Федеральное бюджетное учреждение науки
«Омский научно-исследовательский институт природно-очаговых инфекций»
Федеральной службы по надзору в сфере защиты
прав потребителей и благополучия человека

КЛАССИФИКАЦИЯ ФЛАВИВИРУСНЫХ ГЕНОМОВ ПО РЕЗУЛЬТАТАМ ДИСКРИМИНАНТНОГО АНАЛИЗА ПОКАЗАТЕЛЕЙ ОТНОСИТЕЛЬНОГО ИСПОЛЬЗОВАНИЯ СИНОНИМИЧНЫХ КОДОНОВ

Информационно-методическое письмо

Омск – 2015

Рекомендовано к изданию решением Ученого совета ФБУН «Омский НИИ природно-очаговых инфекций» Роспотребнадзора (протокол № 5 от 20 мая 2015 г.)

К47 Классификация флавивирусных геномов но результатам дискриминантного анализа показателей относительного использования синонимичных кодонов [текст] / В.В. Якименко, Ж.С. Тюлько; ФБУН «Омский НИИ природно-очаговых инфекций» Роспотребнадзора. – Омск: ООО Издательский центр «Омский научный вестник», 2015.-20 с.

ISBN 978-5-91306-076-1

В информационно-методическом письме предложен алгоритм статистической обработки и классификации полноразмерных геномов флавивирусов и их фрагментов большой длины, основанный на дискриминантном анализе показателей относительного использования синонимичных кодонов.

Разработано: ФБУН «Омский НИИ природно-очаговых инфекций» Роспотребнадзора (лаборатория арбовирусных инфекций отдела природно-очаговых вирусных инфекций:(д.б.н., зав. лабораторией *В.В. Якименко*, к.б.н., с.н.с. *Ж.С. Тюлько*).

Предназначено для вирусологов, молекулярных биологов, биоинформатиков.

УДК 578.833.2
ББК 52.639.234

ISBN 978-5-91306-076-1

Содержание

1. Общие представления о геноме флавивирусов и его представленности в генетических банках данных

Семейство флавивирусов (*Flaviviridae*) включает в себя три рода: *Flavivirus, Pestivirus, Hepacivirus*. Род *Flavivirus*объединяет более 70 вирусов, относящихся к 15 антигенным группам. Большинство флавивирусов является арбовирусами и передается хозяевам путем биологической трансмиссии, посредством членистоногих переносчиков, иногда относящихся к разным видам (например вирус Западного Нила). Флавивирусы распространены повсеместно и имеют большое медицинское и ветеринарное значение. Многие из них вызывают тяжелые природно-очаговые заболевания человека и животных (лихорадки, энцефалиты).

Геномная +РНК флавивирусов является инфекционной (≈ 11 тыс. н.о.), и кодирует три структурных (С, М и Е) и семь неструктурных белков (NS1, NS2A, NS2B, NS3, NS4A, NS4B, NS5), которые последовательно считываются в единой рамке считывания и необходимы для размножения вируса в клетках хозяина. Репликация геномной РНК происходит в цитоплазме по полуконсервативному механизму, инициация трансляции осуществляется по кэпзависимому механизму.

Флавивирусный вирион имеет сферическую форму (диаметр ≈ 40-50 нм) состоит из нуклеокапсида и покрывающей его липопротеиновой оболочки. Нуклеокапсид включает в себя структурный элемент капсида - капсидный белок С и геномную +РНК, он окружен липидной мембраной, в которую включены М- и Е- белки (мембранный - масса 8 кДа и оболочечный - 50 кДа), взаимодействующие при сборке вириона. Белок-предшественник prM участвует в начальных этапах фолдинга белка Е. Белок Е, отвечает за сборку вириона, слияние мембран и рецепторное связывание. В случае нейтрального значения pH гликопротеин Е представляет собой димер, чьи мономеры состоят из трех отдельных доменов. Считается, что мутации, влияющие на патогенность вируса, связаны с изменениями кодирующей последовательности этих трех доменов белка Е. NS2A, NS2B, NS4A и NS4B это низкомолекулярные, гидрофобные белки, чьи функции в развитии вируса в клетках пока ясны не до конца, но предполагается, что они могут взаимодействовать с другими вирусными белками и РНК. NS5 это РНК-зависимая РНК-полимераза.

В 5’ и 3’ – некодирующих концевых последовательностях вирусной РНК, обрамляющих рамку считывания, содержатся регуляторные элементы РНК, которые, предположительно, участвуют в формировании вторичной структуры РНК и контролируют многочисленные РНК-РНК и РНК-белковые взаимодействия, необходимые для осуществления репликации [16, 18].

Известно также, что вторичная структура РНК чувствительна к возникновению мутаций [19]. Эффекты, вызываемые появлением синонимичных и несинонимичных замен, определяются местом их возникновения, которое может соответствовать расположению важных регуляторных элементов РНК, а также особенностям вторичной структуры РНК, изменяющейся с их появлением. Поэтому, изучение закономерностей при возникновении таких мутаций (как синонимичных, так и несинонимичных) и их влияния на особенности строения каждого вируса имеет большой интерес для исследователей и требует применения различных статистических методов. Особенно актуальным это становится в связи со значительным ростом количества полноразмерных последовательностей вирусных РНК, накопленных в генетических банках данных за последние годы.

В качестве примеров подобного роста можно привести количество нуклеотидных последовательностей длиной >9000 оснований, депонированных в GenBank (март 2015 г.), относящихся к вирусам Западного Нила (>1000 последовательностей), дэнге (>4000), желтой лихорадки (>600), японского энцефалита (> 200), клещевого энцефалита (>100) и др. Подобный рост данных делает затруднительным применение некоторых традиционных методов сравнения и анализа к большим массивам последовательностей (например, использование множественного выравнивания) и приводит к необходимости автоматизации процесса предварительной систематизации больших массивов вирусных последовательностей.

