*Математическая биология и биоинформатика*
*2018. Т. 13. № 1. С. 13–37. doi: 10.17537/2018.13.13*

========== **БИОИНФОРМАТИКА** ==========

УДК: 575.22

# Структурированные РНК-маркеры для генотипирования вируса клещевого энцефалита

**Гусев В.Д.[*1], Мирошниченко Л.А.[**1], Титкова Т.Н.[‡1], Джиоев Ю.П.[†2] , Козлова И.В.[3], Парамонов А.П.[3]**

[1]*Институт математики им. С.Л. Соболева СО РАН, Новосибирск, Россия*
[2]*НИИ Биомедицинских технологий Иркутского государственного медицинского университета, Иркутск, Россия*
[3]*Научный центр проблем здоровья семьи и репродукции человека, Иркутск, Россия*

***Аннотация.*** Клещевой энцефалит относится к числу опасных природно-очаговых инфекций. Возбудителем заболевания является вирус клещевого энцефалита (ВКЭ), переносимый клещами. Различают три основных подтипа ВКЭ с разными клиническими проявлениями болезни, но допускается и возможность существования других подтипов. Эффективность лечения болезни во многом может зависеть от правильной идентификации генотипа ВКЭ. Исчерпывающая информация о генотипе содержится в полной кодирующей последовательности генома ВКЭ. Из неё можно извлечь ограниченное количество маркеров генотипирования в виде относительно коротких структурированных фрагментов РНК. В данной работе сформулирован достаточно общий подход к выделению структурированных РНК-маркеров для генотипирования ВКЭ. Рассматриваются три типа структур: периодичности, фракталоподобные конструкции и компактно локализованные комбинации разнотипных повторов. Обосновывается выбор этих структур для целей генотипирования и возможной их роли в формировании патогенного потенциала вируса. Подход апробирован на полных кодирующих последовательностях ВКЭ (161 штамм). Приведены примеры наиболее характерных маркеров каждого из трех типов.

***Ключевые слова:*** *вирус клещевого энцефалита, генотипирование, РНК-маркеры, L-граммный анализ, периодичности, фракталоподобные и комбинированные структуры.*

## ВВЕДЕНИЕ

Клещевой энцефалит (КЭ) является опасной природно-очаговой инфекцией с ограниченным ареалом распространения, приводящей во многих случаях к поражению центральной нервной системы [1, 2]. Возбудителем заболевания является вирус клещевого энцефалита (ВКЭ), относящийся к семейству Flaviviridae [3]. Основными переносчиками вируса являются иксодовые клещи. Вирус попадает в организм человека в результате укуса его инфицированным клещом.

Геномная РНК ВКЭ содержит единую открытую рамку считывания длиной 10242–10245 символов (с незначительными вариациями у отдельных штаммов). Она кодирует (без перекрытий) 3 структурных и 7 неструктурных белков [4–6].

[*]gusev@math.nsc.ru
[**]luba@math.nsc.ru
[‡]titkova@math.nsc.ru
[†]alanir07@mail.ru

Кодирующая часть по размерам значительно доминирует над некодирующей. Уровень сходства кодирующих последовательностей у разных штаммов ВКЭ достаточно высок: порядка 90 % и выше. Различия в кодирующих частях геномов носят характер одиночных замен нуклеотидов в определенных позициях, что может привести (либо не привести) к замене аминокислоты, кодируемой искаженным кодоном. Отметим, что замена нуклеотида, даже если она не приводит к изменению аминокислоты, может существенно повлиять на какую-либо фенотипическую характеристику вируса, например, вирулентность [7].

В соответствии с официально принятой классификацией различают три основных типа ВКЭ: дальневосточный (с прототипным штаммом «Софьин»), европейский (прототипный штамм «Найдорф») и сибирский (прототипные штаммы «Васильченко» и «Заусаев») [8–9]. Соответствующие им генотипы обычно нумеруют цифрами 1, 2 и 3, соответственно. Впоследствии были выделены также два предполагаемых генотипа: № 4 (представлен единственным штаммом 178-79) и № 5 (прототипный штамм 886-84) [10–11].

Среди основных критериев аттестации какой-либо группы штаммов в качестве самостоятельного генотипа обычно указывают на достаточно высокий (более 12 %) уровень отличия представителей одного генотипа от представителей других генотипов по первичной структуре их геномов. При этом между представителями одного генотипа эти отличия не превышают нескольких процентов. Другим критерием могут служить особенности структуры геномов. Так, важной особенностью представителей генотипов 4 и 5 является «мозаичность» их геномов, проявляющаяся в наличии внутри них чередующихся фрагментов из геномов трех основных генотипов. Несмотря на столь характерную особенность генотипов 4 и 5 статус их требует дополнительного подтверждения, поскольку в отличие от основных генотипов они пока не признаны официально и их ареалы не имеют широкого распространения на территории Евразии [10]. Указанное обстоятельство наряду с наличием иных точек зрения на проблему генотипирования (см., например, [12]) и существование штаммов, плохо вписывающихся в существующую схему дифференциации ВКЭ (например, штамм Buzuuchuk [10]), дают основание авторам работы [10] полагать, что «официально признанная внутривидовая классификация ВКЭ не является окончательной».

Под *генотипированием* в данной работе мы понимаем задачу отнесения произвольного штамма ВКЭ, представленного полной кодирующей последовательностью, к одному из известных генотипов. Мы не рассматриваем методы генотипирования, основанные на изучении антигенных свойств и других фенотипических маркеров, поскольку они менее точны по сравнению с методами, опирающимися на полногеномное представление, и не всегда приводят к однозначным результатам. Высказываемые порой замечания о значительной трудозатратности полногеномного анализа теряют свою весомость по мере ускорения и удешевления технологий секвенирования.

В работе представлен новый подход к выделению РНК-маркеров для генотипирования ВКЭ. Эти маркеры могут послужить полезным дополнением к используемым на практике дезоксиолигонуклеотидным зондам [13–16, 8] в ситуациях, когда последние не дают однозначного заключения о генотипе конкретного штамма. Конструируемые нами РНК-маркеры отличаются от дезоксиолигонуклеотидных зондов, комплементарных различным участкам генома, своей структурированностью: это тандемные повторы [17], фракталоподобные структуры [18] и комбинации разнотипных повторов, локализованные в ограниченном по размеру фрагменте генома [19]. Структуры подобного типа часто встречаются в регуляторных областях геномов и несут различную функциональную нагрузку. В частности, специфические тандемные повторы используются для проведения ДНК-дактилоскопии [20]. Можно отметить

также, что и сайты рестрикции, используемые для генотипирования в [21], можно трактовать как структурированные РНК-маркеры, представляющие собой короткие комплементарные палиндромы. Однако само генотипирование проводится не по сайтам рестрикции, как таковым, а путем выявления генотипспецифических различий в длинах рестрикционных фрагментов.

В соответствии с вышесказанным можно предполагать, что структурированные РНК-маркеры имеют больше шансов получить содержательную интерпретацию, чем неструктурированные олигонуклеотидные зонды. В частности, комбинированные структуры, включающие в себя палиндромно-шпилечные конструкции, могут иметь отношение к формированию вторичной структуры геномной молекулы вируса. В [8] высказано предположение, что эта структура играет важную роль в репродукции ВКЭ.

Ориентация на структурированные РНК-маркеры позволяет полностью автоматизировать процесс выявления их из исходной (обучающей) подборки штаммов, относящихся к разным генотипам. Вначале с помощью алгоритмов, описанных в [17–19], в текстах подборки выявляются все варианты интересующих нас локальных структур. Затем из них отбираются структуры, представляющие наибольший интерес для генотипирования, т.е. относящиеся лишь к одному (или, преимущественно, к одному) из генотипов. Полезную информацию несет и факт отсутствия структуры в штаммах какого-либо генотипа. Отметим, что в [8] подбор участков генома, используемых для выделения олигонуклеотидных маркеров, осуществляется вручную.

Дальнейший порядок изложения материала таков: вначале даются определения интересующих нас структур и кратко описываются подходы, используемые для их выявления в анализируемой подборке штаммов ВКЭ, представленных своими кодирующими последовательностями; затем приводятся результаты эксперимента (списки потенциально возможных маркеров разного типа) с последующим их обсуждением. Акцент делается на конфигурациях повторов разного типа: периодичностях, фракталоподобных и комбинированных структурах

## ВОЗМОЖНЫЕ ПОДХОДЫ К ВЫДЕЛЕНИЮ РНК-МАРКЕРОВ

Все подходы предполагают наличие достаточно представительной подборки штаммов ВКЭ, относящихся к разным генотипам и представленных полными кодирующими последовательностями. Чем представительнее подборка, тем «устойчивее» полученные результаты (РНК-маркеры) к возможному ее расширению по мере секвенирования новых штаммов. В нашем случае подборка содержит 161 штамм, из них к генотипу 1 относятся 80 штаммов, к генотипу 2 – 46, к генотипу 3 – 28 и к группе 886 (генотип 5) – 7 штаммов. Единственный (на данный момент) штамм 178-79, представляющий генотип 4, в подборке не представлен. Используемые нами подходы к выявлению РНК-маркеров для целей генотипирования не требуют предварительного выравнивания кодирующих последовательностей в исходной подборке. Однако позиционные привязки маркеров во избежание разнобоя делаются все-таки на основе выравнивания (его длина составляет 10248 символов).

### *L*-граммный анализ

*L*-граммный анализ (неструктурированные маркеры) дает представление о том, из каких цепочек символов длины $L$ ($L$-грамм) состоит текст. Фиксируются многообразие и частоты цепочек при каждом значении $L$ ($L = 1,2,\ldots L_{max}$), где $L_{max}$ – длина максимального повтора в тексте (нас интересуют только повторяющиеся $L$-граммы). Фактически, $L$-граммное описание текста – это его представление в терминах повторов.

Применительно к группе текстов (например, относящихся к одному генотипу) полезным является понятие совместного $L$-граммного спектра, фиксирующего многообразие $L$-грамм во всех текстах группы с указанием частот встречаемости

каждой *L*-граммы в каждом тексте. Если *L*-грамма встречается во всех текстах группы, она характеризует генотип в целом, но если она присутствует и в штаммах других генотипов, то не представляет интереса в плане генотипирования. Наиболее перспективным в этом отношении является *L*-граммный анализ, ориентированный на случай нескольких классов текстов, где каждый класс ассоциируется с отдельным генотипом. В рамках этого анализа фиксируется распределение каждой *L*-граммы по разным классам (генотипам), а в каждом классе – по разным текстам (штаммам). *L*-граммы, присутствующие во всех текстах конкретного класса и отсутствующие в текстах других классов, будем называть «контрастными». Именно они претендуют на роль потенциальных генотипспецифических РНК-маркеров.

