АНАЛИЗ ВРЕМЕННЫХ РЯДОВ И СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ
(Вычислительные системы)
1991 год Выпуск 141

УДК 519.764

АНАЛИЗ СЕРИЙ В ГЕНЕТИЧЕСКИХ ТЕКСТАХ

В.Д.Гусев, Л.А.Немытикова

В в е д е н и е

Проявления повторности в генетических текстах чрезвычайно многообразны. К примеру, различные типы повторов в ДНК (пря - мые, инвертированные, комплементарные) имеют отношение к образованию вторичной структуры РНК, участвуют в формировании "знаков пунктуации" (участков, ответственных за регуляцию основных генетических процессов), выступают посредниками в ходе эволю - ционного процесса, реализуемого в виде множественных замен,делеций, вставок и т.д.

Одной из разновидностей повторов являются серии - цепочки из однотипных элементов, ограниченные по краям элементами другого типа. Анализу серий в генетических текстах не уделялось слишком большого внимания, хотя были отмечены отдельные инте - ресные эффекты (например, поли-А-тенденция в геномах эукариот [1]). Причину этого, по-видимому, следует искать в том, что традиционно серии считались малоинформативным объектом по сравнению с повторами общего вида и основные усилия были на - правлены в сторону поиска гомологий - одной из ключевых проб - лем в исследовании генетических текстов.

Цель настоящей работы - подчеркнуть преимущества методо - логического и вычислительного характера, связанные с использо-

ванием серий как языка описания генетических текстов, и продемонстрировать нестандартные возможности использования разработанного аппарата анализа серий.

Эксперимент проводился преимущественно на различных сегментах генома вируса гриппа, что позволило выявить ряд структурных особенностей этого генома, не отмечавшихся ранее в литературе и представляющих интерес в эволюционном и классификационном плане.

1. Обоснование подхода

Сформулируем более подробно соображения, которыми мы руководствовались, выбирая серии в качестве объекта для исследования. Они же легли в основу разработанного нами программного комплекса анализа серий.

1.1. В последних работах по выявлению и интерпретации серий в генетических текстах просматривается весьма перспективная, на наш взгляд, тенденция к агрегированию (группированию) элементов исходного алфавита [2-4]. Алфавит ДНК-молекул состоит из 4 элементов (A - аденин, G - гуанин, C - цитозин, T - тимин), которые допускают различные варианты осмысленного объединения. Так, по типу азотистого основания, входящего в состав нуклеотидов, их можно разделить на пурины (A , G) и пиримидины (C , T); по числу водородных связей, возникающих при комплементарном спаривании, нуклеотиды можно разделить на слабые (A , T - 2 связи) и сильные (C , G - 3 связи) и т.п. (см. для иллюстрации [2,3]).

Алфавит аминокислот состоит из 20 элементов. Вариантов осмысленного группирования здесь значительно больше, чем у нуклеотидов. Можно объединять аминокислоты в группы по таким параметрам, как наличие или отсутствие заряда [4], предрасположенность к вхождению в α-спирали или β-участки (элементы

вторичной структуры белков), степень мутабельности (относительная частота заменяемости остатка в ходе эволюции) и т.п.

Агрегирование подчеркивает функциональную близость отдельных элементов алфавита по некоторому свойству. Если, к примеру, алфавит делится на два подмножества, элементы которых кодируются, соответственно, нулями и единицами, то наличие в тексте аномально длинных серий нулей (единиц) или характерных их чередований легко трактуемо на содержательном уровне и служит хорошей основой для последующего (уже целенаправленного) поиска гомологий.

Идея агрегирования созвучна наметившейся тенденции к выявлению "размытых" гомологий с помощью разного рода "профилей" последовательности, которые количественно (в сильной шкале) оценивают степень проявления какого-либо свойства (гидрофильности, мутабельности, заряженности и т.п.) в каждой позиции последовательности. Если агрегирование алфавита проводится по тому же параметру, по которому строится профиль, то получаемая двоичная последовательность является грубой аппроксимацией (в слабой шкале) кривой профиля. Ее анализ удобно проводить с помощью серий.

В разработанном авторами комплексе программ идея агрегирования доведена до "логического конца": для небольших алфавитов (в частности, для НК-молекул) предусмотрена возможность получения всех возможных разбиений элементов алфавита на два подмножества. Для алфавитов большей мощности (например, для аминокислот) вариант разбиения задает сам пользователь.

1.2. Серии перспективны не только в плане выявления достаточно протяженных функционально близких фрагментов. Анализ коротких серий (длины 1,2) позволяет обнаружить неслучайности в чередовании отдельных элементов, которые подчас носят столь же универсальный характер, как правила динуклеотидного предпочтения Р.Нуссинов [5].

Локализация коротких серий вдоль последовательности представляет интерес при выявлении эффекта кластеризации редких событий. Последние возникают при сильном нарушении баланса между нулями и единицами. Число серий в таких последовательностях невелико,и для хранения информации об их расположении не требуется больших затрат памяти.

1.3. Алгоритм подсчета числа серий разной длины имеет линейную трудоемкость и очень прост в логическом отношении.Заметим,что для выявления повторов общего вида также созданы линейные алгоритмы,но лежащая в их основе техника (префиксные и суффиксные деревья,ациклические графы слов) довольно сложна, что ведет к большой мультипликативной константе в оценке трудоемкости.Более того,эта техника перестает работать при переходе от совершенных повторов к несовершенным,тогда как при переходе от идеальных серий к сериям с дефектами практически мало что меняется.

Отмеченная легкость вычисления серий позволяет использовать имитационный подход к оценке их статистической значимости. Это имеет существенное значение при изменении модели случайной последовательности на входе и при использовании нестандартных статистик, достаточно сложным образом зависящих от числа серий и длины максимальной серии успехов. Заметим, что известные результаты [6-8], за редким исключением [9], касаются последовательности независимых испытаний.

2. Описание программного комплекса

Рассмотрим случай, когда размер алфавита n мал.

Осуществляем все возможные разбиения алфавита на 2 непустых подмножества (порядок подмножеств несуществен). Таких разбиений будет $2^{n-1}-1$. Соотнесение этой величины с вычислительными ресурсами пользователя позволяет определить допустимые размеры алфавита.

Для каждого разбиения проводим агрегирование алфавита. За один просмотр агрегированной (двоичной) последовательности вычисляем следующие характеристики: r_{0j} - число серий нулей длины j ($j=1,2,\ldots,l_{0max}$, где l_{0max} - длина максимальной серии нулей); r_{1j} - число серий единиц длины j ($j=1,2,\ldots,l_{1max}$, где l_{1max} - длина максимальной серии единиц); $r = r_{0j} + r_{1j}$ - общее число серий нулей и единиц, $S = S_{0j} + S_{1j}$ - число разновидностей серий нулей и единиц (в диапазоне от 1 до l_{0max} и соответственно от 1 до l_{1max} могут быть представлены не все серии); $r_0(k) = \sum_{j=1}^{k} r_{0j}$ - число серий нулей с длиной, не превышающей k; $d(k)$ - длина максимального фрагмента, в котором расстояния между соседними единицами не превышают k (две последних статистики используются для выявления кластеризации редких событий, кодируемых единицами).