Кроме того, по мере накопления данных становится возможным не только анализ количества, типа и расположения нуклеотидных замен, но также и выявление взаимосвязей при возникновении точечных мутаций в разных частях вирусной РНК статистическими методами, что уже было показано ранее.

Перспективной представляется разработка новых и применение существующих методов для анализа вирусной РНК и ее нуклеотидного состава, учитывающих особенности строения, отбора и эволюции вирусных геномов.

При этом предпочтение следует отдавать тем методам, которые в дальнейшем могут быть автоматизированы. При загрузке в такую анализирующую программу данных из генетического банка, исследователь, не используя значительный объем «ручной работы», за короткое время, сможет получить развернутый сравнительный анализ статистических свойств большой группы гомологичных последовательностей, с целью их последующего использования для более тщательного рассмотрения выявленных закономерностей у интересующей его подгруппы вирусов.

В настоящее время существует много алгоритмов получения первичных частотных характеристик нуклеотидного состава РНК (смещение нуклеотидного состава у отдельных видов, частота использования кодонов и т.д.). Многие из этих характеристик хорошо исследованы и рекомендованы к использованию. Однако их последующий анализ проводится намного реже. Мы предлагаем использовать при исследовании флавивирусов методы дискриминантного анализа для работы с частотными характеристиками, описывающими использование синонимичных кодонов в вирусных последовательностях.

Для сравнения между собой нуклеотидных последовательностей некоторых флавивирусов, мы вычисляли показатели относительного использования синонимичных кодонов с последующей обработкой, полученных результатов методами дискриминантного анализа. Этот комплексный подход позволил эффективно провести быстрое сравнение и классификацию гомологичных нуклеотидных последовательностей без использования ресурсоемких методов множественного выравнивания. Одновременно он выявил характерные особенности использования кодонов разными группами флавивирусов.

2. Исследование частотных характеристик использования кодонов у флавивирусов методами дискриминантного анализа

Известно, что использование в кодирующих последовательностях синонимичных кодонов не является случайным. Оно зависит от влияния многих факторов, и даже у одного и того же вируса стратегия использования кодонов разными его подтипами может различаться и зависеть от типичных хозяев данного подтипа. Отмечалось также, что тип использования отдельных кодонов может быть связан с наличием отбора и эффективностью трансляции, как у разных организмов, так и в различных тканях одного и того же организма. Были выдвинуты предположения о связи между предпочтениями в использовании кодонов и вирулентностью вируса и о возможности изменения с терапевтической целью состава тРНК хозяина для снижения скорости производства вирусных белков.

Поэтому частотный анализ кодонов, используемых конкретным вирусом, может дать информацию о его уникальных свойствах, позволяющих отличить его от других вирусов и провести корректную классификацию кодирующих вирусных последовательностей, что было показано нами на примере флавивирусов.

Дискриминантный анализ позволяет изучать различия между несколькими группами объектов по многим переменным, а также интерпретировать межгрупповые различия и определять вклад каждой переменной при классификации объектов. Мы применяли его для

анализа различий в частотных характеристиках использования синонимичных кодонов у каждого генотипа.

При анализе выборки из полноразмерных кодирующих последовательностей флавивирусов выполнялись следующие действия:

1. Для каждой кодирующей последовательности из выборки рассчитывались показатели относительного использования синонимичных кодонов, обозначаемые аббревиатурой $RSCU_k$ (Relative Synonymous Codon Usage), где k –тип кодона (стоп кодоны и кодоны, кодирующиеся только одним триплетом, не рассматривались).

$$RSCU_k = \frac{\mathbf{RSCU_{СК}} \cdot n_{\mathbf{k}}}{n_{cK}}$$

$RSCU_{СК}$–количество синонимичных кодонов для данной аминокислоты, n_k – частота использования кодона k, $n_{СК}$ – частота использования всех кодонов для данной аминокислоты в последовательности.

Показатель $RSCU_k$ обычно применяют для проведения корректных сравнений частот использования синонимичных кодонов в различных последовательностях. Он помогает оценить неслучайность появления конкретного кодона k при кодировании аминокислоты, а также сравнить схемы кодирования в разных генах. Большие значения $RSCU_k$ (>1) соответствуют более частому использованию кодона, значения < 1 - более редкому. В результате проведенных расчетов каждая из исследуемых последовательностей получала численное описание посредством значений переменных $RSCU_k$

2. Далее проводилась общая классификация изучаемых последовательностей, методами кластерного анализа, представляемая в форме схемы расстояний (дерева), построенной на основании анализа сходства в использовании синонимичных кодонов по результатам сравнения значений $RSCU_{k..}$

3. При необходимости более детального анализа различий в исследуемой выборке последовательностей, значения показателей $RSCU_k$ для всех типов кодонов нуклеотидной последовательности каждого вируса сравнивалась с аналогичными значениями, полученными для остальных последовательностей. Это осуществлялось посредством дискриминантного анализа, который включает в себя статистические методы анализа многомерных наблюдений в ситуации, когда ранее классифицированные последовательности могут быть использованы для составления обучающих выборок. Нуклеотидные последовательности при этом выступают в роли объектов, которые классифицируются по значениям $RSCU_k$, такая классификация позволяет интерпретировать межгрупповые

различия и определять вклад каждой переменной при классификации объектов.

4. Результаты проведенного анализа представлялись в виде диаграмм рассеяния значений двух выбранных дискриминантных функций (в качестве примера можно привести диаграмму на рис. 3). Из всех рассчитанных дискриминантных функций, отбирались те, которые дают наилучшую дискриминацию. По горизонтали откладывались значения одной из дискриминантных функции f_n, по вертикали значения другой f_m. Каждый маркер на схеме соответствует последовательности из выборки, для которой были рассчитаны значения функций. Расстояния между центроидами выявленных групп и вклад от использования каждого кодона, в рассчитанные значения дискриминантных функций, могут использоваться для описания исследуемых последовательностей.

3. Исследование частотных характеристик использования кодонов у вируса клещевого энцефалита методами дискриминантного анализа

Примером использования данного метода является исследование полноразмерных последовательностей вируса клещевого энцефалита.