Приведем примеры таких маркеров минимальной длины, характеризующих отдельные генотипы. Для генотипа 1 минимальным контрастным фрагментом является 6-грамма (agtaga), присутствующая во всех 80 текстах этого генотипа 189 раз с частотой встречаемости в одном тексте от 1 до 4 (доминирующие позиции – 10137, 1269, 6442). Во втором генотипе минимальная длина контрастных фрагментов составляет 7 символов. Выделено 8 таких 7-грамм: gttattc (поз. 3054, 3110), gcatctg (поз. 3220, 6994), tgcacaa (поз. 3280), gatgcga (поз. 4134), ccggctt (поз. 4379, 4559, 9632), aggattt (поз. 3764, 4546), acagcga (поз. 5431, 8402), agccaat (поз. 6777). Каждая из них встречается во всех 46 текстах второго класса от одного до трех раз. Минимальная длина контрастных *L*-грамм у текстов, представляющих генотип 3, равна 8. Выделено 9 таких 8-грамм: gggtggac, ataagccg, aggagagt, gtgaaaaa, acggagct, gatggcgg, aattgtgg, gggggggag, cggaaccc. Из них 5 первых 8-грамм встречаются по разу в каждом тексте (позиции 5029, 6562, 8020, 2944, 4333, соответственно) за единственным исключением: две последние 8-граммы в одном из текстов (каждая в своем) имеют еще по одному вхождению. Остальные 8-граммы встречаются в текстах этого класса от 1 до 3 раз. И, наконец, в самой малочисленной «группе 886» (генотип 5, всего 7 штаммов) контрастными являются две 7-граммы, имеющие по паре вхождений в каждый из геномов этой группы: cccgctg (поз. 2763 и 5221) и atggcgc (поз. 3859 и 5517).

К достоинствам *L*-граммного подхода следует отнести возможность обнаружения генотипспецифических РНК-маркеров в участках генома с одинаковым аминокислотным составом у всех четырех генотипов. Проиллюстрируем это на примере 9-грамм, расположенных в позициях 3568 и 3571 выравнивания (таблица 1).

**Таблица 1.** РНК-маркеры, выделенные на основе анализа частотно-позиционного распределения 9-грамм

| позиция 3568 | | | | позиция 3571 | | | |
|---|---|---|---|---|---|---|---|
| генотип | 9-грамма (нукл..) | 3-грамма (а.к.) | встреча-емость | генотип | 9-грамма (нукл.) | 3-грамма (а.к.) | встреча-емость |
| 1 | gtg-gca-gtt | VAV | 80 из 80 | 1 | gca-gtt-ggg | AVG | 80 из 80 |
| 2 | gtg-gca-gtg | VAV | 46 из 46 | 2 | gca-gtg-ggg | AVG | 46 из 46 |
| 3 | gtg-gcg-gtg<br>gtt-gcg-gtg | VAV | 22 из 28<br>6 из 28 | 3 | gcg-gtg-ggg | AVG | 28 из 28 |
| 5 | gta-gca-gtc | VAV | 7 из 7 | 5 | gca-gtc-ggg | AVG | 7 из 7 |

Нетрудно видеть, что соответствующие 9-граммам тройки аминокислот совпадают у всех генотипов: VAV (поз. 3568) и AVG (поз. 3571). Генотипспецифические различия наблюдаются только на РНК-уровне. 9-граммные цепочки у разных генотипов отличаются по третьим позициям кодонов. Штаммы генотипа 3 в позиции 3568 представлены двумя 9-граммами, но обе они отсутствуют у других генотипов.

*L*-граммный анализ на РНК-уровне представляет интерес и в плане выявления позиционных привязок, содержащих одинаковые цепочки нуклеотидов во всех штаммах, вне зависимости от генотипа. При $L = 9$ обнаружено 8 таких позиций

(таблица 2). Эти позиции могут представлять интерес в плане дифференциации ВКЭ от других представителей рода флавивирусов.

**Таблица 2.** «Устойчивые» 9-граммы, присущие всем штаммам ВКЭ вне зависимости от генотипа

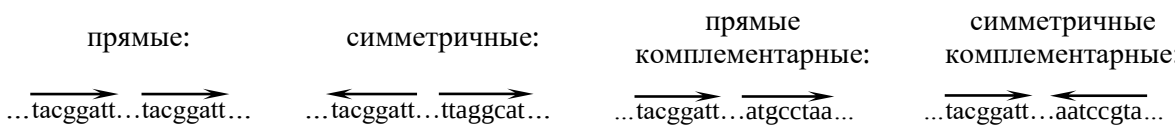
| Позиция | 9-грамма (нукл.) | 3-грамма (а.к.) |
|---|---|---|
| 55 | tcg-aaa-gag | S-K-E |
| 73 | aag-acg-cgt | K-T-R |
| 115 | ttg-atg-cgc | L-M-R |
| 118 | atg-cgc-atg | M-R-M |
| 2269 | atg-tcc-atg | M-S-M |
| 2929 | ctc-tgg-atg | L-W-M |
| 4621 | acg-atg-tgg | T-M-W |
| 9382 | aac-ata-aag | N-I-K |

Описанные в данном разделе РНК-маркеры довольно коротки и в общем случае не являются структурированными в интересующем нас смысле (отсутствуют проявления повторности и фрактальности). Это может затруднить их содержательную интерпретацию. Ниже излагаются подходы к выявлению структурированных РНК-маркеров, обладающих указанными выше свойствами. Удобным аппаратом для этих целей является РНК-ориентированный сложностной анализ.

## Сложностной анализ (структурированные маркеры)

Сложностной анализ основан на предложенной Лемпелем и Зивом в [22] идее оценивания сложности конечной символьной последовательности числом шагов некоторого гипотетического процесса, порождающего данную последовательность. На каждом шаге используется одна из допустимых («порождающих») операций. В качестве таковых Лемпель и Зив предложили: а) операцию генерации «нового» символа и б) операцию копирования любого фрагмента из предыстории, т.е. из уже синтезированной части последовательности. Если по ходу процесса встречается символ, которого не было ранее, используется первая операция. При этом последовательность удлиняется ровно на один элемент. Если же очередной символ или цепочка символов встречались ранее, мы можем их скопировать. При этом последовательность удлиняется не менее чем на один элемент. Из всевозможных процессов порождения выбирается процесс с минимальным числом шагов. Это число и рассматривается в качестве меры сложности последовательности. Минимальность числа шагов обеспечивается выбором при каждом применении операции копирования такого фрагмента-прототипа из предыстории, который максимально удлиняет синтезируемую последовательность.

Авторы данной работы предложили ДНК(РНК)-ориентированную модификацию этой меры, расширив множество допустимых вариантов копирования до четырех. Это было обусловлено стремлением учитывать следующие типы повторов:




В нашем варианте меры сложности на каждом шаге выбирается та операция копирования, которой соответствует максимальный прототип из предыстории. Следует отметить, что симметричные комплементарные повторы проявляют себя в ДНК (РНК)-последовательностях так же ярко, как и повторы в обычном смысле, и их

функциональная значимость не вызывает сомнений. В частности, они составляют основу палиндромно-шпилечных структур, играющих важную роль в регуляции транскрипции и других генетических процессов.

Последовательность шагов процесса порождения последовательности $S$ в виде конкатенации синтезируемых и копируемых цепочек символов мы называем сложностным разложением $S$. Именно из него извлекаются интересующие нас периодичности, а также фракталоподобные и комбинированные структуры [17–19].

## Анализ периодичностей в геномах ВКЭ

Под *периодичностями* мы понимаем тандемно повторяющиеся цепочки символов, например, tacg tacg tacg tacg (или, для краткости, $(tacg)_4$). Здесь базовую цепочку tacg естественно называть *периодом*. Длину периода будем обозначать через $p$, а длину самой периодичности – через $P$. Тогда отношение $k = P/p$ (равное 4 в данном случае) характеризует кратность повторения. Поскольку повторение может быть прервано до завершения очередного периода, кратность не всегда будет целым числом. Длина периода $p$ может меняться от 1 (моносерии – повторы одного элемента) до значений, сопоставимых с длиной анализируемой последовательности. В приведённом выше примере формально можно выделить и другой период tacgtacg с кратностью повторения 2. В дальнейшем, говоря о периодах, будем иметь в виду минимальный из них. Функциональная нагрузка периодичностей часто заключается в дублировании важных структурных элементов (например, терминальных кодонов многих генов или целых генов).

**Таблица 3.** Список и параметры периодичностей, выявленных в кодирующей последовательности штамма Primorie823

| № | Позиция | Периодичность | Длина периодичности | Длина периода |
|---|---|---|---|---|
| 1 | 21 | gaaaggaaagg | 11 | 5 |
| 2 | 1944 | tggtggtggt | 10 | 3 |
| 3 | 2232 | cctggcctgg | 10 | 5 |
| 4 | 2339 | gctgtgctgtg | 11 | 5 |
| 5 | 2593 | gtggtggtgg | 10 | 3 |
| 6 | 2900 | atggcatggca | 11 | 5 |
| 7 | 3162 | aagagaagag | 10 | 5 |
| 8 | 3530 | ctggtctggt | 10 | 5 |
| 9 | 3641 | ttgagttgag | 10 | 5 |
| 10 | 3672 | tgcgctgcgc | 10 | 5 |
| 11 | 4196 | tgctggtgctgg | 12 | 6 |
| 12 | 4344 | gaaagaggaaagag | 14 | 7 |
| 13 | 5260 | gatgtgatgtg | 11 | 5 |
| 14 | 5362 | catagcatagc | 11 | 5 |
| 15 | 5870 | aggaaggaagga | 12 | 4 |
| 16 | 5910 | tgatgatgatgat | 13 | 3 |
| 17 | 6059 | agaaaagaaa | 10 | 5 |
| 18 | 7168 | gctgagctga | 10 | 5 |
| 19 | 7255 | ggagagggagagg | 13 | 6 |
| 20 | 7441 | gtggcctgtggcctg | 15 | 7 |
| 21 | 8179 | aactcaactca | 11 | 5 |
| 22 | 8465 | accgcaccgcacc | 13 | 5 |
| 23 | 8850 | agagagagagaga | 13 | 2 |
| 24 | 9054 | tggagtggag | 10 | 5 |
| 25 | 9791 | ctgcctgcct | 10 | 4 |
| 26 | 10018 | aaagaaaaga | 10 | 5 |



Математическая биология и биоинформатика. 2018. Т. 13. № 1. doi: 10.17537/2018.13.13

Рассмотрим для иллюстрации вариант обработки какого-либо одного штамма (выбран штамм Q825164.1 Primorie823). Учитываются лишь периодичности с порогом по длине $P \geq 10$. Не все из них являются генотипспецифическими. Результаты обработки приведены в таблице 3. Нетрудно видеть, что в этом геноме выявлено 26 периодичностей с длиной $P \geq 10$. Среди них практически нет совпадающих. Исключение составляют лишь пары № 2, № 5 и № 17, № 26, совпадающие с точностью до циклического сдвига. Максимальная длина достигается на № 20 ($P_{\max} = 15$). Минимальная длина совпадает с пороговым значением ($P_{\min} = 10$). Таких периодичностей около половины и в большинстве из них длина периода 5, что связано с выбором порога. В целом, периодичностей с длиной периода $p < 5$ мало, а с длиной 1 и 2 вообще единицы. В силу этого некоторые из них могут даже выступать в качестве маркеров для определения генотипа. Например, периодичность $g_{10}$ ($p = 1$) практически не встречается в штаммах генотипа 1, но представлена во многих штаммах генотипа 2.

Число периодичностей с длиной 10 и выше колеблется в разных штаммах генотипа 1 в пределах от 20 до 30. Исключений мало, но они есть, в частности, в штаммах №№ 39-40 (с географической привязкой к Томску и Новосибирску) этот показатель $\geq 30$.

