ЖУРНАЛ ОБЩЕЙ БИОЛОГИИ, 2003, том 64, № 1, с. 51–63

УДК 577.21

КЛАССИФИКАЦИЯ НУКЛЕОТИДНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ ПО ЧАСТОТНЫМ СЛОВАРЯМ ОБНАРУЖИВАЕТ СВЯЗЬ МЕЖДУ ИХ СТРУКТУРОЙ И ТАКСОНОМИЧЕСКИМ ПОЛОЖЕНИЕМ ОРГАНИЗМОВ

© 2003 г. А. Н. Горбань[1], Т. Г. Попова[1], М. Г. Садовский[2]

1Институт вычислительного моделирования СО РАН 660036 Красноярск, Академгородок

e-mail:gorban@icm.krasn.ru, tanya@icm.krasn.ru

2Институт биофизики СО РАН

e-mail:msad@icm.krasn.ru, uvenal@ktk.ru

Поступила в редакцию 17.01.2002 г.

Цель работы – изучение связи между структурой нуклеотидной последовательности и таксономическим положением носителя. Изучены классификации нуклеотидных последовательностей бактериальных 16S РНК. Показано существование корреляции между таксономическим положением носителей и информационной структурой нуклеотидных последовательностей бактериальных 16S РНК. Две последовательности считались близкими по структуре, если близки их частотные словари в евклидовой метрике. Предложена процедура преобразования частотного словаря, которая выявляет особенности информационной структуры символьной последовательности. Проведено сравнительное исследование классификаций по реальным и преобразованным частотным словарям. Выделены информационно значимые сайты – главные факторы отличия – для полученных классов. Классификация реальных частотных словарей толщины 3 наилучшим образом коррелирует с родом: род, как правило, целиком включен в один класс и исключения редки. В результате иерархической классификации по преобразованным частотным словарям на каждом этапе выделялись одна-две таксономические группы. Структурные различия полученных классов заключены в редком или, наоборот, частом (по сравнению с ожидаемым) появлении некоторых слов, количество которых невелико.

С момента установления механизма передачи наследственной информации и способа ее хранения в живых организмах не прекращается поток работ, связанных с попытками “прочесть”, “расшифровать” либо иным образом выявить те функции, которые закодированы в нуклеотидных последовательностях. В этом направлении достигнуты определенные успехи, хотя первоначальные надежды оказались не реализованными (Гельфанд, 1998). В исследованиях эволюционных процессов и их молекулярного уровня интригующей является проблема связи между структурой нуклеотидной последовательности того или иного организма и его таксономическим положением, определяемым классическим методами по различным морфологическим признакам.

При изучении связи между структурой нуклеотидной последовательности (НП) и функцией, определяемой этой НП, функция, как правило, понимается исследователями одинаково; аналогично не возникает разночтений в понимании исследователями того, что есть таксономическое положение носителя рассматриваемой НП. Сложнее обстоит дело с понятием структуры. Проблема здесь не в том, что различные исследователи воспринимают его по-разному, а в том, что в НП могут быть выделены различные структуры и заранее непонятно, какая из них имеет отношение к функции, а какая к таксономическому положению носителя данной НП. При изучении НП часто говорят об экзон-интронной структуре (Sharp, 1994), о структуре последовательности, определяемой оперонами (Yockey, 1992), и т.п. Возможны и иные понимания того, что есть структура НП. Мы под структурой нуклеотидной последовательности будем понимать ее частотный словарь – реальный (Горбань и др., 1993а, б, 1994б; Гусев и др., 1980), восстановленный (Бугаенко и др., 1996; Bugaenko et al., 1998) или преобразованный (Gorban et al., 1998). Такое понимание структуры НП позволяет весьма просто ввести понятие близости структур. Две (или несколько) НП считаются близкими по структуре, если близки соответствующие им частотные словари.

Изучение структуры НП требует определения каких-либо дополнительных отношений (например, классификаций) на множестве таких последовательностей. Ранее проведенные исследования показали, что большая совокупность НП может быть разбита на несколько классов, в каждом из которых эти последовательности близки друг дру-

гу в смысле близости их реальных частотных словарей в евклидовой метрике. Кроме того, на множестве генов всегда существует как минимум две независимые классификации: по таксономическому положению носителя гена и по функции, определяемой данным геном. В работе (Gorban et al., 1998) было показано, что классификация генов Ca-зависимых белков, полученная по реальным частотным словарям, коррелирует с функцией белков, которые они кодируют, и с таксономическим положением их носителей. Однако полного соответствия не наблюдалось. Это связано с тем, что структурные различия нуклеотидных последовательностей могут быть связаны с различием в кодируемых ими функция, с одной стороны, и с таксономическим положением носителя – с другой. Поэтому целесообразнее из триады "структура – таксономия – (биохимическая) функция" исследовать пары "структура – таксономия" или "структура – функция", т.е. рассматривать НП, которые определяют одну и ту же функцию у различных организмов, или НП, принадлежащие одному организму, но кодирующие различные функции.

Цель настоящей работы – изучение связи между структурой НП (понимаемой как реальный или преобразованный частотный словарь) и таксономическим положением ее носителя. В настоящее время накоплен значительный объем расшифрованных последовательностей, что позволяет провести указанное сравнительное исследование на достаточном множестве генов. В нашей работе использовались последовательности 16S РНК бактерий различных видов. Все последовательности такого типа выполняют одинаковую функцию не только у бактерий, но и у других организмов.

ВОССТАНОВЛЕННЫЕ И ПРЕОБРАЗОВАННЫЕ ЧАСТОТНЫЕ СЛОВАРИ: ВЫДЕЛЕНИЕ ИНФОРМАЦИОННЫХ ХАРАКТЕРИСТИК

Каждая НП является символьной последовательностью из четырехбуквенного алфавита той же длины N – генетический текст. Любую связную подпоследовательность длины q из рассматриваемой последовательности будем называть словом или q-плетом. Сопоставим каждое слово и его частоту (число его копий в изучаемом тексте, отнесенное к общему числу разных слов в данном тексте); такой список всех слов длины q, входящих в данную последовательность, вместе с частотами их встречаемости назовем частотным словарем толщины q и обозначим W_q (Горбань и др., 1993а, 1994а, б; Гусев и др., 1980; Бугаенко и др., 1996; Bugaenko et al., 1998).

Рассмотрим набор частотных словарей, соответствующих одному генетическому тексту: W_1, W_2, …, W_q, …, W_N. Возникает вопрос: какую часть информации о тексте содержит частотный словарь толщины q? Известно, что, начиная с некоторой длины слова d^*, все слова в тексте встречаются по одному разу (Гусев и др., 1980; Горбань и др., 1993а, 1994б). Следовательно, при $q > d^*$ все слова в W_q имеют единственное продолжение и по словарю W_q однозначно восстанавливаются и любой словарь W_k, $k > q$, и весь исходный текст (Гусев и др., 1980; Горбань и др., 1994б). Таким образом, любой частотный словарь W_q при $q > d^*$ содержит всю информацию об исходном тексте.