Вирус клещевого энцефалита, представлен тремя основными генотипами, имеющими широкое географическое распространение, и несколькими, имеющими локальное распространение. Вирусы каждого генотипа, в пределах своего ареала являются сочленами экосистем с различной структурой. Один из важнейших вопросов в экологии вирусов – каким образом возбудитель адаптируется к смене хозяев с различным уровнем организации (холоднокровные членистоногие – теплокровные позвоночные) и различного систематического положения (видам, родам и др.). В последнее время появляется информация о хозяин-зависимых изменениях в структуре генома ВКЭ. Другой важный вопрос – в чем причины разнообразия биологических свойств вируса в разных частях его ареала? Наличие связи патогенных свойств ВКЭ с принадлежностью к различным генотипам спорно, т.к., нет строгой связи между многообразием клинических проявлений заболевания и степенью его тяжести для конкретных генотипов вируса. Способом внести вклад в решение данных вопросов является изучение системы кодирования генома вирусов, часть, которой - это стратегия кодирования белка, понимаемая здесь как закономерность использования кодонов в соответствующих кодирующих последовательностях ДНК и РНК.

Проанализировано закономерности использования кодонов основными генотипами ВКЭ. С этой целью, для полноразмерных кодирующих последовательностей ВКЭ определялись показатели относительного использования синонимичных кодонов, которые в дальнейшем изучались методами дискриминантного анализа.

Было показано, что различные генотипы ВКЭ неодинаково используют значительную часть синонимичных кодонов, т.е. их стратегии кодирования различаются. При этом в пределах одного генотипа наблюдается большее сходство в использовании синонимичных кодонов, чем между разными генотипами ВКЭ, независимо от способа изоляции штамма вируса и его последующего культивирования (Рис.1). Сравнение проводилось с помощью модуля «Групповой анализ» программы STATISTICA (Joining (tree-clustering)) с использованием различных метрик для расчета расстояний (данные не приведены), но топология дерева при сравнении с использованием евклидовых расстояний не менялась**.**

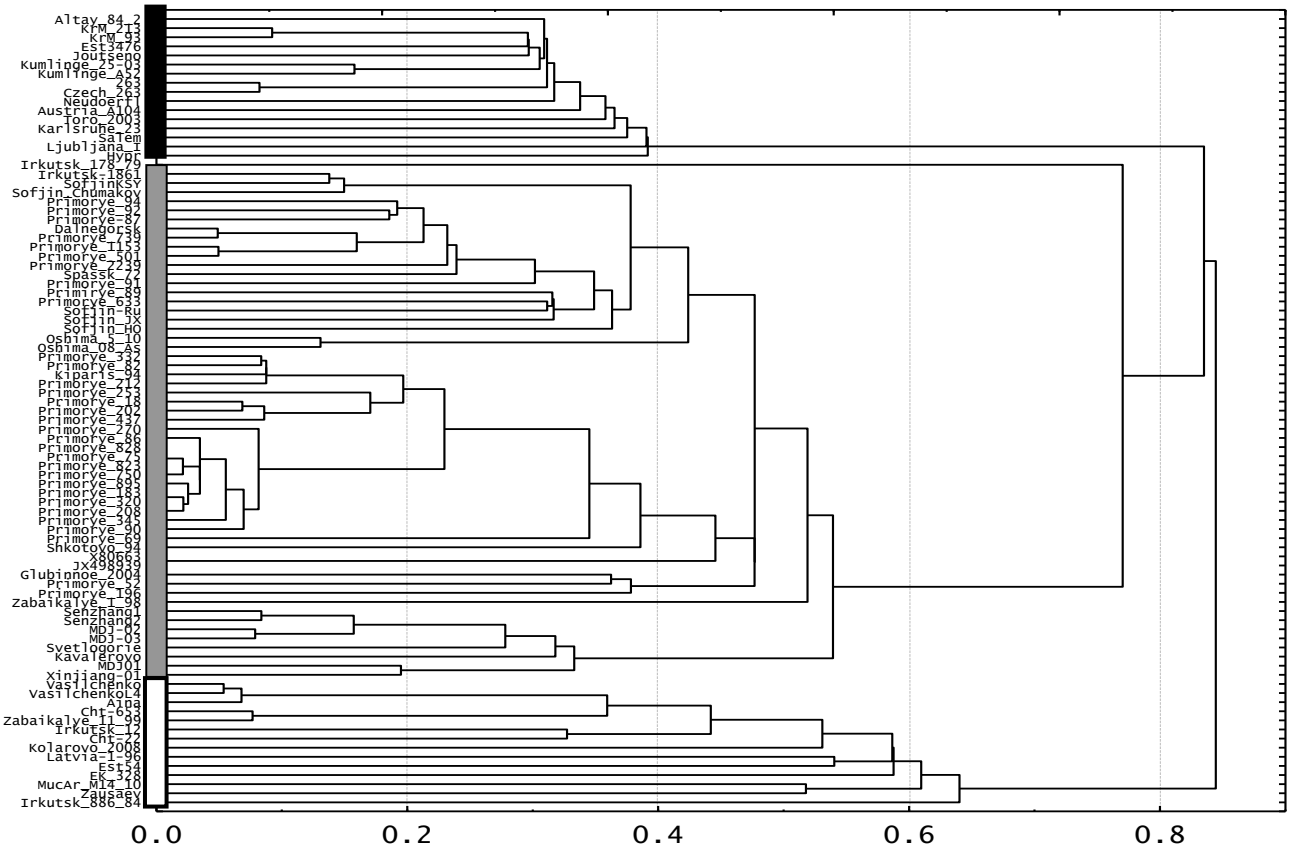

Рисунок 1. Схема, построенная на основании анализа сходства в использовании синонимичных кодонов по результатам сравнения значений $RSCU_k$. По оси абсцисс указаны значения евклидовых расстояний. Вдоль оси ординат черный прямоугольник отмечает изоляты, относящиеся к европейскому генотипу, серый прямоугольник – к дальневосточному, белый к сибирскому.