Трудоемкость в среднем данного этапа - $O(N)$, где N - длина последовательности. Линейность обеспечивается за счет быстрого поиска нужного счетчика при фиксации очередной серии. Трудоемкость поиска не зависит от числа счетчиков S (используется один из вариантов процедуры хеширования). При выборе размера расстановочного поля можно руководствоваться оценкой максимально возможного значения параметра S ($2\sqrt{N}$ для специально сконструированной последовательности с равным балансом 0 и 1).

После вычисления на реальной последовательности указанных выше характеристик проводится имитационное моделирование с целью выявления возможных аномальностей в полученных значениях. При этом предполагается, что имитационным аналогом исходного текста в первом приближении может служить последовательность независимых одинаково распределенных случайных величин с тем же (что у реальной последовательности) частотным составом

элементов. Такие последовательности получаются из исходного текста путем его случайного перемешивания. Более сложные модели, отражающие специфические особенности генетического текста (например, правила динуклеотидного предпочтения, локальные неоднородности в распределении нуклеотидов и т.п.), могут быть получены с помощью более сложных схем перемешивания.

При проведении имитационного моделирования задается параметр m - число повторений эксперимента с новой случайной последовательностью (мы использовали значение $m = 100$). В каждом эксперименте определяются указанные выше характеристики серий. По результатам m экспериментов для каждого параметра α вычисляются его минимальное и максимальное значения

$$\alpha^{сл}_{min} = \min_{i=1-m} \{\alpha_i\}, \qquad \alpha^{сл}_{max} = \max_{i=1-m} \{\alpha_i\},$$

среднее $\bar{\alpha} = \frac{1}{m} \sum_{i=1}^{m} \alpha_i$ и среднеквадратичное отклонение

$$s = \left[\frac{1}{m} \sum_{i=1}^{m} (\alpha_i - \bar{\alpha})^2 \right]^{1/2} .$$

Если значение параметра α, наблюдаемое на реальной последовательности, выходит за границы интервала $\alpha^{сл}_{min} - \alpha^{сл}_{max}$, полагаем его аномальным и предполагаем, что соответствующая аномалия является функционально значимой. Многочисленные подтверждения этому обнаруживаются при анализе геномов с известными структурно-функциональными особенностями.

Трудоемкость эксперимента с полным набором агрегирований и имитационным моделированием составляет $O(2^{n-1} \cdot N \cdot m)$. При заданном варианте агрегирования трудоемкость снижается до $O(N \cdot m)$.

Ниже приводятся примеры использования комплекса для исследования генетических текстов. Рассматриваются три типа задач: 1) выявление аномалий в агрегированных нуклеотидных последова-

тельностях; 2) анализ мутационных замен в гомологичных последовательностях; 3) выявление неравномерностей в распределении аминокислот по длине белковой макромолекулы.

В качестве объекта для исследований выбран геном вируса гриппа. Решение задач 1 и 3 способствует выявлению интегральных признаков, объединяющих различные сегменты генома. Решение задачи 2 способствует прояснению характера эволюции дрейфовых и сдвиговых вариантов генома.

3. Выявление аномалий в агрегированных нуклеотидных последовательностях

Обрабатывались все 8 сегментов РНК-содержащего генома вируса гриппа (штамм PR – Пуэрто-Рико, тип A) и отдельные сегменты генома типа B (штамм Lee). Каждый сегмент вируса гриппа, как правило, кодирует один белок (их условные обозначения: HA - гемагглютинин, NA - нейраминидаза, P1, P2, P3 - полимеразы 1,2 и 3 соответственно, NP - нуклеопротеин, M - мембранный белок, NS - неструктурный белок). Наибольший интерес вызывают белки HA и NA, которые определяют антигенную изменчивость вируса гриппа и по которым проводится разбиение типов на подтипы (по крайней мере, для типа A).

Каждый из сегментов вируса гриппа был обработан по схеме с полным агрегированием алфавита, описанной выше. Фиксировались аномальности трех типов, связанные с параметрами r (общее число серий), l_{1max} и l_{0max} (максимальные длины серий из 0 и 1), r_{ij} (число серий типа i, $i = 0,1$, и длины j). По результатам экспериментов можно сделать следующие выводы.

3.1. Выявлено 3 типа агрегирований с аномальным поведением параметра r для большинства сегментов вируса гриппа. Первое из них соответствует разбиению алфавита $\Sigma = \{A,T,G,C\}$

Таблица 1

Характеристики серий для разных сегментов вируса гриппа при агрегировании: $W = \{A,T\}$ и $S = \{C,G\}$

Сегмент	Длина	r	$\bar{r}$	r^{max}	r_{11}	r^{max}_{11}	r_{12}	r^{max}_{12}	r_{01}	r^{max}_{01}	r_{02}	r^{max}_{02}
HA(PR)	1778	919	865	939	210	210	99	125	282	291	110	124
HA(Lee)	1882	985	915	968	214	220	132	135	287	300	143	131
NP(PR)	1565	875	781	833	224	211	115	114	253	240	116	119
NP(Lee)	1841	960	894	952	200	215	142	128	288	299	125	133
NS(PR)	890	487	437	468	120	121	58	70	147	148	62	70
NS(Lee)	1096	576	533	569	127	130	76	80	180	181	67	85
NA(PR)	1413	711	694	736	166	179	76	107	204	225	89	112
NA(Lee)	1557	815	759	819	179	187	105	121	253	259	100	120
M(PR)	1027	555	512	563	141	149	73	80	148	162	78	81
M(Lee)	1191	605	566	609	126	131	65	86	185	199	83	84
P1(PR)	2341	1239	1150	1204	268	281	157	170	381	367	154	162
P2(PR)	2233	1187	1090	1165	252	268	168	157	361	365	161	161
P3(PR)	2341	1277	1150	1215	305	290	160	173	383	367	166	180

Примечание: характеристики с индексом "max" фиксируют рекордные показатели, достигнутые в имитационном эксперименте со 100 случайными последовательностями ($W \to 1$, $S \to 0$).

на подмножества $W = \{A,T\}$ и $S = \{C,G\}$. При данном разбиении, как правило, выполняется соотношение $r > r^{сл}_{max}$, где $r^{сл}_{max}$ - максимальное (по серии из 100 экспериментов) значение параметра r для случайных последовательностей (относительное исключение составляют сегменты NA и M). Иными словами, элементы множеств W и S хорошо перемешаны друг с другом в исходной последовательности и чередуются более регулярно, чем это обычно имеет место для случайных последовательностей. Подобным свойством в естественных языках обладают элементы множеств "гласные" и "согласные".

