АНАЛИЗ ПОСЛЕДОВАТЕЛЬНОСТЕЙ И ТАБЛИЦ ДАННЫХ
(Вычислительные системы)
1994 год Выпуск 150

УДК 519.237.8:519.764

МЕТОДИКА ОЦЕНКИ БЛИЗОСТИ ГЕНЕТИЧЕСКИХ ТЕКСТОВ

Н.А. Чужанова

В в е д е н и е

Одной из актуальных задач компьютерной генетики является задача выявления сходства различных генетических текстов. Наличие такого сходства позволяет делать выводы о таксономической (эволюционной) и/или функциональной близости последовательностей.

Анализ существующих методов решения указанных задач показывает, что частоты встречаемости олигонуклеотидов являются важной классификационной характеристикой, однако используемые меры близости и основанные на них методы таксономии имеют ряд существенных недостатков, которые будут рассмотрены в п.1.

В настоящей работе рассматривается некоторая единая методика оценки таксономической и функциональной близости генетических последовательностей, основанная на вычислении известной меры близости - коэффициента конкордации, введенного в [1] и адаптированного применительно к символьным последовательностям в [2]. Ранее он использовался для распознавания функциональных сайтов в ДНК-последовательностях [3].Достоинства предлагаемой методики - простота вычисления коэффициентов, отсутствие необходимости выравнивания, известное распределение для случайных упорядочений, возможность сравнения более чем двух последовательностей,причем возможно, сильно различающихся по длине.

Возможности методики демонстрируются на реальных нуклео - тидных и аминокислотных последовательностях, взятых из базы данных EMBL. Для сравнения полученных результатов в работе использовались те же последовательности, что и в [4].

1. Существующие методы

Статистические исследования большого числа нуклеотидных последовательностей, проведенные ранее, показали, что частоты встречаемости олигонуклеотидов являются важной классификационной характеристикой последовательностей.

Нуссинов обнаружила. что в ДНК различных таксономических групп существуют устойчивые и статистически значимые асимметрии в частотах встречаемости динуклеотидов [5]. Например, в большинстве проанализированных ею прокариотических последова - тельностей выполнялось неравенство $f(GC) > f(AT) > f(TA)$, а в большинстве эукариотических - $f(GG)>f(GC)>f(GT)>f(TA) > f(CG)$, где $f(ab)$ - частота встречаемости динуклеотида ab.Важность полученных Нуссинов закономерностей не вызывает сомнений, однако их использование для оценки близости и таксономии последовательностей затруднено, так как неравенства справедливы для совокупности последовательностей и могут не выполняться на конкретных последовательностях. Кроме того, представление закономерностей в виде неравенств не является конструктивным.

Блайсделл в [6] показал, что использование марковских цепей 1-го и 2-го порядков для моделирования генетических последовательностей позволяет применить к ним известные статистические меры однородности для оценки таксономической и эволюционной близости. К достоинствам предложенного метода следует отнести возможность сравнения более чем двух последовательностей, отсутствие выравнивания, наличие аппроксимационных характеристик. Однако весьма спорным является вопрос об адекватности моделирования природных нуклеотидных последовательностей однородными марковскими моделями, так как неравномерное распределение

частот встречаемости нуклеотидов по позициям, например, в ко - дирующих последовательностях, противоречит модели однородной марковской цепи.

Грэнтем и др. [7] сформулировали так называемую "геномную гипотезу", согласно которой все гены одного генома или близкородственных геномов, независимо от функций кодируемых ими белков, придерживаются одной стратегии выбора синонимичных кодонов, но в рамках этой стратегии высокоэкспрессируемые и низкоэкспрессируемые гены могут иметь собственные подстратегии. Частоты встречаемости кодонов в том или ином виде учитывались в мерах, введенных в [7-9] для оценки "геномной" близости. В [8] каждый ген представлялся точкой в 61-мерном пространстве кодонов с координатами, равными частотам встречаемости соответствующего кодона. Далее для обеспечения возможности визуализации многомерное пространство проектировалось на плоскость. Расстояние между последовательностями определялось как

$$d^2(i_1,i_2) = \sum_{j=1}^{64} (f_{i_1 j} - f_{i_2 j})^2,$$

где $f_{i_1 j}$ и $f_{i_2 j}$ - частоты встречаемости j-го кодона в текстах i_1 и i_2 соответственно. В [9] последовательность представля - лась точкой в девятимерном пространстве с координатами, соот - ветствующими частотам встречаемости нуклеотидов A,T,G в первой позиции кодона, этих же нуклеотидов во второй и третьей позициях кодона, а расстояние определялось как

$$d_{km} = \sum_{i=1}^{v} |x_{k_i} - x_{m_i}| ,$$

где d_{km} - расстояние между точками x_k и x_m с координатами x_{k_i} и x_{m_i} в девятимерном пространстве. Алгоритмы кластеризации,ис-

пользуемые в [7-9], имеют один существенный недостаток - результаты кластеризации зависят от выбора "начального" кластера.