Полученный результат продемонстрировал, что стратегии кодирования у различных генотипов ВКЭ различаются, так как при построении дерева мы не закладывали никаких эволюционных моделей при расчете дистанций и анализировали только частотные характеристики использования синонимичных кодонов, а не их расположение в последовательности. Однако, разбиение на кластеры в схеме соответствовало разделению вирусов на генотипы (подтипы): сибирский, дальневосточный и европейский. Кроме того,

штаммы Irkutsk_178-79 и Irkutsk_886-84, являющиеся самостоятельными подтипами тоже выделились в отдельные кластеры (Рис.1). Данный результат позволил предположить, что отбор конкретных синонимичных кодонов может быть важной частью процесса микроэволюции ВКЭ, который отражается в структуре филогении вируса.

Причем, как показал анализ схем, построенных для наборов гомологичных фрагментов, вырезанных из кодирующих последовательностей, общий вид схемы не менялся в зависимости от того какие координаты имел набор использованных гомологичных фрагментов в геноме ВКЭ, если длина фрагментов была не меньше 1000 нуклеотидов, несколько менялись только абсолютные значения расстояний на схеме. При меньших длинах, начинали сказываться особенности аминокислотного состава, кодируемого последовательностями разных генов например, в случае редко используемых аминокислот.

Такой результат тоже предполагает наличие однотипного отбора кодонов, не связанного со структурой конкретного гена, при кодировании всего полипротеина в рамках каждого генотипа.

Следующим шагом был анализ того, как именно использует синонимичные кодоны каждый генотип. Общую картину использования кодонов у ВКЭ мы получили, рассчитав по всем анализируемым последовательностям разных генотипов средние значения $RSCU_k$ для каждого кодона k (Рис.2).

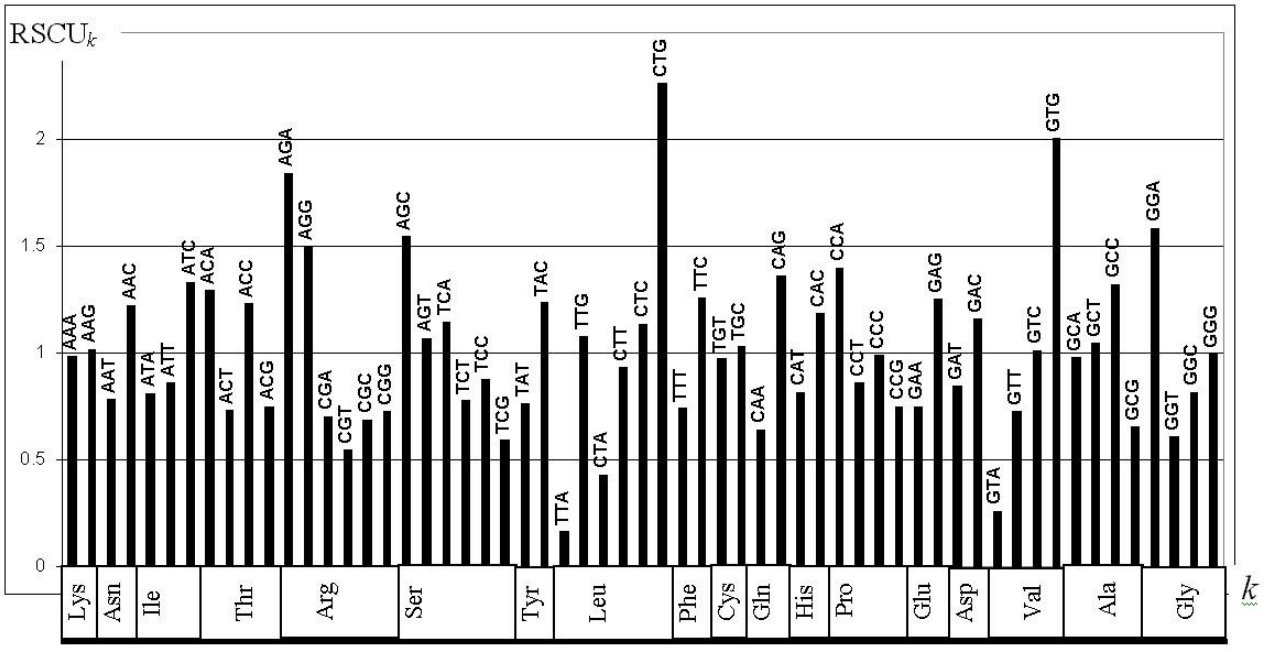

Рисунок 2. Использование синонимичных кодонов у ВКЭ. а) Высота столбцов на рисунке соответствует средним значениям $RSCU_k$ рассчитанным для каждого кодона k по всем

анализируемым последовательностям. У вершины каждого столбца указан нуклеотидный состав кодона k. Вдоль оси абсцисс обозначены аминокислоты, кодируемые каждым кодоном.

Схожая схема использования кодонов, с некоторыми вариациями в случае отдельных кодонов, наблюдалась и при расчете средних значений $RSCU_k$ для каждого генотипа в отдельности. Синонимичные кодоны, кодирующие одну и ту же аминокислоту, используются не одинаково, вне зависимости от того, сколько синонимичных кодонов ей соответствует. Неодинаковое использование синонимичных кодонов может быть вызвано разными причинами, и однозначный вывод о существовании значимых различий в стратегии кодирования между всеми генотипами сделать нельзя. Но важно понять, могут ли различия в использовании конкретных кодонов быть связаны с различиями в общей стратегии кодирования генотипов, или же они вызваны случайными причинами, или действиями независимых факторов.