Поскольку основной вклад в общее число серий

$$r = \sum_{j=1}^{l_{1max}} r_{1j} + \sum_{j=1}^{l_{0max}} r_{0j}$$

составляют серии длины 1 и 2, значения параметров r_{11}, r_{12}, r_{01} и r_{02} также близки к аномальным, т.е. зачастую превышают аналогичные рекордные показатели, достигнутые в имитационном эксперименте (см. для иллюстрации табл. 1). Наряду с этим значения параметров l_{1max} и l_{0max} близки к минимальным, т.е. достаточно длинные кластеры из элементов одного типа (W и S) встречаются редко. Речь может идти лишь о кластеризации на тандемном уровне, там, где велики значения параметров r_{12} и r_{02}.

Второй тип агрегирования соответствует делению алфавита на пурины и пиримидины: $Pu = \{A,G\}$, $Py = \{C,T\}$. Оно характеризуется аномально низким общим числом серий для большинства сегментов ($r < r^{сл}_{min}$, где $r^{сл}_{min}$ - минимальное (по серии из 100 экспериментов) значение параметра r для случайных последовательностей). Это означает, что элементы каждого из множеств проявляют тенденцию к образованию кластеров-

фрагментов, состоящих только из пуринов или пиримидинов. Характерные размеры кластеров - 5-8 символов. Как и в предыдущем случае, параметр r положительно коррелирован с параметрами r_{11}, r_{12} и (в меньшей степени) с r_{01} и r_{02}. Все они, в силу указанного обстоятельства, часто характеризуются аномально низкими значениями (см. для иллюстрации табл. 2).

Третий тип агрегирования соответствует делению алфавита на подмножества $M = \{A,C\}$ и $K = \{G,T\}$. Он также характеризуется аномально низким общим числом серий для большинства сегментов (см. табл. 3). Эффект проявлен в более слабой форме, чем в предыдущем случае (больше исключений (NP, NS, M, P3 - все из штамма PR); параметры r_{12} и r_{02} ближе к средним значениям, чем к минимальным).

Все остальные типы агрегирований не обнаруживают столь устойчивых тенденций к аномальному отклонению общего числа серий от показателей, характеризующих случайные последовательности. Отдельные исключения относятся к категории, описываемой неравенством $r < r^{min}$, и, возможно, неявно вытекают из результатов агрегирования второго и третьего типа.

Обнаруженные закономерности в чередовании различных элементов НК-алфавита носят, по всей видимости, достаточно общий характер. В этом отношении они напоминают известные правила динуклеотидного предпочтения [5], хотя не являются следствием последних, поскольку при одном и том же частотном составе двух последовательностей характер чередования элементов в них может быть совершенно различным.

Как и правила динуклеотидного предпочтения, закономерности чередования, по-видимому, различны для прокариотических и эукариотических геномов. Об этом свидетельствует наш ограниченный эксперимент с вирусными геномами $\phi X174$ и $SV40$ (соответственно прокариотического и эукариотического типа). Закономерности чередования во втором случае были теми же, что и у генома

Таблица 2

Характеристики серий для разных сегментов вируса гриппа при агрегировании: $Pu = \{A,G\}$ и $Py = \{C,T\}$

Сегмент	Длина	r	$\bar{r}$	r^{min}	r_{11}	r^{min}_{11}	r_{12}	r^{min}_{12}	r_{01}	r^{min}_{01}	r_{02}	r^{min}_{02}
HA(PR)	1778	802	868	801	171	148	86	81	213	214	98	84
HA(Lee)	1882	858	933	871	187	178	100	91	224	211	104	92
NP(PR)	1565	670	760	714	123	137	83	71	184	192	73	72
NP(Lee)	1841	798	898	856	145	158	100	84	207	232	105	88
NS(PR)	890	378	439	405	71	76	43	35	97	97	43	37
NS(Lee)	1096	483	536	494	92	88	62	48	127	129	68	50
NA(PR)	1413	666	698	662	124	128	99	63	176	166	83	59
NA(Lee)	1557	736	772	721	169	138	88	73	201	187	86	80
M(PR)	1027	468	509	467	98	88	53	46	117	114	64	45
M(Lee)	1191	502	578	540	99	94	54	51	138	148	53	51
P1(PR)	2341	1056	1150	1088	203	218	131	120	280	291	131	114
P2(PR)	2233	956	1090	1039	193	203	100	107	242	268	117	108
P3(PR)	2341	1050	1130	1079	192	200	147	110	290	293	138	113

Примечание: характеристики с индексом "min" фиксируют нижние границы, полученные в имитационном эксперименте со 100 случайными последовательностями $(Pu \to 1,\ Py \to 0)$.

Таблица 3

Характеристики серий для разных сегментов вируса гриппа при агрегировании: M = {A,C} и K = {G,T}

Сегмент	Длина	r	$\bar{r}$	r^{min}	r_{11}	r^{min}_{11}	r_{12}	r^{min}_{12}	r_{01}	r^{min}_{01}	r_{02}	r^{min}_{02}
HA(PR)	1778	804	885	830	178	170	96	89	182	195	106	87
HA(Lee)	1882	834	931	876	170	177	111	91	211	222	99	95
NP(PR)	1565	758	781	732	168	155	97	70	179	169	104	73
NP(Lee)	1841	841	910	854	167	174	105	91	216	207	106	92
NS(PR)	890	432	444	414	94	83	60	39	110	93	43	34
NS(Lee)	1096	508	549	505	110	101	60	49	121	122	70	47
NA(PR)	1413	626	704	659	134	152	91	67	132	148	75	67
NA(Lee)	1557	718	774	721	161	161	91	73	172	169	89	77
M(PR)	1027	498	513	464	122	106	61	48	125	99	49	50
M(Lee)	1191	550	596	556	115	115	74	55	146	137	58	55
P1(PR)	2341	1034	1175	1106	210	230	121	118	243	287	133	117
P2(PR)	2233	1030	1120	1058	228	236	125	115	224	258	152	111
P3(PR)	2341	1112	1170	1104	235	244	150	121	277	263	136	119

Примечание: обозначения те же, что и в табл.2 (M → 1 , K → 0).

вируса гриппа. В первом случае наблюдали несколько отличную картину.

Сфера действия выявленных закономерностей достаточно широка. Об этом свидетельствует совпадение (в целом) наших результатов с результатами объемного эксперимента, проведенного Блайсделом [3] на множестве кодирующих и некодирующих ДНК-последовательностей эукариотического типа. Блайсдел изучал агрегирования двух типов $(\Sigma = \{W,S\}$ и $\Sigma = (Pu, Py)$ и обнаружил, что кодирующие последовательности устойчиво характеризуются избытком коротких и дефицитом длинных серий из слабых $(W$ - "weak") и сильных $(S$ - "strong") нуклеотидов, классифицируемых по числу образуемых при спаривании водородных связей. Некодирующие же последовательности характеризуются дефицитом коротких и избытком длинных серий из пуринов (Pu) и пиримидинов (Py).

