УДК 577.2:519.23

В. А. Кутыркин, М. Б. Чалей

СТРУКТУРНЫЕ РАЗЛИЧИЯ КОДИРУЮЩИХ И НЕКОДИРУЮЩИХ РАЙОНОВ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ДНК ГЕНОМА ЧЕЛОВЕКА

Проведен количественный анализ регулярных структурных свойств кодирующих и некодирующих районов последовательностей генома человека. На его основе предложен эффективный статистический критерий, позволяющий распознавать кодирующие районы последовательности ДНК генома человека. В них выявлена двухуровневая организация кодирования генетической информации.

E-mail: vkutyrkin@yandex.ru, maramaria@yandex.ru

***Ключевые слова*:** *скрытая периодичность, скрытая профильность, спектрально-статистический подход, распознавание кодирующих районов ДНК.*

Современные технологии секвенирования геномов различных организмов позволили представить полимерные молекулы ДНК в виде текстовых строк в алфавите из четырех букв (A, T, G, C), соответствующих четырем типам мономерных звеньев ДНК – четырем нуклеотидам (нукл.). В этих текстовых строках содержится основная информация о наследуемых свойствах организмов. Скорость накопления таких генетических данных значительно обгоняет экспериментальные и теоретические возможности анализа заключенной в них информации. В текстовых последовательностях ДНК находятся так называемые кодирующие районы, транслируемые в соответствующие последовательности белков. Как известно, белки образуют структурно-функциональную основу всех живых организмов. Выявление кодирующих районов является одной из актуальных проблем генетического анализа последовательностей ДНК. Для ее решения используют методы, обладающие различной степенью достоверности. В результате в базах данных накапливается большой объем последовательностей, относимых к потенциально кодирующим районам и ожидающих в дальнейшем подтверждения достоверности. Поэтому разработка новых эффективных методов выявления кодирующих районов является важной практической задачей. Для создания таких методов необходимо решить про-

блему выявления специфических структурных свойств кодирующих районов ДНК, которой посвящена настоящая работа.

Ранее в литературе отмечалась регулярность структурной организации последовательностей кодирующих районов ДНК, которая эпизодически выявлялась в спектрах корреляционных функций, в спектрах Фурье и т. п. [1—3]. Такая эпизодичность не позволила создать достоверные критерии для характеризации кодирующих районов в последовательностях ДНК. Настоящая работа направлена на достоверное выявление в кодирующих районах последовательностей ДНК характерной регулярности, позволяющей отличить их от некодирующих районов. Для этого используется ранее предложенный спектрально-статистический подход [4—6], разработанный для распознавания нового типа скрытой периодичности в ДНК — профильной периодичности (профильности).

Понятие скрытой профильности [7] расширяет известное понятие размытого тандемного повтора [8], которое применялось ранее для распознавания скрытой периодичности в последовательностях ДНК. Совершенный тандемный повтор является текстовой строкой, которая получена последовательными копиями его подстроки, называемой паттерном периодичности. В размытом тандемном повторе незначительное количество (не более 30 %) букв в копиях паттерна искажаются. При наличии в последовательности ДНК скрытой профильной периодичности искажение букв в каждой позиции копий паттерна обусловлено соответствующим вероятностным распределением. Было показано [4, 5], что скрытая профильность в кодирующих районах может коррелировать с особенностями структуры кодируемых белков. Ранее наряду со скрытой профильной периодичностью в кодирующих районах ДНК была отмечена характерная регулярная неоднородность их структуры [5, 6].

В настоящей работе на примере генома человека проведено количественное исследование наличия в кодирующих и некодирующих районах последовательностей ДНК скрытой профильности и регулярной неоднородности. Исследование было выполнено для кодирующих районов генов человека из базы KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/) [9] и интронов (некодирующих районов ДНК) человека из базы EID (http://www.utoledo.edu/med/depts/bioinfo/database) [10].

Методы. Спектрально-статистический подход опирается на модель скрытой профильности в текстовых строках. Как показано в работе [7], понятие размытого тандемного повтора представляет частый случай общего понятия скрытой профильности.