В [10] предложен способ классификации генов по их экспрессивности, основанный на двух известных биологических фактах: мРНК адаптации и выборе кодонов с пиримидином в третьей позиции. Эксперименты проводились на 83 последовательностях E.coli, некоторые из которых в тот момент еще не были секвенированы полностью, что не позволяет оценить достоинства метода.

Трифонов предложил описывать последовательности словарями характерных олигонуклеотидов длиной от 2 до 6 нуклеотидов и показал, что существует корреляция между таксономической и функциональной близостью последовательностей и составом их словарей [4]. Для предсказания частоты встречаемости слова длиной n использовались марковские цепи (n-2)-го порядка и ожидаемая частота вычислялась следующим образом:

$$E(a_1a_2\ldots a_n) = \frac{f(a_1a_2\ldots a_{n-1})\, f(a_2\ldots a_n)}{f(a_2\ldots a_{n-1})},$$

где E - ожидаемая частота, f - наблюдаемая частота. Слова w отбирались в словарь, если они удовлетворяли условию std(w) = $= |(f(w) - E(w))/E^{1/2}(w)| \geq 3.0$. Мера близости двух словарей определяется как $S_{2-5} = (C_2 + C_3 + C_4 + C_5)/4$ и

$$C_i(A,B) = \frac{\sum_{j=1}^{n_i} q_j(A)q_j(B)}{\left(\sum_{j=1}^{n_i} q_j^2(A) \sum_{j=1}^{n_i} q_j^2(B)\right)^{1/2}}$$

где $C_i(A,B)$ - коэффициент корреляции последовательностей A и B по словарям слов длиной i, n_i - число возможных слов длиной i

(при мощности алфавита в 4 символа число возможных слов равно 4^i), $q_j(X) = std(j)$ в последовательности X.

Как отмечается в [11], к построению и трактовке таких словарей следует относиться весьма осторожно, так как остается неясным вопрос, при каких отклонениях от ожидаемых значений частот можно делать выводы об их биологической значимости. Вероятностные характеристики частот встречаемости слов зависят от структуры самопересечения слов, которая задается автокорреля - ционным многочленом. Учет самопересечения слов, как показано в [11], может привести к совершенно различным словарям.

Основной вывод, который можно сделать из проделанного анализа, состоит в следующем: частоты встречаемости динуклеотидов (и, возможно, тринуклеотидов) являются важной таксономической характеристикой последовательностей, а частоты встречаемости кодонов - характеристикой "геномной" близости.

2. Коэффициент конкордации w_1 как мера таксономической, "геномной" и функциональной близости последовательностей

Коэффициент конкордации как мера согласованности незави - симых упорядочений n объектов m экспертами был введен Кендэ - лом [1] и адаптирован применительно к символьным последователь-ностям в [2] .

Напомним некоторые основные определения, которые понадо - бятся в дальнейшем.

1-грамма (или олигонуклеотид длины 1) - это подцепочка из 1 следующих друг за другом символов. Частотная характеристика 1-го порядка - это совокупность всех 1-грамм последовательно-сти или текста вместе с частотами их встречаемости.

Коэффициент конкордации 1-го порядка текстов $T_1,\ldots,T_m$ - это мера сходства (близости) m текстов по их частотным характе-ристикам 1-го порядка. Каждый текст представляется частотной характеристикой 1-го порядка, упорядоченной по убыванию час-

тот встречаемости 1-грамм. Позиция 1-граммы в упорядочении называется рангом. В случае равных частот 1-граммам присваивается некоторый средний ранг. Коэффициент конкордации 1-го порядка вычисляется по следующей формуле:

$$W_1 = \frac{V_1}{\frac{1}{12} m^2 (A_1^3 - A_1) - m \sum_{i=1}^{m} \Delta_{1i}},$$

где V_1 - сумма (по всем m упорядочениям) квадратов отклонений суммы рангов по всем 1-граммам от среднего значения этой суммы, которое равно $\frac{1}{2} m(A_1+1)$, где A_1 - мощность алфавита 1-грамм, Δ_{1i} - поправка на "связанные" ранги [1].