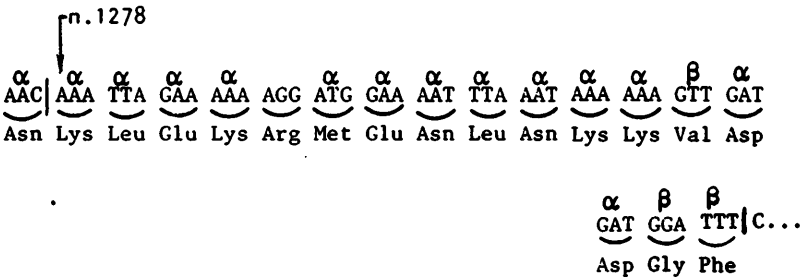
Существенное расхождение наших результатов с выводами Блайсдела заключается в том, что для вируса гриппа обе отмеченные закономерности выполняются на кодирующих последовательностях. Это может быть объяснено различием в исходном материале. Если учесть, что переход от эукариотических последовательностей к прокариотическим также вносит определенную специфику, можно заключить, что неслучайные регулярности в чередовании элементов разных классов в НК-последовательностях безусловно имеют место, но их характер и "сфера действия" подлежат дополнительному уточнению. Аналогичное замечание можно сделать и относительно гипотезы Блайсдела о возможности использования выявленных эффектов для классификации последовательностей по типу: "кодирующая" - "некодирующая".

Возможная трактовка (по Блайсделу) хорошего перемешивания W и S элементов заключается в создании оптимальных условий для расплетания двойной цепи ДНК в процессе репликации (в нашем случае эта трактовка не проходит). Избыток же длинных

серий из пуринов или пиримидинов может влиять на конформацию двойной спирали и способствовать освобождению ее из нуклеосомного комплекса в процессе репликации или транскрипции.

3.2. Большой интерес представляют максимально длинные серии нулей и единиц, выявляемые при различных вариантах агрегирования. Все они, как правило, являются функционально значимыми и расположены в характерных участках генома. Приведем примеры соответствующих серий.

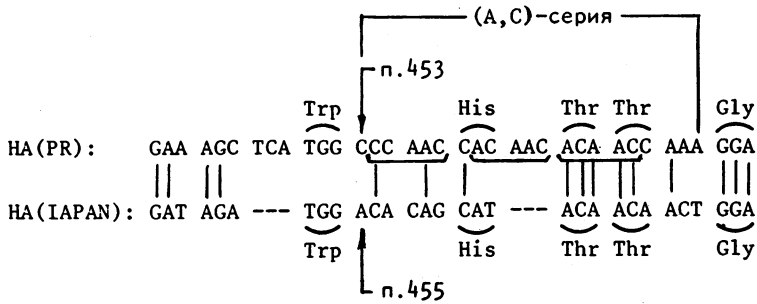
В сегменте HA (штамм PR) в позиции 1278 расположена серия (не C)-элементов длины 51:



Максимальная (не C)-серия, зафиксированная в эксперименте со 100 случайными последовательностями, имела длину $l_{0max}^{сл} = 43$, т.е. обнаруженную в HA серию следует признать аномальной. Анализ аминокислотного состава с точки зрения возможной вторичной структуры обнаруживает явное преобладание аминокислот, характерных для α-спиралей, среди первых 15 остатков (это характерная длина α-спирали). Профиль α/β-структурированности имел бы на данном участке явно выраженный максимум.

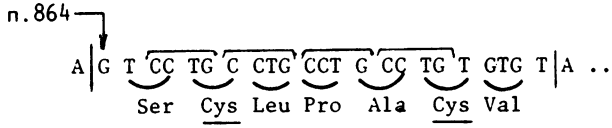
В том же сегменте гемагглютинина в поз.453 расположена аномальная (A,C)-серия длины 21 (соответствующий рекордный показатель для 100 случайных последовательностей равен 17). Дан-

ная серия расположена в зоне локализации антигенных детерми - нант и может рассматриваться в качестве "горячей точки" генома, в которой происходят перестройки блочного типа. Об этом свидетельствует сопоставление двух сдвиговых вариантов гемаг - глютинина: **HA PR** (подтип H1N1) и HA JAPAN (подтип H2N2). Имеющееся у штаммов PR и JAPAN различие в длинах тяжелых цепей гемагглютинина, обозначаемых HA1 , реализовано в виде двух вставок по 3 нуклеотида каждая, локализо - ванных как раз в области расположения аномальной (A,C)-серии:



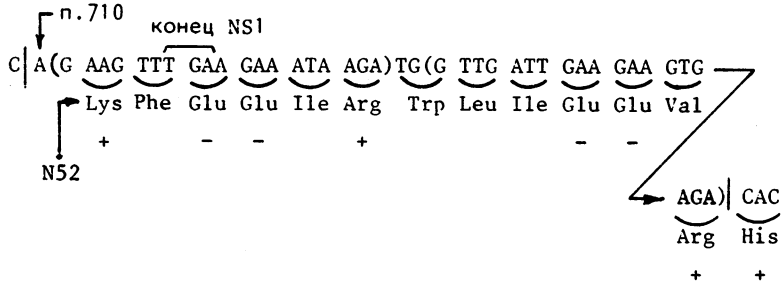
Наличие в (A,C)-серии периодичности вида $(ACAAC)^3$ позволяет предположить, что аномальный размер серии объясняется внутренней дупликацией. О функциональной значимости фрагмента косвенно свидетельствует наличие в его составе редких и структурно важных аминокислот - триптофана (Trp) и гистидина (His) .

В сегменте NP (штамм PR) в позиции 864 расположена (не A)-серия длины 23 (относительная частота появления подобных серий в имитационном эксперименте - 0.04). Серия имеет уникальную внутреннюю структуру в виде периодичности $(CCTG)^4$:



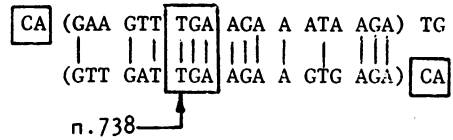
О функциональной значимости фрагмента свидетельствует наличие в аминокислотной последовательности двух (из шести имеющихся в NP) цистеинов, играющих ключевую роль в формировании вторичной структуры белка.

В сегменте NS (штамм PR) в позиции 710 выявлена (не C)-серия длины 41 ($l^{сл}_{omax}$ = 38):



Эта серия расположена в очень характерном месте. Сегмент NS содержит два частично перекрывающихся гена (NS1 и NS2). Начало (не C)-серии соответствует концу зоны перекрытия (в поз. 717 расположен терминальный кодон гена NS1). Остав - шаяся часть серии (п. 720-750) является кодирующей последова - тельностью только для NS2 .