С другой стороны, в последовательности частотных словарей предыдущие словари получаются из последующих простым суммированием: для нахождения частоты некоторого слова длины $q-1$ складываются частоты тех слов длины q, которые содержат в себе данное слово и отличаются только первым (либо только последним) символом. Иными словами, по частотному словарю W_q однозначно восстанавливается любой частотный словарь меньшей толщины. При этом, если $q \le d^*$, то в процессе восстановления часть информации о тексте теряется, и обратный переход (от словаря меньшей толщины к словарю большей толщины) становится, в общем случае, неоднозначным. Для каждого словаря W_q ($q < d^*$) существует множество соответствующих ему частотных словарей W_k одной и той же толщины $k > q$, каждый из которых дает исходный словарь W_q при суммировании. Можно, однако, поставить задачу нахождения наиболее правдоподобного словаря большей толщины, соответствующего словарю меньшей толщины. Иными словами, из множества возможных продолжений $\{W_k\}$ ($k > q$) словаря W_q нужно выбрать один словарь, который является его наиболее вероятным продолжением. Этот словарь назовем *восстановленным словарем* и обозначим $\tilde{W}_k(q)$. Процедура нахождения восстановленного словаря основана на принципе максимума энтропии (Бугаенко и др., 1996; Begaenko et al., 1998), основная идея которого состоит в следующем. Для получения восстановленного словаря $\tilde{W}_k(q)$ из всех возможных продолжений словаря W_q выберем то, которое обладает наименьшей определенностью, т.е. максимальной энтропией. Эта экстремальная задача имеет единственное решение, которое и дает выражения для частот словаря $\tilde{W}_k(q)$ (см. Приложение).

Если бы восстановленный частотный словарь $\tilde{W}_k(q)$ полностью совпал с реальным словарем W_k, то это означало бы, что вся информация о словаре толщины k целиком содержится в словаре W_q. Отличия восстановленного словаря $\tilde{W}_k(q)$ от реального W_k показывают, что нового вносят в систему k-плеты по сравнению с q-плетами ($k > q$).

В связи с этим наибольший интерес представляют отличия реального словаря W_k от наиболее правдоподобной гипотезы о нем – $\tilde{W}_k(q)$. Для измерения этих отличий удобно ввести новый объект – *преобразованный словарь*, в котором каждому слову сопоставлено отношение реальной частоты этого слова и частоты, восстановленной по словарю меньшей толщины, т.е. каждому слову соответствует число, показывающее, насколько его реальная частота отличается от ожидаемой. Такое преобразование частотного словаря выявляет особенности информационной структуры НП.

В данной работе рассматривались частотные словари, восстановленные по словарю предыдущей толщины. Для удобства дальнейшего изложения приведем формулу для частот словаря $\tilde{W}_q(q-1)$ (т.е. словаря толщины q, восстановленного по словарю толщины $q-1$):

$$\tilde{f}_{i_1\ldots i_q} = \frac{f_{i_1\ldots i_{q-1}} f_{i_2\ldots i_q}}{f_{i_2\ldots i_{q-1}}}, \qquad (1)$$

где $i_1\ldots i_q$ – слово длины q, индекс i соответствует некоторому нуклеотиду, а $f_{i_1\ldots i_q}$ – частота соответствующего слова. Полученный таким образом словарь $\tilde{W}_q$ является наиболее вероятным продолжением словаря W_{q-1}. Сравнение восстановленного частотного словаря $\tilde{W}_q$ с реальным W_q позволяет выявить особенности информационной структуры НП, поскольку максимальные отличия реальных частот от восстановленных на данной толщине словаря есть наиболее "невероятные" события при переходе от словарей толщины $q-1$ к словарям толщины q.

Преобразуем частотный словарь нуклеотидной последовательности следующим образом: для каждой длины слова q, начиная с $q = 2$, получим восстановленный по словарю предыдущей толщины словарь – $\tilde{W}_q(q-1)$. Рассмотрим отношения реальных частот слов к восстановленным: $p_j = f_j/\tilde{f}_j$, индекс j пробегает все слова из словаря толщины q. Величины p_j показывают насколько реальные частоты слов отличаются от ожидаемых. Если для некоторого слова в НП $p_j \approx 1$, то информационная ценность этого слова невелика: его наиболее вероятная (в соответствии с принципом максимума энтропии) частота совпадает с наблюдаемой. Слова, для которых отношение реальной и восстановленной частот p_j наиболее отличается от единицы в ту или другую сторону (в рассматриваемых примерах – более, чем на 15–20%), представляют собой информационно значимые сайты данной длины в исследуемой НП. Указание на длину принципиально, поскольку незначимый фрагмент длины q может быть включен в значимый фрагмент длины $q + s$, все дальнейшие продолжения которого вновь незначимы. Для сравнения различных НП по преобразованным словарям важны не отличия отношения частот p_j от единицы (которые могут быть обоюдными), а различие p_j между собой для разных НП. Такое различие, как будет показано в дальнейшем на примерах 16S РНК, также становится значимым при разнице в 15–20%.

Формула (1) совпадает с известным выражением для переходных вероятностей в символьной последовательности, полученной Марковским процессом. Следует, однако, отметить, что получение выражения (1) не связано ни с какими гипотезами о структуре последовательности и представляет собой только наиболее правдоподобную гипотезу о словаре толщины q, которую можно высказать на основе рассмотрения словаря толщины $q-1$, и лишь справедливость данных выражений для реальных (а не восстановленных) частот НП в пределе бесконечной длины означала бы, что данная последовательность является Марковской.

КЛАССИФИКАЦИЯ НП ПО ЧАСТОТНЫМ СЛОВАРЯМ

Задача построения классификации объектов требует задания меры близости объектов. Меры близости для двух или нескольких символьных последовательностей могут устанавливаться многими различными способами (Фукунага, 1979; Yockey, 1992; Bork, 1996; Sankoff, Gedergen, 1983). Мы использовали следующие понятие меры близости. Для некоторой длины слова q генетическому тексту сопоставляется точка в многомерном пространстве, представляющая частотный словарь толщины q (реальный, восстановленный или преобразованный). Две последовательности будут считаться близкими по своей структуре, если близки две точки, соответствующие их частотным словарям толщины q. Расстояние между двумя точками в данном многомерном пространстве задается евклидовой метрикой.

Формально НП записаны в 4-буквенном алфавите, и общее число всех слов длины q составляет 4^q. Понятно, что при достаточно больших q в изучаемом тексте встречаются далеко не все слова этой длины. Дополним частотны словарь НП до полного (т.е. содержащего все слова заданной длины), включив в него слова с нулевой частотой. Тогда каждому частотному словарю W_q можно сопоставить точку $F(f_1, f_2, \ldots, f_{4^q})$ в $4q$-мерном пространстве, координатами которой будут частоты соответствующих слов f_j: $0 \le f_j \le 1$, $j = 1, 2, \ldots, 4^q$. Для преобразованных словарей сопоставим каждой нуклеотидной последовательности точку $P(p_1, p_2, \ldots, p_{4^q})$; координатами точки будут теперь отношения реальных частот слов к восстановлен-

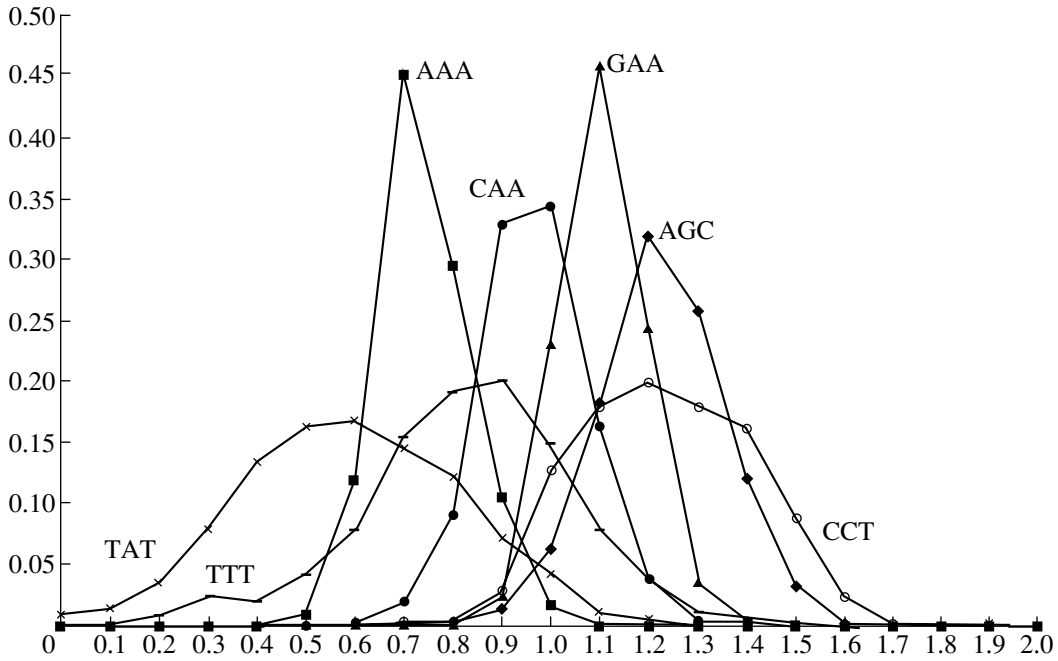

Рис. 1. Плотности распределения преобразованных частот триплетов. По оси абсцисс – значения преобразованных частот *p*, по оси ординат – плотность распределения *p* для данного триплета, построенная по всей совокупности рассмотренных нуклеотидных последовательностей.