Значения коэффициента изменяются от 0 (для случая m = 2 упорядочения противоположны) до 1 (идентичные упорядочения). С увеличением l значения коэффициентов в основном падают, однако в некоторых случаях могут повышаться. Рассмотрим небольшой пример, иллюстрирующий такую возможность. Пусть T_1 = CAAAAAAACAACA длиной в 13 символов и T_2 = AAAAAAAACAACAA длиной в 14 символов. Частотные характеристики 2-го порядка для них имеют вид: $\Phi_2(T_1)$ = {AA,f(AA)=7; AC,f(AC)=2; CA,f(CA)=3} и $\Phi_2(T_2)$ = = {AA,f(AA)=9; AC,f(AC)=2; CA,f(CA)=2}. Коэффициент $W_2 < 1$, так как последовательности рангов не совпадают. Частотные характеристики 3-го порядка $\Phi_3(T_1)$ = {AAA,f(AAA)=5; AAC,f(AAC)=2; CAA,f(CAA)=2; ACA,f(ACA)=2} и $\Phi_3(T_2)$ = {AAA,f(AAA)=6; AAC, f(AAC)=2; CAA,f(CAA)=2; ACA,f(ACA)=2} дают идентичные последовательности рангов и, следовательно, $W_3 = 1$.

С увеличением l частотные характеристики будут представлены, в основном, одиночными 1-граммами и значения коэффициентов конкордации W_1 становятся "статистически незначимыми", так как наибольший вклад в W_1 будут вносить "связанные" ранги. Рекомендуемые значения l - 2,3,4.

Коэффициент конкордации не учитывает порядок следования 1-грамм, однако отражает некоторый их "баланс". Его удобно использовать для сравнения длинных текстов. Коэффициент W_1 легко вычисляется и для более чем двух последовательностей,что очень удобно при определении сходства последовательности с кластером или двух кластеров.

Для оценки значимости полученных значений коэффициента можно использовать аппроксимацию W_1 Z-распределением Фишера [1].

Далее предлагается использовать коэффициент конкордации W_1 как меру таксономической (при 1 = 2,3), "геномной" (при 1 = 3, где каждая триграмма находится в фазе с рамкой считывания) близости нуклеотидных последовательностей и функциональной близости аминокислотных последовательностей при 1 = 1.

3. Анализ поведения коэффициента конкордации на гомологичных последовательностях

На рис.1 представлены значения коэффициента конкордации порядка 1 = 1,2,3,4,5 для последовательностей, близость которых оценивалась ранее с помощью других методик (все цифры взяты из [4]). Так геномы бактериофагов φx174 и s13 являются практически идентичными и имеют только 2.06% несовпадений, а мера S_{2-5} для них равна 0.96. Коэффициенты конкордации 1-го порядка для этих текстов также высоки и близки к 1 (линия A). Различия между двумя env-генами вируса СПИДа (HIV) и гомология HIV и вируса Simian immunodeficiency ($STLV\text{-}III_{agm}$) составляют соответственно 14.59% и 55%, значения S_{2-5} равны приблизительно 0.85 и 0.56 соответственно, значения же коэффициентов конкордации очень высоки (линии B и C). Гомология между аминокислотными последовательностями Myosin heavy chain genes крысы и курицы составляет 83%, нуклеотидные же последовательности демонстрируют более слабое сходство как по лингвистической мере S_{2-5}, так и по коэффициенту конкордации (линия D). Наибольшее