В качестве модели скрытой профильности рассматривается мультиполиномиальная схема испытаний. Эта схема индуцируется L последовательными полиномиальными схемами с четырьмя исходами

(четыре буквы алфавита ДНК). Следовательно, в каждом испытании реализуется случайная буква, которая описывается вероятностным распределением букв алфавита ДНК. В первом испытании этой мультиполиномиальной схемы реализуется первая случайная буква (полиномиальная схема), во втором — вторая, …, в L-м — L-я случайная буква (полиномиальная схема). Затем эта последовательность L испытаний повторяется. На заключительном этапе испытаний может повториться менее L реализаций таких случайных букв (полиномиальных схем). Следовательно, моделью скрытой профильности является совершенный тандемный повтор, паттерн которого представлен случайной строкой из независимых случайных букв. Анализируемая последовательность ДНК рассматривается как реализация такого тандемного повтора. Поэтому скрытая профильная периодичность в текстовой последовательности может быть выявлена только с помощью статистических критериев. При анализе последовательности ДНК в этих критериях используются различные статистики [5, 6], имеющие вид функциональных зависимостей (спектров) от тестируемых периодов последовательности. Тестируемый период (тест-период) — это натуральное число, не превышающее половины длины анализируемой последовательности ДНК.

Для анализируемой последовательности ДНК вычисляются специальные функции с областью определения в диапазоне тест-периодов последовательности и со значениями в области действительных чисел. Такие функции были названы спектрами анализируемой последовательности ДНК. Аналитические формулы для этих спектров приведены в работах [5, 6].

Ранее для оценки периода скрытой периодичности в ДНК был использован так называемый характеристический спектр **H** [5, 6] (рис. 1). Было показано, что в размытых тандемных повторах [11] характеристический спектр имеет первый, ярко выраженный максимум на тестируемом периоде λ (тест-периоде), являющемся периодом этого повтора (рис. 1, *а*). Оказалось, что это свойство характеристического спектра сохраняется и в последовательностях со скрытой профильностью (рис. 1, *б* и *в*).

В характеристических спектрах кодирующего района ДНК, как правило, наблюдается регулярное чередование пиков через два основания (рис. 2, *а*). Такое явление было названо 3-регулярностью характеристического спектра. Как будет показано далее, в характеристических спектрах интронов 3-регулярность обычно не выявляется (см., например, рис. 1, *б* и рис. 2, *б*).

Наличие 3-регулярности характеристических спектров в кодирующих районах обусловлено триплетным законом кодирования аминокислот в белках (аминокислота кодируется тремя последовательными нуклеотидами ДНК).

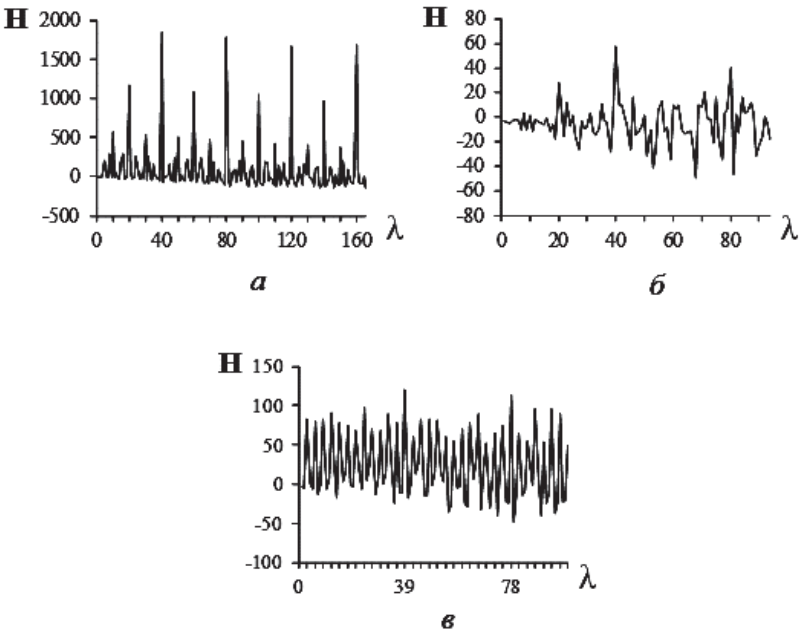

Рис. 1. Характеристические спектры H последовательностей ДНК со скрытой периодичностью:

а — размытый тандемный повтор на хромосоме I человека (TRDB, Indices 1943319 — 1945814, Venter 2007); *б* — интрон гена HMCN2 белка хемицентина (hemicentin 2) человека на хромосоме IX со скрытой профильной периодичностью 40 нукл. (EID, INTRON_36 9864_NT_035014 protein_id:XP_001726994.1, 1 884 нукл.); *в* — кодирующий район поли-А-связывающего цитоплазматического белка человека со скрытой профильной периодичностью 39 нукл. (KEGG, hsa:8761, 1 983 нукл.)

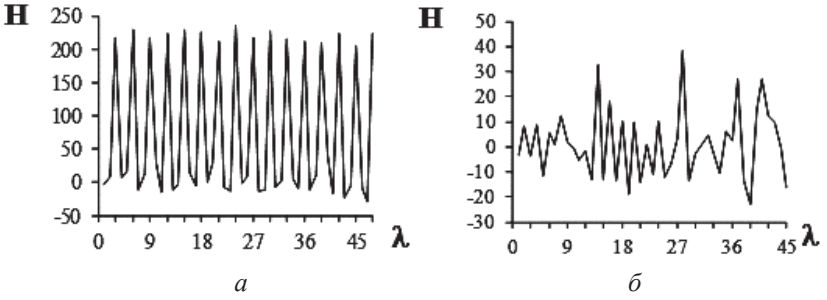

Рис. 2. Характеристические спектры H кодирующих и некодирующих районов последовательностей ДНК:

а — кодирующий район гена трансмембранного белка человека (KEGG, hsa:80757, 960 нукл.); *б* — интрон гена UCHL1 (ubiquitin carboxyl-terminal hydrolase isozyme L1) человека на хромосоме IV (EID, INTRON_4 4383_NT_006238 protein_id: NP_004172.2, 917 нукл.)

В характеристическом спектре кодирующего района может наблюдаться следующая картина (рис. 3). На фоне 3-регулярности проявляется четко выраженный пик (или периодическое повторение выраженных пиков), указывающий на наличие скрытой профильности в этом районе (см., например, рис. 3, *б*). Следовательно, характеристический спектр может проявлять двухуровневую организацию кодирования, когда период скрытой профильности превышает 3 нукл. Первый уровень обусловлен триплетным кодированием аминокислот, второй — скрытой профильностью кодирующего района ДНК.

Характеристический спектр позволяет получить оценку периода скрытой профильной периодичности в анализируемой последовательности ДНК. Первый тест-период, на котором проявляется четко выраженный максимум этого спектра, может служить оценкой периода скрытой профильной периодичности.