Позиционные привязки периодичностей представляют интерес в плане наличия (или отсутствия) аномально длинных зон, не содержащих периодичностей. В штаммах генотипа 1 зона, близкая к аномальной по размерам, расположена в самом начале между позиционными привязками 21 и 1944. В штаммах генотипа 2 привязка 1944 с периодичностью tgg-tgg-tgg-t отсутствует вовсе, а в указанной зоне расположены дополнительно по 2–3 периодичности.

Анализируя нуклеотидный состав самих периодичностей, приведенных в таблице 3, отметим, что он почти во всех периодичностях неполон, т.е. из четырех элементов алфавита {a, c, g, t} присутствуют лишь некоторые из них. Во многом это обусловлено спецификой объекта: все периодичности с длиной периода 1, 2 и 3 будут «недобирать» из алфавита, соответственно, 3, 2 и (1 или 2) элемента. Другой фактор связан с распределением элементов алфавита по частоте в геномах ВКЭ. На первом месте со значительным отрывом идет гуанин, затем аденин, цитозин и тимин. Этим, возможно, обусловлено значительное количество чисто «пуриновых» периодичностей (см. №№ 1, 7, 12, 15, 17, 19, 23, 26).

Итоговая информация по периодичностям, представленным в кодирующих последовательностях штаммов всех четырех генотипов, приведена в таблице 4 для двух пороговых значений по длинам этих структур ($P_1 = 10$ и $P_2 = 12$). Здесь $N$ – число штаммов, $M$ – общее число выявленных периодичностей по всем штаммам одного генотипа, $\overline{m}$ – среднее число периодичностей на один штамм ($\overline{m} = M / N$), $m_{\min}$– минимальное число периодичностей в одном штамме, $m_{\max}$ – максимальное число периодичностей в одном штамме, $Pos$ – число позиций, в которых выявлены периодичности (хотя бы в одном геноме данного генотипа)

**Таблица 4.** Количественные данные по всем периодичностям

| Длина | генотип | $N$ | $M$ | $\overline{m}$ | $m_{\min}$ | $m_{\max}$ | $Pos$ |
|---|---|---|---|---|---|---|---|
| $P \geq 10$ | 1 | 80 | 2034 | 25.4 | 20 | 33 | 76 |
| | 2 | 46 | 1055 | 22.9 | 19 | 26 | 49 |
| | 3 | 28 | 554 | 19.8 | 16 | 24 | 68 |
| | 5 | 7 | 123 | 17.6 | 17 | 18 | 20 |
| $P \geq 12$ | 1 | 80 | 640 | 8.0 | 4 | 12 | 23 |
| | 2 | 46 | 305 | 6.6 | 5 | 8 | 9 |
| | 3 | 28 | 208 | 7.4 | 5 | 11 | 23 |
| | 5 | 7 | 50 | 7.1 | 7 | 8 | 8 |



Математическая биология и биоинформатика. 2018. Т. 13. № 1. doi: 10.17537/2018.13.13

Анализ приведенных результатов компьютерной обработки показывает, что:

– периодичностей в каждом штамме и в сумме по всем штаммам одного генотипа выявляется достаточно много, что позволяет использовать эти структуры для целей генотипирования штаммов ВКЭ;

– с увеличением порога $P$ количество выявляемых периодичностей убывает достаточно быстро;

– тенденция к убыванию параметра $\bar{m}$ с увеличением номера генотипа при $P \geq 10$, по-видимому, не является значимой, поскольку при $P \geq 12$ она себя уже не проявляет (в последнем случае существенно снижается доля периодичностей, которые носят «случайный» характер);

– генотип 2 по показателю Pos существенно уступает генотипам 1 и 3, особенно в варианте анализа с $P \geq 12$. Это означает, что в штаммах генотипа 2 разброс позиционных привязок периодичностей значительно меньше, чем в штаммах двух других генотипов. Данный факт согласуется с высказанным в [10] утверждением о «высокой генетической однородности» представителей генотипа 2.

Важной составляющей анализа периодичностей является этап имитационного моделирования, дающий представление о том, в какой степени изменяются интересующие нас параметры (число периодичностей, их длины, позиционные привязки и др.) для рандомизированного варианта исходных данных. Рандомизация осуществлялась случайным перемешиванием символов в геноме каждого штамма. При этом частотный состав элементов сохранялся, но разрывались связи между ними. Результаты имитационного моделирования, представленные в таблице 5, показали, что существенные различия между исходной и рандомизированной подборками штаммов наблюдаются лишь по количеству выявляемых периодичностей (параметр $\bar{m} = M/N$).

**Таблица 5.** Число периодичностей в исходной и рандомизированной подборке

| Порог по длине | генотип | $N$ | исходная подборка | | рандомизированная подборка | |
|---|---|---|---|---|---|---|
| | | | $M$ | $\bar{m} = M/N$ | $M$ | $\bar{m} = M/N$ |
| $P \geq 10$ | 1 | 80 | 2034 | 25.4 | 1123 | 14.2 |
| | 2 | 46 | 1055 | 22.9 | 655 | 14.2 |
| | 3 | 28 | 554 | 19.8 | 394 | 14 |
| | 5 | 7 | 123 | 17.6 | 105 | 15 |
| $P \geq 12$ | 1 | 80 | 640 | 8 | 258 | 3.2 |
| | 2 | 46 | 305 | 6.6 | 162 | 3.5 |
| | 3 | 28 | 208 | 7.4 | 96 | 3.4 |
| | 5 | 7 | 50 | 7.1 | 23 | 3.3 |

Анализ таблицы 5 показывает, что:

– суммарное число периодичностей, выявляемых в исходной подборке, существенно превышает аналогичный параметр в рандомизированной подборке. Это подразумевает, что весомая доля периодичностей носит неслучайный характер;

– эффект очевидным образом усиливается с увеличением порога $P$.

Еще одно различие (почти очевидное) выявляется лишь при сравнении самих периодичностей и их позиционных привязок в разных штаммах. Как правило, одна и та же периодичность встречается в большом числе штаммов одного генотипа, причем с одинаковой позиционной привязкой. Это естественно ввиду близости геномов конкретного генотипа. В рандомизированных же версиях случаи встречаемости одной и той же периодичности в разных псевдоштаммах редки, не говоря уже о совпадении позиционных привязок.

Все выявленные периодичности можно разделить на несколько групп. В первом приближении исключим из рассмотрения две группы. Одну из них составляют периодичности, присутствующие во всех генотипах. Они представляют интерес в плане

сопоставления ВКЭ с родственными организмами, но не в плане генотипирования. Во вторую группу отнесем редко встречающиеся периодичности, существующие во всей подборке в одном-двух экземплярах. Каждая из них требует индивидуального рассмотрения, но в силу малой частоты встречаемости они, опять же, не представляют особого интереса в плане генотипирования. Все оставшиеся периодичности разобьем на две группы: А и Б. В группу А отнесем периодичности, выявленные только (или преимущественно) в одном из генотипов (см. табл. 6). Они представляют наибольший интерес в плане генотипирования.

В группу Б отнесем периодичности, присутствующие в разных (но не во всех четырех) генотипах (см. табл. 7). Они представляют интерес в плане отсутствия их в том или ином генотипе, что тоже может быть использовано при определении генотипа. Кроме того, периодичности из этой группы указывают на связи между разными генотипами.

В обеих таблицах периодичности упорядочены по мере возрастания их позиционных привязок. Возможны и другие варианты упорядочения, в частности, по генотипам и частоте встречаемости периодичности в штаммах конкретного генотипа. Такие упорядочения можно сделать на основании таблиц 6 и 7.

**Таблица 6**. Периодичности, представленные преимущественно в одном из генотипов

| № | Позиция | Периодичность | Номер генотипа | Частота встречаемости в генотипе |
|---|---|---|---|---|
| 1 | 1944 | tggtggtggt | 1 | 69 из 80 |
| 2 | 2296 | ggtctggtct | 2 | 35 из 46 |
| 3 | 2339 | gctgtgctgtg | 1 | 63 из 80 |
| 4 | 2547 | aactgaactgaa | 5 | 7 из 7 |
| 5 | 2559 | tctggctctggc | 2 | 44 из 46 |
| 6 | 2653 | aaaggaaagga | 2 | 44 из 46 |
| 7 | 3162 | aagagaagag | 1 | 79 из 80 |
| 8 | 3530 | ctggtctggt | 1 | 71 из 80 |
| 9 | 4344 | gaaagaggaaagag | 1 | 47 из 80 |
| 10 | 4469 | gatctgatctg | 5 | 7 из 7 |
| 11 | 5241 | ggtgggtggg | 1 | 7 из 80 |
| 12 | 5241 | gggggggggg | 5 | 7 из 7 |
| 13 | 5269 | tgccatgcca | 3 | 5 из 28 |
| 14 | 5270 | gccacgccac | 3 | 20 из 28 |
| 15 | 5870 | aggaaggaagga | 1 | 35 из 80 |
| 16 | 7114 | gttggtgttggt | 1 | 46 из 80 |
| 17 | 7155 | gtctggtctgg | 2 | 44 из 46 |
| 18 | 7316 | tgctttgctt | 1 | 44 из 80 |
| 19 | 7457 | gtggtgtggt | 2 | 40 из 46 |
| 20 | 7720 | ctggcctggc | 3 | 25 из 28 |
| 21 | 7721 | tggcttggct | 3 | 3 из 28 |
| 22 | 8144 | gggggggggg | 2 | 46 из 46 |
| 23 | 8464 | taccgtaccg | 1 | 7 из 80 |
| 24 | 8465 | accgcaccgc | 1 | 66 из 80 |
| 25 | 9054 | tggagtggag | 1 | 69 из 80 |
| 26 | 9144 | ggctggctgg | 3 | 23 из 28 |
| 27 | 9570 | actttacttt | 2 | 39 из 46 |
| 28 | 9791 | ctgcctgcct | 1 | 72 из 80 |
| 29 | 10018 | aaagaaaaga | 1 | 72 из 80 |
| 30 | 10031 | tggagtggag | 2 | 46 из 46 |
| 31 | 10093 | ggaaggaagga | 3 | 20 из 28 |

Всего в таблице 6 представлено 13 потенциально возможных маркеров генотипа 1, 8 маркеров генотипа 2, 6 маркеров генотипа 3 и 2 маркера «группы 886» (генотип 5).



Математическая биология и биоинформатика. 2018. Т. 13. № 1. doi: 10.17537/2018.13.13

В таблице 7 приведены образцы периодичностей, выявленных только в штаммах двух или трех генотипов из четырех рассматриваемых, при этом, как правило, хотя бы в одном из них она доминирует, т.е. встречается более чем в половине штаммов, характеризующих генотип. Таким образом, эти периодичности ранжируют генотипы по вероятности обнаружения в них указанных структур от очень малых значений до больших. Так, к примеру, периодичность № 3 маловероятно встретить в штаммах генотипов 2 и 5, с большой вероятностью она присутствует в штаммах генотипа 1, но не исключена возможность ее обнаружения и в штаммах генотипа 3.