Структура серии имеет дупликативный характер. Круглыми скобками выделены гомологичные фрагменты:



В гене NS2 им соответствует повтор заряженных аминокислот (Glu Glu – Arg). Вследствие точечных мутаций в терминальном кодоне NS1 он может утратить свои функции. Роль терминального кодона тогда принимает на себя триплет TGA (поз. 738) во втором гомологичном фрагменте, что приводит к удлинению гена NS1 в некоторых штаммах (например, в штамме Udorn (подтип H3N2)).

Дупликативная структура (не C)-серии, возможность мутаций, приводящих к существенным структурным изменениям в гене NS1, наличие коротких фланкирующих повторов (CA) косвенно свидетельствуют о том, что и данную серию потенциально можно отнести к "горячим точкам" генома.

Интересно отметить, что и конец NS2 также характеризуется аномальной (A,T)-серией длины 18 ($l^{сл}_{1max} = 17$), расположенной в поз. 857:

GC|TTATT $\overline{\text{TAA}}$ $\overline{\text{TAA}}$ $\overline{\text{TAA}}$ AAAA|CA

n.857 конец NS2

Терминальный кодон здесь также задублирован ((TAA)^3), и потенциально имеется возможность незначительного удлинения гена NS2 путем точечных мутаций.

Приведенные примеры свидетельствуют о том, что аномально длинные серии выявляются при разных вариантах агрегирования, часто расположены на границах структурных областей (например, на концах генов) и могут содержать важные структурные элементы. Возможны и обратные ситуации, когда максимально длинные серии оказываются аномально короткими. Так, максимально длинная (C, G)-серия в сегменте NS (штамм PR) состоит из 6 элементов, тогда как минимальное (по 100 экспериментам) значение этого параметра для случайных текстов равно 7. Описанные ситуации до-

вольно редки и характерны, в основном, для агрегирования $\{W, S\}$, что объясняется отмеченным выше эффектом регулярной чередуемости W и S элементов.

3.3. Параметры r_{ij} также могут принимать аномально низкие или высокие значения. Чаще всего это имеет место для r_{01}, r_{11}, r_{02}, r_{12} и обусловлено аномальностью параметра r при том или ином типе агрегирования. Аномальности в значениях r_{ij} при $j > 2$ встречаются более редко и носят неустойчивый характер (т.е. выполняются далеко не для всех представителей одного и того же семейства, например, не для всех нейраминидаз или гемагглютининов). Трактовка их затруднена из-за отсутствия доминирующего мотива. В каждом конкретном случае эффект обусловлен сочетанием нескольких факторов.

Так, к примеру, в сегменте NA (Lee, тип B) аномально мало число (не T)-серий длины 3 ($r_{13} = 20$, тогда как $r^{min}_{13} = 30$). Иными словами, очень мало число комбинаций T --- T, где вместо пробела может стоять любой из трех элементов: A, G, C. В качестве объясняющих факторов можно привести следующие:

а) в кодирующей рамке должны отсутствовать комбинации TAA-T, TAG-T, TGA-T, содержащие терминальный кодон. Поскольку сегмент NA(Lee) кодирует два перекрывающихся гена (NA и NB), указанный выше запрет распространяется (в пределах зоны перекрытия, составляющей четверть сегмента) сразу на две кодирующие рамки. Заметим, что в пределах зоны перекрытия имеются лишь две (из 20) комбинаций T---T. Интересно также отметить, что сегмент NA(PR), где указанный эффект отсутствует, кодирует лишь один ген;

б) в силу CG-эффекта (аномально низкая частота встречаемости биграммы CG в эукариотических геномах) должно быть мало число комбинаций TCG-T и T-CGT (фактически тако-

вых не оказалось вовсе). У NA(PR) CG-эффект проявлен слабее, т.е. число биграмм CG больше, чем у NA(Lee), при меньшей длине сегмента. В итоге в NA(PR) имеются 4 комбинации указанного вида;

в) частоты некоторых аминокислот в обеих нейраминидазах существенно отличаются (то же можно сказать об использовании кодонов). Так, в нейраминидазе PR содержание триптофана значительно выше, чем в остальных сегментах штамма PR и в нейраминидазе Lee . Вследствие этого комбинации TGGGT и TGGAT , содержащие (в половине случаев) триптофан в первой кодирующей рамке, по 5 раз встречаются в NA(PR) и ни разу - в NA(Lee) . Аналогично, по содержанию аспарагина (Asn) нейраминидаза PR значительно опережает нейраминидазу Lee (ген NA). В соответствии с этим комбинация TGAAT , содержащая (в 4 случаях) аспарагин в третьей кодирующей рамке, 7 раз встречается в NA(PR) и ни разу - в NA(Lee) .

4. Анализ точечных мутаций с помощью серий

Наличие значительного количества секвенированных вариантов вируса гриппа (дрейфовых и сдвиговых) создает базу для анализа эволюционных замен с помощью серий. Как правило, дрейфовые варианты вируса гриппа, принадлежащие одному подтипу, эволюционируют посредством замен, т.е. длины их остаются неизменными. Это дает возможность сравнивать посимвольно сегменты дрейфовых вариантов и фиксировать результаты сравнения в виде последовательности 0 и 1, где 0 соответствует совпадению символов, а 1 - несовпадению (т.е. замене). Для полученной последовательности вычисляются серийные характеристики, а затем проводится имитационный эксперимент с перемешиванием 0 и 1 для оценки разброса соответствующих характеристик.

Представляют интерес такие показатели, как характер замен (что на что меняется), распределение числа серий нулей и единиц, возможность кластеризации замен, количественные различия между числом замен в сдвиговых и дрейфовых вариантах, локализация аномально длинных серий нулей (консервативных зон).

При сравнении сдвиговых вариантов, незначительно отличающихся по длине, предварительно с помощью алгоритмов динамического программирования осуществлялось выравнивание текстов, на основе которого определялись места вставок и делеций. Затем проводился анализ замен на участках одинаковой длины, не содержащих вставок и делеций.

По итогам данного эксперимента можно сделать следующие выводы.

4.1. По всем сегментам преобладают замены типа $A \longleftrightarrow G$ (пурин переходит в пурин) и $C \longleftrightarrow T$ (пиримидин - в пиримидин). Они составляют в среднем от 2/3 до 3/4 всех замен. Среди них незначительно преобладают замены типа $A \longleftrightarrow G$ (возможно, из-за повышенного содержания аденина во всех сегментах). На третьем месте, значительно уступая первым двум в количественном отношении, фигурируют замены типа $A \longleftrightarrow C$.