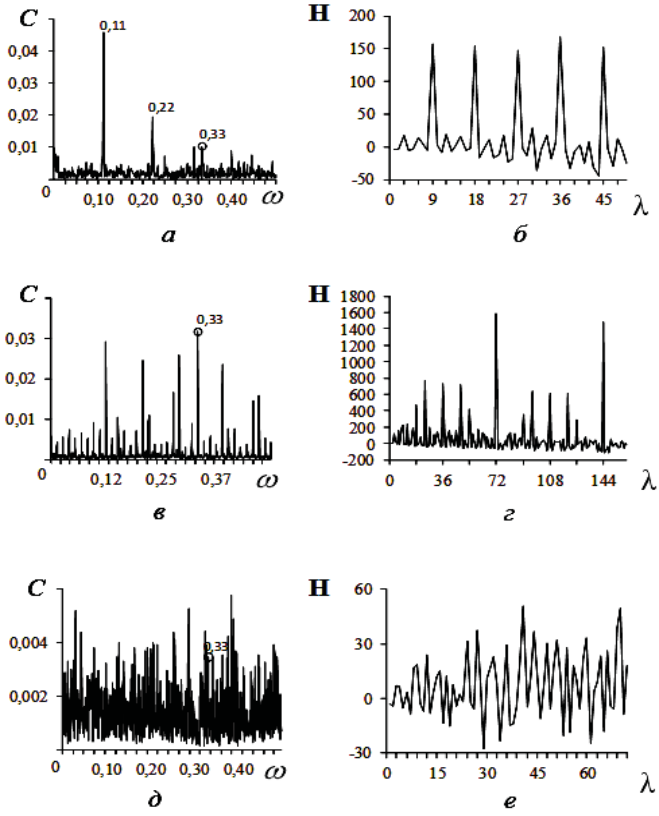

Рис. 3. Спектры Фурье и характеристические спектры H кодирующих районов трех генов человека из базы данных KEGG (*C* — спектральная плоскость; *ω* — частота):

а, *б* — KEGG, hsa:338872, 1 002 нукл., фактор некроза опухоли; *в*, *г* — KEGG, hsa:57055, 1 605 нукл., DAZ протеин 2; *д*, *е* — KEGG, hsa:149998, 1 446 нукл, протеин семейства липаз

Сравнение информативности характеристических спектров и спектров Фурье. Для оценки периода скрытой периодичности традиционно используют спектры Фурье последовательностей ДНК [1—3]. Приведем примеры сравнения характеристических спектров и спектров Фурье в кодирующих районах.

При коротком периоде скрытой периодичности его оценки на основе характеристического спектра и спектра Фурье могут совпадать, что иллюстрирует рис. 3, *а* и *б*. Но в отличие от спектра Фурье, кроме четко выраженных максимумов в характеристическом спектре на рис. 3, *б* наблюдается 3-регулярность, характерная для кодирующих районов. Таким образом, характеристический спектр фиксирует наличие двух уровней организации кодирования. Для кодирующего района (рис. 3, *в* и *г*) гена человека одного из белков семейства DAZ (Deleted in Azoospermia) в спектре Фурье отмечается доминирующий пик на частоте 0,33, соответствующий тест-периоду 3 нукл. В характеристическом спектре этого района наблюдается 3-регулярность (см. рис. 3, *г*) в соответствии с доминирующим пиком спектра Фурье. Период скрытой профильной периодичности в рассматриваемом районе равен 72 нукл., на что однозначно указывают результаты анализа его характеристического спектра. Следует особо отметить, что две трети длины этого района составляет слабо размытый тандемный повтор с паттерном периодичности 72 нукл. Таким образом, в этом случае характеристический спектр выделяет двухуровневую организацию кодирующего района. В спектре Фурье кодирующего района белка из семейства липаз (рис. 3, *д*) пик на частоте 0,33 не является доминирующим. Однако в характеристическом спектре этого района наблюдается 3-регулярность, характерная для кодирующих районов.

Приведенные выше примеры показывают высокую чувствительность характеристического спектра к структурным особенностям последовательности ДНК. В сравнении со спектром Фурье, с точки зрения авторов, характеристический спектр является более информативной характеристикой последовательностей генома. Поэтому в настоящей работе анализ свойств характеристического спектра был положен в основу исследования структурных свойств кодирующих и некодирующих районов (интронов) последовательностей генома человека.

Оценка периода скрытой профильности. Характеристический спектр используют для оценки периода скрытой профильной периодичности (профильности). Получение такой оценки еще не означает существования скрытой периодичности в последовательности ДНК. Для подтверждения гипотезы о периодичности применяют дополнительные методы, выявляющие ее паттерн или указывающие на его существование. Например, при распознавании тандемных повторов принято выводить соответствующий консенсус-паттерн повтора. Для выявления скрытой профильной периодичности сначала проверяют гипотезу о не-

однородности анализируемой последовательности, а затем — гипотезу о статистической неотличимости этой последовательности от профильной строки с соответствующим паттерном периодичности [5, 6].