**Таблица 7.** Периодичности, представленные в штаммах двух или трёх генотипов

| № | Позиция | Периодичность | Геno-тип | частота | Геno-тип | частота | Геno-тип | частота | связи |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 21 | gaaaggaaagg | 1 | 70 из 80 | 3 | 27 из 28 | 5 | 7 из 7 | не 2 |
| 2 | 397<br>398 | gaagggaagg<br>aaggaaagga | 2<br>2 | 10 из 46<br>36 из 46 | 3 | 22 из 28 | | | 2 + 3 |
| 3 | 2232 | cctggcctgg | 1 | 77 из 80 | 3 | 5 из 28 | | | 1 + 3 |
| 4 | 2900 | atggcatggc | 1 | 66 из 80 | 3 | 26 из 28 | 5 | 7 из 7 | не 2 |
| 5 | 3641 | ttgagttgag | 1 | 77 из 80 | 2 | 10 из 46 | | | 1 + 2 |
| 6 | 3672 | tgcgctgcgc | 1 | 56 из 80 | 2 | 9 из 46 | | | 1 + 2 |
| 7 | 4196 | tgctggtgctgg | 1 | 30 из 80 | 2 | 39 из 46 | | | 1 + 2 |
| 8 | 5260 | gatgtgatgtg | 1 | 67 из 80 | 2 | 11 из 46 | 5 | 7 из 7 | не 3 |
| 9 | 5362 | catagcatagc | 1 | 38 из 80 | 3 | 4 из 28 | | | 1 + 3 |
| 10 | 5963 | ataacataaca | 1 | 10 из 80 | 3 | 16 из 28 | | | 1 + 3 |
| 11 | 7120<br>7121 | gttgggttgg<br>ttggcttggc | 3<br>2 | 5 из 28<br>24 из 46 | | | | | 2+3 |
| 12 | 7168 | gctgagctga | 1 | 30 из 80 | 3 | 2 из 28 | | | 1 + 3 |
| 13 | 7255 | ggagagggagag<br>ggggaggggag | 1<br>2 | 77 из 80<br>45 из 46 | 3 | 8 из 28 | | | не 5 |
| 14 | 7441 | gtggcctgtggcctg | 1 | 64 из 80 | 3 | 9 из 28 | 5 | 1 из 7 | не 2 |
| 15 | 8355 | gaaagagaaaga | 1 | 4 из 80 | 5 | 7 из 7 | | | 1 + 5 |
| 16 | 8850<br>8851 | agagagagagaga<br>gagagagagaga | 1<br>3 | 80 из 80<br>5 из 28 | 3<br>5 | 22 из 28<br>7 из 7 | | | не 2 |
| 17 | 9144 | ggctggctgg | 2 | 2 из 46 | 3 | 23 из 28 | 5 | 7 из 7 | не 1 |
| 18 | 9411<br><br>9414 | ggaggggaggg<br>ggaaggggaagggg<br>aggggagggg | 1<br>2<br>2 | 55 из 80<br>41 из 46<br>3 из 46 | 3 | 12 из 28 | | | не 5 |
| 19 | 9715 | gtgccgtgccg | 2 | 42 из 46 | 3 | 28 из 28 | | | 2 + 3 |
| 20 | 10072 | atgttatgtt | 2 | 41 из 46 | 5 | 7 из 7 | | | 2 + 5 |

Здесь периодичности с близкими или совпадающими позиционными привязками (см. №№ 2, 11, 13, 16, 18) объединены в группы.

Завершая раздел об использовании периодичностей в качестве потенциально возможных маркеров генотипирования ВКЭ, отметим, что в исходном материале (161 штамм ВКЭ) их выделено гораздо больше, чем представлено в таблицах 6 и 7 (свыше 2000 при $P \geq 10$). Среди них много таких, которые также обладают генотипспецифическим потенциалом, хотя и выраженным не столь ярко, как у структур из таблиц 6 и 7, т.е. встречающихся в меньшем числе штаммов каждого генотипа. Различные комбинации таких структур могут обеспечить 100 %-ю покрываемость штаммов каждого генотипа, поэтому их можно рассматривать в качестве факультативных признаков при решении задачи генотипирования.

## Фрактальные и фракталоподобные структуры в геномах ВКЭ

*Локальными фракталами* мы называем фрагменты ДНК (РНК), которые характеризуются проявлениями *самоподобия,* основанного на свойстве симметрии или комплементарной симметрии [18]. Элемент самоподобия проявляется в том, что

повторение обычного палиндрома (рис. 1,а) или комплементарного (рис. 1,б) приводит к *усилению конструкции,* т.е. образованию нового палиндрома (соответственно, комплементарного палиндрома) вдвое большей длины.

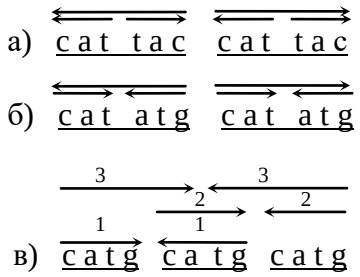



**Рис. 1.** Примеры образования фрактальных структур. Повторяющиеся фрагменты подчёркнуты; расходящиеся стрелки сверху обозначают палиндром, сходящиеся – комплементарный палиндром.

При кратности повторений выше двух (на рис. 1,в – с 3-кратным повторением комплементарного палиндрома catg) возникают множественные структуры.

При наличии незначительных искажений внутри повторяющихся фрагментов, равно как и вставок между ними, используем термины *фракталоподобные структуры* или *несовершенные локальные фракталы*. Предполагается, что размеры вставок могут быть сопоставимы с длинами повторяющихся фрагментов. Именно по такой схеме при кратности повторений 2 возникают «шпилечные» структуры, участвующие в регуляции основных генетических процессов. Необходимость рассмотрения фракталоподобных структур обусловлена тем, что в реальных геномах идеальные «локальные структуры» встречаются довольно редко.

Анализ подборки геномов ВКЭ подтверждает эту эмпирически наблюдаемую закономерность. В таблице 8 приведена информация о всех фракталах длины 12 и выше, выявленных в исходных данных. Структуры упорядочены по частоте встречаемости их в штаммах конкретного генотипа. Квадратные скобки в последнем столбце выделяют цепочки аминокислот, представленных полными триплетами.

**Таблица 8.** Информация о локальных фракталах в геномах ВКЭ

| Генотип | Период | Позиция в выравнивании | Число штаммов, содержащих данный фрактал | Разбивка фрактала на кодирующие триплеты |
|---|---|---|---|---|
| 1 | agga | 5870 | 25 из 80 | ag gaa gga agg a<br>Q [E G R] T |
| | gtg | 805 | 10 из 80 | gtg gtg gtg gtg<br>[V V V V] |
| | agaaaga | 4343 | 1 из 80 | ag aaa gaa gaa aga<br>E [K E E R] |
| 2 | ggaagg | 9411 | 27 из 46 | g gaa ggg gaa ggg g<br>M [E G E G] V |
| 3 | agaga | 6055 | 18 из 28 | gaa gag aag aga a<br>[E E K R] |
| | ggaagg | 9411 | 1 из 28 | g gaa ggg gaa ggg g<br>M [E G E G] V |
| | gaaag | 2652 | 1 из 28 | g aaa gga aag ga<br>K [K G K D |
| 5 | — | — | 0 | — |

Анализ таблицы 8 позволяет сделать следующие выводы:



Математическая биология и биоинформатика. 2018. Т. 13. № 1. doi: 10.17537/2018.13.13

– локальные фракталы длины 12 и выше представлены далеко не во всех штаммах конкретного генотипа (например, в генотипе 1 они присутствуют лишь в 36 штаммах из 80);

– все фрактальные структуры вырожденные, т.е. построены лишь на двух из четырех возможных элементов алфавита, чаще всего это пурины;

– фрактальные структуры на аминокислотном уровне могут быть представлены либо кластером из одноимённых аминокислот (VVVV для gtg- периода, генотип 1), либо повторяющимися цепочками аминокислот (EGEG для ggaagg- периода, генотип 2), либо цепочкой аминокислот, не содержащей повторений (EGR для agga- периода, генотип 1);

– отсутствие фрактальных структур в штаммах генотипа 5 («группа 886»), скорее всего, объясняется малым объёмом выборки по данному генотипу (всего 7 штаммов);

– локальные фракталы (периодичности), представленные в штаммах генотипа 1, встречаются порознь и лишь в штамме № 37 (> GQ825147.1 Shkotovo 94) присутствуют одновременно и $(gtg)_4$, и $(agga)_3$. Аналогично, в генотипе 3 лишь в штамме № 26 (> GU183384.1 Est54) присутствуют одновременно и $(gaaag)_{2.4}$, и $(agaga)_{2.6}$;

– почти все фрактальные структуры специфичны для конкретного генотипа, т.е. могут служить его маркерами. Исключение составляет лишь структура, инициируемая повторением цепочки ggaagg, являющаяся маркером для генотипа 2, но представленная один раз и в генотипе 3 (штамм > KG626343.1 Buzuuchuk).

В связи с двумя последними выводами интересно отметить, что в работе [10] штаммы > GQ825147.1 Shkotovo 94 (генотип 1) и > KG626343.1 Buzuuchuk (генотип 3) признаны «выпадающими» из имеющейся классификации.

**Таблица 9.** Информация о фракталоподобных структурах в кодирующих последовательностях геномов ВКЭ

| № | Ядро | Кратность повторения | Длина структуры | Позиция | Представленность в генотипах | | | | Цепочка аминокислот |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | 1 | 2 | 3 | 5 | |
| 1 | gttg | 5 | 26 | 311 | 23 | 0 | 0 | 0 | GWLLVVVLL |
| 2 | gttg | 5 | 26 | 311 | 5 | 0 | 0 | 0 | GWLLVVALL |
| 3 | gttg | 4 | 18 | 311 | 13 | 0 | 0 | 0 | GWLLVVV |
| 4 | tggt | 4 | 23 | 313 | 8 | 0 | 0 | 0 | WLLVLVLV |
| 5 | tggt | 3 | 18 | 755 | 0 | 0 | 1 | 0 | VVAVVWL |
| 6 | ggcc | 3 | 20 | 2229 | 37 | 0 | 4 | 0 | VALAWLGL |
| 7 | ctag | 3 | 19 | 3068 | 0 | 0 | 1 | 0 | ASLAGPR |
| 8 | agttga | 2 | 18 | 5787 | 1 | 0 | 0 | 0 | EVDGRVE |
| 9 | catg | 3 | 18 | 6107 | 3 | 0 | 5 | 7 | PWLAWHV |
| 10 | ggcc | 3 | 19 | 6993 | 0 | 0 | 1 | 0 | VASGAQA |
| 11 | gtcactg | 2 | 22 | 7065 | 0 | 0 | 6 | 0 | MSLYVVSL |
| 12 | gaccag | 2 | 20 | 8275 | 0 | 2 | 0 | 0 | DQRGPTR |
| 13 | gttg | 3 | 19 | 8317 | 4 | 0 | 0 | 0 | VGTRCVV |
| 14 | agga | 4 | 24 | 8357 | 1 | 0 | 0 | 0 | KEKDVQERI |
| 15 | agga | 3 | 18 | 8363 | 0 | 0 | 2 | 0 | KDVQERI |
| 16 | cggc | 4 | 22 | 8471 | 0 | 41 | 0 | 0 | TAPTGSAA |
| 17 | gagctc | 2 | 20 | 10219 | 75 | 4 | 0 | 7 | ELKLESS |
| 18 | gagctc | 2 | 20 | 10219 | 1 | 40 | 18 | 0 | ELRLESS |
| 19 | gagctc | 2 | 20 | 10219 | 0 | 0 | 1 | 0 | ELRMESS |

В таблице 9 приведена информация о всех фракталоподобных структурах с длиной $P \geq 18$, выявленных в исходных данных. Поскольку это уже нестрогие периодичности, повторяющуюся цепочку символов будем называть «ядром» (рассматриваются ядра длины 3 и выше). Предполагается, что ядра не могут быть искажены, т.е. являются строгими симметриями, либо комплементарными симметриями. Кратность их

повторения – не меньше 2. Варьируют лишь расстояния между вхождениями соседних ядер и количество повторяющихся ядер. Представленные в таблице 8 локальные фракталы не входят в таблицу 9, поскольку их длины не удовлетворяют ограничению $P \geq 18$.