Анализ коэффициентов дискриминантных функций выявил следующие различия в использовании кодонов:

AAA, AAG – лизин (Lys); AAT, AAC – аспарагин (Asn); TAT, TAC – тирозин (Tyr), TTT, TTC – фенилаланин (Phe); TGT, TGC – цистеин (Cys); CAA – глутамин (Gln); GAC, GAT - аспарагиновая кислота (Asp); GAG, GAA - глутаминовая кислота (Glu); CAC, CAT - гистидин (His) - не вносят вклада в дискриминацию генотипов, тогда как различия в использовании остальных кодонов позволяют провести такую дискриминацию. Эти значимые кодоны соответствуют аминокислотам: ATA, ATC, ATT – изолейцин (Ile); ACA, ACC, ACG, ACT - треонин (Thr); AGA, AGG, CGA, CGC, CGG, CGT - аргинин (Arg); AGC, AGT, TCA, TCC, TCG, TCT - серин (Ser); CTA, CTC, CTG, CTT, TTA, TTG - лейцин (Leu); CCA, CCC, CCG, CCT - пролин (Pro); GTA, GTC, GTG, GTT - валин (Val); GCA, GCC, GCG, GCT - аланин (Ala); GGA, GGC, GGG, GGT - глицин (Gly). Как и ожидалось, наибольшее значение для дискриминации имеют кодоны аминокислот с наибольшей вырожденностью. Ранее эти аминокислоты (кроме пролина) уже упоминались в качестве генотипспецифических признаков в отдельных позициях вирусной последовательности. Здесь же указывалось на наличие генотипических отличий на уровне синонимичных мутаций в положении третьего нуклеотида кодона.

Примечательно, что использование одних и тех же кодонов оставалось значимым для классификации последовательностей, как при сравнительном анализе различных подтипов ВКЭ, так и при анализе отдельных геновариантов в рамках одного и того же генотипа (были проанализированы дальневосточный и сибирский генотипы), менялись только значения

стандартизованных коэффициентов, которыми в нашем исследовании описывалась изучаемая стратегия кодирования.

Анализ показал неодинаковый вклад рассчитанных дискриминантных функций: функция f_1 вносит основной вклад в дискриминацию европейского генотипа от дальневосточного и сибирского, а функция f_2 позволяет наиболее достоверно разделить сибирский и дальневосточный генотипы.

Наибольший вклад в дискриминантную функцию f_1 вносят кодоны CTA, CTC, CTG, CTT, TTG - (соответствует Leu); CCA, CCC, CCG - (Pro); AGA, AGG, CGC, CGT - (Arg); GGG - (Gly), присутствие которых изменяет ее значения и вызывает смещение в сторону европейского генотипа, и кодоны GTC, GTT - (Val); AGC, AGT, TCA, TCC, TCG, TCT - (Ser); ATA, ATC, ATT - (Ile), присутствие которых вызывает смещение к сибирскому и дальневосточному генотипам.

В случае функции f_2 наибольший вклад в увеличение её значения (это соответствует смещению к сибирскому генотипу) вносят кодоны: ATA, ATC, ATT - (Ile); AGA, AGG, CGA, CGC, CGG, CGT - (Arg); AGC, AGT, TCC, TCG, TCT - (Ser); ACC, ACG, ACT - (Thr); в уменьшение – кодоны: CTG, TTG - (Leu); GGA, GGC, GGG, GGT - (Gly); CCA, CCC, CCG, CCT - (Pro).

Качество классификации, построенной по результатам, проведенного дискриминантного анализа (Рис.3), оказалось хорошим (лямбда Уилкса = 0.0081), ошибочных классификаций для генотипов 1,2,3 в кросс-проверочной выборке не было.

