

УДК 577.2.519.23

СТРУКТУРА БЕЛКОВ И СКРЫТАЯ ПЕРИОДИЧНОСТЬ В ГЕНАХ

М.Б. Чалей¹, В.А. Кутыркин²

(¹лаборатория биоинформатики Института математических проблем биологии РАН; e-mail: maramaria@yandex.ru; ²кафедра вычислительной математики и математической физики Московского государственного технического университета имени Н.Э. Баумана; e-mail: vkutyarkin@yandex.ru)

Оригинальные методы спектрально-статистического подхода были разработаны для распознавания нового типа скрытой периодичности в ДНК — скрытой профильной периодичности (скрытой профильности). Исследование скрытой профильности в кодирующих районах генов позволяет выявлять различные уровни кодирования генетической информации, в том числе и на локальных участках ДНК.

Ключевые слова: скрытая профильная периодичность, структура гена, структура белка.

Понятие скрытой профильности в последовательностях ДНК было введено в работе [1]. Известные методы, применявшиеся для распознавания размытых тандемных повторов, не могут быть использованы для выявления этого нового типа скрытой периодичности. В настоящей работе предлагаются методы распознавания скрытой профильной периодичности, основанные на спектрально-статистическом подходе (2С подходе) [2]. В работе приведены примеры, когда скрытая профильность, выявленная в кодирующих районах ДНК, транслируется в особенности структуры аминокислотной последовательности белка. Непосредственное выявление таких особенностей является достаточно сложной задачей, поскольку цель поиска неизвестна априори.

Методы

Пусть $Chr(\mathbf{p})$ — случайная буква со столбцом частот $\mathbf{p} = (p^1, \dots, p^K)^T$, которая является случайной величиной, принимающей с вероятностью p^i значение i -й буквы алфавита $A = \langle a_1, \dots, a_K \rangle$. Специальная случайная строка $Str = Str_n(\pi) = Chr(\mathbf{p}_1) \dots Chr(\mathbf{p}_n)$ из n независимых случайных букв индуцируется матрицей $\pi = (\mathbf{p}_1, \dots, \mathbf{p}_n) = (\pi_j^i)_n^K$, называемой n -профильной матрицей. Любое целое число L из диапазона $1, \dots, L_{\max}$, где $L_{\max} : \frac{n}{5K}$ называется тест-периодом строки Str . Букву $a_i \in A$ можно отождествить со случайной буквой, все компоненты столбца частот которой — нулевые, за исключением i -й единичной компоненты. Поэтому любую текстовую строку в алфавите A можно отожде-

ствлять с соответствующей специальной случайной строкой той же длины.

Пусть L — тест-период, $0 \leq M < L$ и $Str = Str_n(\pi) = Str_L(\pi_1) \dots Str_L(\pi_m) Str_M(\pi_{m+1})$ — разложение строки Str на подстроки длины L . Тогда, если $M = 0$, матрица $\Pi_{Str}(L) = \frac{1}{m} \sum_{i=1}^m \pi_i$ называется L -профильной

матрицей строки Str . Если $M \neq 0$, то в матрицу $\Pi_{Str}(L)$ вносятся соответствующие поправки. Таким образом, для строки Str введен профильно-матричный спектр Π_{Str} , определенный на каждом тест-периоде. Если $\pi_1 = \dots = \pi_m = \pi_0$ и $\pi_0 = (\pi_{m+1}, \pi_0)$, то строка Str называется L -профильной строкой со случайным паттерном периодичности $Str_L(\pi_0)$. В этом случае для строки Str используется обозначение $Tdm_L(\pi_0, n)$ и матрица π_0 называется ее главной профильной матрицей, поскольку она индуцирует весь профильно-матричный спектр этой строки.

Пусть str — текстовая строка длины n в алфавите A , анализируемая на наличие скрытой профильности. Для оценки ее предполагаемого скрытого периода используется характеристический спектр H , принимающий на тест-периоде λ значение:

$$H(\lambda) = \psi(\Pi_{str}(\lambda), \Pi_{Tdm_\lambda}(\lambda), n) - M(\chi_{(K-1)\lambda}^2), \quad (1)$$

