2003. **25.** 847–855.
15 Kalita M.K., Ramasamy G., Duraisamy S., Chauhan V.S., Gupta D. ProtRepeatsDB: a database of amino acid repeats in genomes. *BMC Bioinformatics.* 2006. **7**. 336.
16 Turutina V.P., Laskin A.A., Kudryashov N.A., Skryabin K.G., Korotkov E.V. Identification of amino acid latent periodicity within 94 protein families. *J. Comput. Biol.* 2006. **13**. 946-964.
17 Silverman B.D., Linsker R. A measure of DNA periodicity. *J. Theor. Biol.* 1986. **118**. 295-300.
18 Sharma D., Issac B., Raghava G.P., Ramaswamy R. Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation. *Bioinformatics*. 2004. **20**. 1405-1412.
19 Marple S.L. *Digital Spectral Analysis with Applications*. Baltimore. Prentice-Hall. 1987.
20 Altaiski M., Mornev O., Polozov R. Wavelet analysis of DNA sequences. *Genet. Anal*. 1996. **12**. 165-168.
21 Dodin G., Vandergheynst P., Levoir P., Cordier C., Marcourt L. Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences. *J Theor Biol*. 2000. **206**. 323-326.
22 Landau G., Schmidt J., Sokol D. An algorithm for approximate tandem repeats. *J. Comp. Biol*. 2001. **8**. 1-18.

23 Castello A.T., Martins W., Gao G.R. TROLL – tandem repeat occurrence locator. *Bioinformatics*. 2002. **18**. 634-636.
24 Hauth A.M., Joseph D.A. Beyond tandem repeats: complex pattern structures and distant regions of similarity. *Bioinformatics*. 2002. **18**. 31-37.
25 Shulman M.J., Steinberg C.M., Westmoreland N. The coding function of nucleotide sequences can be discerned by statistical analysis. *J. Theor. Biol*. 1981. **88**. 409-420.
26 Korotkov E.V., Korotkova M.A., Kudryashov N.A. Information decomposition method to analyze symbolical sequences. *Phys. Lett. A*. 2003. **312**. 198-210.
27 Korotkova M.A., Korotkov E.V., Rudenko V.M. Latent periodicity in protein sequences. *J. Mol. Model*. 1999. **5**. 103-115.
28 Gatherer D., McEwan N. Analysis of sequence periodicity in *E. coli* proteins. *J. Mol. Evol.* 2003. **57**. 149-158.
29 Shelenkov A., Skryabin K., Korotkov E. Search and classification of potential minisatellite sequences from bacterial genomes. *DNA Res*. 2006. **13**. 89-102.
30 Li W. The study of correlation structures of DNA sequences: a critical review. *Computers Chem*. 1997. **21**. 257-271.
31 Cramer H. *Mathematical methods of statistics*. Stockholm. 1946.
32 Kullback S. *Information theory and statistics*. Dover Publications. 1968.
33 Chaley M.B., Korotkov E.V., Skryabin K.G. Method revealing latent periodicity of the nucleotide sequences modified for a case of small samples. *DNA Res*. 1999. **6**. 153-163.
34 Gribskov M., Lüthy R., Eisenberg D. Profile analysis. *Meth. Enzymol*. 1990. **183**. 146-159.

Материал поступил в редакцию 30.03.2007, опубликован 22.04.2007.