Подобный характер замен объясняется, по-видимому, преобладанием в генетическом коде аминокислот с вырожденностью 2 (их 12 из 20). Пары кодонов, кодирующие каждую из этих аминокислот, различаются лишь по третьим позициям, содержащим либо $\{A, G\}$, либо $\{C, T\}$, но не смесь их. Очевидно, что в этом случае лишь замена типа $A \longleftrightarrow G$ либо $C \longleftrightarrow T$ в третьей позиции не приведет к изменению аминокислоты. Преобладание замен такого типа есть отражение эволюционной стратегии, направленной на минимизацию изменений в первичной структуре.

4.2. Распределение числа серий нулей r_{0j} в зависимости от их длины j характеризуется явно выраженной периодичностью с длиной периода, равной 3. Пики этого распределения со-

ответствуют значениям $j = 2,5,8$ и т.д. При этом величины r_{02}, r_{05} и r_{08} часто намного превышают соответствующие рекордные показатели (r_{02}^{max}, r_{05}^{max} и r_{08}^{max}) имитационного эксперимента. Для иллюстрации в табл. 4 представлены величины r_{0j}, $j = 1\text{-}17$, характеризующие результаты сравнения сегментов полимеразы 3 у штаммов WSN (подтип H1N1) и NT (подтип H3N2). При длине сегментов $N_1 = N_2 = 2341$ зафиксировано замен - 221.

Таблица 4

Число серий нулей (r_{0j}) в (0,1)-последовательности, фиксирующей точечные замены в сегментах P3 (штаммы WSN и NT)

j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
r_{0j}	16	**29**	9	8	**28**	6	5	**25**	5	3	**18**	2	1	**6**	3	3	**10**
r_{0j}^{max}	28	29	26	22	21	20	23	17	19	14	15	12	14	12	11	9	9

Данный эффект объясняется теми же факторами, что были изложены в п. 4.1. Доминирование замен по третьим позициям кодонов приводит к увеличению числа серий нулей длины 2 (замены в третьих позициях i-го и $(i+1)$-го кодонов), длины 5 (замены в третьих позициях i-го и $(i+2)$-го кодонов) и т.д.

Заметим, что эффект наиболее отчетливо проявляется при "умеренном" числе замен (ориентировочно это соответствует десятипроцентному "зашумлению"). При значительно меньшем числе замен (близкие дрейфовые варианты, например, гемагглютинины штаммов NT и AICHI) эффект завуалирован из-за малости выборки (в рассматриваемом примере имеем всего 8 замен). При очень большом числе замен (сдвиговые варианты HA) эффект также частично ослабляется тем, что увеличивается число замен по первой и второй позициям. Значительная часть этих замен,однако,не при-

водит к изменению аминокислот. Такое возможно для аминокислот, кодируемых 6 кодонами (лейцин, серин и аргинин).

Эффект не наблюдается для некодирующих последовательнос - тей,равно как и для кодирующих, но не гомологичных (см.п.4.6). Представляет интерес анализ возможности использования эффекта для целей классификации по упомянутым двум признакам.

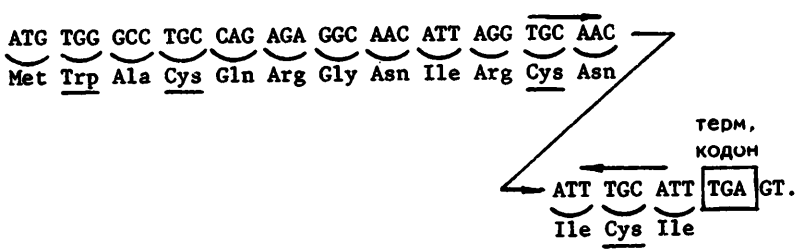
4.3. Замены в целом не имеют тенденции к локальной кластеризации. Нули и единицы перемешаны достаточно равномерно. Параметры r_{1j} находятся в пределах разброса, фиксируемого в имитационном эксперименте. Отсутствуют слишком длинные серии единиц. Общее число серий нулей и единиц, как правило, выше среднего, а иногда превышает максимальное значение, достигае - мое в эксперименте с перемешиванием (в частности, при сравне - нии полимераз у сдвиговых (по HA и NA) вариантов).

4.4. Аномально длинные серии нулей соответствуют эволюционно устойчивым (а, следовательно, функционально важным) зонам. Они обнаружены в некоторых сегментах вируса гриппа,в частности, в полимеразах 1 и 3. Так,при сравнении сегментов P1 у штаммов PR (подтип H1N1) и NT (подтип H3N2) обнаружена консервативная зона длиной в 88 нуклеотидов, расположенная в п. 2248. Она покрывает последние 50 нк в кодирующей части и 35 в некодирующей. Соответствующий рекордный показа - тель, достигнутый в имитационном эксперименте, равен 59. Ана - логично, при сравнении сегментов P3 у этих же штаммов обнаружена консервативная зона длиной в 107 нк, заканчивающаяся терминальным кодоном гена P3 (относительная частота появле - ния серии нулей такой длины в имитационном эксперименте составила 0,03).

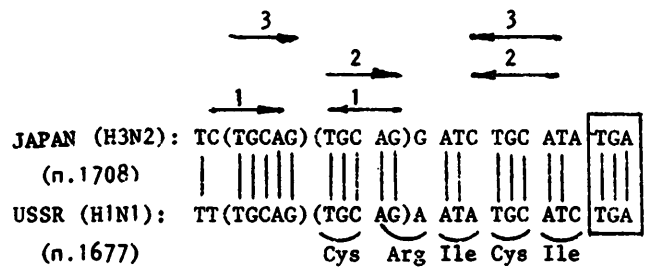
В большинстве случаев, однако, максимально длинные серии нулей не являются аномальными, т.е. не выходят за границы разброса, фиксируемого в имитационном эксперименте. Тем не менее, они представляют интерес по двум причинам: а) часто эти серии

приходятся на начало или конец крупных структурных единиц (генов или отдельных их частей типа HA1- и HA2-субъединиц гемагглютинина); б) обычно они содержат в своем составе важные структурные аминокислоты (цистеин, триптофан и др.).

Так, например, выявленные при сравнении дрейфовых вариантов гемагглютинина у штаммов DUCK (подтип H3N8) и AICHI (подтип H3N2) максимально длинные консервативные зоны расположены в п.1011 (j = 50, конец HA1) и 1683 (j = = 50, конец HA2). Последняя серия интересна по своему аминокислотному составу:



Два цистеина, предшествующие терминальному кодону, входят в состав элементарного палиндрома TG CA , повторение которого приводит к образованию шпилечной структуры. Она устойчиво сохраняется в разных дрейфовых и сдвиговых вариантах гемагглютинина, например, в подтипах H1 и H3 :



В данном случае появляются и альтернативные варианты этой структуры (1,2,3) из-за трехкратного повтора элементарного палиндрома TGCA. Заметим, что максимально длинная серия нулей, полученная при сравнении HA2 (JAPAN и USSR), соответствует как раз периодичности $(TGCAG)^2$, т.е. имеет длину 10.