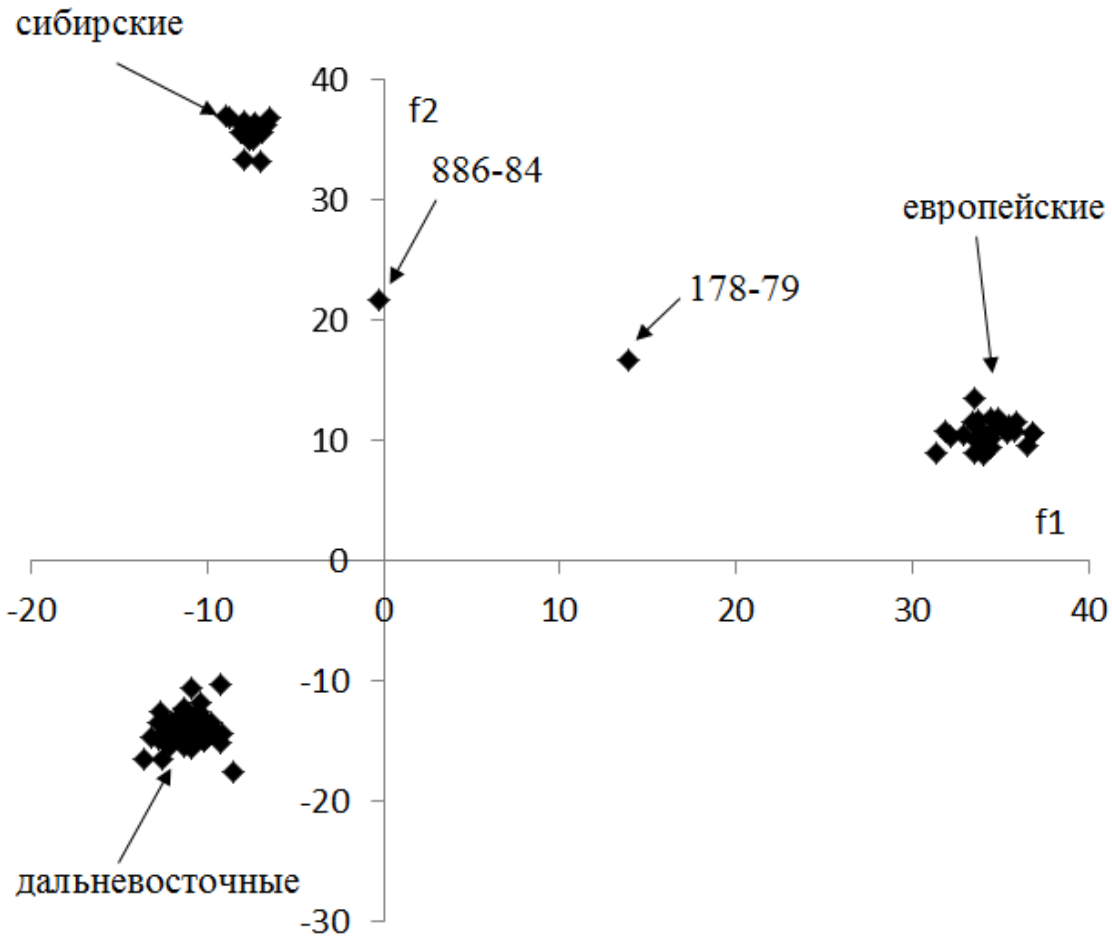

Рисунок 3. Диаграмма рассеяния значений дискриминантных функций, рассчитанных для последовательностей ВКЭ, рассчитана по полноразмерным кодирующим последовательностям ВКЭ. По оси абсцисс отложены значения дискриминантной функции f_1, по оси ординат f_2. Каждый маркер на схеме соответствует определенной последовательности из банка данных. Последовательности генотипов 886_84 и 178_79 не входили в обучающую выборку.

Локализация включенных в кросс-проверочную выборку геновариантов Irkutsk_886-84 и Irkutsk_178-79 подтверждает выделение их в самостоятельные генотипы.

Качество дискриминации при использовании фрагмента РНК, кодирующего только белок Е, оказалось несколько хуже - расстояния между центроидами групп значительно меньше. Полученные классификационные функции можно применять для типирования гомологичных полноразмерных или протяженных (желательно более 1500 нуклеотидов) фрагментов кодирующих нуклеотидных последовательностей ВКЭ, и по мере увеличения их количества уточнять значения коэффициентов функций.

Таким образом, продемонстрировано, что разные генотипы ВКЭ имеют не одинаковые, хотя и схожие стратегии использования кодонов. Поэтому появление одних и тех же синонимичных замен в последовательностях со схожей стратегией можно объяснить только наличием этой стратегии, без дополнительных предположений об их эволюционной истории.

В настоящее время можно считать, что стратегии кодирования у различных генотипов ВКЭ несколько различаются и отбор конкретных синонимичных кодонов, может быть важной частью процесса микроэволюции ВКЭ, который отражается в филогении вируса, т.е. не все синонимичные кодоны используются одинаково разными генотипами.

Наличие внутри генотипа отличной от других геновариантов стратегии кодирования аминокислот может быть началом микроэволюционного процесса, приводящего к выделению новых генотипов ВКЭ.

4. Алгоритм проведения исследования и используемое программное обеспечение

4.1. **Формирование входного файла** из гомологичных вирусных нуклеотидных последовательностей, по поисковому запросу, выполненному в банках генетических данных GenBank, EMBL и т.д. Сохранение сформированного файла в формате, наиболее подходящем для дальнейшего анализа.

4.2. **Расчет значений $RSCU_k$** для каждой изучаемой последовательности поддерживается различными видами программного обеспечения, например:

- RescueNet (The Relative Synonymous Codon Usage Neural Network) свободно распространяющееся программное обеспечение (http://bioinf.nuigalway.ie/RescueNet). Использует для анализа нуклеотидные последовательности в формате FASTA, ограничений на размер входного файла нет. После анализа формирует файл с рассчитанными значениями $RSCU_k$ (k=1,2,…,59) для каждой последовательности.
- seqinr (Biological Sequences Retrievaland Analysis) свободно распространяющееся программное обеспечение (http://seqinr.r-forge.r-project.org/), является дополнительно устанавливаемым пакетом системы статистического анализа данных R. Для эффективного использования требует предварительного изучения языка R. Преимуществом использования является возможность организации всех этапов исследования - от вычисления показателей относительного использования синонимичных кодонов с последующей обработкой, полученных результатов методами дискриминантного анализа, до итоговой визуализации результатов в одной и той же программной среде. Если используется только для расчетов значений $RSCU_k$, то можно сформировать выходной файл, содержащий эти значения в любом требуемом формате.

- Mega 4.0.1 позволяет рассчитать значения $RSCU_k$ (путь в меню окна «Sequence Data Explorer»: Statistics - Codon Usage), как для отдельной последовательности так и по результатам анализа массива последовательностей (в последнем случае значения $RSCU_k$ не рассчитываются для каждой последовательности в отдельности). Использует для анализа нуклеотидные последовательности, записанные в файлах различных форматов. Результат в виде таблицы сохраняется в текстовом файле и содержит, как частоты использования кодонов, так и значения $RSCU_k$, данные в скобках рядом с ними.

4.3. **Кластерный анализ.** Необходимо загрузить или скопировать, полученный файл содержащий, рассчитанные значения $RSCU_k$, в программу STATISTICA, таким образом, чтобы значения $RSCU_k$, рассчитанные для каждого кодона k по всем последовательностям, формировали столбцы в таблице исходных данных, а имена нуклеотидных последовательностей задавали имена строк таблицы. После чего, можно использовать методы кластерного анализа, реализованные в программе STATISTICA для построения классификационных деревьев (дендрограмм) изучаемого массива нуклеотидных последовательностей, описанного с помощью значений $RSCU_k$ (путь в меню: Statistics - Multivariate Exploratory Techniques – Cluster Analysis - Joining (tree-clustering)). Параметры выполнения следующие: «Input file: Raw Data», «Cluster: Cases(rows)», «Horizontal hierarchical tree plot». В этом случае, по горизонтали указываются значения расстояний, а вдоль вертикальной оси размещаются окончания ветвей, соответствующих каждой последовательности на схеме.

Кроме программы STATISTICA, на этом и дальнейших этапах возможно применение и других статистических программ, например: R, SPSS и т.д., способных рассчитывать матрицу расстояний для объектов, описываемых числовыми переменными.