ным: $p_j = f_j / \tilde{f}_j$ при $f_j \neq 0$ и $p_j = 1$ при $\tilde{f}_j = 0, j = 1, 2, \ldots, 4^q$.

В соответствии с таким представлением частотных словарей множество НП задает набор точек в 4^q-мерном пространстве. Изучение распределения точек в данном многомерном пространстве позволяет выделить классы символьных последовательностей, близких в смысле данной интерпретации, либо констатировать отсутствие таковых. Для исследования распределения множества точек в многомерном пространстве использовались алгоритмы автоматической классификации, т.е. алгоритмы разделения множества на классы без учителя. В нашей работе использовался метод динамических ядер (Фукунага, 1979; Горбань, Россиев, 1996).

Таблица 1. Таксономический состав нуклеотидных последовательностей бактериальных 16S РНК

Таксономическая единица	Число НП
Группа Chloroflexaceae/Deinococcaceae	29
Cyanobacteria	9
Cytophagales	117
Fibrobacter	13
Firmicutes; Actinomycetes	335
Firmicutes; грамм-положительные бактерии с низким G+C-содержанием	485
Proteobacteria; α-подгруппа	262
Proteobacteria; β-подгруппа	63
Proteobacteria; δ-подгруппа	47
Proteobacteria; ε-подгруппа	43
Proteobacteria; γ-подгруппа	216
Spirochaetales; Leptospiraceae	14
Spirochaetales; Spirochaetaceae	35
Иные	56

РЕЗУЛЬТАТЫ

Исследования проводили на множестве нуклеотидных последовательностей бактериальных 16S РНК общим числом 1730; все последовательности депонированы в EMBL-банке; адреса доступа хранятся на сервере ftp://ccrv.obs-vlfr.fr/pub/christen/16S. В табл. 1 представлен таксономический состав данного множества НП. Он достаточно разнообразен и неоднороден по количеству НП различной таксономической принадлежности.

I. Преобразованный частотный словарь

Рассмотрим структуру преобразованного частотного словаря на примере множества НП бактериальных 16S РНК. Для каждой НП из данного множества был построен преобразованный частотный словарь толщины 3. Характерные плотности распределений значений преобразованных частот триплетов в среднем по выборке показаны на

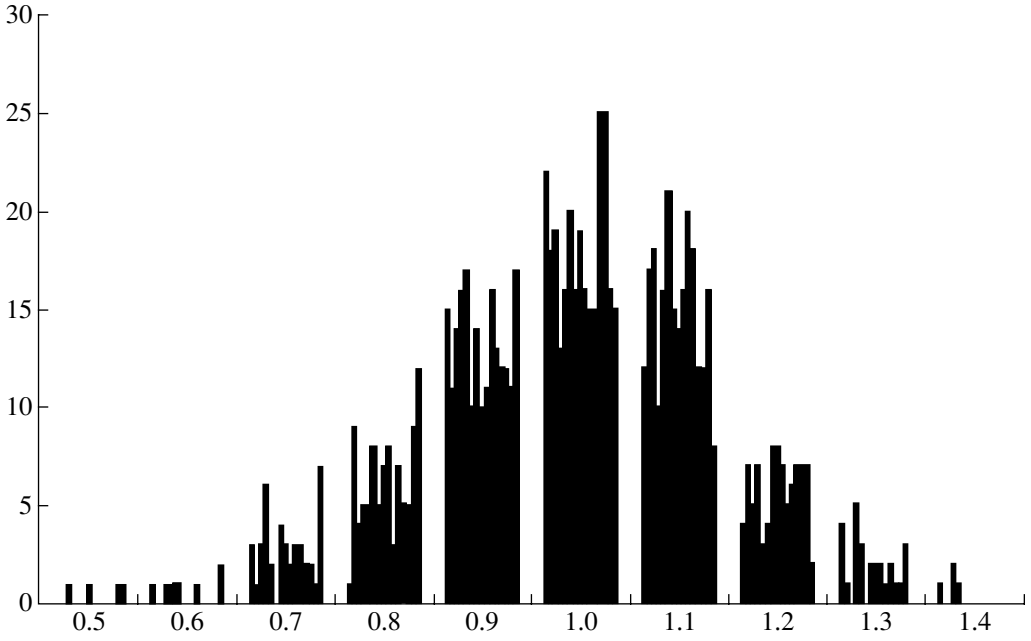

Рис. 2. Гистограмма плотности распределения преобразованных частот триплетов. По оси абсцисс – значения преобразованных частот, по оси ординат – число триплетов (из 64 возможных) с данной преобразованной частотой. Каждый столбец в гистограмме соответствует среднему преобразованному словарю одного из 13 таксонов (см. табл. 1).

рис.1. Понятно, что всего таких распределений 64 по числу триплетов 4-буквенного алфавита. Графики плотностей распределения преобразованных частот для триплетов TAT, CCT, AAA, GAA занимают крайние положения; соответствующие графики для триплетов TTT, CAA, AGC являются наиболее характерными для данной совокупности.

Рассмотрим, как распределены значения преобразованных частот внутри одного словаря (толщины 3). Для этого для каждой из приведенных в табл. 1 таксономических единиц получим преобразованный словарь – среднее по словарям всех НП данной таксономической принадлежности. Гистограмма плотности распределения значений преобразованных частот в среднем по выделенным 13 таксономическим единицам приведена на рис. 2. Числа по вертикальной оси показывают, сколько триплетов из 64 имеют преобразованную частоту из данного диапазона. Каждый узкий столбик соответствует отдельной таксономической единице. Для сравнения на рис. 3 приведены аналогичные данные для реальных частотных словарей. Хорошо видно, что преобразованные словари обладают большей информативностью: с одной стороны, больший размах, с другой – явные различия по таксонам.