Статистический критерий неотличимости от L-профильности в анализируемой последовательности приведен в работах [5, 6]. Этот критерий основан на спектре отклонения $\mathbf{D}_L$ от L-профильной периодичности в анализируемой последовательности ДНК, где L — тестируемый период.

Спектр $\mathbf{D}_1$ характеризует отклонение от однородности (1-профильности) в анализируемой последовательности. Если $N_1/L_{\max} < < 0,05$, где $L_{\max} \sim n/(5K)$ ($K = 4$ — размер алфавита последовательности ДНК) и N_1 — число тест-периодов, на которых значения спектра $\mathbf{D}_1 > 1$, то анализируемая последовательность длины n признается однородной.

Пусть последовательность ДНК признана неоднородной и L — оценка ее периода профильной периодичности, полученная из характеристического спектра этой последовательности. Согласно работам [5, 6], если $N_L/L_{\max} < 0,05$, то в анализируемой последовательности длины n признается наличие скрытой L-профильной периодичности. Например, как видно на рис. 4, *а*, кодирующий район гена фактора некроза опухоли является неоднородным.

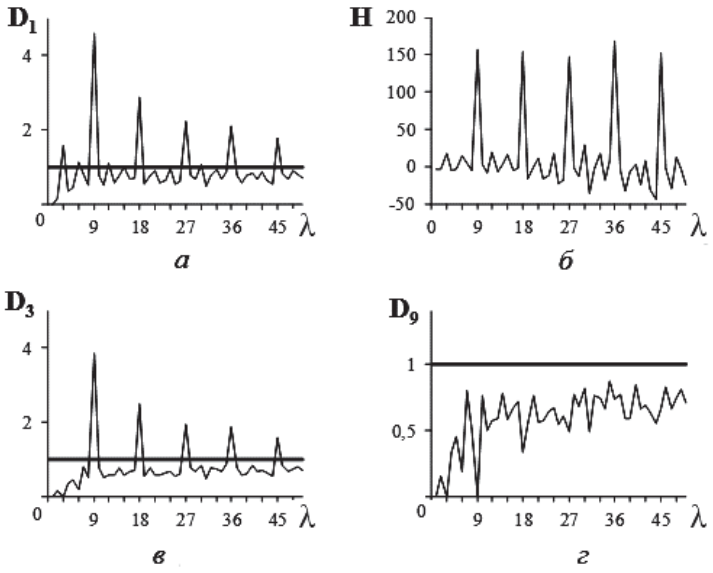

Рис 4. Спектры $\mathbf{D}_L$ отклонения от *L*-профильности при *L* = 1 (*а*), 3 (*в*), 9 (*г*) и характеристический спектр H (*б*) кодирующего района гена фактора некроза опухоли (KEGG, hsa: 338872, 1 002 нукл.)

Из характеристического спектра (рис. 4, *б*) этого района получена оценка периода скрытой профильной периодичности $L = 9$. Гипотеза о наличии скрытой L-профильности в анализируемом районе отвергается, если $L = 3$ (рис. 4, *в*), и принимается, если $L = 9$ (рис. 4, *г*). Отметим, что в этом районе известная программа TRF (Tandem Repeats Finding) [8] не выявила ни одного тандемного повтора. Следовательно, согласно [5, 6], в нем наблюдается всего лишь скрытая 9-профильность.