Для каждой фракталоподобной структуры указываются: повторяющееся в ней *ядро*; *кратность* повторения; *длина* структуры; *количество штаммов* каждого генотипа, в которых она представлена, и *цепочка аминокислот,* соответствующая выделенной структуре. В представленном ниже примере фракталоподобная структура с ядром catg разбита на кодирующие триплеты, дополнена слева одним, а справа – двумя нуклеотидами и подписана внизу соответствующей последовательностью аминокислот:

```
Поз.6107: ccatggctagcatggcatgtc  :  c[ca tgg cta gca tgg cat g]tc
                                      P   W   L   A   W   H   V
```

Структуры, представленные в таблице 9, упорядочены по возрастанию их позиционных привязок.

По итогам анализа фракталоподобных структур можно сделать следующие выводы.

1) Фракталоподобных структур в анализируемой подборке выявлено гораздо больше, чем локальных фракталов. Они присутствуют практически во всех штаммах (кроме первых трёх у генотипа 3), однако такая представительность достигается за счет очень ограниченного количества высокочастотных структур (№№ 16–19). Большая же часть структур относится к разряду низкочастотных. Число структур, представленных в одном штамме, колеблется в диапазоне от 0 до 4. Крайние случаи (0 и 4) встречаются редко (только в генотипе 3), поэтому реальным можно считать диапазон от 1 до 3.

2) Некоторые структуры, представленные в таблице 9, неразличимы по типу ядра и позиционной привязке (см., например, строки 1–3 и 17–19). Но различия между ними проявляются по таким параметрам как кратность повторения ядра, длина структуры (см. строки 14 и 15), а при совпадении этих параметров могут отличаться нуклеотидные цепочки, заполняющие интервалы между вхождениями соседних ядер. Последнее обстоятельство приводит к различию в составе аминокислотных цепочек, соответствующих рассматриваемым структурам. В частности, структуры 1 и 2 отличаются лишь по составу аминокислотных цепочек: валин (V) из структуры 1 заменён на аланин (A) в структуре 2. Ядра могут совпадать и у позиционно разнесённых структур (см., например, 4 и 5 или 3 и 13), но в этом случае аминокислотные цепочки различаются уже существенным образом. Например, структуры 3 и 13 из генотипа 1 образованы повторением одного и того же ядра gttg, но при близости длин структур (18 и 19, соответственно) кратности повторений отличаются (4 и 3), что приводит к принципиально разным аминокислотным последовательностям:

```
                  G  W   L   L   V   V   V
Поз. 311:    gt tgg ttg ctt gtt gtt g
Поз.8317:    gtt ggg act cgt tgt gtt g
                  V  G   T   R   C   V   V ]
```

3) Генотип 3 выглядит не столь однородно по сравнению с генотипами 1 и 2 в плане насыщенности его фракталоподобными структурами. Среди представляющих его штаммов есть такие, которые не содержат ни одной фракталоподобной структуры, и такие, в которых содержится максимальное их количество (четыре). Среди штаммов, относящихся к генотипам 1 и 2, не реализуется ни один из этих «рекордов». Следует также отметить, что среди фракталоподобных структур, характеризующих генотип 3, нет ни одной, которая присутствовала бы в подавляющем числе штаммов этого генотипа (для генотипа 1, например, таковой является структура № 17, для генотипа 2 – структура № 16).

4) В плане *генотипирования* наибольший интерес представляют структуры 17 и 18 (позиция 10219), построенные на двукратном повторении комплементарного палиндрома gagctc. В результате возникает структура gagctcaaactggagagctc, подобная шпилечной, в которой комплементарные палиндромы образуют стебель, а расположенные между ними нуклеотиды формируют «петлю» шпильки:

```
  c   t
a       g
a       g
  a   a
  c - g
  t - a
  c - g
  g - c
  a - t
  g - c
```

Эта структура встречается во всех генотипах, при этом комплементарные палиндромы сохраняются неизменными, а минимальные изменения возникают в петле шпильки и именно они позволяют в большинстве случаев отделить один генотип от другого, в частности, первый от второго и третьего. Ниже приведено выравнивание структур 17 и 18 (доминирующие варианты) для разных генотипов.

```
              E   L   K   L   E   S   S
Поз. 10219:  gag ctc aaa ctg gag agc tc  (генотипы 1 и 5)
             gag ctc aga|ctg gag agc tc  (генотип 2)
             gag ctc aga|ttg gag agc tc  (генотип 3)
              E   L   R   L   E   S   S
```

Нетрудно видеть, что первая строка выравнивания отличается от второй лишь заменой аденина на гуанин в восьмой позиции. Тем не менее, это приводит к замене аминокислоты K (лизин) на R (аргинин). Из таблицы 9 следует, что структура № 17, кодирующая лизин, в подавляющем большинстве случаев (75 из 80) характеризует генотип 1, тогда как структура № 18, кодирующая аргинин, в 40 случаях из 46 характеризует генотип 2. Структуру № 17 можно использовать и для разделения генотипов 1 и 3, поскольку в последнем она отсутствует. Структура № 18 также подходит для этой цели, поскольку встречается лишь в одном штамме генотипа 1, но в 18 из 28 штаммах генотипа 3. Разделить же генотипы 2 и 3 в приведенном выравнивании не удается на аминокислотном уровне, но удается на нуклеотидном (замена цитозина на тимин в 10-й позиции).

Важное значение для идентификации штаммов генотипа 2 имеет структура № 16 (позиция 8471), построенная на 4- кратном повторении палиндрома cggc:

```
                T   A   P   T   G   S   A   A
Поз. 8471:  cg gca cca acc ggc tcg gcg gc
```

Она встретилась у 41-го из 46 штаммов генотипа 2 и не представлена ни в каком другом генотипе. И, наконец, определённый интерес для идентификации штаммов генотипа 1 представляет структура № 6 (позиция 2229), основанная на трёхкратном повторении комплементарного палиндрома ggcc:

```
              ..V   A   L   A   W   L   G   L
Поз. 2229:   g gcc ctg gcc tgg ttg ggc c
```

Она встретилась в 37 из 80 штаммах генотипа 1 и отсутствует в генотипах 2 и 5.

В штаммах генотипа 5 выявлены всего две фракталоподобные структуры (см. № 9 и № 17). Каждая из них присутствует во всех 7 штаммах, которыми представлен генотип 5. Подборка слишком мала, но по структуре № 9 можно в первом приближении



Математическая биология и биоинформатика. 2018. Т. 13. № 1. doi: 10.17537/2018.13.13

отделить генотип 5 от генотипа 2, в котором она отсутствует. Аналогично, по структуре № 17 можно отделить генотип 5 от генотипа 3 по тем же соображениям. Однако ни одна из этих структур не позволяет отделить генотип 5 от генотипа 1. Для этой цели лучше использовать периодичности № 4 и № 10 из таблицы 6 или № 20 из таблицы 7.

**Комбинированные структуры в геномах ВКЭ**

Комбинированными мы называем структуры, состоящие из двух разнотипных повторов (прямой плюс симметричный; прямой плюс комплементарный симметричный и т.п.) с ограничениями снизу на длины повторяющихся цепочек (не меньше, чем $r$) и сверху (не больше, чем $R$). Ограничивается также и расстояние между концом одного компонента и началом следующего за ним (не больше, чем $d$). Ограничения снизу нужны для отсеивания случайных («шумовых») структур, а сверху – для обеспечения компактности структуры. Порядок чередования цепочек, образующих повторы разных типов, произвольный, возможны наложения и совпадения цепочек, относящихся к разным повторам. Алгоритм выявления комбинированных структур в ДНК последовательностях описан в [18]. Приведём для иллюстрации варианты комбинированных структур, обнаруженных в геноме ВКЭ (штамм 1: > JQ825164.1. Primorye 823). В приводимых примерах выбраны следующие значения параметров: $r = 7$, $R = 20$, $d = 14$. Пары фрагментов, образующих симметричный повтор (обычный или комплементарный), будем обозначать Xs и X. Фрагменты, образующие прямые повторы, будем обозначать через Y1и Y2.

*Пример* 1 (комбинация симметричного повтора с прямым). Порядок компонентов: XsYYX. Начальные позиции: Xs – 5866, X – 5876, Y1 – 5870 и Y2 – 5874.

```
                    Xs          X
Поз. 5864:   gaagacaggaaggaaggacagatg
                     Y1    Y2
Поз. 1955:   G  R  Q  E  G  R  T  D
```

Здесь центральную роль играет периодичность $(agga)_3$, являющаяся локальным фракталом. На его основе формально выделяются перекрывающиеся фрагменты Y (подчёркнуты). Фрактал допускает симметричное расширение в обе стороны на 4 символа (см. стрелки сверху), в результате чего формируется симметричный повтор XsX (|Xs| = |X| = 10). Структура 1 – лишь одна из четырёх подобного типа, обнаруженных в штамме № 1.

*Пример* 2 (комбинация прямого повтора с симметричным комплементарным в том же штамме). Порядок компонентов: XsYXY. Начальные позиции: Xs – 7087; X – 7097; Y1 – 7087 и Y2 – 7114.

```
                  Xs          X
Поз. 7085:  tagttggcgcaacgccaacctctttggtggttggcgtt
                 Y1                        Y2
Поз. 2362: L  V  G  A  T  P  T  S  L  V  V  G  V
```

Здесь обращает на себя внимание совпадение фрагментов Xs и Y1 и наличие ещё одного фрагмента Y2 правее X. Как следствие, X может образовывать симметричный комплементарный повтор как с первым (левым) вхождением Y1, так и со вторым (правым – Y2). В итоге, в том же самом фрагменте генома, расширенном вправо на 3 символа, реализуется структура того же типа с порядком следования компонентов: YXsXY (*пример 3*). Начальные позиции: Xs – 7094; X – 7113; Y – 7087 и 7114.

```
                        Xs          X
Поз. 7085:  tagttggcgcaacgccaacctctttggtggttggcgttggt
                 Y1                        Y2
Поз. 2362: L  V  G  A  T  P  T  S  L  V  V  G  V  G
```

Следует отметить наличие кластера из валинов (V) и глицинов (G) в правом конце аминокислотного фрагмента (подчёркнут), а также тот факт, что повторяющиеся

фрагменты Y кодируются в фазе (что не всегда бывает), поэтому повтору на нуклеотидном уровне соответствует повтор (VG) и на аминокислотном уровне. Обратим внимание на значительную длину фрагментов, образующих симметричный комплементарный повтор (|Xs| = |X| = 11). С учетом достаточной их разнесённости вполне возможным представляется образование реальной шпилечной структуры. В целом, этот фрагмент генома характеризуется наличием блочных перестроек. Всего в рассматриваемом штамме № 1 выделено 7 структур, представляющих собой комбинацию прямого повтора с симметричным комплементарным.