4.5. Отношение числа замен к длине сравниваемых текстов (плотность несовпадений ρ) может служить диагностирующим признаком для различения сдвиговых вариантов и дрейфовых, а также для таксономии дрейфовых вариантов HA и NA. В табл. 5 приведены значения ρ, полученные при сравнении различных пар сегментов. Строки 1 и 2 количественно характеризуют плотность замен у сдвиговых вариантов гемагглютинина по каждой из субъединиц в отдельности. Плотность существенно выше в субъединице HA1, где расположены антигенные детерминанты. Строки 3,4,5 характеризуют различия между дрейфовыми вариантами гемагглютинина. Плотность замен в несколько раз ниже, чем у сдвиговых вариантов.

Строки 6-9 фиксируют различия между дрейфовыми вариантами нейраминидазы. Плотность замен, как это видно из сопоставления строк 6 и 9, а также 8 и 9, не зависит существенно от того, к каким подтипам (одинаковым или разным) принадлежат сегменты у сравниваемых штаммов. Иная картина наблюдается для сегментов полимераз (строки 10-14), где плотность замен для пар штаммов с различными подтипами HA и NA в несколько раз выше, чем для штаммов с совпадающими подтипами HA и NA.

Интересно отметить резкое отличие в скоростях эволюции полимераз на нуклеотидном и аминокислотном уровне. Сопоставление строк 10 и 11 показывает, что число нуклеотидных замен в P1 для пары PR и NT примерно в 6 раз выше, чем для пары PR и WSN. Однако число аминокислотных замен в обоих случаях отличается незначительно (26 замен для первой пары и 23 - для

Таблица 5

Значение плотности несовпадений для пар одноименных сегментов из разных штаммов

№	Сегмент	Штаммы	Подтипы	Значение ρ	Примечания
1	HA1	PR JAPAN	H1N1 H2N2	0,385	Удалены 18 нк в PR и 12 нк в JAPAN, где локализованы вставки
2	HA2	USSR JAPAN	H1N1 H2N2	0,286	Длины совпадают
3	HA	DUCK AICHI	H3N8 H3N2	0,091	Дрейфовые варианты подтипа H3
4	HA	NT AICHI	H3N2 H3N2	0,0046	Совпадение подтипов по HA и NA.
5	HA	PR USSR	H1N1 H1N1	0,078	Дрейфовые варианты подтипа H1
6	NA	Tokyo Victoria	H2N2 H3N2	0,052	Дрейфовые варианты подтипа N2
7	NA	Tokyo NT	H2N2 H3N2	0,017	- " -
8	NA	Udorn Victoria	H3N2 H3N2	0,015	- " -
9	NA	NT Victoria	H3N2 H3N2	0,052	- " -
10	P1	PR NT	H1N1 H3N2	0,163	Штаммы с разными подтипами HA и NA
11	P1	PR WSN	H1N1 H1N1	0,028	Штаммы с одинаковыми подтипами HA и NA
12	P2	PR NT	H1N1 H3N2	0,072	Штаммы с разными подтипами HA и NA
13	P3	NT WSN	H3N2 H1N1	0,094	- " -
14	P3	PR WSN	H1N1 H1N1	0.037	Штаммы с одинаковыми подтипами HA и NA

второй). Поддержание на одном уровне числа аминокислотных замен при резком увеличении числа нуклеотидных обеспечивается уже упоминавшимися выше факторами: доминированием замен по третьим позициям кодонов (доля их для пары PR и NT составляет 0,8 от общего числа) и неравномерностью выбора кодонов, в которых происходят замены по 1-й и 2-й позициям (преобладают кодоны, кодирующие аминокислоты с вырожденностью кода 6 - Arg, Ser, Leu ; для них замены в 1-й и 2-й позициях могут не приводить к изменению аминокислоты). Для успешной реализации подобной стратегии необходимо, по крайней мере, чтобы соответствующий полипептид был богат указанными аминокислотами. Полимераза 1 удовлетворяет этому условию в полном объеме: ранги аминокислот Leu, Arg и Ser в частотном упорядочении составляют соответственно 2,3,4. Особенно выделяется в этом плане аргинин, занимающий обычно место во втором десятке в частотных упорядочениях, характеризующих разные классы белков.

4.6. Формальный анализ замен между разными сегментами P1 и P3 (штамм WSN), поводом к которому послужило удивительное совпадение длин ($N = 2341$, $\rho = 0,72$), не обнаружил (за двумя исключениями) значимых отклонений от гипотезы случайности, позволяющих предположить наличие у них общего предка. Первое исключение касается общего числа серий, которое аномально мало. Второе отражает известный факт совпадения концевых фрагментов у разных сегментов вируса гриппа и проявляется в наличии аномально длинных ($j = 12$ и 13) серий нулей в начале и конце (0,1)-последовательности.

5. Выявление регулярностей в локализации аминокислот с помощью серий

Распределение аминокислот по длине последовательности представляет интерес в плане выявления различных функциональных центров белковых молекул, классификации этих молекул, раз-

личения кодирующих и некодирующих частей. Если частотные характеристики белковых молекул исследованы достаточно хорошо, то о закономерностях распределения аминокислот по длине молекулы практически ничего не известно. Серии являются удобным инструментом для проведения соответствующего анализа, если иметь в виду такие их характеристики, как l_{0max}, l_{1max}, $r_0(k)$ и $d(k)$ (см. п. 2).

Агрегирование аминокислотных последовательностей проводилось по следующей схеме. Выделялась интересующая нас аминокислота и все вхождения ее кодировались единицей. Оставшиеся аминокислоты объединялись в один таксон, элементы которого в тексте кодировались нулями. Далее вычислялись серийные характеристики двоичной последовательности, и в ходе имитационного эксперимента оценивалась их значимость.

По описанной схеме были обработаны все сегменты штамма PR (тип A) и отдельные сегменты штамма Lee (тип B). В ходе эксперимента осуществлялся перебор по всем элементам алфавита аминокислот. Ниже сформулированы основные выводы по итогам этого эксперимента.

5.1. Очень редки триплетные вхождения аминокислот во всех сегментах. К примеру, в нейраминидазе (штамм PR) нет ни одного случая трехкратного повторения какой-либо аминокислоты, хотя имитационные оценки вероятности такого события достаточно высоки (для серина - $\hat{p}(Ser^3) > 0{,}5$, для глицина - $\hat{p}(Gly^3) \sim 0{,}38$ и т.д.).

5.2. Максимальные по длине серии нулей (далеко не всегда аномальные - в среднем от 40 до 100 остатков) часто расположены в начале и в конце генов. Это означает, что некоторые аминокислоты редко встречаются в концевых фрагментах.