Критерий наличия 3-регулярности в последовательности ДНК. Разобьем область определения характеристического спектра анализируемого района ДНК на последовательные тройки тест-периодов. Для каждой такой тройки тест-периоду с максимальным значением характеристического спектра ставится в соответствие единица, двум остальным — нули. В результате образуется бинарная строка из нулей и единиц, т. е. текстовая строка *str* в алфавите $A = \langle 0,1 \rangle$ размера $K = 2$. Эта строка сравнивается с совершенной периодической строкой той же длины, паттерн периодичности которой имеет вид 001. Индексом 3-регулярности анализируемой последовательности I_3 называется число, равное отношению количества совпадений компонент совершенной периодической строки и бинарной строки *str* к длине анализируемой последовательности ДНК. Если индекс 3-регулярности $I_3 > 0,7$, то считаем, что в характеристическом спектре наблюдается 3-регулярность. Например, согласно такому критерию, в характеристических спектрах на рис. 2, *а* и рис. 3, *б*, *г*, *е*, соответствующих кодирующим районам, наблюдается 3-регулярность. В характеристическом спектре на рис. 2, *б*, соответствующем последовательности интрона, 3-регулярность не выявляется. Для характеристических спектров на рис. 2, *а* и рис. 3, *б* 3-регулярность очевидна. Наличие 3-регулярности в характеристических спектрах на рис. 3, *г* и рис. 3, *е* подтверждается значениями индексов 3-регулярности $I_3 = 0,87$ и $I_3 = 0,78$ соответственно. Отсутствие 3-регулярности в характеристическом спектре на рис. 2, *б* следует из значения индекса $I_3 = 0,42 < 0,7$ этого спектра.

Результаты и обсуждение. В настоящей работе спектрально-статистический подход [4—6] был использован для выявления структурных различий кодирующих (CDS) и некодирующих районов (интронов) последовательностей ДНК генома человека. Чтобы установить типичные особенности структуры кодирующих районов, из 25 704 CDS базы данных KEGG-54,1 [9] были сформированы две выборки. В первой, специальной, выборке были собраны 17 652 CDS белков с достоверно исследованными физико-химическими свойства-

ми. Во второй выборке были собраны 4 821 CDS, которые соответствуют так называемым гипотетическим (hypothetical) белкам [12]. Последовательности 277 477 интронов человека из базы данных EID [10] составили третью выборку. В эту выборку вошли интроны, длина которых не превышала максимальную длину (26 000 нукл.) анализируемых CDS.

В соответствии со спектрально-статистическим подходом, результаты статистического исследования структурных особенностей перечисленных выше трех выборок последовательностей ДНК представлены в виде соответствующих дендрограмм на рис. 5—7. В вершинах этих дендрограмм показан процентный состав (от общего объема каждой выборки) последовательностей ДНК, обладающих структурными особенностями: однородностью (неоднородностью); 3-регулярностью (отсутствием 3-регулярности); скрытой профильностью (отсутствием скрытой профильности); скрытой 3-профильностью (скрытой профильностью, отличной от 3-профильности); скрытой профильностью, кратной 3 (скрытой профильностью, не кратной 3).

Согласно дендрограмме, представленной на рис. 5, практически все (с учетом статистической погрешности) CDS в специальной выборке являются неоднородными и 3-регулярными. При этом для большинства этих последовательностей (74 %) выявляется скрытая

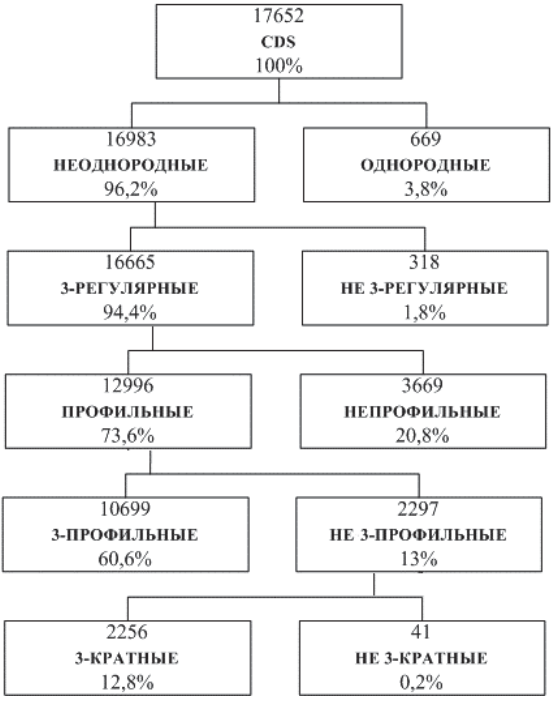

Рис. 5. Дендрограмма разделения специальной выборки 17 652 CDS достоверно исследованных белков из базы KEGG-54,1 на группы в соответствии с выявленными свойствами их последовательностей