Порядок Firmicutes; Actinomycetes представлен в исходной выборке 335 НП. Таблица 2 содержит информационно значимые сайты длины 3 и их преобразованные частоты в среднем по этой группе. Таблица состоит из двух частей: в первой приведены триплеты, частоты встречаемости которых больше, чем предсказанные (p_j), во второй – меньше (p_j). В табл. 3 представлены информационно значимые сайты длины 4. Данная таблица состоит из трех частей: первые две части аналогичны табл. 2; третья содержит 4-плеты с нулевой частотой встречаемости, но с ненулевой восстановленной (ожидаемой) частотой ($\tilde{f}_i \neq 0; f_i = 0$), n – предсказанное число копий данного 4-плета по всей выборке. Сравнение данных таблиц позволяет сделать вывод о том, что информационно значи-

Таблица 2. Информационно значимые сайты длины 3, в среднем по группе Firmicutes; Actinomycetes

Частота больше, чем ожидалась		Частота меньше, чем ожидалась	
триплет	$f/\tilde{f}$	триплет	$f/\tilde{f}$
CCT	1.355	CCA	0.641
AGC	1.338	TAT	0.670
TAA	1.302	AAA	0.709
CTT	1.282	TAG	0.762
TCA	1.194	TCT	0.771
TAC	1.189	TTT	0.784
GAT	1.177	GAC	0.810

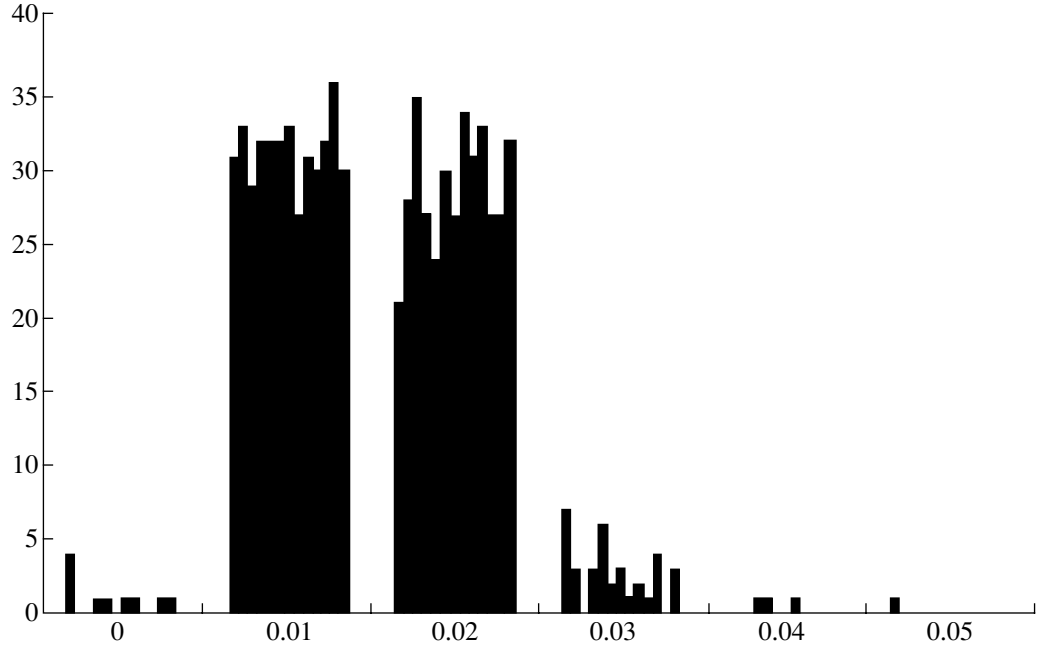

Рис. 3. Гистограмма плотности распределения реальных частот триплетов. Обозначения те же, что и на рис. 2.

мые сайты одной длины могут входить в информационно значимые сайты большей длины, а могут и не входить. И наоборот, информационно значимые 4-плеты могут содержать информационно значимые триплеты, а могут и не содержать.

Таблица 3. Информационно значимые сайты длины 4 в среднем по группе Firmicutes; Actinomycetes

Частота больше, чем ожидалось		Частота меньше, чем ожидалось		Нулевая частота при ненулевой ожидаемой	
4-плет	$f/\tilde{f}$	4-плет	$f/\tilde{f}$	4-плет	n
ATAT	2.035	CAAA	0.600	GTAT	487
TATC	1.918	TTCA	0.595	CATA	346
TTGT	1.874	GTTA	0.587	ATTT	321
ATAC	1.838	CACT	0.568	ACTT	293
ACTC	1.833	TAAG	0.561	TATA	273
ACAC	1.757	AGAG	0.560	ATAG	271
TTAT	1.729	TACT	0.555	TTCA	227
ATTA	1.698	TTAC	0.550	TTTA	138
CACA	1.678	GACA	0.516	CCTC	129
GCGA	1.660	ACAG	0.505	ATAA	125
AGTC	1.657	TGAT	0.483	AAAT	124
CATG	1.653	CGAG	0.482	CCAT	101
ATCA	1.633	AAAT	0.461	ATCT	92
CGAA	1.629	TTTA	0.452	TACT	89
CGCA	1.579	TTCT	0.430	TCTA	66
AGAT	1.565	CCTC	0.422	CCGA	66
TTCC	1.559	GGCA	0.418	TCTT	57
TCCA	1.505	CCGA	0.385	TGAT	55

II. Классификация

Методом динамических ядер были получены разбиения данного множества последовательностей на классы. Классификации исходного множества НП строились отдельно по реальным частотным словарями и отдельно – по преобразованным. Как и следовало ожидать, реальному и преобразованному частотным словарям соответствуют различные классификации исходного множества последовательностей. Тем не менее обе классификации хорошо коррелируют с классификацией по таксономическому положению носителя.

Классификации строились по частотному словарю толщины 3. Это связано с тем, что надежная работа метода динамических ядер возможна, если число классифицируемых объектов многобольше размерности пространства. Кроме того, частотный словарь толщины 3 отражает большее (чем словарь толщины 2 и тем более – толщины 1) число структурных элементов НП; по крайней мере одна из них – структура генетического кода – представлена в нем полностью.

На рис. 4 приведены результаты классификации множества НП бактериальных 16S РНК по реальным частотным словарям. Исходное множество последовательностей разделилось на два класса. На диаграмме по вертикальной оси обозначены таксономические единицы, по горизонтальной оси – количество НП данной таксономической принадлежности в рассматриваемой группе последовательнос-