Анализ комбинированных структур геномов ВКЭ проводился по той же схеме, что использовалась для периодичностей и фракталоподобных структур. А именно: выявлялись комбинированные структуры в каждом из штаммов подборки и отбирались те из них, которые представляли наибольший интерес в плане генотипирования, т.е. имели значимую частоту встречаемости во всей подборке ($F \geq 3$), характеризующуюся неравномерным распределением по отдельным генотипам (отсутствие в одних и максимальная концентрация в других).

**Таблица 10.** Данные о частоте встречаемости и позиционных привязках* комбинированных структур первого типа в геномах ВКЭ

| № | Позиция | Длина структуры | Частоты встречаемости в генотипах | | | | Позиционные привязки компонентов структуры | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 5 | Xs | X | Y1 | Y2 |
| 1 | 89 | 23 | 10 | 1 | 25 | 0 | 89 | 105 | 93 | 100 |
| 2 | 354 | 34 | 24 | 0 | 0 | 0 | 354 | 381 | 366 | 368 |
| 3 | 601 | 45 | 7 | 0 | 0 | 0 | 601 | 625 | 616 | 639 |
| 4 | 754 | 38 | 0 | 10 | 0 | 0 | 754 | 773 | 766 | 785 |
| 5 | 755 | 42 | 0 | 18 | 2 | 0 | 755 | 789 | 766 | 785 |
| 6 | 766 | 35 | 0 | 26 | 3 | 0 | 775 | 794 | 766 | 785 |
| 7 | 1182 | 46 | 0 | 3 | 0 | 0 | 1182 | 1220 | 1185 | 1200 |
| 8 | 1936 | 21 | 13 | 0 | 0 | 0 | 1936 | 1950 | 1944 | 1947 |
| 9 | 2020 | 39 | 3 | 0 | 0 | 7 | 2039 | 2049 | 2020 | 2050 |
| 10 | 2738 | 43 | 0 | 0 | 4 | 0 | 2738 | 2760 | 2749 | 2774 |
| 11 | 3232 | 40 | 60 | 0 | 0 | 7 | 3232 | 3264 | 3242 | 3261 |
| 12 | 3973 | 42 | 14 | 0 | 0 | 0 | 3973 | 3991 | 3992 | 4008 |
| 13 | 5828 | 28 | 0 | 0 | 3 | 0 | 5828 | 5849 | 5838 | 5843 |
| 14 | 5870 | 31 | 9 | 0 | 0 | 0 | 5887 | 5894 | 5870 | 5874 |
| 15 | 5887 | 36 | 9 | 0 | 0 | 0 | 5887 | 5894 | 5910 | 5913 |
| 16 | 6093 | 38 | 0 | 0 | 8 | 0 | 6093 | 6124 | 6113 | 6118 |
| 17 | 6113 | 53 | 0 | 1 | 14 | 0 | 6137 | 6158 | 6113 | 6118 |
| 18 | 6954 | 51 | 0 | 0 | 0 | 7 | 6954 | 6998 | 6970 | 6981 |
| 19 | 6970 | 25 | 49 | 0 | 0 | 0 | 6970 | 6988 | 6970 | 6981 |
| 20 | 6970 | 25 | 49 | 0 | 0 | 0 | 6981 | 6988 | 6970 | 6981 |
| 21 | 7064 | 58 | 0 | 0 | 6 | 0 | 7096 | 7114 | 7064 | 7079 |
| 22 | 7087 | 34 | 23 | 0 | 0 | 0 | 7087 | 7097 | 7087 | 7114 |
| 23 | 7087 | 37 | 23 | 0 | 0 | 0 | 7094 | 7113 | 7087 | 7114 |
| 24 | 7373 | 32 | 0 | 5 | 0 | 0 | 7373 | 7398 | 7387 | 7390 |
| 25 | 7441 | 56 | 11 | 0 | 0 | 0 | 7468 | 7489 | 7441 | 7448 |
| 26 | 8144 | 26 | 0 | 9 | 0 | 0 | 8149 | 8163 | 8144 | 8145 |

*Позиция структуры определяется как минимальное значение позиционных привязок ее компонентов.

Информация о наиболее интересных комбинациях прямого повтора с комплементарным симметричным (структуры первого типа) представлена в таблице 10. Структуры упорядочены по возрастанию позиций, занимаемых ими в выравнивании кодирующих частей геномов ВКЭ. Данные, представленные в таблице, позволяют



*Математическая биология и биоинформатика*. 2018. Т. 13. № 1. doi: 10.17537/2018.13.13

однозначно восстановить достаточно объемную структуру по кодирующим последовательностям исходной подборки.

По итогам анализа комбинированных структур можно сделать следующие выводы:

1) Штаммы генотипа 1 в наибольшей степени насыщены комбинированными структурами. В 80 штаммах выявлена 321 структура, т.е. в среднем на штамм приходится порядка 4 структур. Минимальное число структур в одном штамме – 0, максимальное – 8. Выделено 24 позиции, в которых обнаружена комбинированная структура хотя бы в одном из штаммов.

На втором месте по насыщенности комбинированными структурами стоит «группа 886» (генотип 5). В каждом штамме обнаружено по 3 структуры, т.е. минимальный и максимальный показатели встречаемости структур в одном штамме совпадают. Следует однако учесть, что объем выборки по генотипу 5 слишком мал (7 штаммов), чтобы достоверно судить о приведенных выше показателях.

На третьем месте по насыщенности комбинированными структурами стоит генотип 3 (70/28 = 2.5 структуры на один штамм). Минимальное число структур в одном штамме – 1, максимальное – 5. Выделено 12 позиций, в которых обнаружена комбинированная структура хотя бы в одном из штаммов.

На четвертом месте по насыщенности комбинированными структурами стоит генотип 2 (77/46 = 1.67 структуры на один штамм). Минимальное число структур в одном штамме – 0, максимальное – 3, число позиций, в которых выявлены комбинированные структуры, характеризующие генотип, равно 10.

2) Обращают на себя внимание весьма существенные различия по числу комбинированных структур, представленных в одном штамме. В генотипах 2 и 5 этот показатель не превышает 3, в генотипе 3 он равен 5, но штаммов, содержащих свыше трех структур, всего 3 из 28. Однако в генотипе 1 примерно половина штаммов из 80 имеют по 4 и более структур. Возможно, это говорит о том, что уровень одиночных мутационных замен, способных исказить структуру, в генотипе 1 ниже, чем в других генотипах.

3) Многие структуры, представленные в таблице 10, налагаются друг на друга. Об этом свидетельствует близость или даже совпадение их позиционных привязок. Можно указать, в частности, на структуры №№ 4-6 (позиции 754, 755, 766), №№ 14-15 (позиции 5870, 5887), №№ 19, 20 (позиция 6970) и др. Случай точного совпадения позиционных привязок проиллюстрирован выше (примеры 2 и 3). Из приведенных примеров видно, что могут совпадать не только позиционные привязки, но и отдельные компоненты смежных структур (но не все). Наложение структур часто сигнализирует о том, что зона активных блочных перестроек, фиксируемых конкретной комбинированной структурой, может быть расширена.

4) На текущий момент (т.е. в рамках исходной подборки) представители генотипа 5 (7 штаммов) могут быть идентифицированы с помощью структуры № 18 (поз. 6954), не представленной в других генотипах. Однако уже у генотипа 3 не идентифицируются однозначно штаммы №№ 7, 25-28. Аналогично, у генотипа 2 не идентифицируются штаммы №№ 10, 11, 16, 18, 29. И наконец, у генотипа 1 не идентифицируются штаммы 19, 27, 53, 54, 74-80. Эти данные можно трактовать в плане проявления неоднородности внутри подборок по каждому генотипу.

5) При рассмотрении списков штаммов, содержащих какую-то конкретную комбинированную структуру из таблицы 10, во многих случаях наблюдается кластеризация штаммов с близкими порядковыми номерами, например, структура № 25, встретившаяся 11 раз, только в генотипе 1 (позиция 7441), представлена в штаммах №№ 39-46 и 69-71. Аналогично, структура № 3, встретившаяся 7 раз только в генотипе 1 (позиция 601), представлена в штаммах с номерами 39-46 (за исключением № 41). Приведенные примеры говорят о том, что штаммы в подборку конкретного

генотипа включались не в случайном порядке, а с использованием каких-то сложившихся представлений об их близости.

6) Отдельный интерес вызвало рассмотрение структуры № 11 (позиция 3232), встретившейся в 60 штаммах генотипа 1 и 7 штаммах генотипа 5. По характеристикам, представленным в таблице 10 (позиция, длина структуры, позиционные привязки компонентов), генотипы 1 и 5 неразделимы по данной структуре. Однако различие (причем довольно тонкое) существует: в штаммах генотипа 1 прямой повтор реализуется дублированием цепочки agagtgg (длины 7), тогда как в штаммах генотипа 5 повторяющаяся цепочка на один символ длиннее (agagtggt) при тех же позиционных привязках.

Информация о наиболее интересных комбинациях прямого повтора с симметричным (структуры второго типа) представлена в таблице 11.

Одна из таких комбинаций обсуждалась в примере 1. Мы не будем подробно останавливаться на этом типе структур, поскольку нам неизвестны какие-либо содержательные примеры их функционирования. Однако формально такие структуры выделяются (при значениях параметров $r = 7$, $R = 20$, $d = 14$) и их можно использовать для целей генотипирования. Позиционные привязки компонентов структур ради краткости опущены.

**Таблица 11.** Позиции и частоты встречаемости комбинаций второго типа

| № | Позиция | Частота встречаемости в генотипах | | | | № | Позиция | Частота встречаемости в генотипах | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 5 | | | 1 | 2 | 3 | 5 |
| 1 | 340 | 24 | 0 | 0 | 0 | 9 | 3213 | 0 | 0 | 16 | 0 |
| 2 | 641 | 9 | 0 | 0 | 0 | 10 | 4341 | 48 | 0 | 0 | 0 |
| 3 | 752 | 0 | 0 | 18 | 0 | 11 | 5866 | 25 | 0 | 0 | 0 |
| 4 | 804 | 10 | 0 | 0 | 0 | 12 | 6274 | 0 | 7 | 1 | 0 |
| 5 | 851 | 0 | 0 | 0 | 6 | 13 | 7070 | 0 | 8 | 0 | 0 |
| 6 | 1380 | 8 | 0 | 0 | 0 | 14 | 8273 | 40 | 0 | 18 | 0 |
| 7 | 1916 | 66 | 0 | 0 | 0 | 15 | 8393 | 0 | 0 | 0 | 7 |
| 8 | 2468 | 0 | 0 | 9 | 0 | 16 | 9519 | 0 | 46 | 25 | 5 |

Количество выявленных комбинированных структур с симметричным повтором несколько уступает аналогичному показателю для структур с комплементарным симметричным повтором в генотипах 1, 2, 5. Например, в генотипе 1 их 263 по сравнению с 321 для случая комплементарной симметрии. Однако в генотипе 3 ситуация обратная: комбинированных структур с симметричным повтором – 135, а с комплементарным симметричным – всего 70. Удовлетворительного объяснения этому факту пока не находится.

Генотип 1 наиболее полно представлен в таблице 11. На второе место можно смело поставить генотип 3, тогда как генотип 2, по-прежнему (как и в табл. 8), представлен слабо. Структура № 16 (позиция 9519) позволяет достаточно уверенно отделить генотип 2 от генотипа 1, но не от двух оставшихся генотипов (3 и 5).