4.4. **Дискриминантный анализ.** При проведении анализа с помощью программы STATISTICA, необходимо использовать модуль «Общие модели дискриминантного анализа» (путь в меню: Statistics – Multivariate Exploratory Techniques – General Discriminant Analysis Models), так как он предоставляет более широкие возможности контроля при своем исполнении и не имеет ряда ограничений для анализируемых значений по сравнению с каноническим дискриминантным анализом. Перед его использованием требуется создать дополнительный столбец таблицы, в котором в символьной или цифровой форме будет указана групповая принадлежность ранее классифицированных последовательностей, которые входят в обучающую выборку. При выполнении программы этот столбец должен быть указан в качестве зависимой переменной (Dependent variable). Остальные столбцы указываются в качестве непрерывных предикторов (Continuous predictors). После выполнения

расчёта, в окне «GDA results» можно просмотреть и скопировать все необходимые в дальнейшем результаты. Рассчитывая канонические дискриминантные функции (f_1 , f_2, ...) число которых на единицу меньше числа выделяемых групп, мы оптимально разделяем группы и видим вклад каждой переменной в дискриминацию. Стандартизованные коэффициенты дискриминантных функций позволяют оценить как вклады, так и направления вкладов переменных в каждую каноническую функцию, т.к. они рассчитываются по стандартизованным переменным и принадлежат к абсолютной шкале измерений. Достоинством метода является то, что полученные классификационные функции позволяют производить классификацию новых последовательностей по рассчитанным значениям $RSCU_k$, не проводя полного дискриминантного анализа повторно каждый раз, например, в редакторе текстовых таблиц EXCEL. Используя средства модуля «Общие модели дискриминантного анализа» легко оценить качество и значимость проведенной дискриминации данных.

4.5. **Представление диаграмм рассеяния** значений отобранных дискриминантных функций может осуществляться как средствами программы STATISTICA, так и с помощью любых программ и редакторов, позволяющих строить точечные двумерные или трехмерные графики.

5. Выводы

1. Выявляемые различия в частотах использования кодонов разными флавивирусами и вариантами вирусов одного вида применимы для быстрой классификации и первичного изучения структуры генома возбудителя.

2. Реализация классификации осуществляется методами дискриминантного анализа без процедуры предварительного выравнивания последовательностей, что позволяет избежать времязатратных вычислительных процедур.

3. Анализ частотных характеристик использования кодонов с помощью методов дискриминантного и кластерного анализа эффективен для классификации протяженных (>1500 нуклеотидов) гомологичных фрагментов флавивирусных геномов, как быстрый метод классификации, в тех случаях, когда из-за большого количества и длины анализируемых последовательностей их выравнивание и расчет филогенетических деревьев затруднительны. Таким образом разработанный алгоритм классификации имеет неоспоримое значение для оптимизации и упрощения эпидемиологического мониторинга за КЭ и других флавивирусов в органах Роспотребнадзора.

6. Рекомендации

1. Применять предложенный алгоритм, основанный на дискриминантном анализе показателей относительного использования синонимичных кодонов, для статистической обработки и классификации полноразмерных кодирующих последовательностей флавивирусов.

2. Использовать данный алгоритм для выявления различий в стратегиях кодирования аминокислот у различных групп флавивирусов. Отбор конкретных синонимичных кодонов, может быть важной частью процесса микроэволюции флавивирусов, который отражается в филогении.

3. Можно рекомендовать использование данного алгоритма для классификации и исследования различий в стратегиях кодирования аминокислот в гомологичных кодирующих последовательностях других вирусов, при условии достаточной длины (> 1500 нуклеотидов) вирусной рамки считывания.

7. Литература

1. Руководство по вирусологии: Вирусы и вирусные инфекции человека и животных / Под.ред. Д.К. Львова. М.: ООО «Издательство «Медицинское информационное агентство», 2013. 1200 с.

2. Бутвиловский, А.В. Изучение стратегии кодирования белков / А.В. Бутвиловский, В.Э. Бутвиловский, Е.А.Черноус // Медицинский журнал. 2009. (2). С.24-27.

3. Вотяков, В.И. Клещевые энцефалиты Евразии (вопросы экологии, молекулярной эпидемиологии, нозоологии, эволюции) / В.И. Вотяков, В.И. Злобин, Н.П. Мишаева. Новосибирск: Наука, 2002. 438 с.

4. Демина, Т.В. Вопросы генотипирования и анализ генетической вариабельности вируса клещевого энцефалита: автореф. дисс. …. докт. биол. Наук. Иркутск, 2013. 46 с.

5. Карганова Г.Г. Хозяин-специфические детерминанты в геноме вируса клещевого энцефалита. // В кн.: Фундаментальные и прикладные аспекты изучения паразитических членистоногих в XXI в. СПб.; 2013. С. 71-73.

6. Леонова Г.Н. Клещевой энцефалит в Приморском крае. Владивосток: Дальнаука, 1997.

7. Лукашев, В.В. Молекулярная эволюция и филогенетический анализ / В.В. Лукашев. М.: БИНОМ, 2009. 256 с.

8. Погодина, В.В. Сибирский подтип вируса клещевого энцефалита, доминирующий на территории России. Распространение и патогенность /В.В. Погодина, Л.С. Карань, Н.М. Колясникова, Л.С. Левина, С.Г. Герасимов и др. // Мед. Вирусология. 2013. Т.27(1). С. 36.

9. Тюлько Ж.С. К проблеме изменчивости генома вирусов клещевого энцефалита / Ж.С. Тюлько, В.В. Якименко // Национальные приоритеты России. Современные аспекты природной очаговости болезней. 2011. №2. С. 168-170.

10. Тюлько Ж.С. Оценка внутривидовой изменчивости при анализе использования кодонов у различных подтипов ВКЭ / Ж.С. Тюлько, В.В. Якименко // Национальные приоритеты России. Актуальные аспекты природной очаговости болезней. 2014. №3. С. 117-120.

11. Халафян, А.А. Учебник STATISTICA 6. Статистический анализ данных. М.: Бином, 2007.

12. Belalov, I.S. Causes and Implications of Codon Usage Bias in RNA Viruses / I.S. Belalov, A.N. Lukashev // PLOS ONE. 2013. 8 (2). – URL: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0056642.