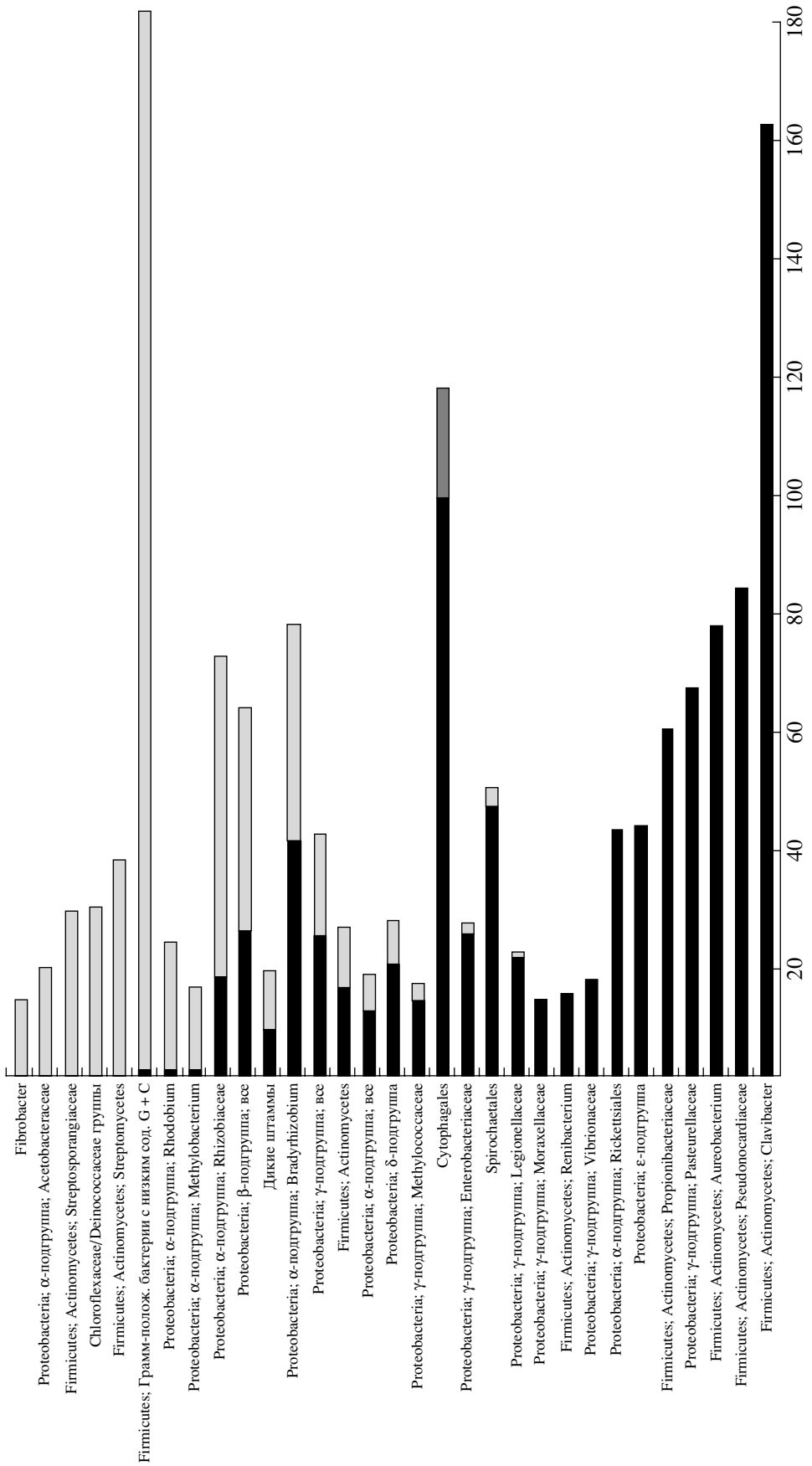

Рис. 4. Классификация множества нуклеотидных последовательностей 16S РНК по реальным частотным словарям толщины 3. По оси абсцисс – число нуклеотидных последовательностей, по оси ординат – таксономическая единица. Черным цветом указано число нуклеотидных последовательностей, принадлежащих первому классу, серым – число нуклеотидных последовательностей, принадлежащих второму классу.

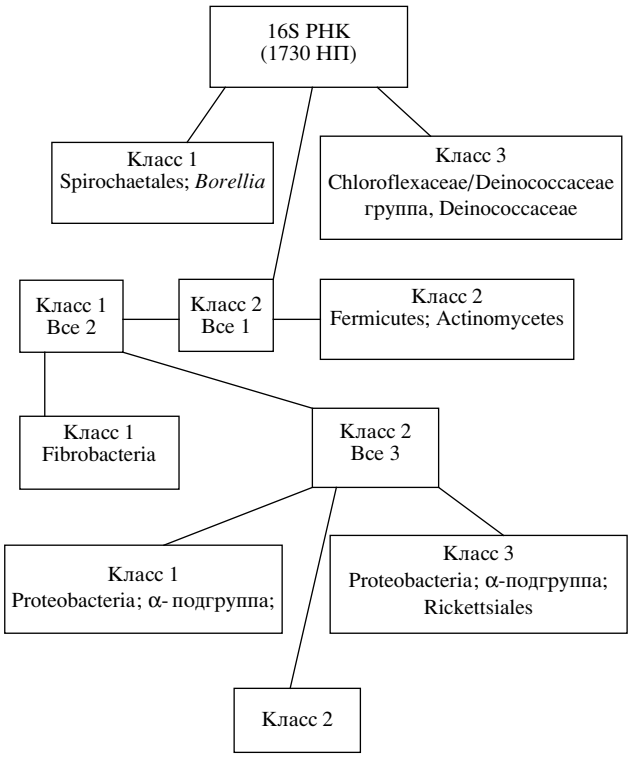

Рис. 5. Иерархическая классификация нуклеотидных последовательностей бактериальных 16S РНК по преобразованным частотным словарям толщины 3. На рисунке показаны четыре уровня классификации.

тей 16S РНК. Серым и черным цветом показано распределение последовательностей каждой таксономической группы по двум классам. Неслучайность распределения таксонов по классам очевидна. Более того, несмотря на то что гены части таксонов встречаются в каждом из классов, хорошо видна неравномерность их распределения между классами.

Построение аналогичной диаграммы для таксономического уровня следующего (более высокого) ранга ведет к тому, что таксономическая принадлежность становится равно представленной в каждой из классов. То обстоятельство, что классификация на основе статистических свойств нуклеотидных последовательностей оказывается чувствительной к таксономическому уровню, является косвенным свидетельством того, что таксономия (особенно высших уровней) прокариот носит весьма искусственный характер (Шлегель, 1993).

Для НП некоторых организмов, и прокариот в особенности, существенно выражено предпочтение в появлении вполне определенных нуклеотидов, при этом предпочтения весьма слабо влияют на структуру белков, кодируемых этими НП. Тогда корреляция между классификацией, построенной по реальным частотным словарям, и таксономией является следствием различий нуклеотидного состава этих НП. Поэтому классификация по преобразованным словарям представляется нам более интересной как с точки зрения метода выделения структур в НП, так и с точки зрения собственно классификации данной конкретной группы НП.

На рис. 5 представлены результаты иерархической классификации множества НП по преобразованным частотным словарям. Хорошо видно, что на каждом уровне классификации из множества НП выделялись отдельные таксономические единицы. Несмотря на немногочисленность выделенных единиц, они содержат практически все последовательности данной таксономической принадлежности из исходной группы НП 16S РНК бактерий.

В результате построения классификации одна из получившихся групп оказалась достаточно многочисленной (см. рис. 2), что позволило перейти к следующему уровню классификации: группа Все 1 в свою очередь также может быть разбита на несколько классов. Однако классы, получившиеся на втором уровне иерархической классификации, не удовлетворяют условию разделимости. Тем не менее при разбиении данного множества на два класса происходит выделение семейства Firmicutes; Actinomycetes, что представляется нам интересным. Отсутствие разбиения, удовлетворяющего условию разделимости классов на втором уровне классификации, может иметь несколько причин. По-видимому, одной из самых важных является таксономическое разнообразие в рассматриваемой группе НП, которое обусловливает постепенное изменение информационных характеристик НП и связанное с ним отсутствие четко выделяемых классов. На третьем уровне классификации оставшееся множество НП разделилось на два класса: отличным по структуре оказалось семейство Fibrobacteria. Таблица 5 (аналогично табл. 4) содержит главные факторы, определяющие различия структур; в ней также приведены координаты полученных классов, определяемые по реальным частотам.

Классы, полученные на четвертом уровне иерархической классификации, не удовлетворяют условию разделимости классов и приведены нами в качестве завершающего этапа. Разделение на три класса здесь в каком-то смысле оптимально: именно для трех классов достигается наилучшее соотношение между средними радиусами классов и расстоянием между центрами этих классов. Кроме того, представляется интересным выделение последовательностей одной определенной таксономической принадлежности в один класс.