Так, начальный фрагмент нейраминидазы PR свободен от таких аминокислот, как Asp, Arg, Glu, Phe, Ala, Tyr, Met (кроме инициирующего кодона), а конец - от Tyr.

Интересно отметить, что первые 3 аминокислоты в этом ряду обладают зарядом. Начало **HA1** (штамм **PR**) свободно от **Phe**, **Glu** и **Met** (с той же оговоркой), начало **HA2** - от **Cys** и **Pro**, конец **HA2** - от **Glu**. Серия, фиксирующая отсутствие пролина в начале **HA2**, аномальна по длине (l_{0max} = 164 > $l^{сл}_{0max}$ = 157); серия по цистеину (l_{0max} = = 160) близка к этому.

5.3. Вхождения некоторых аминокислот кластеризованы. Можно отметить 2 разновидности кластеризации - глобальную и локальную. Первая характеризуется повышенной концентрацией некоторой аминокислоты в пределах какого-либо участка, сопоставимого по размерам с целым геном (к примеру, все 7 метионинов в гене **NS1** (**Lee**) локализованы в первой половине гена). Локальная кластеризация подразумевает скопление аминокислотных остатков одного типа в очень ограниченных по размеру участках. О наличии локальных кластеров сигнализируют аномальные значения статистик $r_0(k)$ и $d(k)$. Частным случаем локальной кластеризации является тандемная, характеризующаяся аномально высоким числом парных вхождений какой-либо аминокислоты (статистика r_{12}).

Примером кластеров, выделенных в **HA(PR)** с помощью статистик $r_0(k)$ и $d(k)$ при значении $k = 6$, являются кластеры лизина в позициях 171-188 (1000001010000001101), 412-427 (1000010010000011), 460-475 (1000010100010001). Для лизина $r_0(6)=18$, в то время как рекордное значение, зафиксированное в имитационных экспериментах, равно 17.

Пример тандемной кластеризации демонстрирует лейцин в **HA(PR)** : значение $r_{12} = 10 > r^{сл}_{12} = 9$. Комбинация **LeuValLeuLeu** (1011) встречается трижды, а ее расширение 1011001 - дважды - в начале и в конце **HA**. Отметим, что в **NA(PR)** также встречается кластер алифатических аминокислот (**Leu** и **Ile**) в начале гена (п.20, $d(4) = 19$). Вхожде-

ния изолейцина в этом фрагменте отличаются большой регулярностью:

Ile--Ile--Ile--IleIle-Ile--

(в первой позиции каждой тройки аминокислот стоит **Ile**).

5.4. Наряду с локальной кластеризацией наблюдается и обратное явление: сверхравномерное распределение некоторых аминокислот по длине цепи. Формально оно характеризуется аномально низкими значениями параметров r_{12} (как правило, $r_{12} = 0$) и l_{0max}. Примером может служить распределение глицина в аминокислотной последовательности нейраминидазы PR ($r_{12} = 0$, т.е. из 44 вхождений нет ни одного тандемного; $l_{0max} = 25 < l^{сл}_{0max} = 27$).

Заключение

Обоснована и реализована в виде программного комплекса схема анализа серий в генетических текстах, отличающаяся следующими особенностями:

а) путем разбиения элементов алфавита осуществляется перевод исходной последовательности в двоичную, для которой вычисляются элементарные и производные от них серийные характеристики; предусмотрена возможность перебора по всевозможным разбиениям алфавита;

б) статистическая значимость полученных характеристик оценивается в ходе имитационного эксперимента с двоичными последовательностями, получаемыми из исходной путем перемешивания нулей и единиц;

в) для последовательностей с сильно нарушенным балансом нулей и единиц предусматривается возможность выявления кластеров редких событий, обычно кодируемых символом 1.

Продемонстрированы нестандартные возможности использования разработанного аппарата для выявления закономерностей че-

редования различных элементов в генетических текстах, анализа мутационных замен, обнаружения регулярностей в локализации аминокислот по длине цепи.

Эксперименты с дрейфовыми и сдвиговыми вариантами вируса гриппа позволили выявить ряд интересных структурных особенностей. Отметим, в частности, что элементы подмножеств $W = \{A,T\}$ и $S = \{C,G\}$ в сегментах вируса гриппа чередуются более регулярно, чем это обычно имеет место для случайных последовательностей. Элементы же подмножеств $Pu = \{A,G\}$ и $Py = \{C,T\}$, наоборот, проявляют тенденцию к образованию кластеров-фрагментов, состоящих только из пуринов или пиримидинов. Аналогичная тенденция наблюдается при разбиении алфавита на подмножества $M = \{A,C\}$ и $K = \{G,T\}$.

Другой интересной особенностью является тяготение максимально длинных серий (не обязательно аномальных) к границам структурных областей (например, к концам генов, зонам их перекрытия и т.п.). Этот факт может быть использован для автоматического выделения соответствующих областей.

Третья особенность связана с обнаружением периодичности в зависимости числа серий от их длины для двоичных последовательностей, фиксирующих расположение мутационных замен при сравнении дрейфовых и сдвиговых вариантов вируса гриппа. Эта закономерность может быть использована для формальной оценки степени гомологии двух последовательностей.

Естественное развитие описанной методики просматривается в направлении выявления серий с дефектами, анализа серий в скользящем окне, разбиения алфавита на большее чем два число подмножеств.

Литература

1. GRANTHAM R. Nucleic acid seguence semilarities: "Poly(A) tendency" //Febs Letters. - 1980. - Vol. 121, N 2.-P.193-199.

2. VASS J.K., WILSON R.H. "ZSTATS" - a statistical analysis for potenzial Z-DNA sequences //Nucleic. Acids. Research.- 1984. - Vol. 12. -P. 825-832.

3. BLAISDELL B.E. A prevalent persistent global nonran - domness that distinguishes coding and non-coging eucaryotic nuclear DNA sequences //J.Mol.Evol. - 1983. -Vol. 19, N 2. - P. 122-133.

4. KARLIN S., BRENDEL V. Charge configurations in Viral proteins //Proc. Natl. Acad. Sci. USA. - 1988. -Vol. 85. - P. 9396-9400.

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

6. УИЛКС С. Математическая статистика. Пер. с англ. - М.: Наука, 1967. - 632 с.

7. GUIBAS L.Y., ODLIJZKO A.M. Long repetitive patterns in random sequences //Z.Wahrscheinlichkeitstheorie verw. Gebiete. - 1980. - Bd. 53, N 3. -P. 241-262.

8. GORDON L., SCHILLING M.F., WATERMAN M.S. An extreme value theory for long head runs //Probability Theory and Related Fields. - 1986. - Vol 72. -P. 279-287.

9. FOULSER D., KARLIN S. Maximal success durations for a semi-Markov process //Stoch. Proc. Appl. - 1987. - Vol. 24, N2. -P. 203-224.

Поступила в ред.-изд.отд.
12 сентября 1991 года