13. Chambers, T.J. Flavivirus genome organization, expression and replication / T. J. Chambers, C.S. Hahn, R. Galler, C.M. Rice // Ann. Rev. Microbiol. -1990. Vol.44. P. 649–688.

14. Dunster, L.M. Attenuation of virulence of flaviviruses following passage on HeLa cells / L.M. Dunster, C.A. Gibson, J.R. Stephenson // J.gen.Virol. 1990. Vol.71. P. 601–607.

15. Frias, D. Human Retrovirus codon Usage from tRnA point of View: Therapeutic Insights / D. Frias, J.P. Monteiro-Cunha, A.C. Mota-Miranda, V.S. Fonseca, T. Oliveira, B. Galvao-Castro, L.C.J. Alcantara // Bioinformatics and Biology Insights. 2013. Vol.7. P.335-45.

16. Grard, G. Genetic characterization of tick-borne flaviviruses: New insights into evolution, pathogenetic determinants and taxonomy / G. Grard, G. Moureau, R.N. Charrel, J. Lemasson, J. Gonzalez, P. Gallian, T.S. Gritsun, E.C. Holmes, E.A. Gould, X. de Lamballerie // Virology. 2007. Vol.361. P.80–92.

17. Lorenz, I.C. Folding and Dimerization of Tick-Borne Encephalitis Virus Envelope Proteins prM and E in the Endoplasmic Reticulum / I.C. Lorenz, S.L. Allison, F.X. Heinz, A. Helenius //J. of Virology. 2002. Vol. 76 (11). P. 5480–5491.

18. Mandl, C.W. Adaptation of Tick-Borne Encephalitis Virus to BHK-21 Cells Results in the Formation of Multiple Heparan Sulfate Binding Sites in the Envelope Protein and Attenuation In Vivo / C.W. Mandl, H. Kroschewski, S.L. Allison, R. Kofler, H. Holzmann, T. Meixner, F.X. Heinz // J. of Virology. 2001. Vol. 75 (12). P.5627–5637.

19. Marin, M.S. Phylogeny of TYU, SRE, and CFA virus: different evolutionary rates in the genus Flavivirus / M.S. Marin, P.M. Zanotto, T.S. Gritsun, E.A. Gould // Virology. 1995. Vol. 206 (2). P.1133–1139.

20. McMinn, P.S. Neurovirulence and neuroinvasivenes of Murrey Valley encephalitis virus mutants selected by passage in a monkey kidney cell line / P.S. McMinn, I.D. Marshall, L. Dalgarno // J.gen.Virol. 1995. Vol.76. P.865–872.

21. Perriere, G. Use and misuse of correspondence analysis in codon usage studies / G. Perriere, J. Thioulouse // Nucleic Acids Research. 2002. Vol. 30(20). P. 4548-4555.

22. Plotkin, J.B. Tissue-specific codon usage and the expression of human genes./J.B. Plotkin, H. Robins, A.J. Levine // Proc. Natl. Acad. Sci. USA. 2004. Vol.101. P.12588-12591.

23. Qian,W. Balanced Codon Usage Optimizes Eukaryotic Translational Efficiency / W. Qian, J-R. Yang, N.M. Pearson, C. Maclean, J. Zhang // PLoS Genet. 2012. Vol. 8(3). URL: http://www.plosgenetics.org/ article/info%3Adoi%2F10.1371%2Fjournal.pgen.1002603.

24. Schubert, A.M. Evolution of the sequence composition of Flaviviruses / A.M. Schubert, C. Putonti // Infection, Genetics and Evolution. 2010. Vol.10 (1). P.129-36.

25. Shah, P. Explaining complex codon usage patterns with selection for translational

efficiency, mutation bias, and genetic drift / P.Shah, M.A.Gilchrist // PNAS. 2011. Vol.108 (25). P.10231-10236.

26. Sharp, P. M. Codon usage in yeast cluster-analysis clearly differentiates highly and lowly expressed genes / P.M.Sharp, T.M.F.Tuohy, K.R.Mosurski // Nucleic Acids Research. 1986. Vol.14. P.5125- 5143.

27. Tello, M. Analysis of the use of codon pairs in the HE gene of the ISA virus shows a correlation between bias in HPR codon-pair use and mortality rates caused by the virus./ M. Tello, J.M. Saavedra, E. Spencer //Virology Journal. 2013. Vol.10. URL: http://www.virologyj.com/content/10/1/180.

28. Wong, E.H.M. Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus / E.H.M. Wong, D.K. Smith, R. Rabadan, M. Peiris, L.L.M. Poon // BMC Evolutionary Biology. 2010. Vol.253(10). - URL: http://www.biomedcentral.com/1471-2148/10/253.

Информационное издание

КЛАССИФИКАЦИЯ ФЛАВИВИРУСНЫХ ГЕНОМОВ ПО РЕЗУЛЬТАТАМ ДИСКРИМИНИНТНОГО АНАЛИЗА ПОКАЗАТЕЛЕЙ ОТНОСИТЕЛЬНОГО ИСПОЛЬЗОВАНИЯ СИНОНИМИЧНЫХ КОДОНОВ

Информационно-методическое письмо

Разработано: ФБУН «Омский НИИ природно-очаговых инфекций»

Роспотребнадзора (д.б.н. В.В.Якименко, к.б.н. Ж.С.Тюлько)

Сдано в набор 20.06.2015 Подписано к печати 30.08.2015
Формат 60x84\16. Бумага офсетная. Гарнитура Times New Roman
Печать оперативная. Усл.-печ.л. 1,2 Уч.-изд.л. 1,2 Тираж 300. Заказ 401

ООО Издательский центр «Омский научный вестник»
Тел.: 8-905-921-98-22. E-mail: evga-18@mail.ru

Отпечатано в РПФ «CMYKART», ИП Гусев С.В.
Г. Омск, пр. Мира , 7, т.ф.: 65-16-27
Тел.8-904-323-38-43