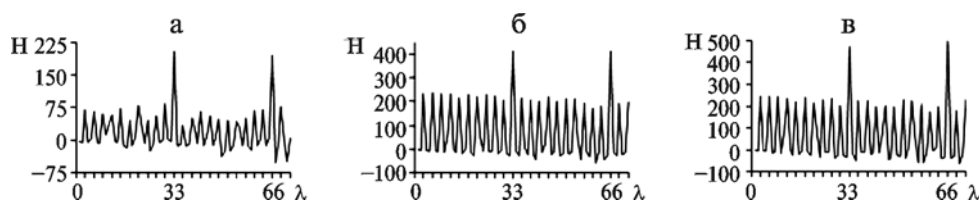


Рис. 1. Характеристические спектры кодирующих районов трех генов из семейства аполипопротеинов: а — APOE рыбы *Danio rerio* (GenBank AJ236882; 1-846 н.о.); б — APOA1 утки *Anas platyrhynchos* (GenBank U86131; 1-795 н.о.); в — APOA4 мыши *Mus musculus* (GenBank M64248; 1-1176 н.о.)

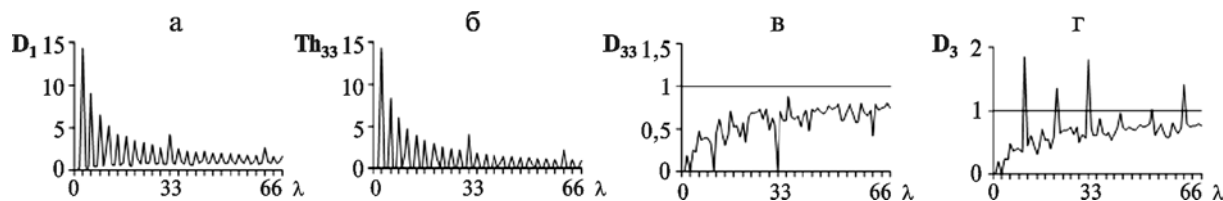


Рис. 2. Распознавание скрытой 33-профильности в кодирующем районе гена APOA1 утки *A. platyrhynchos* (GenBank U86131; 1-795 н.о.): а, в, г — спектры отклонения от 1-, 33-, 3-профильности (см. формулу (3)) соответственно; б — теоретическая реконструкция (см. формулу (4)) спектра D_1 в предположении скрытой 33-профильности гена APOA1

где далее $Tdm_\Lambda = Tdm_\Lambda(\Pi_{str}(\Lambda), n)$, здесь $\Lambda = 1$, $M(\chi^2_N)$ — математическое ожидание χ^2 — распределения с N степенями свободы и если $\Pi_{str}(\lambda) = (\pi^*_{ij})^K_\lambda$ и $\Pi_{Tdm_\Lambda}(\lambda) = (\pi^i_j)^K_\lambda$, то

$$\psi(\Pi_{str}(\lambda), \Pi_{Tdm_\Lambda}(\lambda), n) = \frac{n}{\lambda} \sum_{j=1}^{\lambda} \sum_{i=1}^K (\pi^*_{ij} - \pi^i_j)^2 / \pi^i_j \sim \chi^2_{(K-1)\lambda}, \quad (2)$$

Первый тест-период L с ярко выраженным максимальным значением спектра H служит оценкой скрытого периода в строке str (рис. 1). Он признается периодом скрытой профильности, если строка str статистически неотличима от L -профильной строки Tdm_L . Для этого используется статистика ψ (см. формулу (2)) и спектр D_L отклонения строки str от L -профильности:

$$D_L(\lambda) = \psi(\Pi_{str}(\lambda), \Pi_{Tdm_L}(\lambda), n) / \chi^2_{crit}((K-1)\lambda, \alpha), \quad (3)$$

где $\chi^2_{crit}(N, \alpha)$ — критическое значение χ^2_N -распределения на уровне значимости $\alpha = 0,05$.

Если $L_{max} \sim 100$, то гипотеза о неотличимости принимается, если на каждом тест-периоде $D_L < 1$. Паттерн строки Tdm_L служит оценкой паттерна скрытой L -профильности, если спектр D_1 отклонения от 1-профильности (однородности) строки str (см. формулы (2) и (3)) повторяет спектр Th_L его теоретической реконструкции:

$$Th_L(\lambda) = \psi(\Pi_{Tdm_L}(\lambda), \Pi_{Tdm_1}(\lambda), n) / \chi^2_{crit}((K-1)\lambda, \alpha), \quad (4)$$

Результаты и обсуждение

В характеристических спектрах кодирующих районов наблюдается регулярность пиков в 3 нуклеотидных остатках (н.о.) (см. формулу (1), рис. 1). Таким образом проявляется первый уровень организации кодирования, обусловленный генетическим триплетным кодом. Такой феномен называется регулярной неоднородностью района, если в этом районе от-

сутствует скрытая 3-профильность. При наличии регулярной неоднородности, скрытая профильность выявляет второй уровень организации кодирования. На этот уровень организации указывают ярко выраженные максимумы характеристического спектра района. В качестве примера на рис. 1 приведены характеристические спектры кодирующих районов 3 (из более 150) белков семейства аполипопротеидов [3].