Рассмотрим подробнее результаты классификации по преобразованным частотным словарям на каждом иерархическом уровне, особо обратив внимание на то, чем именно и насколько различаются последовательности из разных классов. Как видно

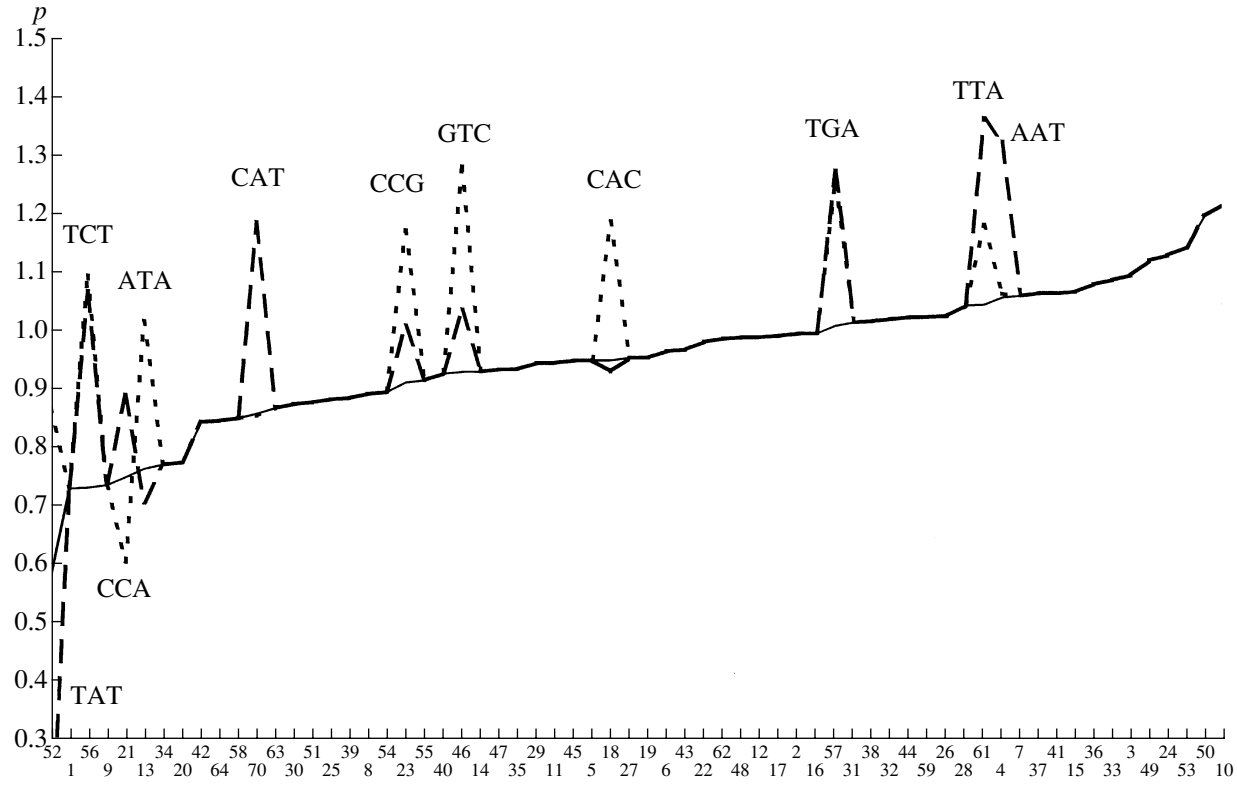

Рис. 6. График наибольших различий в значениях преобразованных частот триплетов, определивших различия между тремя группами на первом уровне классификации. Сплошная линия – группа Все 1, крупный пунктир – группа Chloroflexacaeae/Deinococcaceae, мелкий пунктир – группе Spirochaetales. По оси абсцисс – номера триплетов, упорядоченных по возрастанию значений преобразованных частот группы Все 1; по оси ординат – значения преобразованных частот. Мелкие различия сглажены.

на рис. 5, на первом этапе все исходное множество последовательностей разделилось на три группы: I – Chloroflexaceae/Deinococcaceae; Deinococcaceae; II – Spirocheaetales; Spirochaetaceae; *Borrelia*; III – группа Все 1 (все остальные НП). Рисунок 6 показывает, какие именно триплеты определили различие классов. График построен следующим образом. По вертикальной оси отложены значения

Таблица 4. Главные факторы различия трех классов (отношения реальной и ожидаемой частот) и соответствующие им реальные частоты

Фактор различия	Группа НП					
	Chloroflexaceae/Deinococcaceae; Deinococcaceae		Spirochaetales; Spirochaetaceae; *Borrelia*		все остальные	
	f	$f/\tilde{f}$	f	$f/\tilde{f}$	f	$f/\tilde{f}$
TAT	**0.109**	**0.0008**	0.978	**0.0173**	**0.600**	0.0059
CAT	**1.296**	0.0087	**0.746**	0.0077	0.995	0.0102
GTC	0.998	0.0132	**1.420**	0.0173	1.019	0.0140
TCT	1.060	0.0073	**1.189**	**0.0125**	**0.819**	0.0078
TTG	**0.710**	**0.0075**	0.958	0.0141	0.99	0.0144
ATA	**0.686**	**0.0047**	1.099	**0.0237**	**0.859**	0.0099
TCG	0.916	0.0110	**0.788**	0.0090	1.088	0.0140
TTA	**1.418**	0.0099	1.224	0.0187	1.140	0.0117
CCA	0.952	0.0134	**0.545**	**0.0052**	0.808	0.0102
ATC	1.028	0.0072	**0.737**	0.0081	1.019	0.0098

Примечание. Полужирным шрифтом выделены частоты тех триплетов, которые обусловили наибольшее различие между классами, для преобразованных и реальных частот.

Таблица 5. Главные факторы различия двух классов и соответствующие им реальные частоты

Группа НП	Отношение частот	Фактор различия								
		CAA	TAA	CTA	CAT	CTG	TTC	GTA	TCT	TAC
Все 3	$f/\tilde{f}$	0.973	1.230	0.945	1.001	0.998	0.948	1.048	0.836	1.144
	f	0.0153	0.0182	0.0121	0.0106	0.0173	0.0082	0.0173	0.0080	0.0139
Fibrobacteria	$f/\tilde{f}$	1.337	0.898	0.670	0.756	1.229	1.169	1.268	0.617	1.362
	f	0.0221	0.01	0.0061	0.0091	0.0183	0.0096	0.018	0.0045	0.0129

преобразованных частот. Крупный пунктир соответствует группе I, мелкий – группе II, сплошная линия – группе III. На горизонтальной оси расположены номера триплетов по возрастанию преобразованных частот группы III. Триплеты занумерованы в лексикографическом порядке. Мелкие различия между группами сглажены. В табл. 4 приведены те координаты центров классов (и соответствующие им триплеты), по которым эти классы более всего отличаются друг от друга. Преобразованный частотный словарь отражает вклад неслучайности в распределение нуклеотидов в НП; для сравнения в этой же таблице показаны центры этих классов, определяемые по реальным частотным словарям. Различия преобразованных словарей могут совпадать, а могут и не совпадать с различием реальных частотных словарей. Очевидной корреляции классификаций по преобразованным и по реальным словарям не наблюдается.

ОБСУЖДЕНИЕ

Проблема анализа статистических свойств НП и описания их смысла требует подбора соответствующего представления для рассматриваемой последовательности. Фактически проблема состоит в выборе подходящих “координат”, в которых НП с существенно различающимися свойствами оказываются также наиболее далекими. Говоря здесь о координатах, мы не имеем в виду строгое определение координат в том или ином пространстве; скорее, речь может идти о выборе такого представления, которое позволяет наиболее ясно представить те свойства НП, которые интересуют исследователя. Некоторые представления такого рода (“координаты”) достаточно хорошо известны.

В первую очередь укажем на такое представление НП, как частотные словари. Исследование и сравнение символьных последовательностей при помощи частотных словарей проводилось задолго до открытия ДНК. В настоящее время для анализа генетических текстов используется широкий спектр структур, которые можно назвать частотными словарями: это и традиционно понимаемые полные наборы частот коротких фрагментов – от GC контента до 12-плетов, и всевозможные наборы дальних корреляций, и частоты кодонов, и пр.

Другой способ представлений НП связан с поиском и выделением в них участков (зачастую достаточно протяженных), которые близки по своей структуре тем или иным “образцовым” последовательностям: на деле это представление означает поиск гомологов либо аналогов известных “реперных” последовательностей. Известно также представление НП через химические или физические свойства полинуклеотидов. Примером может служить выделение структур в НП по температуре “поперечного плавления”, т.е. выделение участков в двойной цепочке ДНК, соответствующих разным температурам разрыва комплементарных нитей.