профильность. Следует отметить, что среди таких последовательностей выделяется значительная доля (13 %) последовательностей с двухуровневой организацией кодирования, в которых второй уровень организации кодирования обусловлен скрытой профильностью, кратной 3.

Основные количественные показатели дендрограммы специальной выборки (см. рис. 5) кардинально отличаются от соответствующих показателей дендрограмы интронов, приведенной на рис. 6. Так, бóльшая часть (74 %) интронов представлена однородными последовательностями, в то время как практически все CDS в специальной выборке неоднородны. Кроме того, последовательности интронов (с учетом статистической погрешности) не являются 3-регулярными, тогда как практически все CDS в специальной выборке 3-регулярны. Таким образом, кодирующие районы ДНК характеризуются структурным свойством 3-регулярности, что отличает их от последовательностей интронов. Подчеркнем, что наличие 3-регулярности еще не означает существования в последовательности скрытой триплетной периодичности.

Количественный анализ дендрограммы для выборки CDS гипотетических белков (рис. 7) показывает ее промежуточное положение по отношению к дендрограммам CDS в специальной выборке (см. рис. 5) и интронов (см. рис. 6). Такое промежуточное положение наблюдается как для структурного свойства неоднородности последовательности: 25 % (интроны) < 55 % (CDS гипотетических белков) < < 96 % (CDS достоверно исследованых белков), так и для свойства 3-регулярности: 3 % (интроны) < 40 % (CDS гипотетических белков) < < 94 % (CDS достоверно исследованных белков). Такое промежуточное положение обусловлено тем, что среди CDS гипотетических белков значительную часть могли составлять последовательности, в действительности не кодирующие белки. Это предположение нашло подтверждение при анализе новой версии базы KEGG-60 (январь

Рис. 6. Дендрограмма разделения выборки 277 477 интронов из базы EID на группы в соответствии с выявленными свойствами их последовательностей

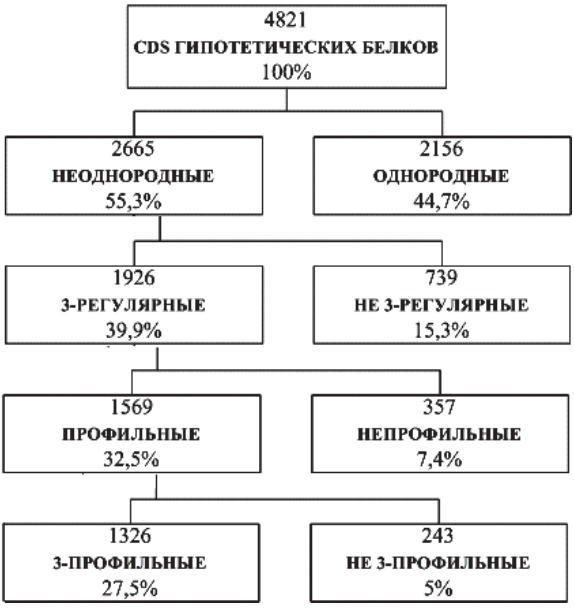

Рис. 7. Дендрограмма разделения выборки 4 821 CDS гипотетических белков из базы KEGG-54,1 на группы в соответствии с выявленными свойствами их последовательностей

2012 г.). В новую версию не вошли многие CDS гипотетических белков, представленные ранее (см. рис. 7) в старой версии KEGG-54,1 (май 2010 г.). Например, более 90 % CDS гипотетических белков из старой версии KEGG, чьи последовательности были признаны однородными, не сохранились в новой версии.