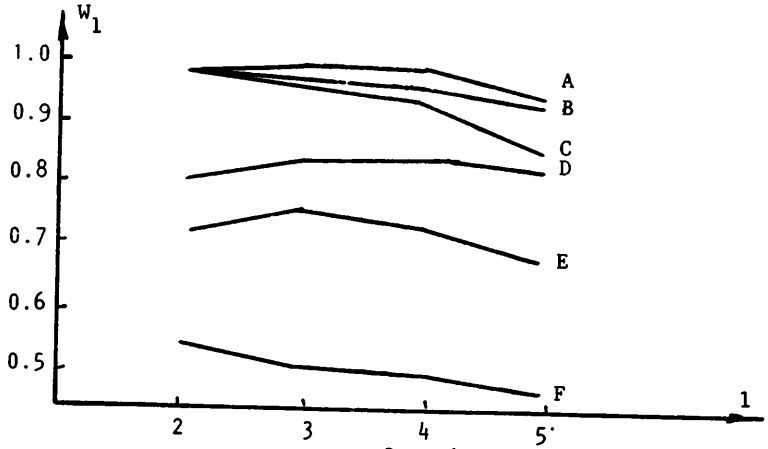

Рис. 1

сходство наблюдается между E.coli unc operon и R.blastica atp operon (линия E) при 1 = 3 и 2,лингвистическая же мера для них невелика и равна 0.42, несмотря на то, что обе эти последова - тельности относятся к одной таксономической группе - граммнега-тивным бактериям.Сходство эукариотической и прокариотической последовательностей Rat myosin heavy chain gene и R. blastica невелико (линия F) и не превышает сходства между случайными последовательностями.

Таким образом,сходство последовательностей, подмеченное с помощью других методик, проявляется и по значениям коэффициен-тов конкордации.

4. Таксономическое сходство последовательностей по W_2

Таксономическое сходство последовательностей предлагается оценивать по коэффициенту конкордации 2-го порядка,соответству-ющего использованию динуклеотидов. С этой целью было проанали-зировано 13 полных геномов из 4 таксономических групп,предста-вляющих растения, прокариоты, митохондрии и вироиды (по 3-4 по-следовательности из каждой группы).Значения коэффициентов пред-ставлены в таблице (нижний треугольник).

Таблица

Попарные значения коэффициентов W_2 и W_3 для последовательностей из разных таксономических групп

Имена последовательностей	Коэффициенты конкордации												
	1	2	3	4	5	6	7	8	9	10	11	12	13
1 Potato patatin	-	.6926	.9574	.4005	.1259	.4275	.4679	.8443	.6585	.9336	.2281	.3127	.4073
2 Maize sucrose	.7561	-	.7793	.5583	.4483	.6801	.5452	.6415	.5655	.6744	.4614	.5892	.6145
3 Soybean uricase	.9868	.8010	-	.4762	.1930	.4952	.5285	.8223	.6341	.9138	.2517	.3571	.4815
4 E.coli unc oper.	.5059	.6229	.5103	-	.7631	.7588	.8795	.3488	.3264	.3774	.5235	.5587	.4448
5 R.blastica atp	.1206	.4478	.1338	.7324	-	.7221	.7065	.1828	.2936	.1359	.7064	.7240	.5643
6 N.gonorrhoea	.4912	.7171	.5456	.7706	.6662	-	.7475	.4526	.4511	.4468	.5729	.6917	.5598
7 Salmonella	.5206	.5795	.5176	.8809	.6941	.8250	-	.3968	.3048	.4310	.5022	.4814	.4270
8 Clawed frog mt.	.8588	.6288	.8618	.2897	.0941	.4912	.3544	-	.9232	.9501	.2997	.4101	.5010
9 Human mt.	.5985	.5125	.6162	.2735	.3029	.4971	.2971	.8838	-	.7976	.3967	.5007	.5332
10 Fruit fly mt.	.9103	.6906	.9221	.3250	.0588	.4809	.3559	.9721	.7853	-	.2770	.3610	.4937
11 Coconut viroid	.1239	.3203	.1379	.6187	.8370	.6327	.5878	.2176	.4167	.1881	-	.7927	.7207
12 Cucumber pale	.2873	.5803	.3359	.5272	.6979	.6413	.4448	.3896	.4882	.4058	.8332	-	.7928
13 Chrysanthemum	.3947	.5667	.4330	.4080	.5390	.6031	.4138	.5007	.5655	.5000	.7884	.8650	-

Как видно из таблицы, значения коэффициента внутри групп в общем случае достаточно высоки и низки для последовательностей из разных таксономических групп. Исключение составляют некоторые последовательности из групп растений и митохондрий. Их сходство можно объяснить тем, что обе эти группы относятся к эукариотам. Высокое значение коэффициента конкордации между R.blastica atp operon и Cadang-cadang coconut viroid можно считать мало достоверным, так как длина вироида слишком мала (287 нуклеотидов).