Мы полагаем, что предложенная нами процедура преобразования частотного словаря и соответственно выделения информационно значимых фрагментов по степени отличия реальной частоты слов от ожидаемой позволяют выявить такие особенности структуры НП, которые являются хорошими координатами для представления НП. Отношения реальных частот к ожидаемым более информативны, чем просто частоты слов. такое отношение показывает “степень неожиданности” наблюдаемой частоты.

Основное значение полученных в настоящей статье результатов состоит в следующем: для группы НП 16S РНК наблюдается четкая связь между “степенями неожиданности” наблюдаемых частот малых фрагментов этих НП и таксономическим положением их носителей. Классификация в пространстве преобразованных частот позволяет выявить более тонкие различия между НП разной таксономической принадлежности.

Одна из парадигм эволюционной молекулярной биологии заключается в том, что изучение эволюционных процессов, а также выводы о степени родства или последовательности смены тех или иных форм в ходе эволюции следует делать на анализе таких молекулярных объектов, которые имеют высокую степень универсальности. В качестве таких объектов исследователь старается выбрать тот или иной (зачастую весьма небольшой)

набор генов (НП), которые фактически в неизменном виде либо с малыми изменениями встречаются в возможно большем числе организмов самого разнообразного таксономического положения. К таким НП традиционно относят 5S РНК, 25S РНК, НП, кодирующие какие-нибудь весьма консервативные белковые системы (например, гены различных цитохромов). С этой точки зрения выбранный нами объект – 16S РНК – не отвечает таким требованиям универсальности. Тем не менее выбор 16S РНК в качестве объекта эволюционного исследования представляется оправданным и целесообразным. Во-первых, эти НП специфичны для царства бактерий; во-вторых, в пределах этого царства они определяют одну и ту же функцию и тем самым позволяют исключить из рассмотрения влияние функциональных различий.

Существование корреляции между статистической структурой НП и таксономическим положение носителя четко проявляется для выбранной группы последовательностей. Одна из важных задач здесь – поиск частотного словаря (реального или преобразованного) такой толщины, который наиболее чувствителен к тем или иным биологическим свойствам, интересным исследователю. Однако последовательное изучение такой связи для сайтов длины большей, чем 3, затруднено тем обстоятельством, что для получения репрезентативных результатов необходимо анализировать очень большую группу НП очень большой длины – число классифицируемых НП и их длина должны быть много больше размерности пространства.

Развитый в настоящей работе подход к анализу связи между значением НП и ее структурой, понимаемой как набор частот ее малых фрагментов, показал свою эффективность на примере построения автоматической классификации 16S РНК последовательностей. Представляются перспективными следующие направления в исследовании связи между структурой и функцией НП: во-вторых, это построение автоматической классификации всех генов того или иного конкретного организма. В настоящее время существует значительное количество полностью расшифрованных и аннотированных (полностью или частично) геномов. Для таких геномов можно взять набор всех генов, обнаруженных в них, и попытаться их расклассифицировать. Поскольку в таком семействе НП все последовательности будут принадлежать одному и тому же организму, постольку можно ожидать, что таксономически обусловленные различия между генами будут максимально нивелированы и определяются разницей их функций. Во-вторых, представляется перспективным изучение набора факторов наибольшего различия НП. Среди сайтов, обладающих высокой информационной ценностью (т.е. заметным различием реальной и наиболее ожидаемой частот), могут найтись такие, которые встречаются в большом числе НП из исследуемой группы. Могут, однако, встречаться и такие, которые весьма специфичны для конкретных последовательностей; именно они представляют интерес для биолога-эволюциониста, поскольку определяют отличия конкретной последовательности от группы. Наконец, в-третьих, еще одним направлением в рамках развитого здесь подхода может стать изучение распределения таких факторов различия по генам одного организма, для которых они были бы определены.

Мы исследовали наличие связи между структурой НП и таксономическим положением ее носителя. На примере группы последовательностей 16S РНК бактерий различных видов изучался вопрос о том, насколько близость по структуре НП соответствует близости по таксономическому положению носителей. Близость структур понималась как близость частотных словарей, реальных или преобразованных, в евклидовой метрике. Наиболее существенно для молекулярных аспектов эволюционной теории то, что классификация реальных частотных словарей толщины 3 наилучшим образом коррелирует с родом: род, как правило, целиком включен в один класс, исключения редки. Объединение НП в таксономические группы по семействам и выше ведет к заметному ухудшению корреляции таксономической классификации с классификацией по статистическим свойствам соответствующих НП.

Преобразование частотного словаря НП (использование восстановленного словаря для выделения неслучайной составляющей в распределении частот k-плетов) позволяет сравнивать НП по их информационным характеристикам. Автоматическая классификация множества НП по преобразованным частотным словарям выделяет классы близких последовательностей. В рассмотренном нами случае для множества бактериальных 16S РНК выделенные классы содержат НП определенной таксономической принадлежности.

Классификации, построенные по реальным и преобразованным частотным словарям, принципиально отличаются друг от друга. Как уже было сказано, классификация по реальным частотным словарям во многом является следствием различий нуклеотидного состава последовательностей. Разделение исследуемого множества НП на две большие группы, хорошо коррелирующие с таксономическим положением носителей, – тому подтверждение. Классификация же по преобразованным словарям отделяет на каждом этапе от всей совокупности одну-две таксономические группы. Структурные различия полученных классов заключены в редком или, наоборот, частом (по сравнению с ожидаемым) появлении некоторых слов, количество которых невелико (см. табл. 4, 5). Выделенные по информационным характеристикам

сайты в НП не обязаны совпадать с теми структурными элементами, которые выделяются иными методами (Gelfand, 1995; Claverie et al., 1990). Представляет интерес сравнение тех функционально значимых участков НП, которые были выделены по их информационным характеристикам, с теми, которые выделяются другими методами; такое сравнение еще предстоит провести.

Полученные выше результаты делают перспективным дальнейшее изучение эволюционных процессов с помощью преобразованных частотных словарей. Возможными объектами для такого исследования могут стать последовательности других классов РНК (упоминавшиеся выше), НП генов поверхностных антигенов бактериальных генов и НП генов HLA человека, НП генов дыхательной цепи (например, гены различных цитохромов), НП генов Са-зависимых белков и некоторые иные. Выбор конкретной системы НП, по которой следует проводить анализ эволюционных изменений, происходивших с носителями этих НП, во многом определяется конкретными задачами исследования.

ПРИЛОЖЕНИЕ

Принцип максимума энтропии

Рассмотрим частотны словарь W_q некоторой НП. Энтропия данного частотного словаря определяется как

$$S_q = -\sum_{j=1}^{4^q} f_j \ln f_j, \qquad (1)$$

где f_j – частоты слов, а индекс j пробегает все слова из словаря. Восстановление словаря $\tilde{W}_{q+s}$ по словарю W_q должно производится без привлечения какой-либо дополнительной информации; следовательно, восстановленный словарь должен иметь максимально возможную энтропию. Экстремальная задача $S_{q+s} \longrightarrow \max$ при условиях связи словарей $W_q \longleftarrow W_{q+s}(q)$ решается методом неопределенных множителей Лагранжа и имеет единственное решение

$$f_{i_1 \ldots i_q i_{q+1} \ldots i_{q+s}} = \\ = \frac{f_{i_1 \ldots i_q} f_{i_2 \ldots i_{q+1}} \cdots f_{i_{q-s+1} \ldots i_{q+s}}}{f_{i_2 \ldots i_q} f_{i_3 \ldots i_{q+1}} \cdots f_{i_{q-s+1} \ldots i_{q+s-1}}} \quad \text{для} \quad q > 1, \qquad (2)$$

$$f_{i_1 \ldots i_{q+s}} = f_{i_1 \ldots i_{q+s}} \quad \text{для} \quad q = 1, \qquad (3)$$

где $i_1 \ldots i_q i_{q+1} \ldots i_{q+s}$ – слово длины $q + s$ и индекс i соответствует некоторому нуклеотиду. В задачах статистической физики соотношения для частот (2, 3) напоминают хорошо известное приближение Кирквуда (Kirkwood, Boggs, 1942).