Дальнейшее применение методов 2С подхода (рис. 2) выявляет скрытую 33-профильность кодирующих районов генов аполипопротеидов. Несмотря на отсутствие видимой гомологии в белках семейства, они обладают общей пространственной структурой (содержат повторы в 22 аминокислоты, каждый из которых образует пару альфа спиралей), коррелирующей со скрытой профильной периодичностью их генов.

В кодирующих районах возможно проявление локальной двухуровневой организации кодирования. В этом случае во всем кодирующем районе наблюдается всего лишь регулярная неоднородность в 3 н.о. Скрытая профильность фиксируется только для отдельных локальных областей кодирующего района. Например, в кодирующем районе гена *суа* бактерии *Bordetella pertussis* (GeneBank Y00545, 981-6101 н.о.) выделяются три области с периодом скрытой профильной периодичности в 27 н.о. (рис. 3).

Согласно описанию белкового продукта гена *суа* бактерии *B. pertussis* — бифункциональной гемолизин/аденилат циклазы (Swiss-Prot P15318), именно в этих областях находятся кальцийсвязывающие сайты гемолизинного типа длиной 18 аминокислот, что соответствует 54 (2×2) н.о.

Последний пример показывает возможность применения метода выявления скрытой профильности для поиска различных функциональных сайтов как в ДНК, так и в белках.

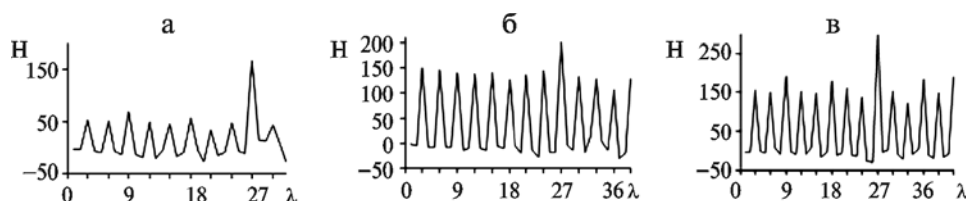


Рис. 3. Характеристические спектры трех локальных областей гена *суа* бактерии *B. pertussis*: а — 4020-4181 н.о.; б — 4443-5036 н.о.; в — 5211-5840 н.о.

Выводы

В работе предложены методы распознавания нового типа скрытой периодичности — скрытой профильности. В кодирующих районах ДНК скры-

тая профильность позволяет различать два уровня организации кодирования генетической информации. Второй уровень может коррелировать с особенностями структуры транслируемых белков.

СПИСОК ЛИТЕРАТУРЫ

1. Chaley M.B., Kutyrkin V.A. Model of Perfect Tandem Repeat with Random Pattern and Empirical Homogeneity Testing Poly-criteria for Latent Periodicity Revelation in Biological Sequences // Mathematical Biosciences. 2008. Vol. 211. N 1. P. 186—204.

2. Chaley M.B., Kutyrkin V.A. Latent Profile Periodicity as New Type of Periodicity in Genome // Dokladi XIV Vse-

rossiiskoi Konferentsii “Matematicheskie Metodi Raspoznaniya Obrazov”. Moskva: Max Press, 2009. P. 614—617 [in Russian].

3. База данных Pfam (URL:<http://pfam.sanger.ac.uk/family/pf01442> 06.04.2010).

Поступила в редакцию
09.04.10

STRUCTURE OF THE PROTEINS AND LATENT PERIODICITY IN THE GENES

M.B. Chaley, V.A. Kutyrkin

Original methods of spectral-statistical approach have been elaborated for recognizing a new type of latent periodicity in DNA, called latent profile periodicity (or latent profility). Search of the latent profility allows revealing different levels of information encoding in the genes, and in local DNA regions included.

Key words: *latent profile periodicity, gene structure, protein structure.*

Сведения об авторах

Чалей Мария Борисовна — канд. биол. наук, ст. науч. сотр. Института математических проблем биологии РАН (ИМПБ РАН). Тел. (4967)73-27-19; e-mail: maramaria@yandex.ru

Кутыркин Владимир Андреевич — канд. физ.-мат. наук, доц. МГТУ им. Н.Э. Баумана. Тел. (495)263-60-18; e-mail: vkutyarkin@yandex.ru