В верхней части таблицы даны для сравнения значения коэффициента конкордации 3-го порядка, соответствующего использованию триплетов. Отметим, что значения коэффициентов при $l = 3$ падают, однако общая картина сходства сохраняется.

Результаты последовательной кластеризации с коэффициентом W_2 в качестве меры близости представлены на дендрограмме рис.2. В скобках приведены длины геномов в нуклеотидах.

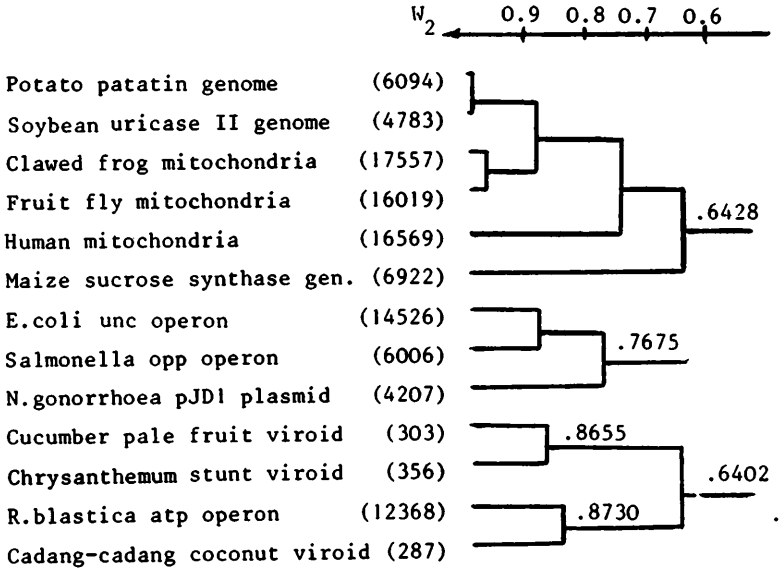

Рис. 2

Как видно из рис.2,на первом этапе при достаточно высоких значениях коэффициента конкордации выделяются группы растений (за исключением Maize sucrose synthase genome),митохондрий (за исключением Human mitochondria), прокариот (без R.blastica atp operon) и вироидов (без Cadang-cadang coconut viroid). Образование смешанного кластера из R.blastica и coconut viroid,как уже отмечалось ранее, можно объяснить большой разницей в дли - нах последовательностей. Группа растений и митохондрий вместе с некластеризованными на предыдущем этапе последовательностями образует группу эукариот. Вироиды же объединяются со смешанным кластером и за исключением R.blastica образуют группу вироидов.

Отметим, что порядок кластеризации последовательностей, промежуточные и результирующие кластеры при использовании в качестве меры близости коэффициента конкордации 3-го порядка полностью совпадают с результатами кластеризации по W_2, что согласуется как с выводами Нуссинов об асимметрии в частотах встречаемости динуклеотидов для различных таксономических групп [5], так и с аналогичным результатом Блайсделла для триграмм [6].

5. "Геномное" сходство последовательностей по $W_{кодон}$

Согласно геномной гипотезе, как отмечалось ранее, все гены одного генома или близкородственных геномов, независимо от функций кодируемых ими белков, придерживаются одной стратегии выбора синонимичных кодонов, но в рамках этой стратегии высокоэкспрессируемые и низкоэкспрессируемые гены могут иметь собственные подстратегии.

Рассмотрим кодирующие области из анализировавшихся ранее геномов E.coli unc operon (13 генов), R.blastica atp operon (7 генов), Salmonella opp operon (4 гена), Clawed frog mito - chondria (12 генов), Human mitochondria (12 генов) и Fruit