Метод динамических ядер

Пусть имеется множество $\{F^i\}$, $i = 1, 2, \ldots, M$, состоящее из M точек пространства, которое нужно разбить на некоторое количество классов. Пусть заданы начальное количество классов и начальное разбиение исходного множества на заданное количество классов. Для каждого k-го класса вычисляется его центр $C^k(c_1^k, \ldots, c_{n^q}^k)$:

$$c_j^k = \frac{1}{l_k} \sum_{i=1}^{l_k} p_j^i, \quad j = 1, 2, \ldots, n^q, \qquad (5)$$

где l^k – количество элементов в k-м классе. Затем для каждой точки множества F вычисляется расстояние от нее до центра каждого класса

$$d_i^k = \rho(C^k, F^i), \quad i = 1, 2, \ldots, M \qquad (6)$$

и принадлежность каждой точки переопределяется. Точка считается принадлежащей тому классу, расстояние до центра которого наименьшее. После перебора всех точек множества центры групп пересчитываются. Такая процедура – вычисление центров и перенос точек – продолжается до тех пор, пока хотя бы одна точка перемещается в другой класс.

Если все полученные классы отличаются друг от друга, то построение классификации завершено; если нет, то два ближайших класса объединяются в один, и вся процедура повторяется сначала. Два класса считаются различными, если расстояние между их центрами больше максимального из двух средних радиусов классов. Средний радиус k-го класса определяется формулой:

$$R^k = \frac{1}{l_k} \sum_i d_i^k, \qquad (7)$$

где d_i^k определено согласно (6), а i пробегает все значения, такие, что точка F^i принадлежит k-му классу.

Количество классов, на которые разбивается данное множество точек, заранее неизвестно. Изначально множество разбивается на достаточно большое число классов. В результате последовательных слияний образуется максимально возможное число классов, при котором выполняется условие разделимости. Примененный здесь алгоритм полностью аналогичен принципу, по которому происходит кластерный анализ нейронными сетями Кохонена (Горбань, Россиев, 1996).

СПИСОК ЛИТЕРАТУРЫ

Бугаенко Н.Н., Горбань А.Н., Садовский М.Г., 1996. Об определении информационного содержания

нуклеотидных последовательностей // Молекуляр. биология. Т. 30. № 3. С. 529–541.

Гельфанд М.С., 1998. Компьютерный анализ последовательностей ДНК // Молекуляр. биология. Т. 32. № 1. С. 103–120.

Горбань А.Н., Россиев Д.А., 1996. Нейронные сети на персональном компьютере. Новосибирск: Наука. 275 с.

Горбань А.Н., Миркес Е.М., Попова Т.Г., Садовский М.Г., 1993а. Новый подход к изучению статистических свойств генетических последовательностей // Биофизика. Т. 38. С. 762–767.

Горбань А.Н., Миркес Е.М., Попова Т.Г., Садовский М.Г., 1993б. Сравнительная избыточность генов различных организмов и их вирусов // Генетика. Т. 29. № 9. С. 1413–1419.

Горбань А.Н., Попова Т.Г., Садовский М.Н., 1994а. Избыточность генетических текстов и мозаичная структура генома // Молекуляр. биология. Т. 28, № 2. С. 313–322.

Горбань А.Н., Попова Т.Г., Садовский М.Г., 1994б. Корреляционный подход к сравнению нуклеотидных последовательностей // Журн. общ. биологии. Т. 55. № 4/5. С. 420–430.

Гусев В.Д., Куличков В.А., Титкова Т.Н., 1980. Анализ генетических текстов. 1. l-граммные характеристики // Эмпирическое предсказание образов (Выч. системы, вып. 83). Новосибирск: ИМ СО АН СССР. С. 11–33.

Фукунага К., 1979. Введение в статистическую теорию распознавания образов. М.: Глав. ред. физ.-мат. лит. 367 с.

Шлегель Г., 1993. Общая микробиология. М.: Мир. 682 с.

Bork P., 1996. Go hunting in sequence databases but watch out for the traps location // Trends Genet V. 12. P. 425–427.

Bugaenko N.N., Gorban A.N., Sadovsky M.G., 1998. Maximum entropy method in analysis of genetic text and measurement of its information content // Open System & Information Dynamics. V. 5. № 3. P. 265–278.

Claverie J.-M., Sauvaget I., Bougueleret L., 1990. k-tuple frequency analysis: from intron/exon discrimination to T-cell epitope mapping // Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences / Ed. Doolittle R.F. (Meth. Enzymol. V. 183). N.Y.: Acad. Press. P. 252–281.

Gelfand M.S., 1995. Prediction of function in DNA sequence analysis // J. Comput. Biol. V. 2. P. 87–115.

Gorban A.N., Popova T.G., Sadovsky M.G., 1998. Automatic classification of nucleotide sequences and its relation to natural taxonomy and protein function // Proc. of 1st Int. Conf. on Bioinformatics of Genome Regulation and Structure, Novosibirsk, Aug., 24–27., 1998. V. 2. Novosibirsk: Citology & Genetics Institute of SD of RAS. P. 314–317.

Kirkwood J., Boggs E., 1942. The radial distribtuion function in liquids // J. Chem. Phys. V. 10. № 6. P. 394.

Sankoff D., Gedergen R.J., 1983. Simultaneous comparison of three or more sequences related by a tree // Strings and Macromolecules: The theory and Practice of Sequence Comparison. Reading, MA: Addison-Wesley. P 253–263.

Sharp Ph.A., 1994. Split genes and RNA splicing // Cell. V. 77. № 6. P. 805–815.

Yockey H.P., 1992. Information Theory and Molecular Biology. N.Y.: Cambridge Univ. Press. 472 p.

Classification of Nucleotide Sequences Over Their FrequEncy Dictionaries Reveals A Relation Between the Structure of Sequences and Taxonumy of Their Bearers

A. N. Gorban[1], T. G. Popova[1], M. G. Sadovsky[2]

[1]Institute of computing modelling of SD of RAS; 660036 Russia, Krasnoyrsk, Akademgorodok,
e-mail:gorban@icm.krasn.ru, tanya@icm.krasn.ru
[2]Institute of biophysics of SD of RAS; 660036 Russia, Krasnoyrsk, Akademgorodok,
e-mail:msad@icm.krasn.ru, uvenal@ktk.ru

Classification of 16S RNA sequences over their frequency dictionaries, both real ones, and transformed ones was studied. Two entities were considered to be close each other from the point of view of their structure, if their frequency dictionaries were close, in Eucledian metric. A transformation procedure of a frequency dictionary has been implemented that reveals the peculiarities of information structure of a nucleotide sequence. A comparative study of two classification developed over the real frequency dictionary vs. that one developed over the transformed frequency dictionary was carried out. The strong correlation is revealed between the classification and the taxonomy of 16S RNA bearer. For the classes isolated, the information valuable words were identified. These words are the main factors of a difference between the classes. The frequency dictionaries containing the words of the length 3 exhibit the best correlation between a class and a genus. A genus, as a rule, is included into the same class, and the exclusion are sporadic. A development of hierarchy classification over the transformed frequency dictionaries separated one or two taxonomy groups, as each stage of classification. The unexpectedly frequent, or contrary, unexpectedly rare occurred of words (of the length 3) in entities under consideration make the structure difference between the classes of the nucleotide sequences.