Результаты эксперимента с рандомизированной подборкой вновь (как и в случае периодичностей) показали, что принципиальных различий по длинам выявляемых структур и составу их компонентов не обнаружено. Просто в реальных данных их примерно вдвое больше, но часть из них носит случайный характер. Распределение комбинированных структур по отдельным штаммам проиллюстрируем на примере генотипа 1. Среди 80 штаммов генотипа 1 лишь 3 не содержат ни одной комбинированной структуры. В рандомизированной подборке таких «псевдоштаммов» около 30. Максимальное число структур, выявляемых в одном «псевдоштамме», равно 4. В реальных данных много штаммов, имеющих большее количество структур



Математическая биология и биоинформатика. 2018. Т. 13. № 1. doi: 10.17537/2018.13.13

(в генотипе 1 их 27). В целом, для всех четырех «генотипов» рандомизированной подборки среднее число структур, приходящихся на один «штамм», близко к 1 (соответствующие показатели для реальных данных гораздо выше и существенно различаются в зависимости от генотипа).

Резко различаются распределения частоты встречаемости конкретной структуры в конкретной позиции для разных генотипов. В реальных данных (см. табл. 8) выделялся генотип, в котором эта структура доминировала, т.е. встречалась многократно. В рандомизированных данных практически всегда структура лишь один раз встречается в конкретной позиции (перемешивание разрушает выравнивание, поэтому крайне маловероятно, чтобы у выделенной структуры при двукратном (хотя бы) ее повторении совпали позиционные привязки). Вследствие указанного обстоятельства число позиций в генотипе, содержащих хотя бы одну комбинированную структуру, в рандомизированной подборке гораздо выше, чем в реальной (в рандомизированных штаммах генотипа 1 таких позиций 80 по сравнению с 24 в реальных данных); для генотипов 2, 3 и 5 соответствующие пары показателей выглядят следующим образом: 39 (против 10), 26 (против 12) и 5 (против 3).

## ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Изложенный в данной работе новый подход к проблеме выявления РНК-маркеров для генотипирования ВКЭ основан на использовании полных кодирующих последовательностей для описания штаммов исходной (достаточно представительной) подборки. Высказываемые порой замечания о значительной трудозатратности полногеномного анализа [8] нивелируются в связи с непрерывным усовершенствованием и удешевлением технологий секвенирования.

Выявляемые нами на РНК-уровне маркеры отличаются от используемых на практике олигонуклеотидных зондов [15, 8] своей структурированностью, характеризующейся различными проявлениями повторности. Это либо тандемные повторы (периодичности), либо локальные фракталы и фракталоподобные структуры, либо компактно локализованные комбинации разнотипных повторов. Предполагается, что структурированность выявляемых РНК-маркеров может способствовать их содержательной интерпретации. Ориентация именно на эти структуры обусловлена тем, что:

– тандемная повторность часто ассоциируется с дублированием наиболее важных в функциональном отношении фрагментов генома (например, рибосомных сайтов связывания, терминальных кодонов и др. [17]);

– локальные фракталы и фракталоподобные структуры реализуют универсальный механизм усиления закономерности: повторение короткой структуры приводит к образованию более сильной структуры, например, более длинной симметрии или комплементарной симметрии. Различные аспекты фрактальности применительно к биологическим последовательностям обсуждаются, в частности, и в других работах [23, 24];

– комбинации разнотипных повторов обычно фигурируют в участках генома с пониженной сложностью [17], последние же, в свою очередь, нередко ассоциируются с регуляторными областями. Дополнительным аргументом, подтверждающим правомерность выбора рассматриваемых структур в качестве маркеров генотипирования, являются результаты имитационного эксперимента: в реальных геномах ВКЭ выделяется примерно в 2 раза больше интересующих нас структур, чем в их рандомизированных аналогах.

Формально выделяемые РНК-маркеры для генотипирования ВКЭ могут использоваться независимо от олигонуклеотидных зондов, подбираемых вручную [8, 15], либо дополнять их в затруднительных случаях. В значительной мере эти

дополнительные возможности возникают благодаря наличию структурированных генотипспецифических РНК-маркеров даже в участках генома с одинаковым аминокислотным составом у всех генотипов.

Из трех типов рассматриваемых структур наибольшую представленность имеют маркеры – периодичности. Каждый (каждая) из них позволяет типировать подавляющее большинство штаммов какого-либо генотипа (см. табл. 4). РНК-маркеров, построенных на фракталоподобных и комбинированных структурах, меньше (см. таблицы 6–9), чем маркеров-периодичностей. Каждый из них покрывает уже не подавляющую, а лишь значимую долю штаммов конкретного генотипа. Эти структуры более длинные и допускают позиционные разрывы («гэпы») между составляющими их компонентами. На нуклеотидный состав гэпов не накладывается никаких ограничений, поэтому формально совпадающие структуры с одинаковыми компонентами (повторами) могут иметь отличия на уровне таких гэпов. Характер этих отличий требует дополнительных исследований. Представляет интерес и характер изменений самого маркера, ориентированного на конкретный генотип, в штаммах других генотипов (в силу близости всех геномов ВКЭ этот маркер, пусть и в искаженной форме, в них присутствует).

Любые маркеры, включая и рассматриваемые в данной работе, носят относительный характер, поскольку выявляются на основе анализа ограниченной («обучающей») подборки. При её расширении маркер может в значительной мере утратить свою «генотипспецифичность» (соответствующие примеры применительно к олигонуклеотидным зондам приведены в [8]). Поэтому при любом расширении подборки требуется корректировка множества маркеров, которую легче осуществлять формальными методами типа используемых в данной работе.

В основе нашего подхода лежат легко трактуемые теоретико-множественные понятия «пересечения» и «дополнения», применяемые к структурам, извлекаемым из кодирующих последовательностей разных генотипов. Потенциально возможные маркеры должны отбираться из генотипспецифических дополнений и входить в максимально возможное число штаммов типируемого генотипа. Используемые алгоритмы гарантируют просмотр всех потенциально возможных кандидатов на роль маркера. Схема носит общий характер и может быть использована для дифференциации других классов объектов, например, разных представителей рода флавивирусов.

## ЗАКЛЮЧЕНИЕ

В работе рассматривается актуальная в практическом аспекте задача генотипирования штаммов ВКЭ, т.е. отнесения их к одному из известных классов. Каждый класс характеризуется подборкой соответствующих штаммов. Каждый штамм представлен полной кодирующей РНК-последовательностью. Предполагаемый способ решения сводится к формированию набора генотипспецифических маркеров в виде цепочек РНК, обладающих определенной структурой. Рассматриваются три типа структур: тандемно повторяющиеся цепочки символов (периодичности), фракталоподобные конструкции и компактно локализованные комбинации разнотипных повторов. Приводятся аргументы, обосновывающие указанный выбор.

Исходная подборка содержит 161 штамм с указанием принадлежности каждого из них к определенному генотипу (генотипы 1, 2, 3 – соответственно, 80, 46 и 28 штаммов и генотип 5–7 штаммов). Единственный на данный момент штамм 178-79, представляющий генотип 4, в подборку включен не был. Алгоритм формирования структурированных РНК-маркеров основывается на построении ДНК (РНК)-ориентированных сложностных разложений, компонентами которых являются интересующие нас структуры. Приведены характерные маркеры, типирующие

значительное количество штаммов каждого класса. Наиболее сильные из них могут использоваться по отдельности. Менее сильные целесообразно объединять в группы, позволяющие покрывать все (или подавляющее большинство) штаммов каждого класса. Возможны разные варианты формирования таких групп.

Исследования выполнены при поддержке программы фундаментальных научных исследований СО РАН № I.5.1., проект № 0314-2016-0015, а также в рамках проекта «Разработка комплекса программного обеспечения для моделирования физико-химических и биологических свойств антигенных эпитопов белков разных генотипов вируса клещевого энцефалита» Программы Президиума РАН №I.33П «Фундаментальные проблемы математического моделирования. Фундаментальные проблемы факторизационных методов в различных областях. Алгоритмы и математическое обеспечение» (*Часть «Фундаментальные проблемы математического моделирования»*).