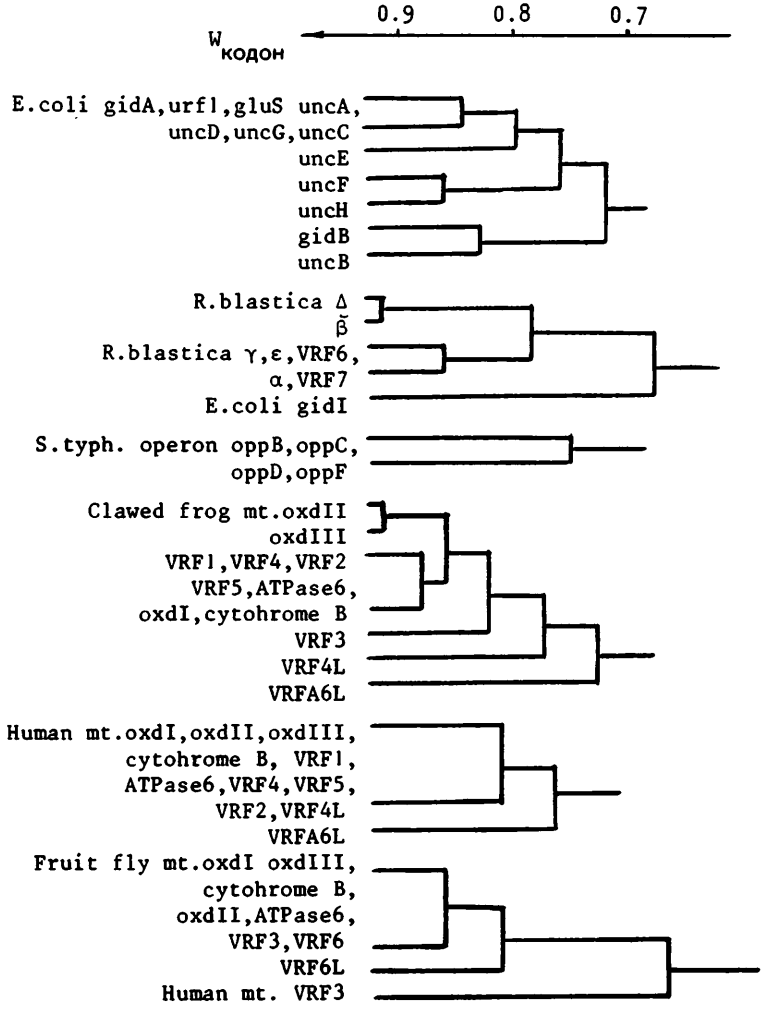

Рис.3. Дендрограмма последовательной кластеризации генов из разных геномов

fly mitochondria (8 генов). В качестве меры близости будем использовать коэффициент конкордации 3-го порядка по 1-граммам, находящимся в фазе с рамкой считывания (такие 3-граммы называются кодонами).

Как видно из результатов кластеризации по $W_{кодон}$, представленных на дендрограмме рис.3, гены одного генома образуют устойчивые кластеры при достаточно высоких значениях коэффициента конкордации. Исключения составляют ген uncI E.coli unc operon (длина 390 нуклеотидов) и ген VRF3 Human mitochondria (длина 345 нуклеотидов), интерпретация этого факта выходит за рамки данной статьи.

Отметим тот факт, что гены одного генома на промежуточных этапах последовательной кластеризации могут оказаться в разных кластерах, что может свидетельствовать о разной экспрессии генов.

6. Функциональная близость белковых последовательностей

Для оценки функциональной близости белковых последовательностей предлагается использовать коэффициент конкордации 1-го порядка на аминокислотных последовательностях. Для анализа использовались рибосомальные протеины, репрессоры и рецепторы E.coli с длиной более 200 аминокислот.

Как следует из рис.4, явно выделяется группа рецепторов. Исключение составляет E.coli aspartate chemoreceptor, демонстрирующий бóльшее сходство с репрессором E.coli Gal ETK. Данный рецептор демонстрировал наибольшее сходство с группой репрессоров и по методике Трифонова [4], который отметил, что в настоящее время отсутствуют объяснения этому факту. С достаточно высоким значением коэффициента конкордации выделяется группа рибосомальных протеинов E.coli, причем большие единицы (L3, L4,L2) и малые (S4,S3,S1) не образуют отдельных кластеров. Два