Выводы. Предложенный ранее спектрально-статистический подход [4—6] позволил выработать статистический критерий для распознавания кодирующих последовательностей в геноме человека на основании их характерного свойства 3-регулярности. Этот критерий базируется на пороговом значении индекса 3-регулярности анализируемой последовательности ДНК. Если значение индекса 3-регулярности не менее 0,7, то последовательность признается кодирующей. Количественные исследования генома человека (см. рис. 5 и 6) показали высокую эффективность (надежность и мощность критерия не менее 95 %) распознавания кодирующих районов последовательностей.

Спектрально-статистический подход выявил новый тип скрытой периодичности — скрытую профильность в последовательностях ДНК генома человека. Скрытую профильность, отличную от 3-профильности, следует рассматривать как проявление второго уровня организации кодирования генетической информации. Первый уровень обусловлен триплетным кодированием аминокислот (универсальным генетическим кодом), следствием которого является 3-регулярность характеристического спектра кодирующего района ДНК.

Ранее было показано [4—6], что наличие скрытой профильной периодичности в гене коррелирует со структурно-функциональными свойствами кодируемого белка. Поиск таких свойств является актуальной и сложной биологической задачей. В настоящей работе спектрально-статистический подход позволил выявить наличие скрытой профильной периодичности в большинстве (75 %) кодирующих районов последовательностей ДНК генома человека. Двухуровневая организация кодирования была обнаружена в 13 % CDS генов человека.

СПИСОК ЛИТЕРАТУРЫ

1. Tsonis A. A., Elsner J. B., Tsonis P. A. Periodicity in DNA coding sequences: Implications in gene evolutions // J. Theor. Biol. – 1991. – Vol. 151. – P. 323–331.
2. Fickett J. W., Tung C.-S. Assessment of protein coding measures // Nucleic Acids Res. – 1992. – Vol. 20. – P. 6441–6450.
3. Lobzin V. V., Chechetkin V. R. Order and correlations in genomic DNA sequences. The spectral approach // Physics – Uspekhi. – 2000. – Vol. 43(1). – P. 55–78.
4. Chaley M. B., Kutyrkin V. A. Structure of proteins and latent periodicity in their genes // Moscow Univ. Biol. Sci. Bull. – 2010. –Vol. 65. – P. 133–135.
5. Chaley M., Kutyrkin V. Profile-statistical periodicity of DNA coding regions // DNA Res. – 2011. –Vol. 18. – P. 353–362.
6. Кутыркин В. А., Чалей М. Б. Распознавание различных уровней в организации кодирования генетической информации // Вестн. МГТУ им. Н.Э. Баумана. Сер. Естеств. науки. Спец. вып. Мат. моделирование. – 2011. – С. 200–215.
7. Chaley M., Kutyrkin V. Model of perfect tandem repeat with random pattern and empirical homogeneity testing poly-criteria for latent periodicity revelation in biological sequences // Math. Biosci. – 2008. –Vol. 211. – P. 186–204.
8. Benson G. Tandem repeats finder: a program to analyze DNA sequences // Nucleic Acids Res. – 1999. – Vol. 27. – P. 573–580.
9. Kanehisa M., Goto S., Sato Y., et al. KEGG for integration and interpreta–tion of large-scale molecular data sets // Nucleic Acids Res. – 2011. – Vol. 1–6 (doi:10.1093/nar/gkr988).
10. Shepelev V., Fedorov A. Advances in the Exon-Intron Database (EID) // Brief. Bioinform. – 2006. –Vol. 7(2). – P. 178–185.
11. Gelfand Y., Rodriguez A., Benson G. 2006. TRDB – The tandem repeats database // Nucleic Acids Res. Database issue. – D1–D8 (doi:10.1093/nar/gkl1013).
12. Lubec G., Afjehi-Sadat L., Yang J. W., John J. P. Searching for hypothetical proteins: theory and practice based upon original data and literature, Prog. Neurobiol. – 2005. – Vol. 77. – P. 90–127.

Статья поступила в редакцию 03.07.2012.