# ПРИЛОЖЕНИЕ 1. СПИСОК ШТАММОВ ВКЭ

| | Генотип 1 | Генотип 2 | Генотип 3 | Генотип 5 |
|---|---|---|---|---|
| 1 | >JQ825164.1_Primorye823 | >U27495.1_Neudoerfl | >JN003209.1_Irkutsk12 | >617-90 |
| 2 | >JQ825163.1_Primorye750 | >GQ266392.1_AS33 | >KF826916.1_Sakhalin_611 | >711-84 |
| 3 | >JQ825162.1_Primorye437 | >KP938507.1_Sorex_1810 | >KC422663.2_Zabaikalye_68B00 | >740-84 |
| 4 | >JQ825161.1_Primorye345 | >KP331443.1_IrkutskBR_145609 | >KF826914.1_Zabaikalye_109 | >606-90 |
| 5 | >JQ825160.1_Primorye320 | >KP331442.1_IrkutskBR_143409 | >KC414090.1_Zabaikalye_1199 | >608-90 |
| 6 | >JQ825159.1_Primorye274 | >KP331441.1_IrkutskBR_9908 | >JN003208.1_Cht22 | >EF469662.1_88684 |
| 7 | >JQ825158.1_Primorye208 | >KF151173.1_A104 | >KF823822.1_Irkutsk_BR_68311 | >KJ633033.1_88684 |
| 8 | >JQ825157.1_Primorye202 | >HM535611.1_KrM_93 | >LC017693.1_MGLSelenge1314 | |
| 9 | >JQ825156.1_Primorye739 | >HM535610.1_KrM_213 | >LC017692.1_MGLSelenge1312 | |
| 10 | >JQ825154.1_Primorye52 | >KP716978.1_HyprVs_str | >FJ968751.1_Kolarovo2008 | |
| 11 | >JQ825155.1_Primorye196 | >KP716977.1_HyprVs_prME | >KM019545.1_TomskPT122 | |
| 12 | >JQ825153.1_Primorye183 | >KP716976.1_HyprVs_E | >JN003207.1_Cht653 | |
| 13 | >JQ825152.1_Primorye75 | >KP716975.1_Hypr_IC | >JN003206.1_Aina | |
| 14 | >JQ825150.1_Primorye91 | >KP716974.1_Hypr_IC | >AF069066.1_Vasilchenko | |
| 15 | >JQ825149.1_Primorye87 | >KJ922514.1_Skrivanek | >L40361.3 | |
| 16 | >HM859895.1_Primorye2239 | >KJ922515.1_Tobrman | >KP345889.1_SibXJX5 | |
| 17 | >HQ901366.1_Primorye1153 | >KJ922516.1_Vlasaty | >KP716973.1_VsHypr_str | |
| 18 | >HQ901367.1_Primorye501 | >KJ922513.1_Petracova | >KP716972.1_VsHypr_prME | |
| 19 | >HM859894.1_Primorye633 | >KJ922512.1_Kubinova | >KP716971.1_VsHypr_E | |
| 20 | >EU816452.1_Primorye270 | >KC835596.1_285 | >KJ626343.1_Buzuuchuk | |
| 21 | >EU816451.1_Primorye253 | >KC835597.1_CGl223 | >JQ429588.1_MucAr_M14/10 | |
| 22 | >EU816450.1_Primorye212 | >DQ401140.3_Toro2003 | >LC017691.1_IR9922f7 | |
| 23 | >AY169390.3_Primorye332 | >KC835595.1_114 | >KP644245.1_C1113 | |
| 24 | >HQ201303.1_Primorye92 | >U39292.1\|TEU39292_Hypr | >AF527415.1_Zausaev | |
| 25 | >FJ997899.1_Primorye90 | >FJ572210.1_Salem | >KJ701416.1_Lesopark_11 | |
| 26 | >GQ228395.1_Primorye18 | >KF991107.1_Mandal2009 | >GU183384.1_Est54 | |
| 27 | >FJ906622.1_Primirye89 | >KF991106.1_Saringe2009 | >GU183382.1_Latvia196 | |
| 28 | >EU816455.2_Primorye86 | >JQ654701.1_Ljubljana_I | >DQ486861.1_EK328 | |
| 29 | >EU816454.1_Primorye94 | >GU183383.1_Est3476 | | |
| 30 | >EU816453.1_Primorye69 | >GU183381.1_Joutseno | | |
| 31 | >JQ825146.1_Kiparis94 | >GU183380.1_Kumlinge_A52 | | |
| 32 | >JQ825148.1_Primorye82 | >GU183379.1_Kumlinge_2503 | | |
| 33 | >JQ825145.1_Primorye895 | >HM120875.1_84.2 | | |
| 34 | >JQ825144.1_Primorye828 | >U27491.1_263 | | |
| 35 | >AB062063.2_Oshima_510 | >DQ153877.1_variant_of_263 | | |
| 36 | >AB753012.1_Oshima_08As | >KJ000002.1_Absettarov | | |
| 37 | >JQ825147.1_Shkotovo94 | >U39292.1_Hypr | | |
| 38 | >KF880803.1_9024 | >AM600965.1_K23 | | |
| 39 | >KM019546.1_TomskPT12 | >1G-98 | | |
| 40 | >KJ914683.1_TomskM202 | >118-71 | | |
| 41 | >KJ914682.1_TomskPT14 | >126-71 | | |
| 42 | >KJ739731.1_TomskM83 | >163-74 | | |
| 43 | >KJ739730.1_TomskK6 | >262-74 | | |
| 44 | >KJ739729.1_NovosibirskL2008 | >Zmeinogorsk-1 | | |
| 45 | >DQ989336.1_205 | >Zmeinogorsk-5 | | |
| 46 | >JX498939.1_205 | >Zmeinogorsk-9 | | |
| 47 | >KF880804.1_8696 | | | |
| 48 | >KP869172.1_Nikolaevsk_855 | | | |
| 49 | >FJ402886.1_Dalnegorsk | | | |
| 50 | >JQ825151.1_Spassk72 | | | |
| 51 | >JN003205.1_Irkutsk1861 | | | |
| 52 | >JF819648.2_SofjinKSY | | | |
| 53 | >JX498940.1_Sofjin | | | |
| 54 | >JN229223.1_SofjinRu | | | |
| 55 | >AB062064.1_SofjinHO | | | |
| 56 | >KC806252.1_SofjinChumakov | | | |
| 57 | >KF951037.1_4072 | | | |
| 58 | >KP844725.1_Chichagovka_1223 | | | |
| 59 | >KP844724.1_Chichagovka_1222 | | | |
| 60 | >KJ744034.1_Malishevo | | | |
| 61 | >KF880805.1_1230 | | | |
| 62 | >KT001072.1_Khekhtzir_1713 | | | |
| 63 | >KT001071.1_Khekhtzir_1013 | | | |
| 64 | >KT001070.1_Khekhtzir_913 | | | |
| 65 | >KP844727.1_Birobidzhan_1357 | | | |
| 66 | >KP844726.1_Birobidzhan_1354 | | | |
| 67 | >DQ862460.1_Glubinnoe/2004 | | | |
| 68 | >KT001073.1_Lazo_MP36 | | | |
| 69 | >JX968560.1_Zabaikalye198 | | | |
| 70 | >KC422667.2_Zabaikalye_3000 | | | |

| | Генотип 1 | Генотип 2 | Генотип 3 | Генотип 5 |
|---|---|---|---|---|
| 71 | >KF826915.1_Zabaikalye_609 | | | |
| 72 | >GU121642.1_Svetlogorie | | | |
| 73 | >FJ402885.1_Kavalerovo | | | |
| 74 | >JF316708.1_MDJ03 | | | |
| 75 | >JF316707.1_MDJ02 | | | |
| 76 | >JQ650522.1_MDJ01 | | | |
| 77 | >JX534167.1_Xinjiang01 | | | |
| 78 | >KJ755186.1_WH2012 | | | |
| 79 | >JQ650523.1_Senzhang | | | |
| 80 | >AY182009.1_Senzhang | | | |

## СПИСОК ЛИТЕРАТУРЫ


1. Зильбер А.А. *Эпидемические энцефалиты*. М.: Медгиз, 1945. 255 с.
2. Вотяков В.И., Злобин В.И., Мишаева Н.П. *Клещевые энцефалиты Евразии*. Новосибирск: Наука, 2002. 438с.
3. *Virus Taxonomy*: VIIIth Report of the ICTV. Eds. Fauquet C.M., Mayo M.A., Maniloff J., Desselberger U., Ball L.A. London: Elsevier/Academic Press, 2005. 1162 p.
4. Плетнёв А.Ш., Ямщиков В.Ф., Блинов В.М. Нуклеотидная последовательность генома и полная аминокислотная последовательность полипротеина вируса клещевого энцефалита. *Биоорганическая химия*. 1989. Т. 15. № 11. С 1504–1521.
5. Mandl C.W., Heinz F.X., Stöckl E., Kunz C. Genome sequence of tick-borne encephalitis virus (Western subtype) and comparative analysis of nonstructural proteins with other flaviviruses. *Virology*. 1989. 173. P. 291–301.
6. Плетнёв А.Г. *Структура, организация и детекция генома вируса клещевого энцефалита*. Автореферат дис. … д-ра хим. наук. М., 1990 – 48 стр.
7. Беликов С.И., Гусев В.Д., Мирошниченко Л.А., Титкова Т.Н. *Сравнительный анализ геномов вируса клещевого энцефалита: дифференциация по степени вирулентности:* доклады IV Международной конференции «Математическая биология и биоинформатика» (ICMBB12) (14–19 октября 2012 года, г. Пущино). С. 52–53.
8. Дёмина Т.В. *Вопросы генотипирования и анализ генетической вариабельности вируса клещевого энцефалита*: дис. … д-ра биол. наук: 03.02.02- вирусология. Иркутск, 2013. 248 с.
9. Ecker M., Allison S.L.,Meixner T., Heinz F.X. Sequence analysis and genetic classification of tick-borne encephalitis viruses from Europe and Asia. *Journal of general virology*. 1999. V. 80. № 1. P. 179-85. doi: 10.1099/0022-1317-80-1-179
10. Демина Т.В., Злобин В.И., Верхозина М.М., Очирова Л. А., Батомункуев А.С., Чхенкели В.А. Анализ кодирующей части генома вируса клещевого энцефалита. *Живые и биокосные системы*. 2014. № 9. URL: http://jbks.ru/archive/issue-9/article-20 (дата обращения: 26.02.2018).
11. Демина Т.В., Джиоев Ю.П., Козлова И.В., Верхозина М.М., Ткачев С.Е., Дорощенко Е.К., Лисак О.В., Парамонов А.И., Злобин В.И. Генотипы 4 и 5 вируса клещевого энцефалита: особенности структуры геномов и возможный сценарий их формирования. *Вопросы вирусологии*. 2012. № 4. С. 13–19.
12. Grard G., Moureau G., Charrel R.N., Lemasson J.-J., Gonzalez J.-P., Gallian P., Gritsun T.S., Holmes E.C., Gould E.A., de Lamballerie X. Genetic characterization of tick-borne flaviviruses: New insights into evolution, pathogenic determinants and taxonomy. *Virology*. 2007. V. 361. P. 80–92.
13. Kerschner J.H., Vonrdam A.V., Monath T.P., Trent D.W. Genetic and epidemiological studies of Dengue type 2 viruses by hybridization using synthetic deoxyoligonucleotides as probes. *J. Gen. Virol*. 1986. V. 67. P. 2645–2661
14. Shamanin V.A., Pletnev A.G., Rubin S.G., Zlobin V.I. Differentiation of strains of tick-borne encephalitis virus by means of RNA-DNA hybridization. *J. Gen. Virol*. 1990. V. 71. P. 1505−1515.
15. Козлова И.В., Злобин В.И., Беликов С.И., Верхозина М.М., Демина Т.В., Джиоев Ю.П., Адельшин Р.В., Газо М.Х., Дорощенко Е.К. Молекулярные зонды для генетического типирования вируса клещевого энцефалита. *Вопросы вирусологии*. 2001. № 4. С. 43−47.
16. Шаманин В.А. *Детекция штаммов вируса клещевого энцефалита методом молекулярной гибридизации нуклеиновых кислот*: дис. … канд. биол. наук. Новосибирск, 1990. 24 с.

17. Гусев В.Д., Куличков В.А., Чупахина О.М. Сложностной анализ геномов. I Меры сложности и классификация выявляемых закономерностей. *Молекулярная биология*. 1991. Т. 25. № 3. С. 825−833.
18. Гусев В.Д., Мирошниченко Л.А., Чужанова Н.А. Выявление фракталоподобных структур в ДНК-последовательностях. В: *Classification, Forecasting, Data Mining*. № 8 Sofia: ITHEA, 2009. (Information Science & Computing. International Book Series) С. 117−123.
19. Гусев В.Д., Мирошниченко Л.А. Поиск комбинированных структур в ДНК-последовательностях. В: *Математические методы распознавания образов*: доклады всероссийской конференции ММРО-13 (Ленинградская обл., г. Зеленогорск. 30 сентября-6 октября 2007г.). М.: Макс-Пресс, 2007. С. 473−476.
20. Jeffreys A., Wilson V., Thein S. Individual-specific “fingerprints” of human DNA. *Nature*. 1985. V. 316. № 6023. P. 76–79.
21. Карань Л.С., Погодина В.В., Бочкова Н.Г. Маленко Г.В., Левина Л.С., Фролова Т.В., Мязин А.Е., Платонов А.Е. Определение генотипа вируса клещевого энцефалита с использованием метода анализа полиморфизма длин рестрикционных фрагментов. В: *Генодиагностика инфекционных болезней*: сборник трудов 5-й всероссийской научно-практической конференции. Том 2, раздел 6: Генодиагностика особо опасных инфекций. 2004. С. 177−179.
22. Lempel A., Ziv J. On the complexity of finite sequences. *IEEE Trans. Inform. Theory*. 1976. V. IT-22. No. 1. P. 75–81.
23. Соловьев В.В., Королев С.В., Туманян В.Г., Лим Х.А. Новый подход к классификации участков ДНК, основанный на фрактальном представлении набора функционально сходных последовательностей. *ДАН*. 1991. Т. 319. № 6. С. 1496–1500.
24. Jeffrey H.J. Chaos game representation of gene structure. *Nucl. Acids Res*. 1990. V. 18. P. 2163–2170.

Рукопись поступила в редакцию 22.01.2018.
Дата опубликования 16.03.2018.