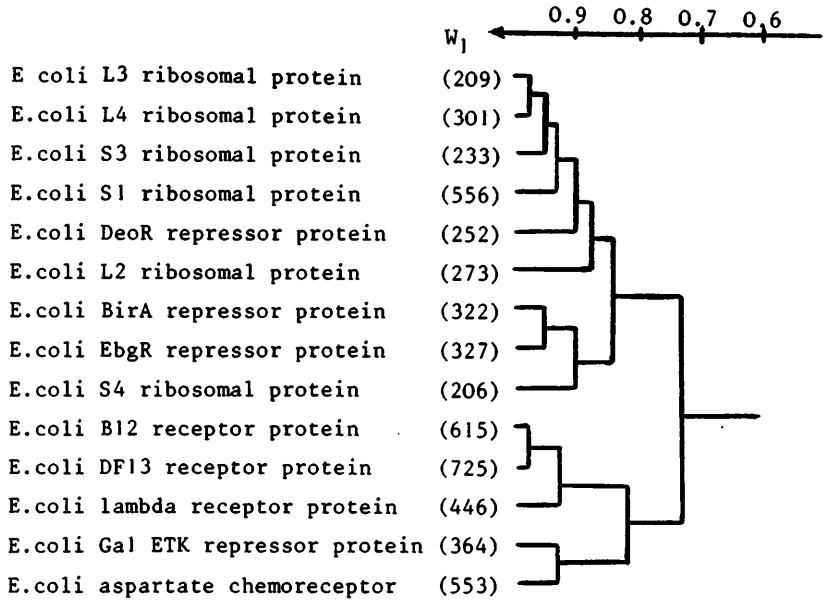

Рис. 4

репрессора E.coli BirA и EbgR объединяются в кластер с высоким значением коэффициента (0.9463). Неправильную кластеризацию репрессора E.coli DeoR (длина 252 аминокислот) и протеина E.coli S4 (длина 206) можно объяснить их малой длиной.

З а к л ю ч е н и е

Рассмотренная методика оценки близости (сходства) генетических текстов свободна от необходимости текстуального сравнения последовательностей (поиска гомологий, максимально длинных подпоследовательностей и т.п.), представляющего нетривиальную в вычислительном отношении задачу. Методика проста, не требует выравнивания последовательностей и позволяет сравнивать после-

довательности, сильно различающиеся по длине. Возможности методики продемонстрированы на задачах оценки таксономической, "геномной" и функциональной близости нуклеотидных и аминокислотных последовательностей. Полученная кластеризация в основном соответствует существующей биологической. Методика может использоваться для предварительной спецификации последовательности и оценки ее сходства с другими последовательностями.

Автор благодарит В.Д.Гусева за плодотворные обсуждения и Институт антропологии и генетики человека (г.Фрайбург, ФРГ) за предоставленную возможность работы с базой данных EMBL.

Литература

1. КЕНДЭЛ М. Ранговые корреляции. - М.: Статистика,1975.

2. ВЫСОЦКАЯ Г.С., ГУСЕВ В.Д., КУЛИЧКОВ В.А. Метод выявления информативных зон в генетических знаках пунктуации //Теоретические исследования и банки данных по молекулярной биологии и генетике: Сб. науч. тр. - Новосибирск, 1986. -С.54-58.

3. GUSEV V., CHUZHANOVA N. The algorithms of recognition of the functional sites in genetic texts //Proc. of the Workshop on Algorithmic Learning Theory. 1990, Japan. - P.109-119.

4. PIETROKOVSKI S., HIRSHON J., TRIFONOV E.N. Linguistic measure of taxonomic and functional relatedness of nucleotide sequences //J.Biomolec. Struct. Dyn.- 1990. - Vol.7.- P.1251-1268.

5. NUSSINOV R. Strong duplet preferences in nucleotide sequences and DNA geometry //J.Mol. Evol. - 1984. - Vol. 20. - P.111-119.

6. BLAISDELL B.E. A measure of similarity of sets of sequences not requiring sequence alignment //Proc. Natl. Acad. Sci. USA. - 1986. - Vol. 83. P. 5155-5159.

7. GRANTHAM R. et al. Codon catalog usage and genome hypothesis //Nucl. Acids Res. - 1980. - Vol. 8. -P. 49-62.

8. GRANTHAM R. et al. Codon catalog usage is a genome strategy modulated for gene expressivity //Nucl. Acids. Res. - 1981. - Vol. 9. - P. r43-r74

9. ROWE G.W., SZABO V.L., TRAINOR L.E.H. Cluster analysis of genes in codon space //J.Molec. Evol. - 1984. - Vol. 20. - P. 167-174.

10. GOUY M., GAUTIER C. Codon usage in bacteria: correlation with gene expressivity //Nucl.Acids Res. - 1982.-Vol. 10. – P. 7055-7074.

11. PEVZNER P.A., BORODOVSKY M.Yu., MIRONOV A.A. I. The significance of deviation from mean statistical characteris - tics and prediction of the frequency of occurrence of words //J. Biomolec. Struct. Dyn. - 1989. - Vol. 6. -P. 1013-1026.

Поступила в ред.-изд.отд.
21 ноября 1994 года