Статистические методы выявления паттернов скрытой периодичности биологических последовательностей в условиях недостаточного объема выборки

Чалей М. Б., Назипова Н. Н., Кутыркин В. А.
maramaria@yandex.ru
Пущино, Институт Математических Проблем Биологии РАН,
Московский Государственный Технический Университет им. Н. Э. Баумана

В настоящее время выявление скрытой периодичности в биологических последовательностях основано на понятии размытого тандемного повтора [1, 2], представленного линейным списком поврежденных копий исходного текстового фрагмента (паттерна). Внутренние повреждения копий паттерна могут быть обусловлены не только заменами его отдельных символов, но и вставками или выпадением букв. Тандемные (т. е. идущие один за другим без перерывов) повторы играют существенную роль в процессах регуляции, функционирования и структурирования геномной ДНК, связаны с рядом наследственных болезней.

Для выявления тандемных повторов широко применяются комбинаторные методы, в том числе и методы динамического программирования [1, 2, 3]. Альтернативные методы выявления скрытой периодичности используют статистические критерии проверки однородности строк [4, 5] и различные методы спектрального анализа [6].

Как показано в настоящей работе, комбинаторные методы не всегда оптимально выявляют паттерн периодичности тандемного повтора [7]. Альтернативные методы, строго говоря, фиксируют неоднородности в биологических последовательностях. Но одна только фиксация неоднородностей еще не означает, что последовательность является размытым тандемным повтором. Кроме того, на практике, для достоверной фиксации неоднородности объем статистических данных, как правило, оказывается недостаточным. В результате возможны не только значительные погрешности при оценке размера паттерна периодичности, но и ошибочное признание наличия в последовательности размытого тандемного повтора.

В работе для выявления скрытой периодичности нестандартным образом (подробнее см. [7]) используются статистики стандартных критериев проверки однородности текстовых строк: критерий Пирсона (P-критерий), нормализованный критерий Пирсона (NP-критерий) и информационный критерий (IC-критерий [8]). Для текстовой строки длины n, записанной в алфавите из K букв, анализируемой на тест-периоде L (предварительная оценка длины паттерна периодичности), эти

стандартные статистики имеют вид:

$$\nu_P = R_L \sum_{j=1}^{L} \sum_{i=1}^{K} (\pi_j^i - p^i)^2 / p^i;$$

$$\nu_{NP} = R_L \sum_{j=1}^{L} \sum_{i=1}^{K} (\pi_j^i - p^i)^2 / p^i (1 - p^i);$$

$$\nu_{IC} = 2R_L \sum_{j=1}^{L} \sum_{i=1}^{K} \big(\pi_j^i \ln(\pi_j^i) - p^i \ln(p^i)\big); \qquad (1)$$

где $R_L = \frac{n}{L}$ — число копий тест-периода, π_j^i — частота встречаемости i-й буквы алфавита в j-й позиции тест-периода, $p^i = \frac{1}{L} \sum_{j=1}^{L} \pi_j^i$ — оценка частоты встречаемости i-й буквы алфавита во всем тексте.

Следует отметить, что статистика ν_{NP} ранее не использовалась для выявления скрытой периодичности в биологических последовательностях, а статистика ν_{IC} представлена здесь в виде, отличном от того, как она была введена в работе [8] и использована в работе [4]. Формула (1) показывает, что IC-статистика суммирует по позициям тест-периода отклонения энтропии реального распределения букв алфавита в каждой позиции тест-периода от ожидаемой энтропии.

В настоящей работе при недостатке объема статистического материала проблема достоверного выявления неоднородности в биологических последовательностях решается на основе модели дополнительных статистических экспериментов, использующих метод Монте-Карло. Эта модель адекватно описывает проявление неоднородностей при использовании статистических критериев проверки однородности текстовых строк. На основе этой модели предлагаются нестандартные двухэтапные поликритерии, удобные для предварительного автоматизированного выявления скрытой периодичности в условиях недостаточного объема выборки [7]. На первом этапе этих критериев используются результаты большого числа предварительных статистических экспериментов. В процедуре второго этапа используется метод Монте-Карло, основанный на дополнительных статистических экспериментах. Объем дополнительных экспериментов сокращается за счет совместного использования статистик различных критериев для проверки однородности текстовых строк. Предлагаемые в работе нестандартные поли-критерии обеспечивают уровень значимости выявляемой неоднородности порядка 10^{-6}. Введение второго этапа существенно повышает мощность поли-критерия.

Во многих случаях применение разработанных поли-критериев позволило более оптимально оценить размер и состав паттерна периодич-

ности для известных тандемных повторов из базы данных TRDB (`http://tandem.bu.edu/cgi-bin/trdb/trdb.exe`). Наиболее характерные примеры такой переоценки паттерна приведены в работе [7].

Работа выполнена при поддержке РФФИ, проекты № 06-07-89274, № 06-01-08039.

Литература

[1] *Benson G.* Tandem repeats finder: a program to analyze DNA sequences // Nucl. Acids Res. — 1999. — V. 27. — P. 573–580.

[2] *Kolpakov R., Bana G., Kucherov G.* mreps: efficient and flexible detection of tandem repeats in DNA // Nucl. Acids Res. — 2003. — V. 31. — P. 3672–3678.

[3] *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. — V. 22. — P. 676–684.

[4] *Korotkov E. V., Korotkova M. A., Kudryashov N. A.* Information decomposition method to analyze symbolical sequences // Phys. Lett. A. — 2003. — V. 312. — P. 198–210.

[5] *Gatherer D., McEwan N.* Analysis of sequence periodicity in E. coli proteins // J. Mol. Evol. — 2003. — V. 57. — P. 149–158.

[6] *Li W.* The study of correlation structures of DNA sequences: a critical review // Computers Chem. — 1997. — V. 21. — P. 257–271.

[7] *Чалей М. Б., Назипова Н. Н., Кутыркин В. А.* Совместное использование различных критериев проверки однородности для выявления скрытой периодичности в биологических последовательностях // Мат. биол. и биоинф. (эл.журнал) — 2007. — Т. 2. — С. 20–35. — `www.matbio.org/downloads/Chaley2007(2_20).pdf`

[8] *Кульбак С.* Теория информации и статистика // М.: Наука. — 1967.